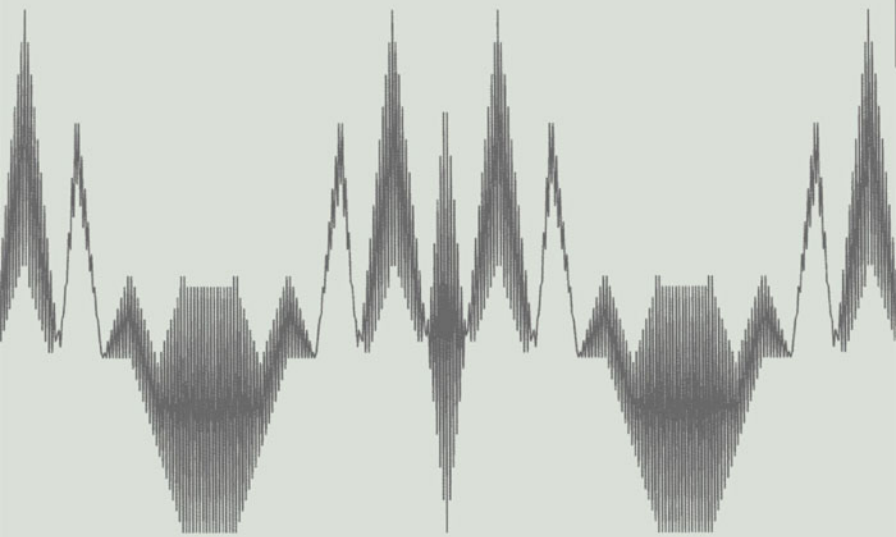
# Signal Processing for Telecommunications and Multimedia

Edited by
Tadeusz A. Wysocki
Bahram Honary
Beata J. Wysocki



Springer

# SIGNAL PROCESSING FOR TELECOMMUNICATIONS AND MULTIMEDIA

# MULTIMEDIA SYSTEMS AND APPLICATIONS SERIES

*Consulting Editor*

**Borko Furht**
*Florida Atlantic University*
borko@cse.fau.edu

Recently Published Titles:

**ADVANCED WIRED AND WIRELESS NETWORKS** edited by Tadeusz A. Wysocki, Arek Dadej and Beata J. Wysocki; ISBN: 0-387-22847-0; e-ISBN: 0-387-22928-0

**CONTENT-BASED VIDEO RETRIEVAL:** ***A Database Perspective*** by Milan Petkovic and Willem Jonker; ISBN: 1-4020-7617-7

**MASTERING E-BUSINESS INFRASTRUCTURE,** edited by Veljko Milutinović, Frédéric Patricelli; ISBN: 1-4020-7413-1

**SHAPE ANALYSIS AND RETRIEVAL OF MULTIMEDIA OBJECTS** by Maytham H. Safar and Cyrus Shahabi; ISBN: 1-4020-7252-X

**MULTIMEDIA MINING:** ***A Highway to Intelligent Multimedia Documents*** edited by Chabane Djeraba; ISBN: 1-4020-7247-3

**CONTENT-BASED IMAGE AND VIDEO RETRIEVAL** by Oge Marques and Borko Furht; ISBN: 1-4020-7004-7

**ELECTRONIC BUSINESS AND EDUCATION:** ***Recent Advances in Internet Infrastructures,*** edited by Wendy Chin, Frédéric Patricelli, Veljko Milutinović; ISBN: 0-7923-7508-4

**INFRASTRUCTURE FOR ELECTRONIC BUSINESS ON THE INTERNET** by Veljko Milutinović; ISBN: 0-7923-7384-7

**DELIVERING MPEG-4 BASED AUDIO-VISUAL SERVICES** by Hari Kalva; ISBN: 0-7923-7255-7

**CODING AND MODULATION FOR DIGITAL TELEVISION** by Gordon Drury, Garegin Markarian, Keith Pickavance; ISBN: 0-7923-7969-1

**CELLULAR AUTOMATA TRANSFORMS:** ***Theory and Applications in Multimedia Compression, Encryption, and Modeling,*** by Olu Lafe; ISBN: 0-7923-7857-1

**COMPUTED SYNCHRONIZATION FOR MULTIMEDIA APPLICATIONS,** by Charles B. Owen and Fillia Makedon; ISBN: 0-7923-8565-9

**STILL IMAGE COMPRESSION ON PARALLEL COMPUTER ARCHITECTURES** by Savitri Bevinakoppa; ISBN: 0-7923-8322-2

**INTERACTIVE VIDEO-ON-DEMAND SYSTEMS:** ***Resource Management and Scheduling Strategies,*** by T. P. Jimmy To and Babak Hamidzadeh; ISBN: 0-7923-8320-6

**MULTIMEDIA TECHNOLOGIES AND APPLICATIONS FOR THE 21st CENTURY:** ***Visions of World Experts,*** by Borko Furht; ISBN: 0-7923-8074-6

# SIGNAL PROCESSING FOR TELECOMMUNICATIONS AND MULTIMEDIA

**edited by**

**Tadeusz A. Wysocki**
*University of Wollongong, Australia*
**Bahram Honary**
*Lancaster University, UK*
**Beata J. Wysocki**
*University of Wollongong, Australia*

**Springer**

eBook ISBN: 0-387-22928-0
Print ISBN: 0-387-22847-0



Created in the United States of America

Visit Springer's eBookstore at: http://ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

# CONTENTS

## PART II: ERROR-CONTROL CODING, CHANNEL ACCESS, AND DETECTION ALGORITHMS



## PART III: HARDWARE IMPLEMENTATION

*This page intentionally left blank*

# PREFACE

The unprecedented growth in the range of multimedia services offered these days by modern telecommunication systems has been made possible only because of the advancements in signal processing technologies and algorithms. In the area of telecommunications, application of signal processing allows for new generations of systems to achieve performance close to theoretical limits, while in the area of multimedia, signal processing the underlying technology making possible realization of such applications that not so long ago were considered just a science fiction or were not even dreamed about. We all learnt to adopt those achievements very quickly, but often the research enabling their introduction takes many years and a lot of efforts. This book presents a group of invited contributions, some of which have been based on the papers presented at the 7th International Symposium on DSP for Communication Systems held in Coolangatta on the Gold Coast, Australia, in December 2003.

Part 1 of the book deals with applications of signal processing to transform what we hear or see to the form that is most suitable for transmission or storage for a future retrieval. The first three chapters in this part are devoted to processing of speech and other audio signals. The next two chapters consider image coding and compression, while the last chapter of this part describes classification of video sequences in the MPEG domain.

Part 2 describes the use of signal processing for enhancing performance of communication systems to enable the most reliable and efficient use of those systems to support transmission of large volumes of data generated by multimedia applications. The topics considered in this part range from error-control coding through the advanced problems of the code division multiple

access (CDMA) to multiple-input multiple-output (MIMO) systems and space-time coding.

The last part of the book contains seven chapters that present some emerging system implementations utilizing signal processing to improve system performance and allow for a cost reduction. The issues considered range from antenna design and channel equalisation through multi-rate digital signal processing to practical DSP implementation of a wideband direct sequence spread spectrum modem.

The editors wish to thank the authors for their dedication and lot of efforts in preparing their contributions, revising and submitting their chapters as well as everyone else who participated in preparation of this book.

*Tadeusz A. Wysocki*
*Bahram Honary*
*Beata J. Wysocki*

# PART 1:

# MULTIMEDIA SOURCE PROCESSING

*This page intentionally left blank*

Chapter 1

# A CEPSTRUM DOMAIN HMM-BASED SPEECH ENHANCEMENT METHOD APPLIED TO NON-STATIONARY NOISE

Mikael Nilsson, Mattias Dahl and Ingvar Claesson
*Blekinge Institute of Technology, School of Engineering, Department of Signal Processing, 372 25 Ronneby, Sweden*

**Abstract:** This paper presents a Hidden Markov Model (HMM)-based speech enhancement method, aiming at reducing non-stationary noise from speech signals. The system is based on the assumption that the speech and the noise are additive and uncorrelated. Cepstral features are used to extract statistical information from both the speech and the noise. A-priori statistical information is collected from long training sequences into ergodic hidden Markov models. Given the ergodic models for the speech and the noise, a compensated speech-noise model is created by means of parallel model combination, using a log-normal approximation. During the compensation, the mean of every mixture in the speech and noise model is stored. The stored means are then used in the enhancement process to create the most likely speech and noise power spectral distributions using the forward algorithm combined with mixture probability. The distributions are used to generate a Wiener filter for every observation. The paper includes a performance evaluation of the speech enhancer for stationary as well as non-stationary noise environment.

**Key words:** HMM, PMC, speech enhancement, log-normal

## 1. INTRODUCTION

Speech separation from noise, given a-priori information, can be viewed as a subspace estimation problem. Some conventional speech enhancement methods are spectral subtraction [1], Wiener filtering [2], blind signal separation [3] and hidden Markov modelling [4].

Hidden Markov Model (HMM) based speech enhancement techniques are related to the problem of performing speech recognition in noisy

environments [5,6]. HMM based methods uses a-priori information about both the speech and the noise [4]. Some papers propose HMM speech enhancement techniques applied to stationary noise sources [4,7]. The common factor for these problems is to the use of Parallel Model Combination (PMC) to create a HMM from other HMMs. There are several possibilities to accomplish PMC including Jacobian adaptation, fast PMC, PCA-PMC, log-add approx-imation, log-normal approximation, numerical integration and weighted PMC [5,6]. The features for HMM training can be chosen in different manners. However, the cepstral features have dominated the field of speech recognition and speech enhancement [8]. This is due to the fact that the covariance matrix, which is a significant parameter in a HMM, is close to diagonal for cepstral features of speech signals.

In general, the whole input-space, with the dimension determined by the length of the feature vectors, contains the speech and noise subspaces. The speech subspace should contain all possible sound vectors from all possible speakers. This is of course not practical and the approximated subspace is found by means of training samples from various speakers and by averaging over similar speech vectors. In the same manner the noise subspace is approximated from training samples. In non-stationary noise environments the noise subspace complexity increases compared to a stationary subspace, hence a larger noise HMM is needed. After reduction it is desired to obtain only the speech subspace.

The method proposed in this paper is based on the log-normal approximation by adjusting the mean vector and the covariance matrix. Cepstral features are treated as observations and diagonal covariance matrices are used for hidden Markov modeling of the speech and noise source. The removal of the noise is performed by employing a time dependent linear Wiener filter, continuously adapted such that the most likely speech and noise vector is found from the a-priori information. Two separate hidden Markov models are used to parameterize the speech and noise sources. The algorithm is optimized for finding the speech component in the noisy signal. The ability to reduce non-stationary noise sources is investigated.

## 2. FEATURE EXTRACTION FROM SIGNALS

The signal of concern is a discrete time noisy speech signal $x(n)$, found from the corresponding correctly band limited and sampled continuous signal. It is assumed that the noisy speech signal consists of speech and additive noise

$$x(n) = s(n) + w(n) \tag{1.1}$$

where $s(n)$ is the speech signal and $w(n)$ the noise signal.

The signals will be divided into overlapping blocks of length $L$ and windowed. The blocks will be denoted

$$\mathbf{x}_t^{\text{time}} = \mathbf{s}_t^{\text{time}} + \mathbf{w}_t^{\text{time}}, \quad \in \mathbb{R}^L \tag{1.2}$$

where $t$ is the block index and "time" denotes the domain. Note that the additive property still holds after these operations.

The blocks are represented in the linear power spectral domain as

$$\mathbf{x}_t^{\text{lin}} = \left|\mathbf{F}\mathbf{x}_t^{\text{time}}\right|^2, \quad \in \mathbb{R}^D \tag{1.3}$$

where $\mathbf{F} \in \mathbb{R}^{D \times L}$ is the discrete Fourier transform matrix and $D = L/2 + 1$ due to the symmetry of the Fourier transform of a real valued signal. Further, $|\bullet|$ denotes absolute value and "lin" denotes the linear power spectral domain. In the same manner $\mathbf{s}_t^{\text{lin}}$ and $\mathbf{w}_t^{\text{lin}}$ are defined. Hence the noisy speech in linear power spectral domain will be found as

$$\mathbf{x}_t^{\text{lin}} = \mathbf{s}_t^{\text{lin}} + \mathbf{w}_t^{\text{lin}} + 2\cos(\boldsymbol{\theta})\sqrt{\operatorname{diag}\left(\mathbf{s}_t^{\text{lin}}\right)\mathbf{w}_t^{\text{lin}}} \tag{1.4}$$

where $\boldsymbol{\theta}$ is a vector of angles between the individual elements in $\mathbf{F}\mathbf{s}_t^{\text{time}}$ and $\mathbf{F}\mathbf{w}_t^{\text{time}}$. The cosine for these angles can be found as

$$\begin{aligned}\cos(\boldsymbol{\theta}) = \operatorname{diag}\Big(\operatorname{diag}\Big(\Re\left\{\operatorname{diag}\left(\mathbf{F}\mathbf{s}_t^{\text{time}}\right)\cdot\left(\mathbf{F}\mathbf{w}_t^{\text{time}}\right)^*\right\}\Big)\cdot \\ \operatorname{diag}\Big(\operatorname{diag}\left(\left|\mathbf{F}\mathbf{s}_t^{\text{time}}\right|\right)\cdot\left|\mathbf{F}\mathbf{w}_t^{\text{time}}\right|\Big)^{-1}\Big)\end{aligned} \tag{1.5}$$

The speech and the noise signal are assumed to be uncorrelated. Hence, the cross term in Eq. (1.4) is ignored, and the approximation

$$\mathbf{x}_t^{\text{lin}} = \mathbf{s}_t^{\text{lin}} + \mathbf{w}_t^{\text{lin}} \tag{1.6}$$

is used.

Further, the power spectral domain will be transformed into the log spectral domain

$$\mathbf{x}_t^{\log} = \log\left(\mathbf{x}_t^{\text{lin}}\right), \quad \in \mathbb{R}^D \tag{1.7}$$

where the natural logarithm is assumed throughout this paper and "log" denotes the log spectral domain. The same operations are also applied for the speech and the noise. Finally the log spectral domain is changed to the cepstral domain

$$\mathbf{x}_t^{\text{cep}} = \mathbf{C}\mathbf{x}_t^{\log}, \quad \in \mathbb{R}^D \tag{1.8}$$

where "cep" denotes the cepstral domain and $\mathbf{C} \in \mathbb{R}^{D\times D}$ is the discrete cosine transform matrix defined as

$$C_{ij} = \begin{cases} \sqrt{\frac{1}{M}}\cos\left((i-1)(j-0.5)\pi/D\right) & \text{if } i = 1 \\ \sqrt{\frac{2}{M}}\cos\left((i-1)(j-0.5)\pi/D\right) & \text{otherwise} \end{cases} \tag{1.9}$$

where $i$ is the row index and $j$ the column index.

## 3. ERGODIC HMMS FOR SPEECH AND NOISE

Essential for model based speech enhancement approaches is to get reliable models for the speech and/or the noise. In the proposed system the models are found by means of training samples, which are processed to feature vectors in the cepstral domain, as described in previous section. These feature vectors, also called observation vectors in HMM nomenclature, are used for training of the models. This paper uses k-means clustering algorithm [9], with Euclidian distance measure between the feature vectors, to create the initial parameters for the iterative expectation maximation (EM) algorithm [10]. Since ergodic models are wanted, the clustering algorithm divides the observation vectors into states. The observation vectors are further divided into mixtures using the clustering algorithm on the vectors belonging to each individual state. Using these initial segmentation of vectors, the EM algorithm is applied and the parameters for the HMM are found. The model parameter set for an HMM with $N$ states and $M$ mixtures is

$$\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}) \tag{1.10}$$

where $\boldsymbol{\pi} = \{\pi_i\}$ contains the initial state probabilities, $\mathbf{A} = (a_{ij})$ the state transitions probabilities and $\mathbf{B} = \{b_j\} = \{c_{jk}, b_{jk}\} = \{c_{jk}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}\}$ the parameters for the weighted continuous multidimensional Gaussian functions for state $j$ and mixture $k$. For an observation, $\mathbf{o}_t$, the continuous multidimensional Gaussian function for state $j$ and mixture $k$, $b_{jk}(\mathbf{o}_t)$, is found as

$$b_{jk}(\mathbf{o}_t) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_{jk}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})^T \boldsymbol{\Sigma}_{jk}^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})} \tag{1.11}$$

where $D$ is the dimension of the observation vector, $\boldsymbol{\mu}_{jk}$ is the mean vector and $\boldsymbol{\Sigma}_{jk}$ is the covariance matrix. The covariance matrix is in this paper chosen to be diagonal. This implies that the number of parameters in the model is reduced and the computable cost of the matrix inversion is reduced. The weighted multidimensional Gaussian function for an observation $\mathbf{o}_t$, $b_j(\mathbf{o}_t)$, is defined as

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{M} c_{jk} b_{jk}(\mathbf{o}_t), \quad j = 1, 2, \ldots, N \tag{1.12}$$

where $c_{jk}$ is the mixture weight.

## 4. ERGODIC HMM FOR NOISY SPEECH USING PARALLEL MODEL COMBINATION

Given the trained models for speech and noise, a combined noisy-speech-model can be found by PMC $\lambda_x = \lambda_s \otimes \lambda_w$, where $\lambda_s$ and $\lambda_w$ are the model parameters for the speech and the noise HMM respectively and $\otimes$ denotes the operations needed to create the composite model.
This paper uses a non-iterative model combination and log-normal approximation to create the composite model parameters for the noisy speech. The compensation for the initial state is found as

$$\overset{x}{\pi}_{[iu]} = \overset{s}{\pi}_i \bullet \overset{w}{\pi}_u \tag{1.13}$$

In the same manner the transition probabilities, are given by

$$a^{x}_{[iu][jv]} = a^{s}_{ij} \bullet a^{w}_{uv} \tag{1.14}$$

where the state $[iu]$ represents the noisy speech state found by clean speech state $i$ and the noisy state $u$, and similar for $[iu]$.

The compensated mixture weights are found as

$$c^{x}_{[jv][kl]} = c^{s}_{jk} \bullet c^{w}_{vl} \tag{1.15}$$

where $[kl]$ is the noisy speech mixture given the clean speech mixture $k$ and the noise mixture $l$.

Since the models are trained in the cepstral domain, the mean vector and the covariance matrix are also in cepstral domain. Hence the mean vector and the covariance matrix in Eq. (1.11) are in the cepstral domain. Since the uncorrelated noise is additive only in the linear spectral domain, transformations of the multivariate Gaussian distribution are needed. These transformations are applied both for the clean speech model and the noise model. The first step is to transform the mean vectors and the covariance matrices from cepstral domain into the log spectral domain (the indices for state $j$ and mixture $k$ are dropped for simplicity)

$$\begin{aligned} \boldsymbol{\mu}^{\log} &= \mathbf{C}^{-1}\boldsymbol{\mu}^{\text{cep}} \\ \boldsymbol{\Sigma}^{\log} &= \mathbf{C}^{-1}\boldsymbol{\Sigma}^{\text{cep}}\left(\mathbf{C}^{-1}\right)^{T} \end{aligned} \tag{1.16}$$

Equation (1.16) is the standard procedure for linear transformation of a multivariate Gaussian variable. Equation (1.17) defines the relationship between the log spectral domain and the linear spectral domain for a multivariate Gaussian variable[6]

$$\begin{aligned} \mu_m^{\text{lin}} &= e^{\mu_m^{\log} + \Sigma_{mm}^{\log}/2} \\ \Sigma_{mn}^{\text{lin}} &= \mu_m^{\text{lin}}\mu_n^{\text{lin}}\left(e^{\Sigma_{mn}^{\log}} - 1\right) \end{aligned} \tag{1.17}$$

where $m$ and $n$ are indices in the mean vector and the Gaussian covariance matrix for state $j$ and mixture $k$. Now the parameters for the clean speech and the noise are found in the linear spectral domain. The mean vectors for the speech and the noise in linear spectral domain are stored to be used in the

enhancement process. In Eq. (1.17) it can be seen that the linear spectral domain is log-normal distributed. Given the assumption that the sum of two log-normal distributed variables are log-normal distributed, the distorted speech parameters can be found as

$$\begin{aligned} \overset{x}{\boldsymbol{\mu}}^{lin} &= \overset{s}{\boldsymbol{\mu}}^{lin} + g\,\overset{w}{\boldsymbol{\mu}}^{lin} \\ \overset{x}{\boldsymbol{\Sigma}}^{lin} &= \overset{s}{\boldsymbol{\Sigma}}^{lin} + g^2\,\overset{w}{\boldsymbol{\Sigma}}^{lin} \end{aligned} \tag{1.18}$$

where $g$ is a gain term introduced for signal to noise discrepancies between training and enhancement environment. The noise parameters are subsequently inverse transformed to the cepstral domain. This is done by first inverting Eq. (1.17)

$$\begin{aligned} \overset{x}{\mu}_m^{log} &= \log\left(\overset{x}{\mu}_m^{lin}\right) - \frac{1}{2}\log\left(\frac{\overset{x}{\Sigma}_{mm}^{lin}}{\left(\overset{x}{\mu}_m^{lin}\right)^2} + 1\right) \\ \overset{x}{\Sigma}_{mn}^{log} &= \log\left(\frac{\overset{x}{\Sigma}_{mn}^{lin}}{\overset{x}{\mu}_m^{lin} \bullet \overset{x}{\mu}_n^{lin}} + 1\right) \end{aligned} \tag{1.19}$$

and then transform the log spectral domain expression into the cepstral domain

$$\begin{aligned} \overset{x}{\boldsymbol{\mu}}^{cep} &= \mathbf{C}\,\overset{x}{\boldsymbol{\mu}}^{log} \\ \overset{x}{\boldsymbol{\Sigma}}^{cep} &= \mathbf{C}\,\overset{x}{\boldsymbol{\Sigma}}^{log}\mathbf{C}^T \end{aligned} \tag{1.20}$$

yielding all parameters are found for the compensated model.

## 5. CLEAN SPEECH SIGNAL ESTIMATION

The enhancement process, see Fig. *1-1,* uses information gained from training of speech and noise models to estimate the clean speech signal. The information needed are the stored mean vectors in linear spectral domain, $\overset{s}{\boldsymbol{\mu}}_{jk}^{lin}$ and $\overset{w}{\boldsymbol{\mu}}_{jk}^{lin}$, for the speech and the noise respectively, the compensated

model, $\lambda_x$, and the gain difference, $g$, between training and enhancement environment.

The stored mean vectors are first restored from length $D$ to the full block length, $L$, according to

$$\begin{aligned} \overset{s}{P}_m &= \begin{cases} \overset{s}{\boldsymbol{\mu}}{}^{\text{lin}}_{m} & \text{if } m \le D \\ \overset{s}{\boldsymbol{\mu}}{}^{\text{lin}}_{(2D-m)} & \text{otherwise} \end{cases} \\ \overset{w}{P}_m &= \begin{cases} \overset{w}{\boldsymbol{\mu}}{}^{\text{lin}}_{m} & \text{if } m \le D \\ \overset{w}{\boldsymbol{\mu}}{}^{\text{lin}}_{(2D-m)} & \text{otherwise} \end{cases} \end{aligned} \tag{1.21}$$

where $m$ = 1,2,..., $L$ is the index in the full length vector, which can be interpreted as an unfolding operation of the result from Eq. (1.3). The vectors, $\overset{s}{\mathbf{P}}_{jk}$ and $\overset{w}{\mathbf{P}}_{jk}$, are the prototypes for power spectral densities of clean speech and noise respectively.

Given the compensated model $\lambda_x$ the scaled forward variable, $\hat{\alpha}_t(j)$, can be found by employing the scaled forward algorithm [10]. The scaled forward variable yields the probability vector for being in state $j$ for an observation at time $t$. Given the scaled variable and the mixture weights, it is possible to find the probability of being in state $j$ and mixture $k$ at observation time $t$

$$\hat{\vartheta}_t(j,k) = \hat{\alpha}_t(j) \bullet c_{jk} \tag{1.22}$$

where $c_{jk}$, in this case, are the mixture weights for the compensated model.

Given the probability and the spectral prototypes, the most probable, according to the models, clean speech vector and noise vector at observation time $t$ can be found by calculating the average for all states and mixtures

$$\begin{aligned} \overset{s}{\bar{P}}{}^{a}_{t} &= \sum_{j=1}^{N}\sum_{k=1}^{M} \hat{\vartheta}_t(j,k)\, \overset{s}{\bar{P}}{}^{a/2}_{jk} \\ \overset{w}{\bar{P}}{}^{a}_{t} &= \sum_{j=1}^{N}\sum_{k=1}^{M} \hat{\vartheta}_t(j,k)\, \overset{w}{\bar{P}}{}^{a/2}_{jk} \end{aligned} \tag{1.23}$$

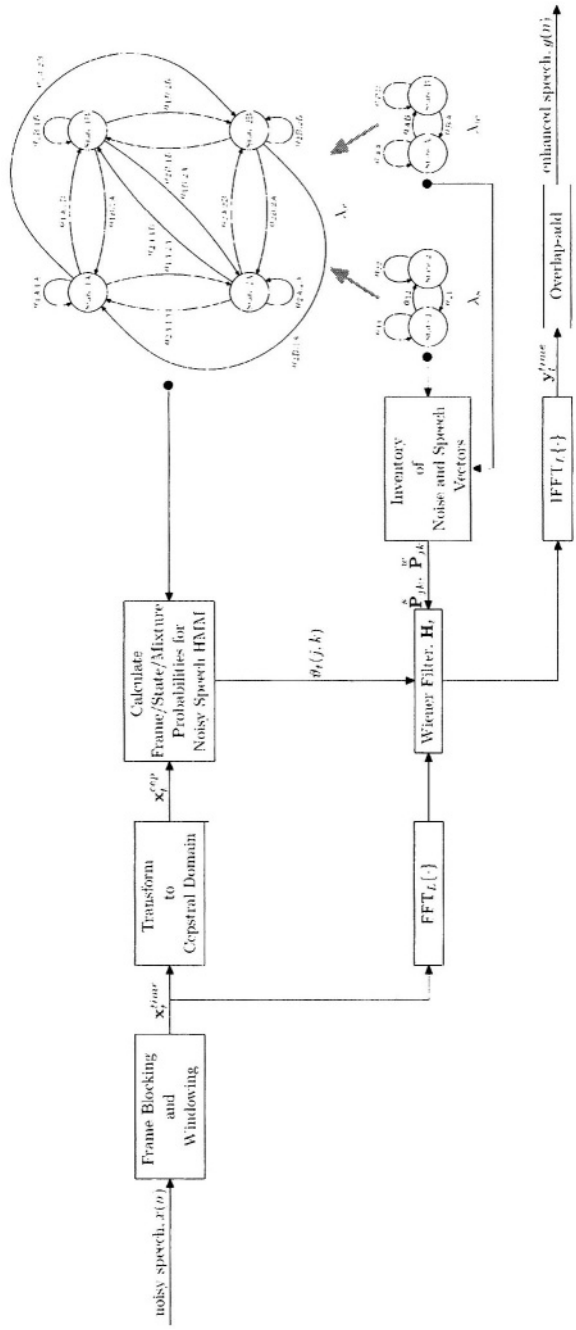



*Figure 1-1.* The enhancement process.

where $a$ determines whether magnitude ($a = 1$) or power spectrum ($a = 2$) is used. Given the most likely clean speech and noise vector, a linear Wiener filter

$$\mathbf{H}_t = \left( \frac{\overline{P}_t^{s\,a}}{\overline{P}_t^{s\,a} + g\,\overline{P}_t^{w\,a}} \right)^{\frac{1}{a}} \tag{1.24}$$

is created.

In order to control the noise reduction a noise reduction limit, $H_{floor}$, can be selected in the interval [0,1]. The floor is applied for the filter vector at every observation time and is defined as

$$H_m = \begin{cases} H_m & \text{if } H_m > H_{floor} \\ H_{floor} & \text{otherwise} \end{cases} \tag{1.25}$$

where $m = 1,2,\ldots,L$ is the index in the full length filter at observation time $t$.

A filter is applied to the $L$-point fast Fourier transform, $FFT_L$, of $\mathbf{x}_t^{\text{time}}$ followed by the filtering and the inverse fast Fourier transform, $IFFT_L$, of the filtered signal.

Given the filtered blocks, the discrete time enhanced speech signal, $y(n)$, is reconstructed using conventional overlap-add [11].

## 6. EXPERIMENTAL RESULTS

In this section the proposed speech enhancer is evaluated on both stationary and non-stationary noise. During the training phase the speech and the noise signals are divided and windowed (Hamming) into 50% overlapping blocks of 64 samples. The ergodic speech model used is trained on all sentences from district one in the TIMIT database [12] (380 sentences from both female and male speakers sampled at the rate 16 kHz). The speech model consists of $N = 5$ states and $M = 5$ mixtures.

The stationary noise is recorded in a car, and is modeled by $N = 1$ state and $M = 1$ mixture.

The non-stationary noise source is a machine gun noise from NOISEX-92 database [13]. This noise is modeled with $N = 3$ states and $M = 2$ mixtures.

Given the trained speech and noisy model the enhancement is performed using $a = 1$, i.e. a filter created in the magnitude domain with the noise reduction limit set to $H_{floor} = -24$ dB. Here the floor is given in decibel scale (dB-scale). The speech signals during enhancement were selected from the testing set of the TIMIT database. Note that the evaluation sentences are not included in the training phase.

The speech enhancement evaluation of stationary noise contaminated speech can be found in Fig. *1-2.*

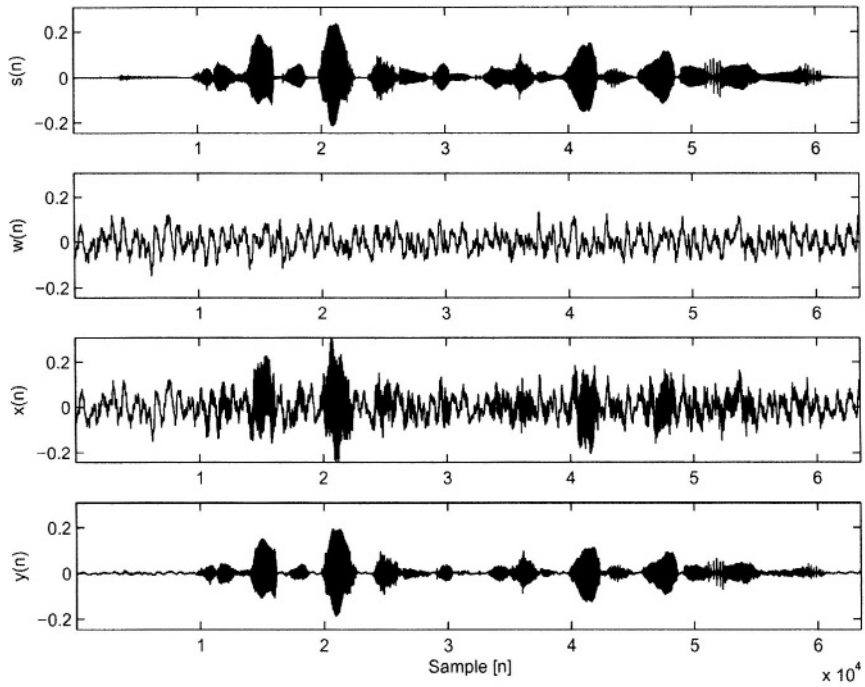



*Figure 1-2.* Clean speech - *s(n),* car noise - *w(n),* noisy speech - *x(n)* and enhanced speech - *y(n)* using proposed HMM method.

In this particular case the signal to noise ratio is improved from -5 dB to 10.8 dB. The signal to noise ratio is calculated for the whole sequence.

The result of reducing such a powerful intermittent noise source, such as machine gun noise, can be found in Fig. *1-3.* In the non-stationary noise source case, the signal to noise ratio is calculated for the whole sequence. The calculated SNR before and after was -5 dB and 9.3 dB, respectively.

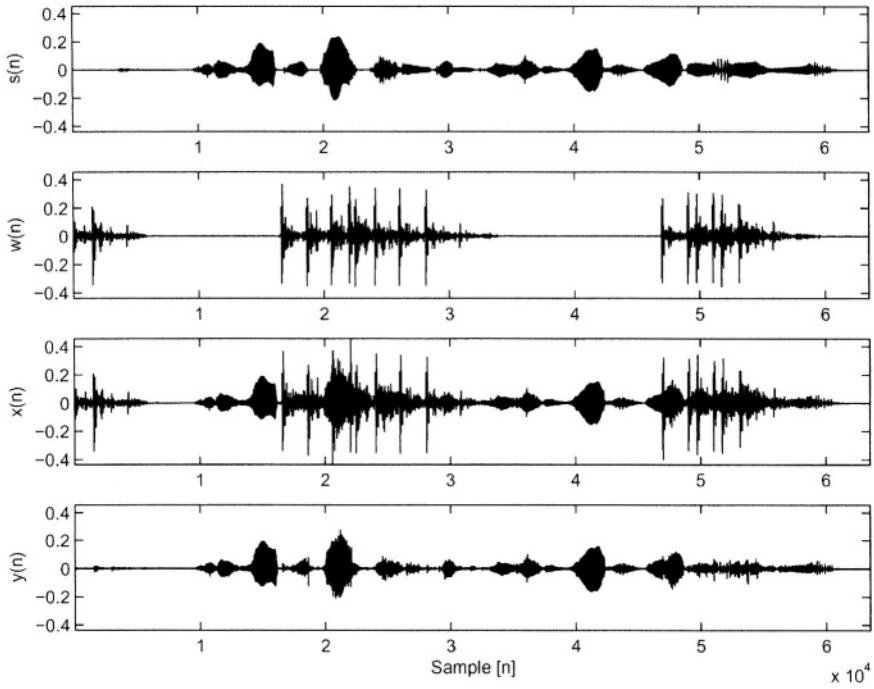



*Figure 1-3.* Clean speech - s(n), machine gun noise - w(n), noisy speech - x(n) and enhanced speech - y(n) using proposed HMM method.

## 7. CONCLUSIONS

This paper presents a cepstral-domain HMM-based speech enhancement method. The method is based on a-priori information gathered from both the speech and the noise source. Given the a-priori information, which is collected in ergodic HMMs, a state dependent Wiener filter is created at every observation. Parameters for the Wiener filter can be chosen to control the filtering process. The proposed speech enhancement method is able to reduce non-stationary noise sources. In enhancement problems, where speech is degraded by an impulsive noise source, such as a machine gun noise, the proposed method is found to substantially reduce the influence of the noise.

# REFERENCES


1. S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-27(2), pp. 113-120, April 1979.
2. Deller John R. Jr., Hansen John J. L., and Proakis John G., *Discrete-time processing of speech signals* (IEEE Press, 1993, ISBN 0-7803-5386-2)
3. C. Jutten and J. Heuralt, Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture, *Signal Processing,* vol. 24, pp. 1-10, June 1991.
4. Y. Ephraim, D. Malah, and B. H. Juang, On the application of hidden markov models for enhancing noisy speech, *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, no. 12, pp. 1846-1856, December 1989.
5. H. K. Kim and R. C. Rose, Cepstrum-domain model combination based on decomposition of speech and noise for noisy speech recognition, *in Proceedings of ICASSP,* May 2002, pp. 209-212
6. S. J. Young and M. J. F. Gales, Cepstral parameter compensation for hmm recognition in noise for noisy speech recognition, *Speech Communication,* vol. 12, no. 3, pp. 231-239, July 1993.
7. C. W. Seymour and M. Niranjan, An hmm-based cepstral-domain speech enhancement system, *in Proceedings ICSLP,* pp. 1595-1598, 1994.
8. Y. Ephraim and M. Rahim, On second order statistics and linear estimation of cepstral coefficients, *IEEE Transactions on Speech and Audio Processing,* vol. 7, no. 2, pp. 162-176, March 1999.
9. A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering; a review, *ACM Computing Surveys,* vol. 31, no 3, pp. 264-323, 1999.
10. L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, *in Proceedings of the IEEE,* vol. 77, no. 2, pp. 164-171, February 1989.
11. Proakis John G. and Manolakis Dimitris G., *Digital Signal Processing* (Prentice-Hall, 1996, ISBN 0-13-394289-9)
12. W. Fisher, G Doddington, and K. Goudie-Marshall, The DARPA speech recognition research database: specifications and status, *in Proceedings DARPA Speech Recognition Workshop,* pp. 93-99, February 1986.
13. A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, The NOISEX-92 study on the effect of additive noise on automatic speech recognition, Tech. Rep., DRA Speech Research Unit, 1992.

*This page intentionally left blank*

Chapter 2

# TIME DOMAIN BLIND SEPARATION OF NONSTATIONARY CONVOLUTIVELY MIXED SIGNALS

Iain T. Russell,[1] Jiangtao Xi,[2] and Alfred Mertins[3]

[1] *Telecommunications and Information Technology Research Institute, University of Wollongong, Wollongong, N.S.W 2522, Australia.*

[2] *School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, N.S.W 2522, Australia.*

[3] *Signal Processing Group, Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany.*

**Abstract** We propose a new algorithm for solving the Blind Signal Separation (BSS) problem for convolutive mixing completely in the time domain. The closed form expressions used for first and second order optimization techniques derived in [1] for the instantaneous BSS case are extended to accommodate the more practical convolutive mixing scenario. Traditionally convolutive BSS problems are solved in the frequency domain [2–4] but this requires additional solving of the inherent frequency permutation problem. Where this is good for higher order systems, systems with a low to medium number of variables benefit from not being subject to a transform such as the DFT. We demonstrate the performance of the algorithm using two optimization methods with a convolutive synthetic mixing system and real speech data.

**Key words:** Blind source separation, joint diagonalization, multivariate optimization, MIMO systems, Newton method, nonstationarity, steepest gradient descent

## 1. INTRODUCTION

Blind Signal Separation (BSS) [5, 6] has been a topic which attracted many researchers in recent years. With the advent of more powerful processors and the ability to realize more complex algorithms BSS has found useful applications in the areas of audio processing such as speech recognition, audio interfaces, and hands free telephony in reverberant environments. In view of

the exponential growth of mobile users in the wireless-communications world together with the limited capacity of resources available for data transmission, modern communication systems increasingly require training-less adaptation, to save on bandwidth capacity or to accommodate unpredictable channel changes. Future systems must utilize spatial diversity multiple access techniques that obtain their channel information exclusively from the received signal. These systems fit the instantaneous and convolutive BSS models. Blind algorithms are useful here as they can be self-recovering and do not require *a priori* knowledge of any training sequence [7]. For example communication systems such as GSM can devote up to 22% of their transmission time to pilot tones which could be otherwise used for data transmission [8]. BSS has also found a fruitful application in multimedia modelling, and recent work on modelling combined text/image data for the purpose of cross-media retrieval has been made using ICA [9].

There is an abundance of various methods used to solve BSS problems and these are often application dependent, however; this paper investigates an algorithm which demonstrates the convolutive mixing model which is relevant to the applications mentioned above and provides a method that avoids the frequency domain permutation problem. The most prevalent of the aforementioned applications suitable for this particular BSS criterion is in the area of speech processing as it exploits the nonstationarity assumption of the algorithm.

We extend approaches in [1] to the convolutive mixing cases. Section 2 gives a brief description of modelling BSS in a convolutive mixing environment. In Section 3 the approaches in [1] are briefly reviewed. The extended approach to convolutive mixing cases is given in Section 4. Section 5 presents the simulation results giving the performance of two optimization methods: Gradient, and Newton optimization with speech data. Finally, a conclusion is provided in Section 6.

The following notations are used in this chapter. We use bold upper and lowercase letters to show matrices and vectors, respectively in the time, frequency and $z$ domains, e.g., $\mathbf{A}(t), \mathbf{A}(\omega), \mathbf{A}(z)$ for matrices and $\mathbf{a}(t)$ for vectors. Matrix and vector transpose, complex conjugation, and Hermitian transpose are denoted by $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H \triangleq ((\cdot)^*)^T$, respectively. $E\{\cdot\}$ means the expectation operation. $\|\cdot\|_F$ is the Frobenius norm of a matrix. $\otimes$ is the Kronecker product and *Trace*($\mathbf{A}$) is the trace of matrix $\mathbf{A}$. With $\mathbf{a} = diag(\mathbf{A})$ we obtain a vector whose elements are the diagonal elements of $\mathbf{A}$ and *diag*($\mathbf{a}$) is a square diagonal matrix which contains the elements of $\mathbf{a}$. *ddiag*($\mathbf{A}$) is a diagonal matrix where its diagonal elements are the same as the diagonal elements of $\mathbf{A}$ and

$$off(\mathbf{A}) \triangleq \mathbf{A} - ddiag(\mathbf{A}). \tag{2.1}$$

$\mathbf{1}_{N\times N}$ is an $N \times N$ matrix of ones, $\mathbf{0}_{N\times N}$ is an $N \times N$ matrix of zeros, and $\mathbf{I}_N$ is the $N \times N$ identity matrix. $vec(\mathbf{A})$ forms a column vector by stacking the columns of $\mathbf{A}$. The operator $mat_{N,MQ}(\mathbf{a})$ reshapes a vector a of length $NMQ$ to an $N \times MQ$ matrix. The matrices $\mathbf{P}_{off}$, $\mathbf{P}_{diag}$, and $\mathbf{P}_{vec}^{(N,L)}$ in Table 2-1 are mainly defined in accordance with [1]. $\mathbf{P}_{off}$ and $\mathbf{P}_{diag}$ are given by

$$\mathbf{P}_{off} = diag(vec(off(\mathbf{1}_{N\times N}))), \tag{2.2}$$

$$\mathbf{P}_{diag} = diag(vec(\mathbf{I}_N)). \tag{2.3}$$

The matrix $\mathbf{P}_{vec}^{(N,L)}$ is the permutation matrix defined by

$$\mathbf{P}_{vec}^{(N,L)} vec(\mathbf{A}^T) = vec(\mathbf{A}), \tag{2.4}$$

for $N \times L$ *matrices* $\mathbf{A}$. Note that for $N \neq L$ the matrix $\mathbf{P}_{vec}^{(N,L)}$ is, in general, not self-inverse like the one that occurs in [1].

## 2. BSS MODEL

The main issue of BSS is that neither the signal sources nor the mixing system are known *a priori.* The only assumption made is that the unknown signal sources are statistically independent. Assume there are $N$ statistically independent sources, $\mathbf{s}(t) = [s_1(t), ..., s_N(t)]^T$. These sources are mixed in a medium providing $M$ sensor or observed signals, $\mathbf{x}(t) = [x_1(t), ..., x_M(t)]^T$, given by

$$\mathbf{x}(t) = \mathbf{H}(t) * \mathbf{s}(t), \tag{2.5}$$

where $\mathbf{H}(t)$ is a $M \times N$ mixing matrix with its element $\mathbf{h}_{ij}(t)$ being the impulse response from $j$th source signal to $i$th measurement. $*$ defines the convolution of corresponding elements of $\mathbf{H}(t)$ and $\mathbf{s}(t)$ following the same rules for matrix multiplication.

Assuming that the mixing channels can be modelled as FIR filters with length $P$, Equation (2.5) can be rewritten as

$$\mathbf{x}(t) = \sum_{\tau=0}^{P-1} \mathbf{H}(\tau)\mathbf{s}(t-\tau). \tag{2.6}$$

The $M$ observed signals $\mathbf{x}(t)$ are coupled to the $N$ reconstructed signals $\hat{\mathbf{s}}(t)$ via the demixing system. The demixing system has a similar structure to the mixing system. It contains $N \times M$ FIR filters of length $Q$, where $Q \geq P$. The de-mixing system can also be expressed as an $N \times M$ matrix $\mathbf{W}(t)$, with its element $\mathbf{w}_{ij}(t)$ being the impulse response from $j$th measurement to $i$th output. The reconstructed signal can be obtained as

$$\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{Q-1} \mathbf{W}(\tau)\mathbf{x}(t-\tau), \tag{2.7}$$

where $\hat{\mathbf{s}}(t) = [\hat{s}_1(t), ..., \hat{s}_N(t)]^T$. A straight forward approach for BSS is to identify the unknown system first and then to apply the inverse of the identified system to the measurement signals in order to restore the signal sources. This approach can lead to problems of instability. Therefore it is desired that the demixing system be estimated based on the observations of mixed signals.

The simplest case is the instantaneous mixing in which matrix $\mathbf{H}(t) = \mathbf{H}$ is a constant matrix with all elements being scalar values. In practical applications such as hands free telephony or mobile communications where multipath propagation is evident, mixing is convolutive, in which situation BSS is much more difficult due to the added complexity of the mixing system. The frequency domain approaches are considered to be effective to separate signal sources in convolutive cases, but another difficult issue, the inherent permutation and scaling ambiguity in each individual frequency bin, arises which makes the perfect reconstruction of signal sources almost impossible [10]. Therefore it is worthwhile to develop an effective approach in the time domain for convolutive mixing systems that don't have an exceptionally large amount of variables. Joho and Rahbar [1] proposed a BSS approach based on joint diagonalization of the output signal correlation matrix using gradient and Newton optimization methods. However the approaches in [1] are limited to the instantaneous mixing cases whilst in the time domain.

## 3. OPTIMIZATION OF INSTANTANEOUS BSS

This section gives a brief review of the algorithms proposed in [1]. Assuming that the sources are statistically independent and non-stationary, observing the signals over $K$ different time slots, we define the following noise free instantaneous BSS problem. In the instantaneous mixing cases both the mixing and demixing matrices are constant, that is, $\mathbf{H}(t) = \mathbf{H}$ and $\mathbf{W}(t) = \mathbf{W}$. In this case the reconstructed signal vector can be expressed as

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t). \tag{2.8}$$

The instantaneous correlation matrix of $\hat{\mathbf{s}}(t)$ at time frame $k$ can be obtained as

$$\mathbf{R}_{\hat{s}\hat{s},k} = \mathbf{W}\mathbf{R}_{xx,k}\mathbf{W}^T \tag{2.9}$$

$$\mathbf{R}_{xx,k} = E\{\mathbf{x}(k)\mathbf{x}^T(k)\}. \tag{2.10}$$

For a given set of $K$ observed correlation matrices, $\{\mathbf{R}_{xx,k}\}_{k=1}^{K}$, the aim is to find a matrix $\mathbf{W}$ that minimizes the following cost function

$$\mathcal{J}_1 \triangleq \sum_{k=1}^{K} \beta_k \| off(\mathbf{W}\mathbf{R}_{xx,k}\mathbf{W}^T) \|_F^2, \tag{2.11}$$

where $\{\beta_k\}$ are positive weighting *normalization* factors such that the cost function is independent of the absolute norms and are given as

$$\beta_k = \left(\sum_{i=1}^{K} \|\mathbf{R}_{xx,i}\|_F^2\right)^{-1}. \tag{2.12}$$

Perfect joint diagonalization is possible under the condition that $\{\mathbf{R}_{xx,k}\} = \{\mathbf{H}\mathbf{\Lambda}_{ss,k}\mathbf{H}^T\}$ where $\{\mathbf{\Lambda}_{ss,k}\}$ are diagonal matrices due to the assumption of the mutually independent unknown sources. This means that full diagonalization is possible, and when this is achieved, the cost function is zero at its global minimum. This constrained non-linear multivariate optimization problem can be solved using various techniques including gradient-based steepest descent and Newton optimization routines. However, the performance of these two techniques depends on the initial guess of the global minimum, which in turn relies heavily on an initialization of the unknown system that is near the global trough. If this is not the case then the solution may be sub-optimal as the algorithm gets trapped in one of the local multi-minima points.

To prevent a trivial solution where $\mathbf{W} = \mathbf{0}$ would minimize Equation (2.11), some constraints need to be placed on the unknown system $\mathbf{W}$. One possible constraint is that $\mathbf{W}$ is unitary. This can be implemented as a penalty term such as given below

$$\mathcal{J}_2 \triangleq \|\mathbf{W}\mathbf{W}^T - \mathbf{I}\|_F^2, \tag{2.13}$$

or as a hard constraint that is incorporated into the adaptation step in the optimization routine. For problems where the unknown system is constrained to be unitary, Manton presented a routine for computing the Newton step on the manifold of unitary matrices referred to as the *complex Stiefel manifold.* For further information on derivation and implementation of this hard constraint refer to [1] and references therein.

The closed form analytical expressions for first and second order information used for gradient and Hessian expressions in optimization routines are taken from Joho and Rahbar [1] and will be referred to when generating results for convergence. Both the Steepest gradient descent (SGD) and Newton methods are implemented following the same frameworks used by Joho and Rahbar. The primary weakness of these optimization methods is that although they do converge relatively quickly there is no guarantee for convergence to a global minimum which provides the only true solution. This is exceptionally noticeable when judging the audible separation of speech signals. To demonstrate the algorithm we assume a good initial starting point for the unknown separation system to be identified by setting the initial starting point of the unknown system in the region of the global trough of the multivariate objective function.

## 4. OPTIMIZATION OF CONVOLUTIVE BSS IN THE TIME DOMAIN

As mentioned previously and as with most BSS algorithms that assume convolutive mixing, solving many BSS problems in the frequency domain for individual frequency bins can exploit the same algorithm derivation as the instantaneous BSS algorithms in the time domain. However the inherent *frequency permutation problem* remains a major challenge and will always need to be addressed. The tradeoff is that by formulating algorithms in the frequency domain we can perform less computations and processing time falls, but we still must fix the permutations for individual frequency bins so that they are all aligned correctly. This chapter aims to provide a way to utilize the existing algorithm developed for instantaneous BSS and apply it to convolutive mixing but avoid the permutation problem.

Now we extend the above approach to the convolutive case. We still assume that the demixing system is defined by Equation (2.7), which consists of $N \times M$ FIR filters with length Q. We want to get a similar expression to those in the instantaneous cases. It can be shown that Equation (2.7) can be rewritten in the following matrix form

$$\hat{\mathbf{s}}(n) = \mathcal{W}\mathcal{X}(n), \tag{2.14}$$

where $\mathcal{W}$ is a $(N \times QM)$ matrix given by

$$\mathcal{W} = [\mathbf{W}(0), \mathbf{W}(1), ..., \mathbf{W}(Q-1)], \tag{2.15}$$

and $\mathcal{X}(n)$ is a $(QM \times 1)$ vector defined as

$$\mathcal{X}(n) = \begin{bmatrix} \mathbf{x}(n) \\ \mathbf{x}(n-1) \\ \vdots \\ \mathbf{x}(n-(Q-1)) \end{bmatrix}. \tag{2.16}$$

The output correlation matrix at time frame $k$ can be derived as

$$\mathbf{R}_{\hat{s}\hat{s},k}(0) = \mathcal{W}\mathbf{R}_{\mathcal{X}\mathcal{X},k}(0)\mathcal{W}^T, \tag{2.17}$$

where,

$$\mathbf{R}_{\mathcal{X}\mathcal{X},k}(0) = E\{\mathcal{X}(k)\mathcal{X}^T(k)\}. \tag{2.18}$$

Correlation matrices for the recovered sources for all necessary time lags $\tau$ can also be obtained as

$$\begin{aligned} \mathbf{R}_{\hat{s}\hat{s},k}(\tau) &= \mathcal{W}E\{\mathcal{X}(k)\mathcal{X}^T(k+\tau)\}\mathcal{W}^T \\ &= \mathcal{W}\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\mathcal{W}^T. \end{aligned} \tag{2.19}$$

Using the joint-diagonalization criterion in [1] for the instantaneous modelling of the BSS problem we can formulate a similar expression for convolutive mixing in the time domain. Consider the correlation matrices with all different time lags we should have the following cost function

$$\mathcal{J}_3 \triangleq \sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \beta_k \| off(\mathcal{W}\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\mathcal{W}^T)\|_F^2. \tag{2.20}$$

The only difference between $\mathcal{J}_1$ and $\mathcal{J}_3$ is that we now take into account all the different time lags $\tau$ for the correlation matrices for each respective time epoch $k$ where the SOS are changing. Also $\beta_k$ is now defined as

$$\beta_k = ( \sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \|\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)\|_F^2)^{-1}, \tag{2.21}$$

and we note the new structure of $\mathcal{W}$. In the ideal case where we know the exact system $\mathcal{W}_{ideal}$, all off-diagonal elements would equal zero and the value of the objective function would reach its global minimum where $\mathcal{J}_3 = 0$. Each value of $k$ represents a different time window frame where the Second Order Statistics (SOS) are considered stationary over that particular time frame. In adjacent non-overlapping time frames $k$, the SOS are changing due to the nonstationarity assumption. As this is a non-linear constrained optimization problem with $NQM$ unknown parameters we can rewrite it as

$$\begin{aligned} \mathcal{W}_{opt} = & \quad \arg\min_{\mathcal{W}} \ \mathcal{J}_3(\mathcal{W}) \\ s/t \quad & \mathcal{J}_4(\mathcal{W}) = \|ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I})\|_F^2 = 0. \end{aligned} \tag{2.22}$$

Due to the structure of the matrices and with the technique of matrix multiplication to perform convolution in the time domain, optimization algorithms similar to those performed in the instantaneous climate can be utilized. Notice also that in the instantaneous version the constraint used to prevent the trivial solution $\mathbf{W} = \mathbf{0}$ was a unitary one. In the convolutive case a different constraint is used where the row vectors of $\mathcal{W}$ are normalized to have length one. Again referring to the SGD and Newton algorithms closed form analytical expressions of the gradient and Hessian deduced by Joho and Rahbar [1] are extended slightly to accommodate the time domain convolutive climate of the new algorithm. These expressions are shown in Table 2-1. $\mathbf{R}_{\mathcal{X}\mathcal{X},k}(\tau)$ will be denoted as $\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}$. With these expressions the SGD and Newton methods are summarized in the Tables 2-2 and 2.3 respectively. Table 2-2 is relatively easy to interpret as it is a simple iterative update or learning rule with a fixed step size. As an alternative to a constant step-size $\mu$ the natural gradient method

proposed by Amari [11] could be used instead of the absolute gradient although faster convergence can be expected from second-order methods. Table 2-3 gives the general Newton update with penalty terms incorporated to ensure that the Hessian of the constraint, denoted as $\mathbf{H}_4$, and the gradient of the constraint, denoted as $\mathbf{G}_4$, are accounted for in the optimization process. Note the $\mathcal{J}_4$ defines the constraint given in Equation (2.22) and expresses the unit energy of the rows of $\mathcal{W}$.

*Table 2-1.* Closed form analytical expressions for the gradient and Hessian of the cost function and constraints.

| Cost function - $\mathcal{J}_3$ |
|---|
| $\mathcal{J}_3 \triangleq \sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \beta_{k,\tau} \| off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T) \|_F^2$ |
| Gradient - $\mathbf{G}_3$ |
| $\mathbf{G}_3 = 2\sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \beta_{k,\tau} \{ off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T)\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T + off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T)\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \}$ |
| Hessian - $\mathbf{H}_3$ |
| $\mathbf{H}_3 = 2\sum_{\tau=-\tau_{min}}^{\tau_{max}} \sum_{k=1}^{K} \beta_{k,\tau} \{ (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T)) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes off(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T)) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}^{(N,N)}_{vec}\mathbf{P}_{off}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k} \otimes \mathbf{I}_N) + (\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{off}\mathbf{P}^{(N,N)}_{vec}(\mathcal{W}\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}{}^T \otimes \mathbf{I}_N) \}$ |
| Row-normalized Constraint $\mathcal{J}_4$ |
| $\mathcal{J}_4 = \| ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N) \|_F^2$ |
| Constraint Gradient $\mathbf{G}_4$ |
| $\mathbf{G}_4 = 4 ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)\mathcal{W}$ |
| Constraint Hessian $\mathbf{H}_4$ |
| $\mathbf{H}_4 = 4(\mathbf{I}_{MQ} \otimes ddiag(\mathcal{W}\mathcal{W}^T - \mathbf{I}_N)) + 4(\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{diag}(\mathcal{W} \otimes \mathbf{I}_N) + 2\mathbf{P}^{(N,MQ)}_{vec}(\mathbf{I}_N \otimes \mathcal{W}^T)\mathbf{P}_{diag}(\mathcal{W} \otimes \mathbf{I}_N) + 2(\mathcal{W}^T \otimes \mathbf{I}_N)\mathbf{P}_{diag}(\mathbf{I}_N \otimes \mathcal{W})(\mathbf{P}^{(N,MQ)}_{vec})^T$ |

*Table 2-2.* Gradient descent subband BSS algorithm for the joint-diagonalization task with a weighted constraint.

Initialization $(r = 0) : \mathcal{W}_0$

For $r = 1, 2, ...$

$\mathbf{w}_r = \mu\{\text{vec}(\mathbf{G}_3 + \alpha\mathbf{G}_4)\}$
$\triangle\mathcal{W}_r = \text{mat}_{N,MQ}(\mathbf{w}_r)$
$\mathcal{W}_{r+1} = \mathcal{W}_r - \triangle\mathcal{W}_r$

*Table 2-3.* Newton-type subband BSS algorithm for the joint-diagonalization task with a weighted constraint.

Initialization $(r = 0) : \mathcal{W}_0$

For $r = 1, 2, ...$

$\mathbf{w}_r = \mu(\mathbf{H}_3 + \alpha\mathbf{H}_4)^{-1}\text{vec}(\mathbf{G}_3 + \alpha\mathbf{G}_4)$
$\triangle\mathcal{W}_r = \text{mat}_{N,MQ}(\mathbf{w}_r)$
$\mathcal{W}_{r+1} = \mathcal{W}_r - \triangle\mathcal{W}_r$

## 5. SIMULATION RESULTS

To investigate the performance of the extended instantaneous BSS algorithm to the convolutive case in the time domain the SGD and Newton algorithm implementations in [1] were altered to the learning rules given in Tables 2-2 and 2-3 respectively. As the constraint no longer requires the unknown system $\mathcal{W}$ to be unitary the constraint was changed to that given in Equation (2.22). The technique of weighted penalty functions was used to ensure the constraints preventing the trivial solution were met. No longer performing the optimization on the Stiefel manifold as in [1] the SGD and Newton algorithms were changed to better reflect the row normalization constraint for the convolutive case. Using the causal $z$-transform

$$\mathbf{H}_{ij}(z) = \sum_{n=0}^{\infty} \mathbf{h}_{ij}(n) z^{-n}, \quad \forall i, \forall j, \tag{2.23}$$

a first-order two-input-two-output (TITO) two tap FIR known mixing system was chosen and is given below in the $z$ domain as

$$\mathbf{H}(z) = \begin{bmatrix} 1+z^{-1} & -1+z^{-1} \\ -1+z^{-1} & 1+z^{-1} \end{bmatrix}. \tag{2.24}$$

The corresponding known un-mixing system which would separate mixed signals which are produced by convolving the source signals with the TITO mixing system $\mathbf{H}(z)$ given above is

$$\mathbf{W}_{ideal}(z) = \frac{1}{4}\begin{bmatrix} 1+z^{-1} & 1-z^{-1} \\ 1-z^{-1} & 1+z^{-1} \end{bmatrix}. \tag{2.25}$$

This is the exact known inverse multiple-input-multiple-output (MIMO) FIR system of the same order. The convolution of these two systems in cascade would ensure the global system $\mathbf{G}(z) = \mathbf{W}_{ideal}(z)\mathbf{H}(z)$ would be a delayed version of the identity, i.e. $z^{-1}\mathbf{I}$. Using matrix multiplication to perform convolution in the time domain, Equation (2.15) can be used to represent the equivalent structure of Equation (2.24),

$$\frac{1}{4}\mathcal{W}_{ideal} = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix}. \tag{2.26}$$

Through empirical analysis we set the parameters $\mu = 0.6$ and $\alpha = 0.2$ and solve the constrained optimization problem given in Equation (2.22) using the SGD and Newton methods. A set of $K = 15$ real diagonal square uncorrelated matrices for the unknown source input signals were randomly generated. Using convolution in the time domain a corresponding set of correlation matrices $\mathbf{R}^{\tau}_{\mathcal{X}\mathcal{X},k}$ for each respective time instant $k = 1, ..., 15$ at multiple time lags $\tau$ were generated for the observed signals. Each optimization algorithm was run ten independent times and convergence graphs were observed and are shown in Figure 2-1. The various slopes of the different convergence curves of the gradient method depends entirely on the ten different sets of randomly generated diagonal input matrices. Poor initial values for the unknown system lead to convergence to local minima as opposed to the desired global minimum. The initialization of the SGD and Newton algorithms plays an important role in the convergence to either a local or global minimum. Initial values for the estimated unmixing system $\mathcal{W}$ were generated using a perturbed version of the true unmixing system. This was done by adding Gaussian random variables with standard deviation $\sigma = 0.1$ to the coefficients of the true system. As a possible alternative strategy, a global optimization routine *glcCluster* from TOMLAB [12], a robust global optimization software package, can be used where no initial value for the unknown system is needed. This particular solver uses a global search to approximately obtain the set of all global solutions and then

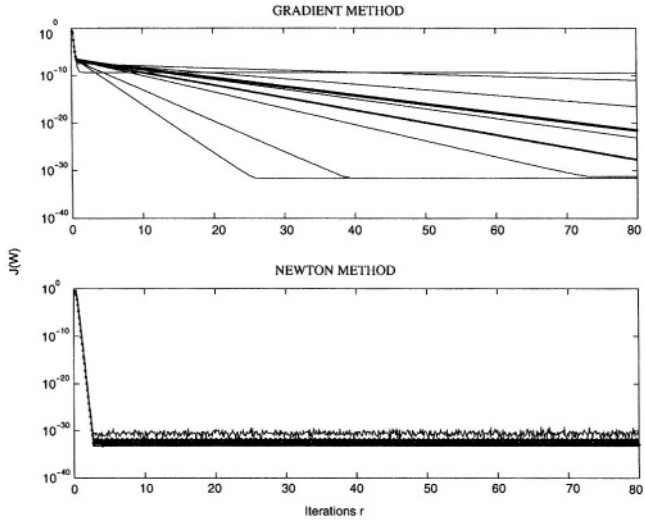



*Figure 2-1.* Convergence of gradient descent and Newton algorithms for a first order TITO FIR demixing system over 10 trials.

uses a local search method which utilizes the derivative expressions to obtain more accuracy on each global solution. This method will be further analyzed as a future alternative to obtaining additional information on the initial system value.

After convergence of the objective function to an order of magnitude approximately equal to $10^{-34}$ the unknown demixing FIR filter system $\mathcal{W}$, in cascade with the known mixing system $\mathbf{H}(z)$, resulted in a global system which was equivalent to a scaled and permuted version of the true global system $z^{-1}\mathbf{I}$ as can be seen by the following example,

$$\mathbf{G}(0) = \begin{bmatrix} -0.17 & 0.17 \\ 0.19 & -0.19 \end{bmatrix} \times 10^{-14},$$

$$\mathbf{G}(1) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \tag{2.27}$$

$$\mathbf{G}(2) = \begin{bmatrix} -0.23 & -0.23 \\ -0.14 & -0.14 \end{bmatrix} \times 10^{-14}.$$

A first order system has been identified up to an arbitrary global permutation and scaling factor. The TITO system identified above using the optimization algorithms has only 8 unknown variables to identify. We now examine a

MIMO FIR mixing system with a higher dimension. Again we have chosen an analytical MIMO multivariate system whose exact FIR inverse is known. The 3rd order mixing system is given below in the $z$ domain

$$\mathbf{H}_{11}(z) = -4 - 4z^{-1} + z^{-2} + z^{-3}, \tag{2.28}$$

$$\mathbf{H}_{12}(z) = -7 - 7z^{-1} + z^{-3}, \tag{2.29}$$

$$\mathbf{H}_{21}(z) = 7 - 7z^{-1} + z^{-3}, \tag{2.30}$$

$$\mathbf{H}_{22}(z) = 9 - 9z^{-1} - z^{-2} + z^{-3}. \tag{2.31}$$

The corresponding known inverse FIR system of the same order is given below also in the $z$ domain as

$$\mathbf{W}_{11}^{ideal}(z) = \frac{1}{13}\mathbf{H}_{22}(z), \tag{2.32}$$

$$\mathbf{W}_{12}^{ideal}(z) = -\frac{1}{13}\mathbf{H}_{12}(z), \tag{2.33}$$

$$\mathbf{W}_{21}^{ideal}(z) = -\frac{1}{13}\mathbf{H}_{21}(z), \tag{2.34}$$

$$\mathbf{W}_{22}^{ideal}(z) = \frac{1}{13}\mathbf{H}_{11}(z). \tag{2.35}$$

The convolution of the mixing and unmixing MIMO FIR systems given in Equations (2.28-2.35) gives the identity matrix $\mathbf{I}$ exactly. A comparison of the convergence behaviour for the more efficient Newton method is given in Figure 2-2 using the same methods described for the first order systems above, keeping the learning factor and weighting terms the same. We see from the figure that with twice as many unknown variables to solve for the demixing system, the third order unknown system takes longer to converge by roughly a factor of two. Both systems converge to their global minimums due to good initialization at approximately $10^{-34}$. For the third order system, one trial produced an outlying convergence curve that takes more iterations $r$ than the other trials. This is dependent on the randomly generated set of diagonal correlation matrices $\{\mathbf{R}_{ss,k}\}$ where $k = 1, 2, ..., 15$ for each trial.

To test the performance of the algorithm on real speech data two independent segments of speech were used as input signals to the MIMO FIR mixing system given in Equation (2.24). These signals were both 4 seconds long and sampled at 8kHz. The signals were convolutively mixed with the synthetic mixing system to obtain 2 mixed signals. With the assumption that speech is quasi-stationary over a period of approximately 20ms, the observed mixed signals were buffered and segmented into 401 frames each having 160 samples in length. The nonstationarity assumption assumes that the SOS in each frame does not change. The correlation matrices $\mathbf{R}_{\mathcal{X}\mathcal{X},k}^{\tau}$ can be found via Equations

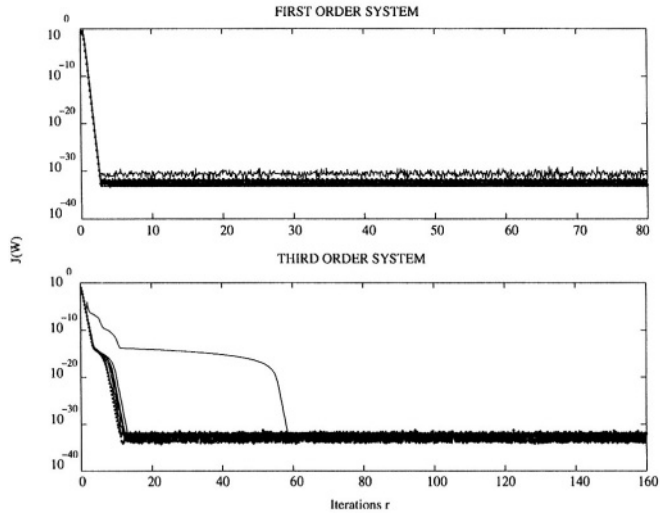



*Figure 2-2.* Convergence of Newton algorithms for first and third order TITO FIR demixing systems over 10 trials.

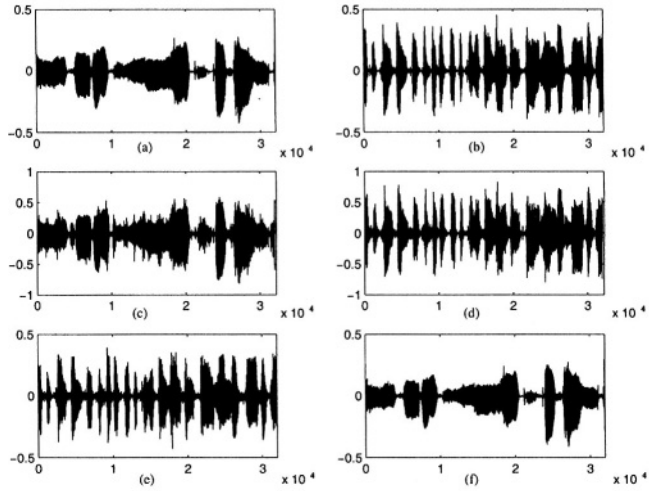



*Figure 2-3.* (a) and (b) are the two original signals, (c) and (d) are the convolutively mixed signals, (e) and (f) are the permuted separated results.

(2.18,2.19) for $K = 401$ frames of the two mixed signals. This allows the method of joint diagonalization by minimizing the off-diagonal elements of the correlation matrices of the recovered signals at each respective time lag $\tau$

as defined in Equations (2.20,2.22). Figure 2-3 shows the input, mixed and recovered speech signals. A good qualitative recovery is confirmed by subjective listening to the recovered audio signals and inspection of graphs (e) and (f) in Figure 2-3.

# 6. CONCLUSION

A new method for convolutive BSS in the time domain using an existing instantaneous BSS framework has been presented. This method avoids the inherent permutation problem when dealing with solving the convolutive BSS problem in the frequency domain. Optimization algorithms including SGD and Newton methods have been compared for convolutive mixing environments. Future work will be directed at implementing the simulations with recorded data such as speech in real reverberant environments where the orders of the mixing and unmixing MIMO FIR systems are very high.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and suggestions. Iain also wishes to thank the support of his mother and father, Diana and Barry Russell, as well as the patience of his partner Sarah.

# REFERENCES

1. M. Joho and K. Rahbar, "Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation," in *Proc. Sensor Array and Multichannel Signal Processing Workshop (SAM),* Rosslyn, VA, USA, Aug. 2002, pp. 403–407.

2. K. Rahbar and J. Reilly, "A New Frequency Domain Method for Blind Source Separation of Convolutive Audio Mixtures," *Submitted to IEEE Trans. on Speech and Audio Processing,* January 2003.

3. S. Ikeda and N. Murata, "A method of ICA in Time-Frequency Domain," in *Proc. ICA,* Aussois, January 1999, pp. 365–361.

4. N. Murata, "An Approach to Blind Source Separation of Speech Signals," *Proceedings of the 8th International Conference on Artificial Neural Networks,* vol. 2, pp. 761–766, September 1998.

5. K. J. Pope and R. E. Bogner, "Blind signal separation I: Linear, instantaneous combinations," *Digital Signal Processing,* vol. 6, no. 1, pp. 5–16, Jan. 1996.

6. K. J. Pope and R. E. Bogner, "Blind signal separation II: Linear, convolutive combinations," *Digital Signal Processing,* vol. 6, no. 1, pp. 17–28, Jan. 1996.

7. M. Feng and K.-D. Kammeyer, "Blind source separation for communication signals using antenna arrays," in *Proc. ICUPC-98,* Florence, Italy, Oct. 1998.

8. T. Petermann, D. Boss, and K. D. Kammeyer, "Blind GSM Channel Estimation Under Channel Coding Conditions," Phoenix, USA, December 1999, pp. 180–185.

9. J. Larsen, L. Hansen, T. Kolenda, and F. Nielsen, "Independent Component Analysis in Multimedia Modelling," in *Proc. ICA,* Nara, Japan, April 2003, pp. 687–696.

10. M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in Blind Separation of Speech Signals in a Reverberant Environment," in *Proc. ICASSP,* Instanbul, Turkey, June 2000, pp. 1041–1044.

11. S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," Paris, France, April 1997, pp. 101–104, Proceedings First IEEE Workshop on Signal Processing Advances in Wireless Communications.

12. Kenneth Holmström, "User's Guide for TOMLAB v4.0," URL: http://tomlab.biz/docs/tomlabv4.pdf, Sept. 2 2002.

*This page intentionally left blank*

Chapter 3

# SPEECH AND AUDIO CODING USING TEMPORAL MASKING

Teddy Surya Gunawan, Eliathamby Ambikairajah, and Deep Sen
*School of Electrical Engineering and Telecommunications, The University of New South Wales, UNSW Sydney 2052, Australia*

**Abstract:** This paper presents a comparison of three auditory temporal masking models for speech and audio coding applications. The first model was developed based upon the existing forward masking psychoacoustic data with an assumption of an approximately 200 ms. The model's dynamic parameters were derived from this data. The previously developed second model was based upon the principle of an exponential decay following higher energy stimuli, where the masking effects have a relatively short duration. The existing third model best matches the previously reported forward masking data using an exponential curve but the effects of the forward masking are restricted to 100-200ms. Objective assessments employing the PESQ measure reveal that these three temporal models have potential for removing perceptually redundant information in speech and audio coding applications. Results show that the incorporation of temporal masking along with simultaneous masking into a speech/audio coding algorithm results in a further bit rate reduction of approximately 17% compared with simultaneous masking alone, while preserving perceptual quality

**Key words:** Temporal masking model, Simultaneous masking model, Gammatone filters, Wavelet Packet, PESQ, Subjective listening test

## 1. INTRODUCTION

The use of auditory models in speech and audio coding is by no means new, and their applications include low bit rate speech coding [1] through to MPEG audio compression [2]. Conventional audio coding algorithms do not exploit knowledge of the temporal properties of the human auditory system, relying solely on simultaneous masking models. Simultaneous masking is a frequency domain phenomenon in which a low-level signal can be rendered

inaudible by a simultaneously occurring stronger signal if both signals are sufficiently close in frequency.

Temporal masking is a time domain phenomenon in which two stimuli occur within a small interval of time [3] This time domain phenomenon plays an important role in human auditory perception. Post-masking occurs when a masker precedes the signal in time, while pre-masking occurs when the signal precedes the masker in time. Post-masking is the more important effect from a coding perspective since the duration of the masking effect can be much longer, depending on the duration of the masker.

In this work, we have developed a temporal masking model and compared its performance using two existing temporal masking models. The first model developed is based on [4, 5], the second model is based upon [6], and the third model is based on [7]. The developed temporal masking model combined with the simultaneous masking model [8] is then used to calculate the combined masking thresholds in the time-frequency domain. These models were first incorporated into a critical band based gammatone auditory filter bank analysis/synthesis system [6] in order to validate the effectiveness of the model. The models were also included in a wavelet packet based audio coding algorithm [9] to quantify the improvement for coding purposes. Results show that the incorporation of temporal masking along with simultaneous masking into a speech/audio coding algorithm results in a further reduction of bit rate of approximately 17% while preserving perceptual quality.

The transparent quality is evaluated using PESQ (Perceptual Evaluation of Speech Quality) measure [10]. PESQ was recently adopted as an ITU-T recommendation P.862. PESQ is able to predict subjective quality with good correlation in a very wide range of conditions, includes coding distortions, errors, noise, filtering, delay and variable delay. Also subjective experiments using informal listening tests were carried out in order to assess the quality of the coded speech and audio signals.

The paper is organized as follows. Section 2 describes the filter bank analysis for speech and audio coding applications. The temporal masking models used in this research are explained in section 3. Masking model performance is evaluated in section 4, while section 5 concludes this paper.

# 2. FILTER BANK ANALYSIS

## 2.1 Gammatone Analysis/Synthesis Filter Bank

Fig. 3-1 shows the gammatone front end processing for speech coding applications that is applicable to audio coding as the number of filter bank

can be increased accordingly. Gammatone filters are implemented using FIR filters to achieve linear phase filters with identical delay in each critical band. To achieve linear phase filters, the synthesis filters $g_m(n)$ is the time reverse of its analysis filters $h_m(n)$. The analysis filter for each subband $m$ is obtained using the following expression:

$$h_m(n) = a(nT)^{N-1} e^{-2\pi b ERB(fc_m)nT} \cos(2\pi fc_m nT + \varphi) \tag{3.1}$$

where $fc_m$ is the center frequency for each subband $m$, $T$ is the sampling period, and $N$ is the gammatone filter order ($N$ = 4). For a sampling period of 8000 Hz, the total number of subband is $M$ = 17, so $m$ = 1...17. The parameter $n$ is the discrete time sample index and $n = 0 \ldots Nf_m$ where $Nf_m$ is the length of each filter for each subband. $ERB(fc_m)$ is the equivalent rectangular bandwidth of an auditory filter. At a moderate power level, $ERB(fc_m) = 24.7 + 0.108\, fc_m$. The parameter $b$ is set to 1.14 while the parameter $a$ is set for each subband to normalize the filter gain to 0 dB.

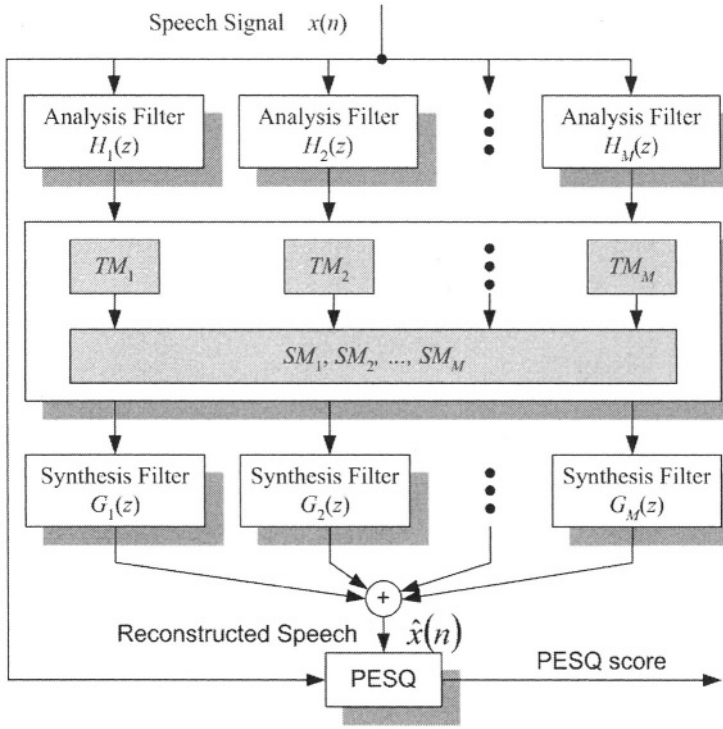



*Figure 3-1.* Gammatone analysis and synthesis filter bank

The analysis filter bank output is followed by a half-wave rectifier to simulate the behavior of the inner hair cell. Moreover, the nature of the neuron firing allows a simple peak-picking implementation. This process results in a series of critical band pulse trains, where the pulses retain the amplitudes of the critical band signals from which they were derived The masking operation is then applied to the pulses, in order to remove the perceptually irrelevant peaks.

## 2.2 The PESQ measure

Formal listening tests must meet several conditions especially the characteristics of listening rooms (ITU-R BS.1116) that require special equipment. Therefore, in this research work we use an informal listening test confirmed by the PESQ measurement system [10] as the tools for evaluating the speech quality.

PESQ has recently been approved as ITU-T recommendation P.862 in February 2001 as a tool for assessing speech quality. The input to the PESQ software tool is the reference speech signal and the processed speech signal. The PESQ then rates the speech quality between 1.0 (bad) and 4.5 (no distortion). However, the informal listening tests revealed that the PESQ score of 3.5 provides transparent speech quality.

## 2.3 Optimum Number of Filter Coefficients

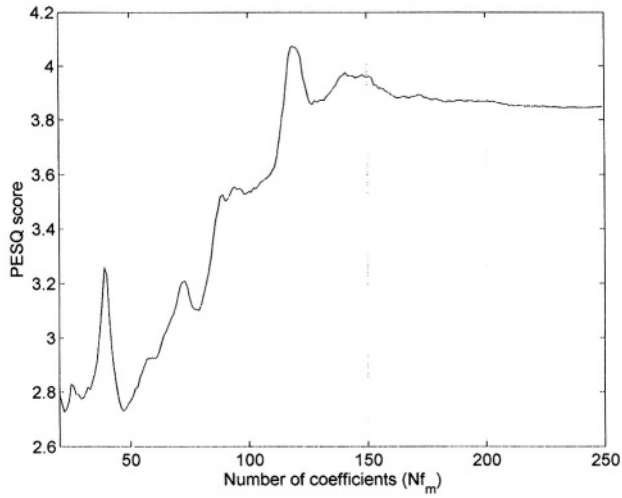



*Figure 3-2.* Variation of PESQ score with number of filter coefficients ( $Nf_m$ )

The number of coefficients required to implement the analysis/synthesis filter bank depends on the impulse response of the gammatone filters. The low frequency filters need more coefficients as compared to the high frequency filters. It is possible to estimate the optimum number of coefficients required for the analysis/synthesis filter bank with peak picking operation, using the PESQ software tool. It is assumed that filters with a constant delay across all bands are required in the analysis stage for time-aligning critical band pulses across different bands. Fig. 3-2 shows PESQ measure against the number of filter coefficients and it can be seen that $Nf_m = 120$ (corresponds to 15 ms delay in 8 kHz sampling rate) provides maximum PESQ score. Hence, the optimum value is used in this paper.

# 3. TEMPORAL MASKING MODELS

In this section, we present the equations required to implement our temporal forward masking model along with the other existing temporal forward masking models. Backward masking models are not considered here, as their effects in coding applications are somewhat limited.

## 3.1 Model 1 (TM1)

Jesteadt et al. [4] describe temporal masking as a function of frequency, masker level, and signal delay. Based on the forward masking experiments carried out by [4], the amount of temporal masking can be well-fitted to psychoacoustic data using the following equation:

$$M_f(t,m) = a(b - \log_{10}\Delta t)(L(t,m) - c) \qquad (3.2)$$

where $M_f(t,m)$ is the amount of forward masking (dB) in the $m$th band, $\Delta t$ is the time difference between the masker and the maskee in milliseconds, $L(t,m)$ is the masker level (dB), and $a$, $b$, and $c$, are parameters that can be derived from psychoacoustic data.

Najafzadeh et al. [5] incorporated the temporal masking model above to the MPEG psychoacoustic model in which they achieve a significant coding gain. Nevertheless, the temporal masking model 1 proposed is based on [4, 5] with several modifications of the parameters $a$, $b$, and $c$.

The parameter $a$ is based upon the slope of the time course of masking, for a given masker level. We have approximated $a$ by curve-fitting of the psychoacoustic data in [4], as follows:

$$a = k_2 fc_m^{\ 2} + k_1 fc_m + k_0 \tag{3.3}$$

where $fc_m$ is the center frequency of the critical band $m$, and $k_0$, $k_1$, $k_2$ have values 0.5806, -0.0357 and 0.0013, respectively.

Assuming that forward temporal masking has duration of 200 milliseconds, and thus $b$ may be chosen as $\log_{10}(200)$ [5]. Similarly to the calculation of parameter $a$, the parameter $c$ is chosen by fitting a curve to the masker level data provided in [4]:

$$c = p_2 fc_m^{\ 2} + p_1 fc_m + p_0 \tag{3.4}$$

where $p_0$, $p_1$,$p_2$ values are 6.6727, 2.979 and -0.11226 respectively. The final value of $c$ is obtained by adding the threshold of hearing [11]. This means that any signal components below the threshold of hearing will automatically be masked.

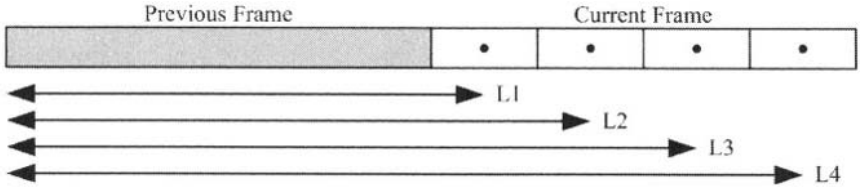



*Figure 3-3.* Efficient masking threshold calculation

The calculation of the masker level can be performed over many frames to accumulate a reliable estimate. However the number of frames may depend on the coding application. In this instance, where experiments on speech were performed, we have developed an efficient method for masking calculation as shown in Fig. 3-3.

The current frame of 16 ms was subdivided into four sub-frames, and the forward masking level $L_j(t,m)$ was calculated for the $j$th sub-frame using the energy accumulated over the previous frame and all sub-frames up to the current sub-frame. The amount temporal of masking $TM1$ is then chosen as the average of $M_f(t,m)$ over $j$ calculated using $L_j(t,m)$. This masking calculation is more efficient than the original method proposed in [5] as they calculated the masking threshold for every sample in a frame, while we calculate the threshold only four times per frame. Calculation of a temporal masking threshold every 4 ms was considered adequate since this provides a good approximation to the decay effect that lasts around 200 ms.

## 3.2 Model 2 (TM2)

The second model developed by Ambikairajah et al. [6] is based on the fact that temporal masking decays approximately exponentially following each stimulus. The masking level calculation for the $m$th critical band signal $M_f(t,m)$ is

$$M_f(t,m) = \begin{cases} L(t,m), & L(t,m) > c_0 L(t-\Delta t,m) \\ c_0 L(t-\Delta t,m), & otherwise \end{cases} \tag{3.5}$$

where $c_0 = \exp(-\tau_m)$. The amount of temporal masking $TM2$ is then chosen as the average of $M_f(t,m)$ for each sub-frame calculation.

Normally first order IIR low-pass filters are used to model the forward masking [6, 12]. We have modified the time constant, $\tau_m$, of these filters as follows, in order to model the duration of forward masking more accurately.

$$\tau_m = \tau_{\min} + \frac{100\,Hz}{fc_m} \cdot (\tau_{100} - \tau_{\min}) \tag{3.6}$$

The time constants $\tau_{\min}$ and $\tau_{100}$ used were 8 ms and 30 ms, respectively. The time constants were verified empirically by listening to the quality of the reconstructed speech, and were found to be much shorter than the 200 ms post-masking effect commonly seen in the literature.

## 3.3 Model 3 (TM3)

Novorita [7] incorporated temporal masking effects into bark spectral distortion measure used for automatic speech quality measurement. Novorita analyzed four masking filters, including exponential, linear, second power and half power, and concluded that temporal masking models conforming to the exponential responses achieved the best performance.

The temporal masking filter with exponential decay used by [7] is as follows:

$$M_m^n(\lambda) = (v^n[m] - \text{au\_min}) \cdot e^{\frac{n-\lambda}{eq}} \tag{3.7}$$

where $n$ is short-time frame index, $\lambda$ is time offset index, $m$ is the critical band number in Barks, $v$ is the value in phons for a given bark and time point, au_min is the convergence point for threshold response decay, and $eq$ is a factor to normalize the time constant.

The main drawback of the above model is that it is difficult to calculate the parameters au_min and *eq.* Therefore in this paper we approximate the above exponential model as follows:

$$M_f(t,m) = L(t,m) \cdot e^{-\Delta t / \tau_m} \tag{3.8}$$

where $M_f(t,m)$ is the amount of forward masking in dB in the *m*th band, $\Delta t$ is the time difference between the masker and the maskee in milliseconds, $\tau_m$ is time constant chosen as shown in equation (3.6), and *L(t,m)* is the masker level (dB). To obtain the masking threshold, *TM*3, we calculate $M_f$ for sub frames ($j$=1,2,3,4) and we take the average of $M_f(j)$.

## 3.4 Combined Masking Threshold

A simultaneous masking model similar to that used in MPEG [1] was employed to calculate the masking threshold, *SM,* for each critical band. It has been shown that simultaneous masking removes redundant pulses in the structure shown in Fig. 3-1.

In our experiment, a combined masking threshold is calculated by considering the effect of both temporal and simultaneous masking. We use the power law method for combining the thresholds [12]

$$MT = \left(TM^p + SM^p\right)^{1/p}, \ 1 \le p \le \infty \tag{3.9}$$

where $MT$ is the total masking threshold, $TM$ is temporal masking threshold, and $SM$ is the simultaneous masking threshold. The parameter $p$ defines the way the masking thresholds add. Setting $p = 5$ provides an accurate masking threshold.

# 4. MASKING MODEL EVALUATION

We evaluated the performance of the temporal masking models using the analysis/synthesis filter bank shown in Fig. 3-1. A wide range of speech materials has been selected with total of six speech files sampled at 8 kHz as shown in Fig. 3-4. The input speech signal was decomposed into 17 bands. The output of each filter was half-wave rectified, and the positive peaks of the critical band signals were kept (every 16 ms). This process results in a series of critical band pulse trains, where the pulses retain the amplitudes of the critical band signals from which they were derived. Then temporal and simultaneous masking models were then applied to these pulses, where

pulses with amplitude below the masking threshold were removed. The overall effect of simultaneous and temporal masking is to represent the input signal using the minimum number of pulses. The critical band signals can then be reconstructed from the masked pulse trains by means of bandpass filtering and summing the outputs.

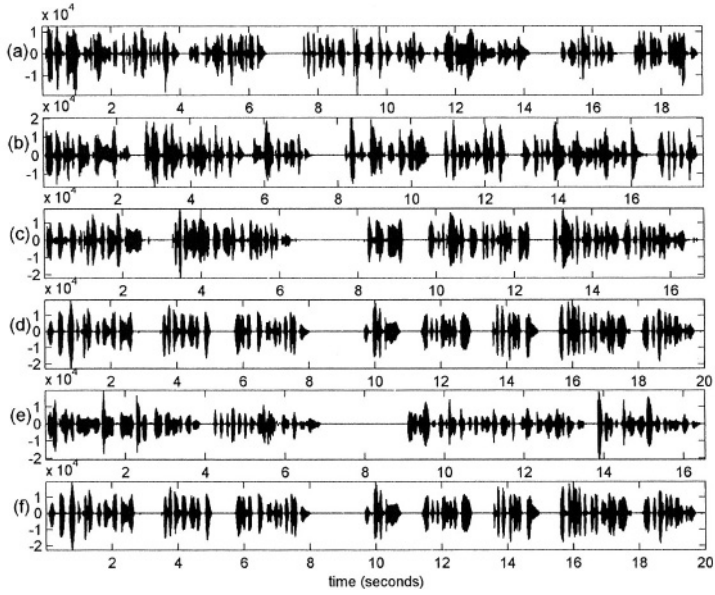



*Figure 3-4.* Speech materials used for performance evaluation, (a) Female Speech (English), (b) Male Speech (English), (c) Female Speech (French), (d) Male Speech (French), (e) Female Speech (German), (f) Male Speech (German)

Firstly, we evaluated the quality of the reconstructed speech using each of the temporal masking models. We then repeated the experiment using the simultaneous masking model. Finally, we tested the combined effect of each temporal masking model with simultaneous masking. We evaluated six speech files and the average PESQ score results and the average percentage pulse reduction for these experiments are shown in Table 3-1.

From Table 3-1, it is seen that the temporal masking effects of all models produce a transparent quality signal (defined here as PESQ > 3.5) with significant pulse reduction. Of the models tested, we found that our proposed temporal masking model combined with simultaneous masking model (TM1 and SM) produces a compromise between the quality (PESQ) and pulse reduction.

*Table 3-1.* Evaluation of Various Masking Models for Speech

| Masking Model | Average PESQ Score | Average Pulses Removed |
|---|---|---|
| No masking | 4.00 | 0 % |
| TM1 | 3.95 | 41.95 % |
| TM2 | 3.86 | 41.43 % |
| TM3 | 3.88 | 46.91 % |
| SM | 3.90 | 37.48 % |
| TM1 and SM | 3.87 | 51.71 % |
| TM2 and SM | 3.81 | 53.90 % |
| TM3 and SM | 3.85 | 48.15 % |

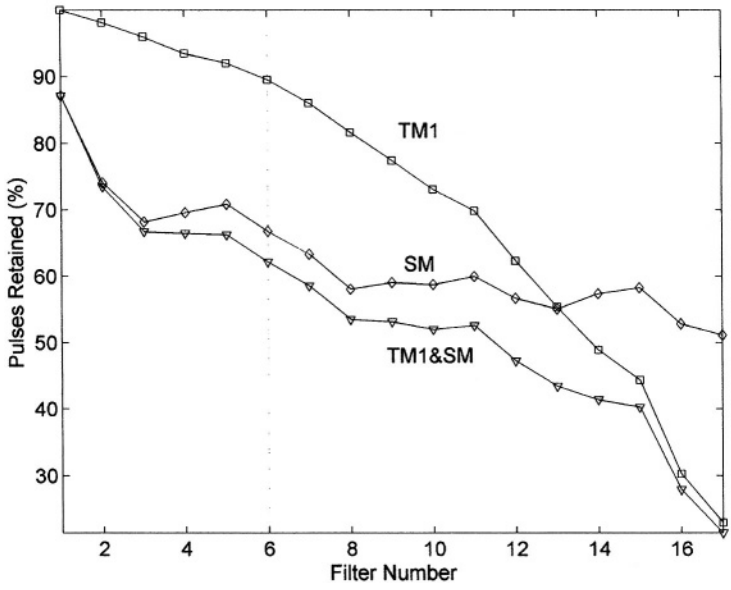



*Figure 3-5.* Pulse Retained (%) using TM1, SM and TM1&SM for German male speech

Fig. 3-5 shows the number of pulses retained in each critical band for male speech (German), averaged over an entire speech utterance (20 sec). The temporal masking model TM1 appears to have the greatest effect for higher frequency bands.

The combined model TM1+SM was also tested in a wavelet packet based audio coder [13] to quantify the improvement in terms of bit rate. We carried out subjective tests according to the ITU-R BS.1116 [14] using various audio materials sampled at 44.1 kHz. We train the listener to become familiar with coding artifacts, such as pre-echo, aliasing, birdies, and speech reverberation as described in AES CD-ROM [15]. The ABC/Hidden

Reference Software Audio Comparison Tool [16] was then used to evaluate our coder performance for various audio materials, in which the listener grades the audio material. Furthermore, the high quality headphone, i.e. Sony MDR-V700DJ, was used in the listening test.

The comparison of required average bitrate for various audio materials sampled at 44.1 kHz is described in Table 3-2. It can be seen that the high quality of the compressed sounds could be maintained while the bit rate was reduced by more than 17% on average if we include the proposed temporal masking method. Note that the bit rates shown in Table 3-2 have included the side information.

*Table 3-2.* Required Average Bitrate for Various Audio Materials

| Audio Material | SM (kbit/s) | TM1& SM (kbit/s) | Bitrate Saving (%) |
|---|---|---|---|
| Mariah Carey | 133.95 | 104.95 | 21.65 |
| Eric Clapton | 123.57 | 98.63 | 20.18 |
| Susan Vega | 93.21 | 76.31 | 18.13 |
| Tracy Chapman | 110.54 | 86.06 | 22.14 |
| Hani Anggraini | 107.40 | 89.21 | 16.94 |
| Castanets | 88.72 | 71.24 | 19.71 |
| Jazz | 145.36 | 108.58 | 25.30 |
| Male Speech | 104.24 | 85.44 | 18.04 |

## 5. CONCLUSION

Speech and audio coding using three temporal masking models has been presented. The model developed in this paper shows an improvement over other previously developed temporal masking models. This has been demonstrated both in terms of PESQ and pulse reduction for speech coding, and in terms of bit rate reduction in a wavelet packet based audio coder. Combined with simultaneous masking models, the developed temporal masking model appears to produce transparent quality of speech and audio. Future investigations will concentrate on furthering the theoretical development of temporal masking models.

## REFERENCES

1. M. Black and M. Zeytinoglu, "Computationally efficient wavelet packet coding of wide-band stereo audio signals," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing,* pp. 3075-3078, 1995.
2. K. Bradenburg, G. Stoll, Y. Dehery, J. Johnston, L. kerkhof, and E. Schroeder, "The ISO/MPEG audio codec: A generic standard for coding of high quality digital audio," *Journal of Audio Engineering Society,* vol. 42, pp. 780-791, 1994.

3. E. Zwicker and T. Zwicker, "Audio engineering and psychoacoustics, matching signals to the final receiver, the human auditory system," *Journal of Audio Engineering Society,* vol. 39, pp. 115-126, 1991.
4. W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *Journal of Acoustic Society of America,* vol. 71, pp. 950-962, 1982.
5. H. Najafzadeh, H. Lahdidli, M. Lavoie, and L. Thibault, "Use of auditory temporal masking in the MPEG psychoacoustics model 2," in *Proc. of the 114th Convention. Audio Engineering Society,* 2003.
6. E. Ambikairajah, J. Epps, and L. Lin, "Wideband speech and audio coding using Gammatone filter banks," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing,* pp. 773-776, 2001.
7. B. Novorita, "Incorporation of temporal masking effects into bark spectral distortion measure," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing,* pp. 665-668, 1999.
8. M. Lynch, E. Ambikairajah, and A. Davis, "Comparison of auditory masking models for speech coding," in *Proc. of 5th European Conference on Speech Communication and Technology,* pp. 1495-1498, 1997.
9. P. Philippe, F. M. d. Saint-Martin, and M. Lever, "Wavelet packet filterbanks for low time delay audio coding," *IEEE Transactions on Speech and Audio Processing,* vol. 7, pp. 310-322, 1999.
10. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "PESQ-The new ITU standard for end-to-end speech quality assessment," in *Proc. of the 109th Convention, Audio Engineering Society,* 2000.
11. E. Terhardt, "Calculating virtual pitch," *Hearing Research,* vol. 1, pp. 155-182, 1979.
12. M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards:* Kluwer Academic Publishers, 2003.
13. F. Sinaga, T. S. Gunawan, and E. Ambikairajah, "Wavelet Packe Based Audio Coding using Temporal Masking," in *Proc. of 4th Int. Conf. on Information, Communication and Signal Processing,* Singapore, 2003.
14. ITU, "ITU-R BS.1116.1, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva 1997.
15. AES, "Perceptual Audio Coders: What to Listen For," Audio Engineering Society, 2001.
16. ff123, "ABC/Hidden Reference: Tool for Comparing Multiple Audio Samples," [http://ff123.net/abchr/abchr.html], 2002.

Chapter 4

# OBJECTIVE HYBRID IMAGE QUALITY METRIC FOR IN-SERVICE QUALITY ASSESSMENT

Tubagus Maulana Kusuma and Hans-Jürgen Zepernick
*Western Australian Telecommunications Research Institute, Perth, WA 6907, Australia*

**Abstract** User-oriented image quality assessment has become a key factor in multimedia communications as a means of monitoring perceptual service quality. However, existing image quality metrics such as Peak Signal-to-Noise Ratio (PSNR) are inappropriate for in-service quality monitoring since they require the original image to be available at the receiver. Although PSNR and others are objective metrics, they are not based on human visual perception and are typically designed to measure the fidelity. On the other hand, the human visual system (HVS) is more sensitive to perceptual quality than fidelity. In order to overcome these problems, we propose a novel objective reduced-reference hybrid image quality metric (RR-HIQM) that accounts for the human visual perception and does not require a reference image at the receiver. This metric is based on the combination of several image artifact measures. The result is a single number, which represents overall image quality.

**Keywords:** Objective image quality metric, perceptual image quality assessment, in-service quality monitoring, reduced-reference system, JPEG

## 1. INTRODUCTION

Transmission of multimedia services such as image and video over a wireless communication link can be expected to grow rapidly with the deployment of third and future generation mobile radio systems [1]. These mobile radio systems are also expected to offer higher data rates, improved reliability and spectrum efficiency. This will allow transmission of a variety of multimedia services over a hostile radio channel while maintaining a satisfactory quality of service. However, experiments have shown that the existing quality measures, such as Bit Error Rate (BER) are not adequate for image and video transmission. Therefore, user oriented quality metrics that incorporate human visual perception have become of great interest in image and video delivery

services [2, 3]. Although the best and truest judge of quality is human (through subjective tests), continuous monitoring of communication systems quality by subjective methods is tedious, expensive and impossible in a real-time environment. Therefore, objective quality measurement methods which closely approximate the subjective test results are sought after.

Another incentive for the search after user oriented quality metrics is the fact that the commonly used image fidelity and quality metrics such as Mean-Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are inappropriate for in-service quality monitoring since the reference image is unavailable at the receiver. These existing metrics fall into the category of full-reference (FR) metrics [3]. In order to overcome this problem, we propose a novel reduced-reference (RR) objective perceptual image quality metric that does not require a reference image and which is based on the human visual system (HVS). Since the overall quality is represented by a single number, the overhead needed to communicate this information can be kept minimum.

This contribution is organized as follows. Different types of image artifacts and metrics are outlined in Section 2. The proposed Hybrid Image Quality Metric (HIQM) is explained in Section 3. In Section 4, application of HIQM for quality monitoring is discussed, followed by experimental results in Section 5. Finally, concluding remarks are given in Section 6.

# 2. IMAGE ARTIFACTS AND METRICS

## 2.1 Image Artifacts

Image artifacts are caused by impairments such as transmission errors and depend on the image compression scheme used. A received data packet may have header information and/or data segments corrupted. In some image formats a single corrupted bit might lead to an incomplete or even undecodable image. In case of Joint Photographic Experts Group (JPEG) images, for example, the bit error location can have a significant impact on the level of image distortion. A bit error that occurs at a marker segment position can severely degrade the image quality. An image may be even completely lost because the decoder fails to recognize the compressed image.

In this contribution, we consider five types of artifacts that have been observed during our simulations. These artifacts are smoothness, blocking, ringing/false edge, masking, and lost block/pixel and will be briefly described below [4–6]:

***Smoothness,*** which appears as edge smoothness or texture blur, is due to the loss of high frequency components when compared with the original image. Blurring means that the received image is smoother than the original (see Fig. 4-1a).

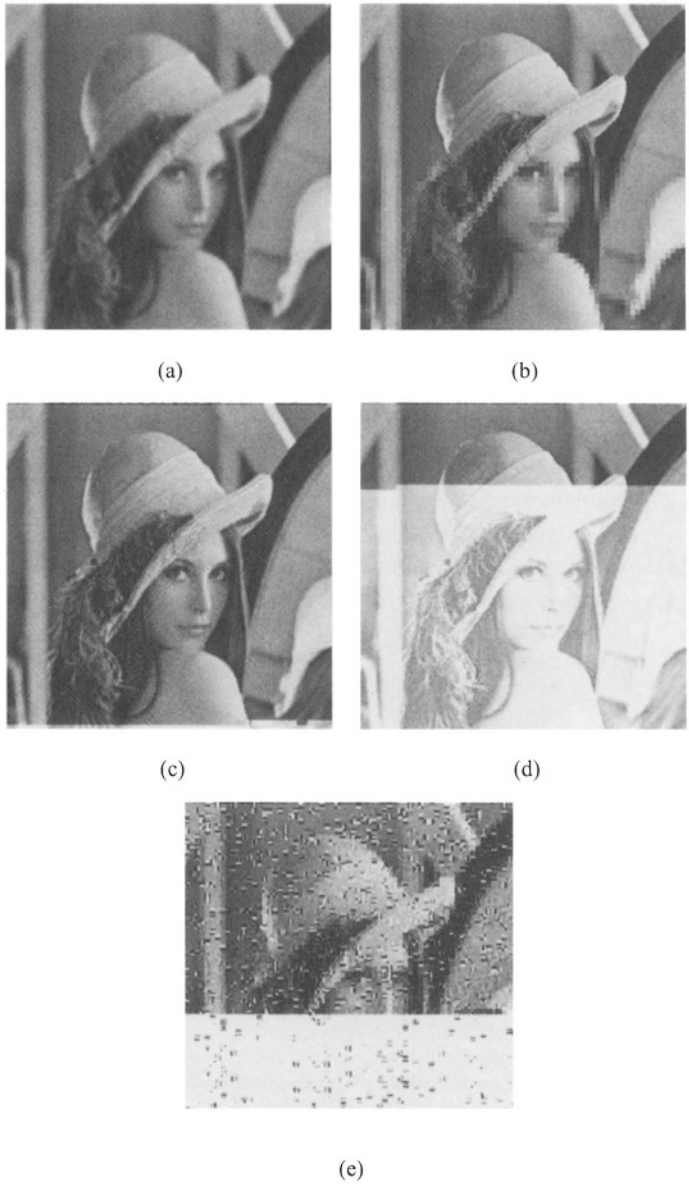



*Figure 4-1.* Samples of image artifacts: (a) Smoothness, (b) Blocking, (c) Ringing/false edge, (d) Masking, and (e) Lost block

***Blocking*** appears in all block-based compression techniques and is due to coarse quantization of frequency components. It can be observed as surface dis-

continuity (edge) at block boundaries. These edges are perceived as abnormal high frequency components in the spectrum (see Fig. 4-1b).

***Ringing*** is observed as periodic pseudoedges around original edges. It is due to improper truncation of high frequency components. This artifact is also known as the Gibbs Phenomenon or Gibbs effect (see Fig. 4-1c). For the worst case, the edges can be shifted far away from the original edge locations. This effect is observed as false edge.

***Masking*** is the reduction in the visibility of one image component (the target) due to the presence of another (the masker). There are two kinds of masking effects. The first is called luminance masking, also known as light adaptation. The second is texture masking, which occurs when maskers are complex textures or masker and target have similar frequencies and orientations (see Fig. 4-1d).

***Lost block/pixel*** is a loss of a block of pixels or an alteration of a pixel value. In common operation of still image compression standards like JPEG, the encoder tiles the image into blocks of $n \times n$ pixels, calculates a 2-D transform, quantizes the transform coefficients and encodes them using Huffman coding. In common wireless scenario, the image is transmitted over wireless channel block-by-block. Due to severe fading, entire image blocks can be lost (see Fig. 4-1e).

## 2.2 Image Metrics

Image metrics may be divided into two categories:

***Image fidelity metrics*** indicate image differences by measuring pixel-by-pixel closeness between images. MSE and PSNR fall into this category.

***Image quality metrics*** define quality based on individual image features; these metrics also incorporate HVS. Much research is being carried out on the image quality metrics, but they mostly concentrate on single artifacts [7–10].

There are three approaches of measuring image fidelity and image quality as follows:

***Full-Reference (FR)*** is a method that compares a distorted image with its undistorted original.

***Reduced-Reference (RR)*** is a method that does not require to store the entire original image but extracts important features from the distorted image and compares them with corresponding stored features of the original image.

***No-Reference (NR)*** techniques do not require prior knowledge about the original image but perform assessment of the distorted image to search for the presence of known artifacts.

## 3. HYBRID IMAGE QUALITY METRIC

The proposed Hybrid Image Quality Metric (HIQM) [11] employs several quality measurement techniques and is calculated as weighted sum of respective quality metrics. It is designed to detect and to measure different image artifacts. The result is a single number that correlates well with perceived image quality. HIQM does not require a reference image at the receiver to measure the quality of a target image.

The proposed RR approach considers five different artifact measurements relating to blocking, blur, image activities/busyness, and intensity masking.

The blocking measurement is based on the algorithm proposed by Wang et. al. [7, 8]. This algorithm extracts the average difference across block boundaries, averages absolute differences between in-block image samples, and calculates the zero-crossing rate. The system has been trained using subjective test results in order to comply with human visual perception. The final blocking measure is calculated using statistical non-linear curve fitting techniques. This metric is classified as an NR type because only the received image is needed to measure the blocking. In our approach, this metric is also used to detect and to measure lost blocks.

The blur measurement algorithm is based on the work of Marziliano et. al. [9]. This algorithm accounts for the smoothing effect of blur on edges by measuring the distance between edges from local maximum and local minimum, the so-called local blur. A Sobel filter is used to detect the edges. Once the edges are detected, the distance between local maximum and local minimum can be measured for both horizontal and vertical directions. The final blur measure is obtained by averaging the local blur values over all edge locations in both directions.

Another important characteristic of an image relates to the activity measure that indicates the 'busyness' of the image. The active regions of an image are defined as those with strong edges and textures. Due to distortion, a received image normally has more activity compared to the original image. The technique used by HIQM is based on Saha and Vemuri's algorithm [10]. We use this metric to detect and to measure ringing and lost blocks. Especially, two types of Image Activity Measures (IAM) are deployed, edge and gradient-based IAM. For an $M \times N$ binary image, the edge activity measure is given by [10]:

$$IAM_{edge} = \left[ \frac{1}{MN} \sum_{i=1}^{M \cdot N} B(i) \right] \cdot 100 \qquad (4.1)$$

where $B(i)$ denotes the value of the detected edge at pixel location $i$, $M$ is the number of image rows and $N$ is the number of image columns.

The gradient-based activity measure for an $M \times N$ image is given by [10]:

$$IAM_{grad} = \frac{1}{MN}\left[\sum_{i=1}^{M-1}\sum_{j=1}^{N}|I(i,j)-I(i+1,j)| - \sum_{i=1}^{M}\sum_{j=1}^{N-1}|I(i,j)-I(i,j+1)|\right] \tag{4.2}$$

where $I(i,j)$ denotes the intensity value at pixel location $(i,j)$ and $|\cdot|$ denotes absolute value. Fig. 4-2 shows an original image and its edge representation whereas Fig. 4-3 shows the distorted version of this image. Clearly, the distorted image appears to have higher activity compared to the original image.

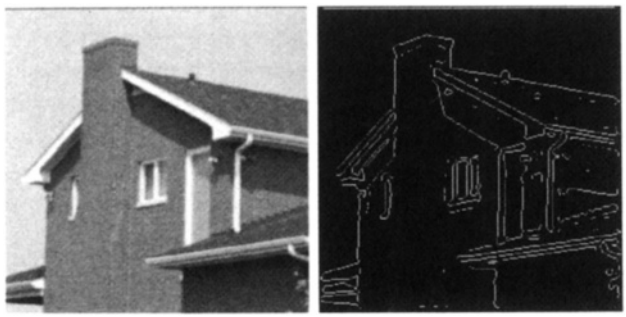


*Figure 4-2.* Sample original image and its edge representation

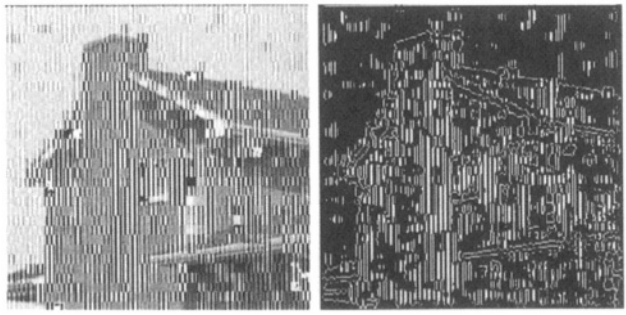


*Figure 4-3.* Sample distorted image with increased activity and its edge representation

Finally, the intensity masking detection is based on the standard deviation of the first-order image histogram, which indicates the distribution of the image data values. The shape of the histogram provides many insights into the character of an image [12]. The histogram of an $M \times N$ image is defined as the percentage of pixels within the image at a given gray level. For a 256 gray level image, the histogram $h_i$ is given by:

$$h_i = \frac{n_i}{MN}, \quad \text{for } 0 \leq i \leq 255 \tag{4.3}$$

where $n_i$ denotes the number of pixels at gray level $i$ and $MN$ is the total number of pixels within the image.

From the histogram information, we can measure the image perceived brightness by calculating the average gray level that is given by:

$$brightness = \sum_{i=0}^{255} i \cdot h_i \tag{4.4}$$

where $h_i$ denotes the image histogram at gray level $i$.

A low average value means dark image, whereas large average implies a bright image. The image contrast can now be measured by estimating the average gray level variation within the image (standard deviation), which is given by:

$$contrast = \sqrt{\sum_{i=0}^{255} i^2 \cdot h_i - brightness^2} \tag{4.5}$$

A small standard deviation indicates a low contrast image, while large value implies a high contrast image. The intensity masking ($mask_{int}$) detection is based on contrast measurement. Therefore, the same expression as mentioned in (4.5) is applied. Fig. 4-4 shows the example of images and their respective histogram information.

The proposed overall quality measure is a weighted sum of all the aforementioned metrics. The weight allocation for individual metrics was based on the impact of the metric on the overall perceptibility of images by human vision. The fine-tuning of the weights was done empirically and was justified by requesting opinion from a group of unbiased test persons. Initially, all the metrics were given the same weight $w \in [0 \cdots 1]$ and were then adjusted based on the contribution of each metric to the perceptibility of image by human eye. In JPEG, for example, blocking is given a higher weight compared to other metrics because it is the most frequently observed artifact in this particular image format and can be easily perceived by human vision. The overall quality is given by:

$$\begin{aligned} HIQM &= w_1 \cdot blockMetric \\ &+ w_2 \cdot blurMetric \\ &+ w_3 \cdot IAM_{edge} + w_4 \cdot IAM_{grad} \\ &+ w_5 \cdot mask_{int} \end{aligned} \tag{4.6}$$

where $w_i$ denotes the weight of a particular metric.

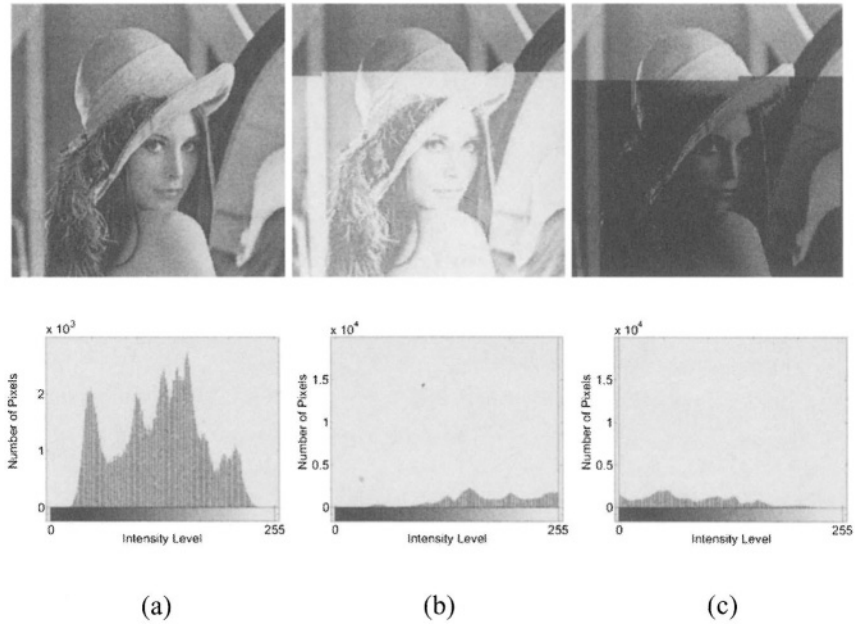



*Figure 4-4.* Samples for test image "Lena" with histogram information: (a) Original Image, (b) Image with White Mask, (c) Image with Black Mask

## 4. QUALITY MONITORING USING HIQM

In image transmission systems, HIQM can be used for continuous in-service quality monitoring since it does not require a reference image to measure the quality of a received image. However, different original images differ in activity and other characteristics. Therefore, each image has its individual HIQM value, which we will refer to as quality baseline.

To obtain a proper measure, we need to normalize or calibrate the measurement system to the quality baseline. Therefore, the quality baseline of an original image needs to be communicated to the receiving end of the system under test. Obviously, this constitutes an RR approach and may replace conventional FR quality techniques. As correct reception of the quality baseline is vital for image quality assessment at the receiver, error control coding is recommended to protect this important parameter.

The basic steps of in-service quality monitoring using HIQM can be summarized as follows (Fig. 4-5):

1 At the transmitter, measure the quality baseline of the original image in terms of its HIQM value.

2 Concatenate quality baseline and image file to form the overall packet format (see Fig. 4-6) before transmission of respective packets.

3 At the receiver, provide a reference quality by extracting quality baseline from received packet and, if necessary, adding a tolerable degradation value to it.

4 Measure HIQM of the received image and compare it with the reference quality.

The total length of the RR-HIQM related quality value may be chosen as 17 bits. It consists of sign information (1 bit), quality (8 bits for the integer, 4 bits for each the $1^{st}$ and the $2^{nd}$ decimal). The HIQM of -0.57, for example, will be concatenated to the header part of the image file as $00000000001010111_2$.

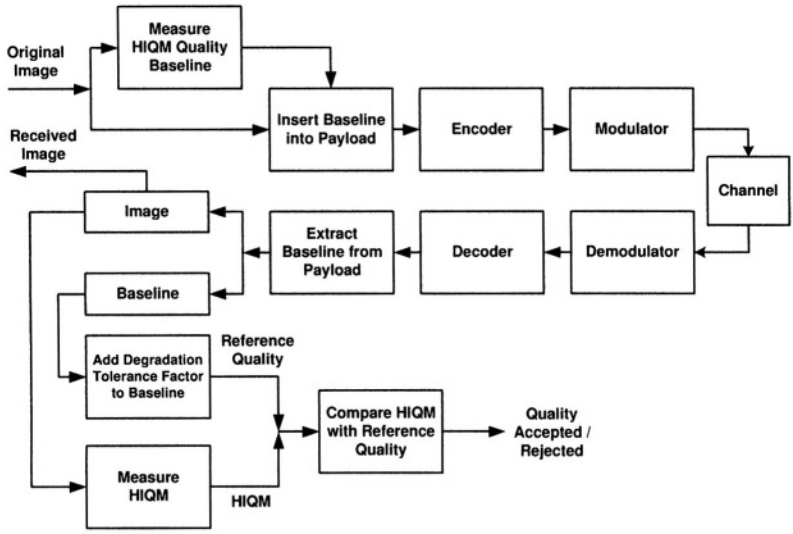



*Figure 4-5.* Reduced reference in-service quality monitoring using the proposed HIQM

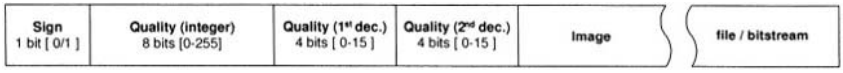



*Figure 4-6.* RR-HIQM packets format

In practice, there is a number of possible applications for RR-HIQM, which include system testing, continuous in-service quality monitoring, link adaptation and performance characterization of image transmission system.

## 5. EXPERIMENTAL RESULTS

In this section, we provide experimental results for test images "Goldhill" (Quality baseline = 1.34), "Lena" (Quality baseline = –1.76) and "Tiffany" (Quality baseline = –0.57). The test scenario for these standard test images was chosen as a Rayleigh flat fading channel. A simple (31, 21) BCH code was applied for error protection purposes along with a soft-combining scheme using a maximum of two retransmissions. The average Signal-to-Noise Ratio (SNR) was set to 5*dB*. We did not use a JPEG restart marker in this experiment in order to gain extreme artifacts.

For this test scenario, the weights of the various metrics were finally obtained from involving a group of test persons as: $w_1 = 1, w_2 = 0.5, w_3 = 1, w_4 = 0.5,$ and $w_5 = 0.3$. From the experimental results, it can be concluded that HIQM inversely relates to PSNR (see Figs. 4-7, 4-8 and 4-9). In other words, a better quality image is represented by a higher PSNR value or a lower HIQM value (see Figs. 4-10, 4-11 and 4-12).

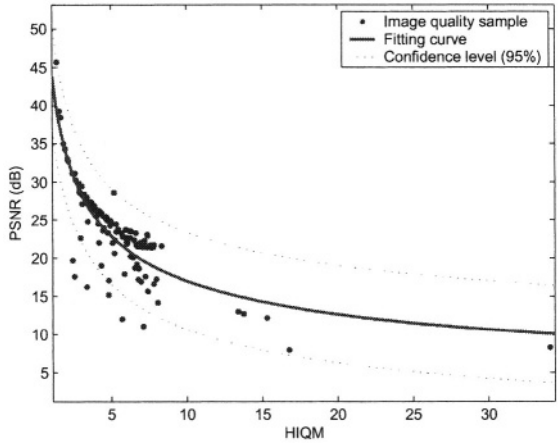



*Figure 4-7.* Metric comparison for image "Goldhill"

Figs. 4-7, 4-8 and 4-9 show some outliers which are due to disagreement of HIQM with PSNR. These disagreements are mostly due to the misjudgement of the PSNR in relation to the opinion of the human subjects. Two samples of misjudgement are presented in Figs. 4-11 and 4-12. For justification, we interviewed 10 persons to give their opinion about the quality when disagreements occurred between HIQM and PSNR. The average opinion of these people were that HIQM provides a better judgement than PSNR.

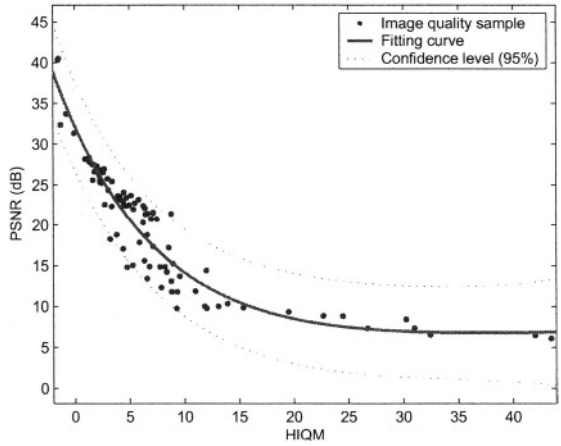



*Figure 4-8.* Metric comparison for image "Lena"

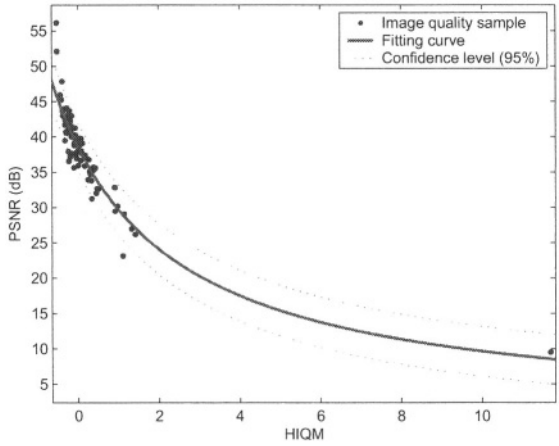



*Figure 4-9.* Metric comparison for image "Tiffany"

## 6. SUMMARY

A new reduced-reference image quality measurement technique was presented. It was shown by way of experiment that the proposed HIQM outperforms PSNR with respect to quantifying user perceived quality. The introduced HIQM may be used for reduced-reference in-service image quality monitoring.

(a) (b) (c)

*Figure 4-10.* Quality samples for test image "Goldhill": (a) $PSNR = 8.28dB$ and $HIQM = 34.09$,(b)$PSNR = 22.05dB$ and $HIQM = 6.95$, (c) $PSNR = 45.70dB$ and $HIQM = 1.43$

(a) (b)

*Figure 4-11.* Quality samples for test image "Lena": (a) $PSNR = 9.76dB$ and $HIQM = 9.28$, (b) $PSNR = 9.78dB$ and $HIQM = 11.99$

## ACKNOWLEDGEMENTS

This work has been partly supported by the Commonwealth of Australia through the Cooperative Research Centre program.

## REFERENCES

1. V. Weerackody, C. Podilchuk, and A. Estrella, "Transmission of JPEG-Coded Images Over Wireless Channels," *Bell Laboratories Technical Journal,* Autumn 1996.

2. M. Knee, "The Picture Appraisal Rating (PAR) - A Single-ended Picture Quality Measure for MPEG-2," *Technical Report,* Snell and Wilcox, Jan. 2000.

3. A. Webster, "Objective and Subjective Evaluation for Telecommunications Services and Video Quality," *National Telecommunications and Information Administration (NTIA)/Institute for Telecommunication Science (ITS), Rapporteur Q21/9, 2002.*

(a) (b)

*Figure 4-12.* Quality samples for test image "Tiffany": (a) $PSNR = 23.17dB$ and $HIQM = 3.30$, (b) $PSNR = 24.15dB$ and $HIQM = 7.78$


4. S. Winkler, "Vision Models and Quality Metrics for Image Processing Applications," *Ph.D. Thesis, École Polytechnique Fédéralé De Laussane (EPFL),* Dec. 2000.
5. A. Jakulin, "Baseline JPEG and JPEG2000 Artifacts Illustrated," <http://ai.fri.uni-lj.si/~aleks/jpeg/artifacts.htm>, accessed on 8 May 2003.
6. S.D. Rane, J. Remus, and G. Sapiro, "Wavelet-Domain Reconstruction of Lost Blocks in Wireless Image Transmission and Packet-Switched Networks," in *Proc. of IEEE International Conference on Image Processing,* pp. 309-312, 2002.
7. Z. Wang, A.C. Bovik, and B.L. Evans, "Blind Measurement of Blocking Artifacts in Images," in *Proc. of IEEE International Conference on Image Processing,* pp. 981-984, Sep. 2000.
8. Z. Wang, H.R. Sheikh, and A.C. Bovik, "No-Reference Perceptual Quality Assessment of JPEG Compressed Images," in *Proc. of IEEE International Conference on Image Processing,* pp. 477-480, Sep. 2002.
9. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A No-Reference Perceptual Blur Metric," in *Proc. of IEEE International Conference on Image Processing,* pp. 57-60, Sep. 2002.
10. S. Saha and R. Vemuri, "An Analysis on the Effect of Image Activity on Lossy Coding Performance," in *Proc. of IEEE International Symposium on Circuits and Systems,* pp. 295-298, May 2000.
11. T. M. Kusuma and H.-J. Zepernick, "In-Service Image Monitoring Using Perceptual Objective Quality Metrics,"*Journal of Electrical Engineering,* vol. 54, no. 9-10, pp. 237-243, Dec. 2003.
12. A.R. Weeks, "Fundamentals of Electronic Image Processing," *SPIE/IEEE Series on Imaging Science and Engineering,* 1998.

*This page intentionally left blank*

Chapter 5

# AN OBJECT-BASED HIGHLY SCALABLE IMAGE CODING FOR EFFICIENT MULTIMEDIA DISTRIBUTION

Habibollah Danyali[1] and Alfred Mertins[2]

[1] *Department of Electrical Engineering, Faculty of Engineering, University of Kurdistan, Sanandaj, Iran.*

[2] *Signal Processing Group, Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany.*

**Abstract** In this chapter a highly scalable image coding algorithm for texture coding of arbitrarily shaped visual objects is proposed. The proposed algorithm is based on the Set Partitioning in Hierarchical Trees (SPIHT) algorithm and is called Object-Based Highly Scalable SPIHT (OBHS-SPIHT). It adds the spatial scalability feature to the SPIHT algorithm through the introduction of multiple resolution-dependent lists for the sorting stage of the algorithm, while retaining the important features of the original algorithm such as compression efficiency and full SNR (rate) scalability. The idea of bitstream transcoding to obtain different bitstreams for various spatial resolutions and bit rates, all from a single bitstream, is completely supported by the algorithm. The proposed algorithm efficiently facilitates the distribution of visual information especially over heterogeneous networks such as the Internet.

**Key words:** Image coding, object-based coding, scalability, SPIHT, OBHS-SPIHT

## 1. INTRODUCTION

In traditional image and video coding systems a picture is represented and processed as a rectangular frame of pixels. Such a structure is able to provide only frame-based functionalities. By contrast, a new generation of image and video coding called *object-based coding* deals with a video scene as a composition of different *video objects* and processes each object individually. As a key advantage, this object-based representation can facilitate an object interactivity functionality. MPEG-4 [1], the emerging standard for audio-visual

information coding has addressed object-based coding. On the other hand, a scalable coding scheme provides a bitstream that consists of embedded parts to offer increasingly better signal-to-noise ratio (SNR) or/and greater spatial resolution for each object in the scene. Therefore an object-based scalable coding system could provide a very convenient coding scheme for the transmission of visual information over heterogenous networks such as the Internet.

The multiresolution image representation provided by the two-dimensional discrete wavelet transform (DWT) gives wavelet based image coding schemes a great potential to support both SNR and spatial scalability. Modifications of the DWT, called shape adaptive DWTs (SA-DWTs), like the one in [2], enable wavelet based coding algorithms to be extended for coding of arbitrarily shaped objects. In recent years, a family of very efficient zerotree based wavelet image coders has been developed and emerged as one of the most promising techniques to meet the challenges for image coding. Based on the idea of grouping wavelet coefficients at different scales and predicting zero coefficients across scales, Shapiro [3] introduced the Embedded Zerotree Wavelet (EZW) coding scheme. The EZW algorithm was improved by Said and Pearlman [4] in their work called Set Partitioning in Hierarchical Trees (SPIHT). SPIHT provides very efficient compression, supports progressive image transmission, and is considered as a benchmark for the state-of-the-art image coding algorithms.

Motivated by the success of embedded zerotree wavelet coding in frame-based schemes, some researchers have extended this framework for coding of arbitrarily shaped objects [5–10]. These works employ a SA-DWT approach for decomposing the object texture followed by a modified version of a zerotree based method which only encodes the wavelet coefficients that belong to the decomposed object. A shape-adaptive extension of the EZW coding technique was proposed by Li *et al.* in [8]. A modification of ZTE [6] for texture coding of still object was reported in [5]. A SA-DWT with even length filters and an extension of EZW for coding of arbitrarily shaped still textures was presented by Mertins and Singh [7]. In [9, 10] an object-based DWT and modified versions of SPIHT algorithm for coding of arbitrarily shaped objects were presented. All of these object-based coding methods provide embedded bitstreams and support SNR scalability but do not provide any sort of spatial scalability. Moreover their bitstreams cannot be reordered according to desired resolutions and fidelity.

In [11–13] we introduced modified versions of the SPIHT algorithm that provide both spatial and SNR scalability features for rectangular (frame-based) images. In this chapter, the method of [12, 13] is further developed for efficient highly scalable texture coding of arbitrarily shaped visual objects. The developed algorithm, called Object-based Highly Scalable SPIHT (OBHS-SPIHT), adds the spatial scalability feature to the SPIHT bitstream without sacrificing

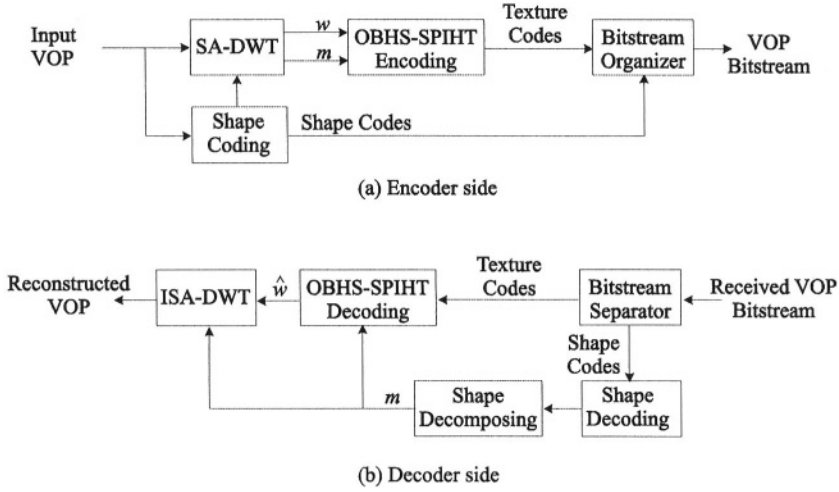



*Figure 5-1.* Block diagram of the OBHS-SPIHT coding system.

compression efficiency and SNR scalability in any way. The OBHS-SPIHT bitstream can be easily reordered to achieve different levels of spatial resolution and quality requested by the decoder.

The rest of this chapter is organized as follows. Section 2 reviews the OBHS-SPIHT coding system and explains the OBHS-SPIHT algorithm. The structure of the OBHS-SPIHT bitstream and the parsing process are presented in Section 3. In Section 4, simulation details are explained and some experimental results are presented, and finally, Section 5 concludes the chapter.

# 2. THE OBJECT-BASED HIGHLY SCALABLE SPIHT (OBHS-SPIHT)

## 2.1 System Overview

The proposed OBHS-SPIHT coding system is depicted in Figure 5-1. The input of the system is a video object plane (VOP) that may come directly from a bank of already known objects existing in some applications or is extracted by a segmentation algorithm from a rectangular image scene. The shape information of the object is assumed to be a binary alpha plane (mask) so that each input pixel is considered either inside or outside the object.

On the encoder side (Figure 5-1 (a)), the shape mask and the object texture are decomposed by a shape adaptive DWT approach. The decomposed texture coefficients $w$ and the decomposed shape mask $m$ are then consigned to the OBHS-SPIHT encoding block. The OBHS-SPIHT texture encoder only

encodes the coefficients that belong to the decomposed object. To recognize these coefficients it uses the decomposed shape mask. Any shape coding algorithm can be utilized to code the shape information. If a lossy shape coding is used, the reconstructed shape must be used in the SA-DWT and the OBHS-SPIHT coding algorithm. The bitstreams from shape coding and texture coding are assembled in the bitstream organizer to generate the final output bitstream for the VOP.

On the decoder side (Figure 5-1(b)), the bitstream separator first extracts the shape and the texture bitstreams from the received VOP bitstream. The shape mask is then reconstructed by decoding the shape bitstream. The decomposed mask, which is required by the OBHS-SPIHT decoder, is provided by applying the same level of decomposition as used by the encoder to the shape mask. The OBHS-SPIHT decoder then decodes the texture bitstream, and the inverse SA-DWT will be applied to the decoded wavelet coefficients to reconstruct the VOP texture.

## 2.2 The OBHS-SPIHT Algorithm

The SPIHT [4] algorithm, considers groups of coefficients in different scales of the wavelet image pyramid together as sets through the parent-offspring dependency depicted in Figure 5-2. At the beginning, the roots of the sets are located at the lowest frequency subband of the wavelet pyramid. It follows a bitplane-manner coding, and in each bitplane coding process, the algorithm deals with the wavelet coefficients as either a root of an insignificant set, an individual insignificant pixel, or a significant pixel. It sorts these coefficients in three ordered lists: the list of insignificant sets (LIS), the list of insignificant pixels (LIP), and the list of significant pixels (LSP). The main concept of the algorithm is managing these lists in order to efficiently extract insignificant sets in a hierarchical structure and identify significant coefficients.

In general, applying $N$ levels of wavelet decomposition to an image allows at most $N + 1$ levels of spatial resolution. To distinguish between different resolution levels, we denote the lowest spatial resolution level as level $N + 1$. The full image then becomes resolution level 1. Thus the actual spatial resolution related to level $k$ is $1/2^{k-1}$ of the resolution of the original image. The three subbands ($\mathrm{HL}_k$, $\mathrm{LH}_k$, $\mathrm{HH}_k$) that need to be added to increase the spatial resolution from Level $k + 1$ to Level $k$ are called spatial subband set level $k$ and referred to as $B_k$ (see Figure. 5-2).

The OBHS-SPIHT algorithm proposed in this chapter solves the spatial scalability problem through the introduction of multiple resolution-dependent lists and a resolution-dependent sorting pass. For each spatial subband set ($B_k$) we define a set of LIP, LSP and LIS lists, therefore we have $\mathrm{LIP}_k$, $\mathrm{LSP}_k$, and $\mathrm{LIS}_k$ for $k = k_{max}, k_{max} - 1, \ldots, 1$ where $k_{max}$ is the maximum number of

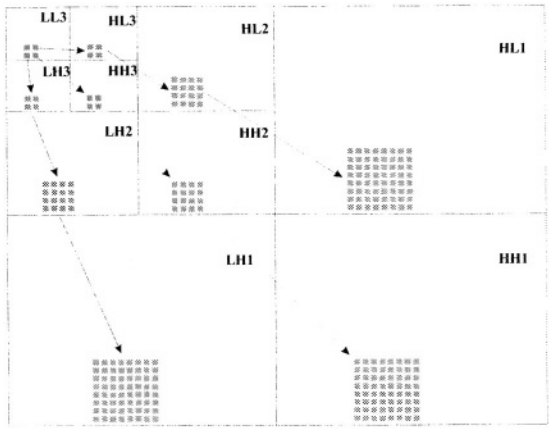



*Figure 5-2.* Orientation of SPIHT sets across wavelet subbands of an object.

spatial resolution levels that is supported by the encoder. The parent-offspring relationship in our algorithm is the same as with SPIHT, but we only consider and process those coefficients and sets which belong to the decomposed object (see Figure 5-2), similar to the modified algorithms in [10, 9]. In each bitplane coding process, the OBHS-SPIHT coder starts encoding from the maximum resolution level ($k_{max}$) and proceeds to the lowest level (level 1). During the resolution-dependent sorting pass for the lists that belong to level $k$, the algorithm first does the sorting for the coefficients in the $\mathrm{LIP}_k$, in the same way as SPIHT, to find and output significance bits for all list entries and then processes the $\mathrm{LIS}_k$. During processing the $\mathrm{LIS}_k$, sets that lie outside the resolution level $k$ are moved to their appropriate $\mathrm{LIS}_{k-1}$. After the algorithm has finished the sorting and refinement passes for resolution level $k$ it will do the same procedure for the next finer resolution level until all spatial resolution levels are finished. The total number of bits belonging to a particular bitplane is the same as for an object-based modification of SPIHT like the methods in [10, 9], but OBHS-SPIHT arranges them in the output bitstream according to their spatial resolution dependency.

Note that the total storage requirement for the $\mathrm{LIP}_k$, $\mathrm{LSP}_k$, and $\mathrm{LIS}_k$ for all resolutions is the same as for the LIS, LIP, and LSP used by the SPIHT algorithm.

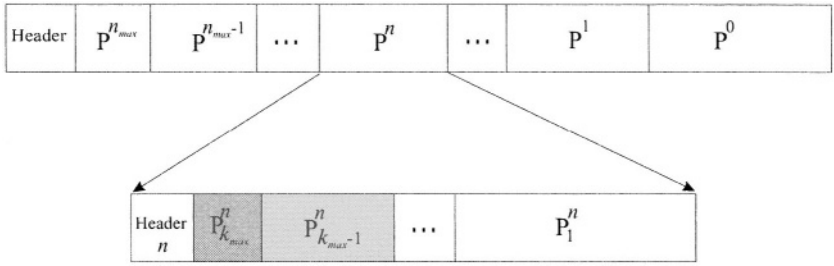



*Figure 5-3.* Structure of the OBHS-SPIHT encoder bitstream. $P_k^n$ is related to the codepart of spatial subband set level $k$ ($B_k$) at bitplane level $n$.

## 3. BITSTREAM FORMATION AND PARSING

The structure of the bitstream generated by the OBHS-SPIHT encoder is shown in Figure 5-3. The bitstream is constructed of different parts according to the different bitplane levels ($P^n$). Inside each bitplane codepart, the bits that belong to the different spatial subband sets, $P_k^n$, are separable. A header at the beginning of the bitstream identifies the number of spatial resolution levels supported by the encoder, as well as information such as the number of wavelet decomposition levels, and the maximum bitplane level required for coding. To support bitstream parsing by an image server/transcoder, at the beginning of each bitplane codepart there is an additional header that provides the information required to identify the different resolution codeparts.

The encoder needs to encode the object texture only once at a high bit rate. Different bitstreams for different spatial resolutions and fidelity can be easily generated from the encoded bitstream by selecting the related resolution codeparts. Figure 5-4 illustrate an example of multicasting a visual object for different users with different capabilities. The parsing process is a simple reordering of the original bitstream and can be carried out by the image server that stores the encoded bitstreams or by an individual parser as a simple part of an active network. The parser does not need to decode any part of the bitstream. For example, to provide a bitstream for resolution level $r$, in each bitplane codepart, only the spatial parts that belong to the spatial resolution levels greater or equal to $r$ are kept and all other parts are removed. As a distinct feature, the reordered bitstreams for each spatial resolution are completely rate-embedded (fine granular at bit level) which means that it can be truncated at any point to obtain the best reconstruction of the object at that bit rate. Note that the headers at the beginning of bitplane codeparts are only used by the parser and do not need to be sent to the decoder.

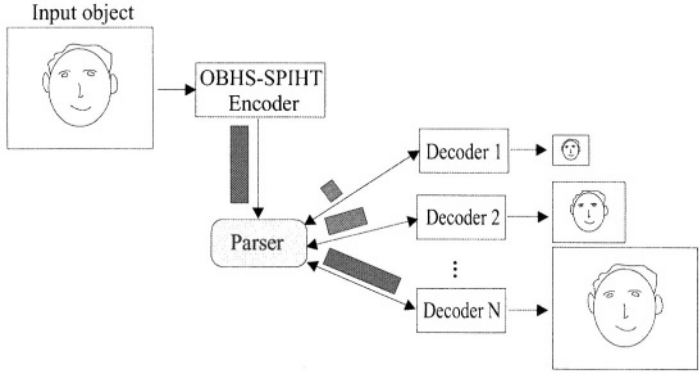



*Figure 5-4.* Illustration of parsing a single OBHS-SPIHT bitstream for decoding at different qualities and resolutions.

The decoder required for decoding the reordered bitstreams exactly follows the encoder, similar to the original SPIHT algorithm. It needs to keep track of the various lists only for spatial resolution levels greater or equal to the required one. Thus, the proposed algorithm naturally provides computational scalability as well.

## 4. SIMULATION DETAILS AND EXPERIMENTAL RESULTS

The OBHS-SPIHT encoder and decoder were fully software implemented. An efficient, non-expansive SA-DWT approach based on the method introduced in [2] was also implemented. The first frames of two MPEG-4 CIF colour (in YUV format) test sequences, Akiyo and Foreman, were selected for the test. Only the foreground objects of the test images were considered for coding. The shape/segmentation masks for these test sequences are supplied by MPEG. On the encoder side, four levels of 2D SA-DWT by 9/7-tap filters [14] with symmetric extension at the boundaries of the objects were first applied. The OBHS-SPIHT encoder was then set to progressively encode the decomposed objects from the maximum required bitplane to bitplane zero with five levels of spatial scalability support.

After encoding, the OBHS-SPIHT bitstream was fed into a parser to produce progressive (by quality) bitstreams for different spatial resolutions requested by a decoder. The OBHS-SPIHT decoder uses the reordered bitstream to decode only the required spatial subbands that are necessary for reconstructing the requested spatial resolution of the object. The inverse SA-DWT is then applied to the decoded spatial subbands to reconstruct the object in the re-

quested resolution. Reference frames for lower resolutions were defined by taking the lowest frequency subband frames after applying appropriate levels of SA-DWT to the original objects, and the fidelity was measured by the peak signal-to-noise ratio (PSNR). The bit rates for all levels were calculated according to the number of pixels in the foreground of the original full size image.

*Table 5-1.* PSNR results for the foreground objects of the first frames of the Akiyo and Foreman MPEG-4 CIF sequences at different spatial resolutions and bit rates.

| Spatial resolution | Rate (bpp) | OB-SPIHT | | | OBHS-SPIHT | | |
|---|---|---|---|---|---|---|---|
| | | Y | U | V | Y | U | V |
| Akiyo Object | | | | | | | |
| Full (level 1) | 0.25 | 27.59 | 38.26 | 37.44 | 27.51 | 38.26 | 37.44 |
| | 0.50 | 32.16 | 40.39 | 39.54 | 31.98 | 40.39 | 39.48 |
| | 1 | 37.77 | 42.99 | 42.88 | 37.71 | 43.13 | 42.94 |
| Half (level 2) | 0.125 | 4723 | 34.57 | 32.48 | 24.97 | 35.40 | 36.14 |
| | 0.25 | 29.70 | 39.04 | 38.25 | 30.06 | 39.04 | 38.25 |
| | 0.5 | 35.41 | 42.57 | 42.01 | 37.08 | 42.76 | 42.32 |
| | 1 | 41.42 | 48.48 | 48.50 | 47.11 | 48.83 | 48.84 |
| Quarter (level 3) | 0.0625 | 21.78 | 20.91 | 23.81 | 22.52 | 30.53 | 31.31 |
| | 0.125 | 27.83 | 35.15 | 32.92 | 28.13 | 36.00 | 37.04 |
| | 0.25 | 32.30 | 41.11 | 40.39 | 38.01 | 42.40 | 41.47 |
| | 0.5 | 38.25 | 48.94 | 49.10 | 54.18 | 53.54 | 53.87 |
| Foreman Object | | | | | | | |
| Full (level 1) | 0.25 | 32.16 | 37.98 | 36.58 | 32.21 | 38.46 | 36.91 |
| | 0.50 | 37.72 | 41.73 | 41.12 | 37.62 | 41.73 | 41.09 |
| | 1 | 43.23 | 45.65 | 45.75 | 42.94 | 45.61 | 45.70 |
| Half (level 2) | 0.125 | 27.59 | 31.70 | 30.38 | 27.78 | 35.35 | 33.41 |
| | 0.25 | 33.36 | 39.32 | 38.05 | 34.10 | 39.79 | 38.52 |
| | 0.5 | 40.01 | 44.35 | 43.59 | 41.57 | 44.34 | 43.68 |
| | 1 | 46.94 | 50.24 | 50.04 | 50.90 | 51.82 | 52.14 |
| Quarter (level 3) | 0.0625 | 21.31 | 20.31 | 21.17 | 22.95 | 23.05 | 23.12 |
| | 0.125 | 28.92 | 32.03 | 30.82 | 29.34 | 35.96 | 34.21 |
| | 0.25 | 36.43 | 41.79 | 41.11 | 39.34 | 41.90 | 41.90 |
| | 0.5 | 42.50 | 51.06 | 48.69 | 55.63 | 55.35 | 55.49 |

Table 5-1 compares PSNR results of OBHS-SPIHT and OB-SPIHT obtained for all colour components (i.e., Y, U and V) of the test objects at various spatial resolutions and bit rates. The OB-SPIHT results refer to our implementation of the original SPIHT algorithm for object-based coding, similar to [10, 9]. The results for spatial resolution level 1 clearly show that the OBHS-SPIHT does not sacrifice the compression efficiency of the OB-SPIHT. The small deviation between OBHS-SPIHT and OB-SPIHT is due to the different scanning order used by these algorithm in their coding process of wavelet co-

efficients. For resolution levels 2 and 3, as the results show, the performance of OBHS-SPIHT is much better than for OB-SPIHT. For these resolutions, the OBHS-SPIHT decoder decodes the bitstreams that were properly tailored by the parser for the given resolution level. For the OB-SPIHT case, there is no possibility of reordering the bitstreams for the requested resolution level. Therefore, the bitstreams were first decoded to obtain the coefficients in the wavelet pyramid for all subbands at the given bit rate. Then the requested spatial resolution was reconstructed by applying the inverse SA-DWT only to the required subbands for that resolution level. The reason for the superior performance of OBHS-SPIHT over OB-SPIHT for resolution levels higher than level 1 is clear. In the transcoded OBHS-SPIHT bitstream for a particular resolution all bits belong to that resolution, while in the OB-SPIHT bitstream, some portion of bits are wasted because they belong to the subbands which are located outside the requested resolution.

For the Y component, which is the most important component and consumes most of the coding budget, the results show more improvement than for the U and V components. The reason is that the U and V components are more correlated than Y and most of their energy (high coefficients in the wavelet decomposed image) are located in the lowest frequency band in the decomposed image. Only a few coefficients are, therefore, considered out of the required resolutions (level 2 and level 3) during the resolution-dependent sorting pass of the OBHS-SPIHT. However, the coding performance for these components (i.e., U and V) are high enough, even for low bit rates for both OB-SPIHT and OBHS-SPIHT. As the resolution level increases, the difference between OBHS-SPIHT and OB-SPIHT results becomes more and more significant. Moreover, for spatial resolutions lower than the full resolution, by increasing the bit rate, the OBHS-SPIHT shows more improvement. This is due to the fact that at higher bit rates, the bitplane coding reaches the low bitplane levels where the significance test threshold is decreased. As a consequence, more coefficients become significant, therefore, more information in the OB-SPIHT bitstream can be found that are not related to the required resolution, while the parsed OBHS-SPIHT bitstream only includes the information that belongs to the resolution.

Note that all the results are obtained without extra arithmetic coding of the encoder output bits. As shown in [4], an improved coding performance (about 0.3-0.6 dB) for SPIHT and consequently for OBHS-SPIHT can be achieved by further compressing the binary bitstreams with an arithmetic coder. Figure 5-5 shows some examples of multiresolution decoding of Akiyo and Foreman objects at three bit rates (0.125 bpp, 0.25 bpp and 0.5 bpp) obtained by the OBHS-SPIHT decoder.

Table 5-2 compares the compression efficiency of the OBHS-SPIHT algorithm with some state-of-the-art object-based coding methods for coding the

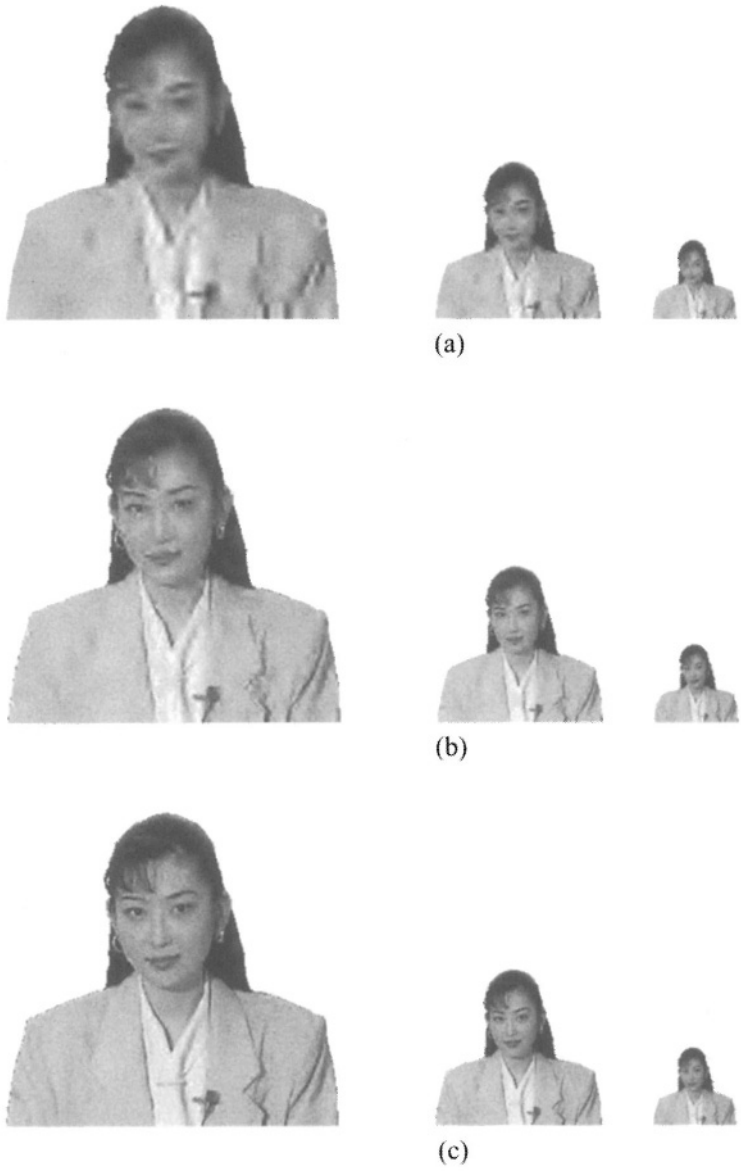



*Figure 5-5.* Luminance of the decoded Akiyo still object using OBHS-SPIHT at full (left), half (middle) and quarter (right) spatial resolutions and (a) 0.125 bpp, (b) 0.25 bpp and (c) 0.5 bpp.

foreground of the first frame of the Akiyo CIF sequence. The OB-SPECK coder [15] is an extension of a binary version of SPECK [16]. The SA-ZTE and SA-DCT methods are implemented in the MPEG-4 verification model reference software [2]. Egger's codec [17] uses a shape-adaptive wavelet transform and employs EZW [3]. The results for OB-SPIHT were obtained from our implementation of an extended version of the original SPIHT algorithm for object-based coding. Both the OB-SPIHT and OBHS-SPIHT results in this table were obtained from decoding the binary bitstreams without applying extra arithmetic coding. In addition to providing efficient compression performance, OBHS-SPIHT fully supports resolution scalability, while the other coders reported in this table are not resolution scalable.

*Table 5-2.* PSNR comparison for the foreground object of the first frame of the Akiyo CIF sequence at full spatial resolution.

| Coding Algorithm | Rate (bpp) | PSNR (dB) | | |
|---|---|---|---|---|
| | | Y | U | V |
| OBHS-SPIHT | 1 | 37.71 | 43.16 | 42.94 |
| OB-SPIHT | 1 | 37.77 | 42.99 | 42.88 |
| OB-SPECK [15] | 1 | 37.55 | 42.55 | 42.25 |
| MPEG-4 SA-ZTE [2] | 0.9538 | 38.06 | 43.43 | 43.25 |
| SA-DCT [2] | 1.0042 | 37.09 | 42.14 | 42.36 |
| Egger's SAWT [17] | 1.0065 | 36.40 | 42.53 | 42.40 |
| OWT [18] | 0.875 | 34.13 | - | - |

# 5. CONCLUSIONS

In this chapter, an object-based Highly scalable SPIHT algorithm (OBHS-SPIHT) for texture coding of arbitrarily shaped still objects has been presented. The proposed algorithm adds the spatial scalability feature to the SPIHT bitstream while keeping important features of the original SPIHT algorithm such as high compression efficiency and rate-embeddedness (very fine granular SNR scalability) of the bitstream. The flexible scalable bitstream of OBHS-SPIHT is easily reorderable (transcodable) without any need of decoding, to obtain different bitstreams tailored for different spatial resolutions and bit rates requested by the decoder. OBHS-SPIHT is a good candidate for multimedia applications such as object-based information storage and retrieval systems, and transmission of visual information especially over heterogenous networks.

# REFERENCES


1. ISO/IEC, "MPEG-4 video verification model version 18.0," *ISO/IEC JTC1/SC29/WG11 N3908,* Jan. 2001.

2. S. Li and W. Li, "Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding," *IEEE Trans. Circ. and Syst. for Video Technology,* vol. 10, no. 5, pp. 725–743, Aug. 2000.

3. J. M. Shapiro, "Embedded image coding using zerotree of wavelet coefficients," *IEEE Trans. Signal Processing,* vol. 41, pp. 3445–3462, Dec. 1993.

4. A. Said and W. A. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circ. and Syst. for Video Technology,* vol. 6, pp. 243–250, June 1996.

5. S. Li and W. Li, "Shape adaptive discrete wavelet transform for arbitrarily shaped texture," in *Proc. SPIE Conf. Visual Communications and Image Processing (VCIP'1997),* Feb. 1997, vol. 3024, pp. 1046–1056.

6. S. A. Martucci, I. Sodagar, T. Chiang, and Y.-Q. Zhang, "A zerotree wavelet video coder," *IEEE Trans. Circ. and Syst. for Video Technology,* vol. 7, no. 1, pp. 109–118, Feb. 1997.

7. A. Mertins and S. Singh, "Embedded wavelet coding of arbitrarily shaped objects," in *Proc. SPIE Conf. Visual Communications and Image Processing (VCIP'2000),* 2000, vol. 4067, pp. 357–367.

8. S. Li, W. Li, H. Sun, and Z. Wu, "Shape-adaptive wavelet coding," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'1998),* Monterey, CA., May 1998, pp. 281–284.

9. Y. Yuan and C. W. Chan, "Coding of arbitrarily shaped video objects based on SPIHT," *IEE Electronics Letters,* vol. 36, no. 13, pp. 1105–1106, Jun. 2000.

10. G. Minami, Z. Xiong, A. Wang, and S. Mehrotra, "3-D wavelet coding of video with arbitrary regions of support," *IEEE Trans. Circ. and Syst. for Video Technology,* vol. 11, no. 9, pp. 1063–1068, Sept. 2001.

11. H. Danyali and A. Mertins, "Highly scalable image compression based on SPIHT for network applications," in *Proc. IEEE Int. Conf. Image Processing (ICIP'2002),* Rochester, NY, USA, Sept. 2002, vol. 1, pp. 217–220.

12. H. Danyali and A. Mertins, "Fully scalable wavelet-based image coding for transmission over heterogeneous networks," in *Proc. 1st Workshop on the Internet, Telecommunications and Signal Processing (WITSP'2002),* Wollongong, NSW, Australia, Dec. 2002, pp. 173–178.

13. H. Danyali and A. Mertins, "Fully spatial and SNR scalable, SPIHT-based image coding for transmission over heterogenous networks," *Journal of Telecommunications and Information Technology, ,* no. 2, pp. 92–98, 2003.

14. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing,* vol. 1, pp. 205–220, Apr. 1992.

15. Z. Lu and W. A. Pearlman, "Wavelet coding of video objects by object-based SPECK algorithm," in *Picture Coding Symposium (PCS'2001),* Seoul, Korea, Apr. 2001, pp. 413–416.

16. A. Islam and W. A. Pearlman, "An embedded and efficient low-complexity hierarchical image coder," in *Proc. SPIE Conf. Visual Communications and Image Processing (VCIP'1999),* San Jose, CA, Jan. 1999, vol. 3653, pp. 294–305.

17. O. Egger, P. Fleury, and T. Ebrahimi, "Shape adaptive wavelet transform for zerotree coding," in *Proc. European Workshop Image Anal. Coding for TV, HDTV and Multimedia Applications,* Rennes, France, Feb. 1996, vol. 1, pp. 201–208.

18. H. Katata, N. Ito, T. Aono, and H. Kusao, "Object wavelet transform for coding of arbitrarily shaped image segments," *IEEE Trans. Circ. and Syst. for Video Technology,* vol. 7, no. 1, pp. 234–237, Feb. 1997.

*This page intentionally left blank*

Chapter 6

# CLASSIFICATION OF VIDEO SEQUENCES IN MPEG DOMAIN

Warwick Gillespie and Thong Nguyen
*School of Engineering University of Tasmania- Australia*

**Abstract:** This chapter describes a method for automatic classification of video shots from a video database by using distance metrics derived from motion information only. The classification serves as the first step of the indexing process of a video scene and its retrieval from a large database in order to partition the database into more manageable sub-units according to the types of scenes, e.g. sport, drama, scenery, news reading. The method is intended for web-based and telecommunication applications and therefore the processing is carried out in the MPEG (compressed) domain making use of the spatio-temporal data already available in MPEG video files. The confidence of the MPEG motion vectors estimated by the block matching algorithm is evaluated using a *block activity factor,* for retaining or discarding the vectors from the classification distance measure by a filtering process of the MPEG motion vector fields. The chapter presents a robust regression technique, based on Least Median-of-Squares, to deal with the situation. A novel metrics called activity power flow is introduced to effectively capture the spatio-temporal evolution of scenes through the video sequence.

**Key words:** video database, video signal processing, multimedia signal processing, pattern recognition.

## 1. INTRODUCTION

The use of visual information is a part of every day life, and with the advent of technology such as digital cameras, the amount of digital image and video data in the world is growing rapidly. Also, technologies such as the Internet and G3 and G4 cellular phones mean the ability to share and access that content is also rapidly increasing. With this rapid growth comes

the need for efficient management and storage of this digital media, to enable users to index and retrieve the video efficiently, reliably and quickly, which is the main challenge for designers of large multimedia databases. The proliferation of such large databases in recent times will continue with applications such as video on demand and Internet video browsing, as well as the need for media or advertising agencies to manage their visual data assets. In large video databases, particularly for access over a medium such as the Internet, it is likely that video sequences will be stored in a compressed format (such as the MPEG standards). As a result, the ability to use information contained in these formats without the need for decoding of the video sequences (a time consuming process) is desirable.

## 2. VIDEO DATABASE INDEXING AND RETRIEVAL

Much of the research into content based video indexing and retrieval (CBVIR) systems is concentrated in a few common areas, namely video summaries, similarity-based retrieval, and classification. One major hurdle in all these research efforts is the ability to transform low level features into high level semantics to enable key frames, key events, or similar scenes to be indexed and searched.

One of the main concepts when dealing with a CBVIR system (as opposed to a still image system) is how a video sequence can be broken into manageable segments. It is unsuitable to index full video sequences (eg a two hour long movie), hence the video sequence needs to be broken down into smaller semantic units, based on the video content. A common video structure used in CBVIR systems is shown in Fig. 6-1. In this structure, a whole video clip/sequence (such as a movie or television program) is broken down into scenes containing similar settings, story lines, or events. Each scene is then further segmented into shots, which are simply the portion of video between two camera breaks (such as a cut, a fade, a wipe, or a dissolve). The shot generally shows a single-camera perspective of a scene. A shot is made up of a number of frames - the basic temporal unit in a video, which are simply still images of a scene at a given time.

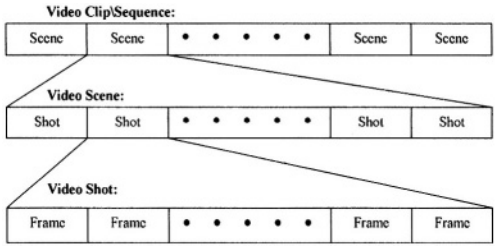


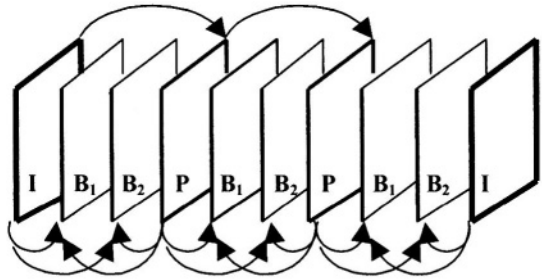

*Figure 6-1.* Structure of a video clip.

Generally, a video sequence is first broken down into shots, using any number of techniques [1], and each detected shot is then processed separately. In our system, shot detection is carried out in the MPEG domain using a modified version of an algorithm [1], which uses the distribution of macroblock types (see Section 3 of this chapter) to detect shot changes.

Once all shots are detected, often a key frame is selected to represent that shot, and features such as colour, object shape, texture, or motion can be used to characterise or index that shot [2]. Similar shots can be grouped together into scenes for semantic temporal segmentation. When searching or browsing video content, often a video sequence can be represented by key frames, which are either similar to a user query, or represent key events in a video sequence.

## 3. MPEG VIDEO COMPRESSION STANDARDS

As all video analysis techniques described in this chapter are performed in the MPEG (compressed) domain, it is important to have an understanding of some of the compression techniques used by the MPEG standards [3]. The MPEG-1 file has a hierarchical structure, like many computing based protocols, with each layer providing a different level of abstraction. In terms of a single video sequence, the top-most layer provides random access to a group of pictures (GOP) or frames with a common frame sequence (see Fig. 6-2). Each GOP in a video stream has the same frame sequence, starting with an I-frame, followed by a pattern of B and P-frames (the sequence in Fig. 6-2 is a common, but not necessarily standard sequence). I-frames are intra-coded (i.e. with no reference to other frames) so provide time references for random access to a video sequence, and prevent coding errors from propagating through a sequence. P and B-frames are inter-coded (i.e. with reference to other frames) so provide greater compression efficiency.




A typical Group of Pictures (GOP) in MPEG-1. **I:** Intra-frame coded, **P:** Predictive inter-frame coded, **B:** Bidirectional inter-frame coded.

*Figure 6-2.* MPEG Frame sequence.

MPEG is a block-based compression scheme, so frames are first divided into 16x16 pixel macroblocks (see Fig. 6-3). Blocks are coded using the YCrCb colour scheme where the Y (luminance) component represents the intensity information, and the Cr and Cb components represent the chrominance (or colour). The human visual system is more sensitive to the luminance component than the chrominance components, so they can be subsampled at Y:Cr:Cb = 4:1:1 (so each macroblock is coded using 4 8x8 blocks of Y and 1 8x8 block for each of the Cr and Cb components). For the case of I-frames, each 8x8 block is transformed into frequency domain using the Discrete Cosine Transform (DCT), and each coefficient is quantised and low energy components discarded.

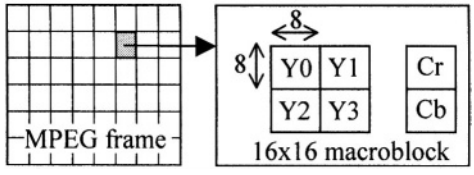



*Figure 6-3.* MPEG Block Structure.

The main function of the inter-coded frames (P and B-frames) is to take advantage of temporal redundancies. The MPEG-1 coding process uses motion vectors in order to do this. For each macroblock in a P or a B-frame, a search window in a reference frame (or frames) is examined to find a similar block. If a similar macro block is found, a reference to it (the motion

vector) as well as the difference between the actual and the reference macroblock is coded. The reference frames for P and B-frames are shown in Fig. 6-2, note that P-frames only reference previous frames, whereas B-frames reference both past and future frames. The motion vector search follows any of a number of common block matching algorithms (BMA) which search a reduced set of all possible blocks within the search window. The search attempts to minimise the mean square error (MSE) in Eq. (6.1), or mean absolute difference (MAD) between two blocks.

$$MSE(i,j) = \frac{1}{NM} \sum_m \sum_n \left| I_t(i+m, j+n) - I_{t-\tau}(i+m+u, j+n+v) \right|^2 \qquad (6.1)$$

The search usually terminates when a block is found with an MSE below a set threshold T. Some common BMAs include the two dimensional logarithmic search, the three step search, hierarchical motion estimation, and the signature based algorithm. Note that the motion search technique, the cost function, and the threshold T are not specified in the MPEG standard and as such will vary from encoder to encoder.

## 4. FILTERING MPEG MOTION VECTOR FIELDS

The goal of motion vectors in MPEG is to maximise compression, not to describe the true motion of a frame, and any prediction error will be compensated at the decoder. As a result there is a possibility that the BMA will not find a suitably matched reference block within the search window. The block will then be coded as an I-block using intra-coding techniques, hence motion flow information (i.e. motion vectors) is not available in the MPEG bit-stream for such blocks. This will occur if the motion is too fast such that the new location is beyond the search window; when the motion of an object occludes or uncovers an image segment (i.e. near the object boundaries); or when object or camera motion causes image segments to leave or enter the screen (i.e. near the screen boundaries).

While the presence of I-blocks limits the amount of useful data describing motion within a frame, the main weakness of a BMA technique is the possibility of the search being *prematurely terminated* at a wrong location. This can occur because the search stops at the first pixel that the MSE is smaller than the set threshold *T,* and if this first pixel is not where the MSE is at a minimum, then the search gives an erroneous motion vector. This premature termination can arise in a *dark region* where the image grey level everywhere is low resulting in a very small MSE below the threshold

*T;* in a *low activity uniform region* where the grey level everywhere is almost the same resulting also in a small MSE; or in a *regular region* where there are similar patterns, resulting in unpredictable termination of the LMSE search.

## 4.1 Activity measure to detect erroneous motion vectors

As low activity blocks may be coded with unreliable motion vectors, these motion vectors should be discarded from further motion based processing [4]. The AC power can be used as a measure of the activity of a block, and can also be easily calculated in frequency domain from DCT coefficients contained in MPEG stream. Hence the relationship between the MSE of a BMA and the AC power of a block must be investigated. Now we express the MSE in Eq. (6.1) in terms of expectations to better show the power in the image, i.e.

$$MSE\ (i, j) = \mathrm{E}[(dc_t + ac_t - dc_{t-\tau} - ac_{t-\tau})^2] \qquad (6.2)$$

From the brightness constancy assumption, we can assume the DC level does not change much between consecutive frames hence the relation in Eq. (6.2) can be simplified to:

$$MSE(i, j) \approx \mathrm{E}[(ac_t - ac_{t-\tau})^2] \le \mathrm{E}[(2ac_t)^2] \le 4 \times Power_{ac_t} \qquad (6.3)$$

Hence, if the AC power is used as a measure of the activity of a block, then the activity threshold, $T_a$, is simply a quarter of the threshold $T$ used in the block matching MSE calculation. For blocks with activity below the threshold $T_a$ there is a potential for the motion vector search to be prematurely terminated, so these blocks should be discarded from further processing. Unfortunately, the threshold $T$ is not defined in the MPEG standard thus it can vary, and is often not possible to determine (especially in proprietary video codecs). As a result we need to determine a best estimate for the activity threshold, $T_a$.

The motion vector fields from a large number of frames from various MPEG sequences were studied, and motion vectors that did not correspond to object or camera motion within the frame were marked as unreliable. Then for varying levels of $T_a$, the percentage of unreliable motion vectors discarded, and the percentage of the correct motion vectors surviving the thresholding was calculated. We found the range for $T_a$ that provided the optimum number of unreliable motion vectors deleted compared with the

number of correct motion vectors surviving the thresholding process was from 2.5 to 3.5. An example of this filtering process can be seen in Fig. 6-5.

## 4.2 Estimating Activity for MPEG P-blocks

The activity of a block can be very simply calculated for an I-block using the DCT coefficients present in the MPEG bitstream i.e.

$$Activity = \sum_{i\&j\neq 0} c_{ij}^2 \qquad (6.4)$$

where $c_{ij}$ is the ac coefficient corresponding to position (i,j) in the DCT block.

Inter-frame coded blocks however, do not have DCT coefficients available, hence the activity of those blocks has to be estimated from the blocks in the previous frames using the motion vector reference. For each block in the first P-frame of the GOP (see Fig. 6-2), the reference block from the previous I-frame is found by projecting the motion for that block back on to the reference I-frame. The reference block may not overlay an exact macro-block in the previous frame. As a result, the activity measure is estimated from the (up to four) macroblocks that it does overlay. This is done using a simple weighted average (i.e. contribution of a block is proportional to the area overlaid in that block). In our system, the activity of a block is calculated solely from the four 8x8 luminance blocks of a 16x16 macroblock because chrominance signals carry no extra activity information.

# 5. CAMERA MOTION ESTIMATION

When analysing the motion within a video sequence, an important step is being able to separate the motion of the camera (global motion) and the motion of objects (local motion). In order to do this we first calculate the global motion within a frame using a robust Least Median of Squares estimator to estimate the parameters in a 2-D global motion model.

We use the 4-parameter simplified affine global motion model [5,6] in Eq. (6.5) where $[u_{ix}\ u_{iy}]$ is the motion vector at the pixel $(x_i, y_i)$ in a frame, and *a1, a2, Px, Py* are the model parameters, where *a1* describes the zoom, *a2* the rotation, *Px* the pan (in the x-direction), and *Py* the tilt (in the y-direction) of the camera.

$$\begin{bmatrix} u_{ix} \\ u_{iy} \end{bmatrix} = \begin{bmatrix} a1 & a2 \\ -a2 & a1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} Px \\ Py \end{bmatrix} \tag{6.5}$$

We use the motion vectors from an MPEG file as the inputs, after discarding unreliable blocks (Section 4), hence for a frame of $N$ reliably predicted macroblocks, we have $N$ sets of linear equations, as in Eq. (6.5), to describe the frames global motion, which in matrix form becomes

$$Y = HX \tag{6.6}$$

where

$$Y = \begin{bmatrix} u_{1x} \\ u_{1y} \\ \vdots \\ u_{nx} \\ u_{ny} \end{bmatrix}, \qquad H = \begin{bmatrix} x_1 & y_1 & 1 & 0 \\ y_1 & -x_1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & 1 & 0 \\ y_n & -x_n & 0 & 1 \end{bmatrix}, \qquad X = \begin{bmatrix} a1 & a2 & Px & Py \end{bmatrix}$$

In order to estimate the model parameters the classical Least Squares (LS) regression technique is usually used and the pseudo-inverse solution for $X$ is

$$X = (H^T H)^{-1} H^T Y \tag{6.7}$$

This technique, whilst being optimal for data contaminated by Gaussian noise, is extremely inaccurate in the presence of *outlier* data. In terms of global motion, the motion vector field used in the estimation (i.e. after filtering) could still contain many outliers. These outliers are due to either object motion, or blocks at the borders of low activity regions in a frame.

We use a robust Least Median-of-Squares (LMedS) technique proposed by Rousseeuw [7] in which the optimal estimate is obtained by solving the *non-linear* minimisation problem

$$\min_i . med\{r_i^2\} \tag{6.8}$$

in which $r_i$ is the estimation error or residual. If we let the MPEG motion vector of macroblock $\mathbf{MB}_i$ be $\hat{\mathbf{u}}_i = [\hat{u}_{ix}\ \hat{u}_{iy}]$, and the global motion vector be $\mathbf{u}_i = [u_{ix}\ u_{iy}]$, then

$$r_i^2 = \left(\hat{\mathbf{u}}_\mathbf{i} - \mathbf{u}_\mathbf{i}\right)^2 = \left(\hat{u}_{ix} - u_{ix}\right)^2 + \left(\hat{u}_{iy} - u_{iy}\right)^2 \tag{6.9}$$

that is, the optimal estimate gives the smallest value for the median of the squared residuals of the entire data set. The concept behind a LMedS estimator is that for a *p*-parameter model (for global motion model in Eq. (6,5), *p*=4), a *p*-tuple sample is a good fit to the estimator model if the sample contains *p* good data points (i.e. without outliers). As a result we can choose *m* random *p*-tuple subsamples to fit the estimator model, and then use the median of the residuals to select the best fit. If $\varepsilon$ is the fraction of outliers in the data, then the probability of error that can be tolerated by taking only *m* randomly chosen *p*-tuples instead of all possible combination of tuples is [7]:

$$P = \left[1 - \left(1 - \varepsilon\right)^p\right]^m \tag{6.10}$$

or

$$m \geq \frac{\log P}{\log[1 - \left(1 - \varepsilon\right)^p]} \tag{6.11}$$

Using conservative values of $P$=0.01 and $\varepsilon = 0.4$ in Eq. (6.11) gives $m$ = 33 sub-samples. The value of $\varepsilon$ depends primarily on the extent of object motion in the frame.

For each *p*-tuple subsample *j,* we use Eq. (6.7) to calculate the model parameters $\boldsymbol{X_j}$ and from which we can calculate the global motion field using Eq. (6.5) for every macroblock *i* in the frame that contains a reliable motion vector. Now, for each subsample *j*, the median $M_j$ of the squared residuals $r_i$ is calculated, i.e.

$$M_j = med\left\{r_i^2\right\} \tag{6.12}$$

We need to determine a weight $w_i$ for the *i*th observation using a scale estimate [7] $\sigma^*$ of the dispersion from the median $Mj$, such that outliers are effectively discarded (having zero weight) from data i.e.

$$w_i = \begin{cases} 1 & if \quad r_i^2 \leq \left(2.5\sigma^*\right)^2 \\ 0 & otherwise \end{cases} \tag{6.13}$$

where

$$\sigma^* = \sqrt{\frac{\sum_i w_i r_i^2}{\left(\sum_i w_i\right) - p}} \tag{6.14}$$

It is obvious that to start the outliers identification process in the LMedS technique, we need an *initial* value for the scale estimate $\sigma^*$. This is proposed as [7]

$$\sigma^0 = 1.4826\left(1 + \frac{5}{n-p}\right)\sqrt{M_j} \tag{6.15}$$

in which $M_j$ in Eq. (6.15) is calculated from residuals $r_i$ of the LS- estimate in Eq. (6.7) using *all* data points. The factor 1.4826 is for consistent estimation in the presence of Gaussian noise, the term 5/(n-p) is the finite sample correction factor. Each iteration of Eq. (6.13) and Eq. (6.14) discards some more outliers and the iteration stops either when $\sigma^*$ converges to less than 1% of its previous value or when the set limit of say 20 or 30 iterations is reached. It is suggested [8] that once the above LMedS technique has discarded most outlier data points, the technique can be refined by a LS procedure on the remaining mainly inlier data points, an example of which can be seen in Fig. 6-5.

# 6. MOTION BASED CLASSIFICATION OF VIDEO SEQUENCES

In this Section, we describe a method that can successfully classify video shots into broadly defined video genres [9], S:Sport, N:News, D:Drama and O:Outdoor or Scenery. We use only low-level motion based metrics to characterise a video shot, and then use a clustering network to group shots from similar genres together.

## 6.1 Spatio-temporal evolution metrics

The motion in a video sequence can be simply defined as how the spatial content of the frames change in time, and a simplified technique for describing this, called the *activity power flow,* is presented [9] (Section 6.1.1). We also use the average object motion intensity (Section 6.1.2) in order to characterise the motion in a video shot.

### 6.1.1 Activity power flow in a video sequence

In Section 4 of this chapter we present a technique of detecting macroblocks of an MPEG frame with unreliably coded motion vectors using the activity of a block. We thus divide the macroblocks of an MPEG frame into three categories: those with reliable predicted motion vectors (C), those with unreliable predicted motion vectors (W), and those that are intra-frame coded (I). The total number of macro-blocks in a frame is therefore

$$N=I+C+W \tag{6.16}$$

We next define the total *activity power* in one frame in the three respective categories as $P_I = I\text{x}AF_I$, $P_C = C\text{x}AF_C$, and $P_W = W\text{x}AF_W$, respectively, where $AF_b$, $b=I,C,W$, is the average *activity factor,* which is defined as the AC power in the block [4], in that category. We use the second-order statistics, i.e. the mean $\mu_b$ and the variance $\sigma_b^2$, as metrics for the activity power flow. The mean gives a coarse measure of the spatial content of a sequence, and the variance represents camera and object motion.

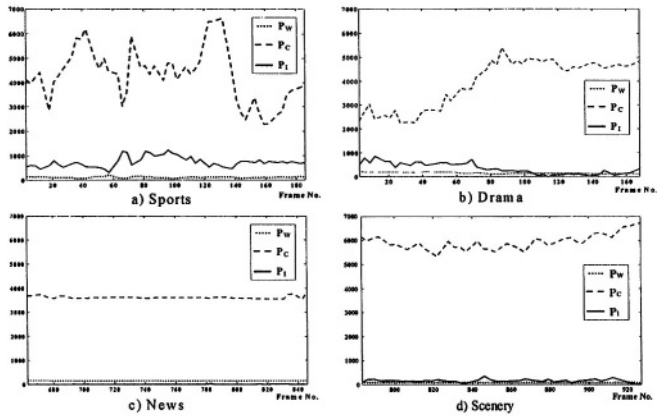



*Figure 6-4.* Example of activity power flow in the three categories of macro-blocks in an MPEG video shot, (a) Sport, (b) Drama, (c) News, (d) Scenery. [9]

It can be seen from Fig. 6-4 (a) that there are large variations in $P_C$ and $P_I$, showing that sports shots contain highly irregular camera and object motion. In contrast, Fig. 6-4 (b) shows that Drama shots can contain variations in camera and object motion, but the variations are generally much smoother. Video shots of News (Fig. 6-4 (c)) and Scenery (Fig. 6-4 (d)) contain very low $P_I$ due to the lack of object motion, but scenery clips generally contain some camera motion resulting in a slight variation in $P_C$.

### 6.1.2 Motion Intensity

The motion intensity is a measure of the average magnitude of motion in a frame and is separated into two categories. The *global motion intensity, G,* is calculated from the global motion field (see Section 5) using Eq. (6.17), and the *object motion intensity, O,* is calculated using Eq. (6.18) by first subtracting the global motion from the MPEG motion vectors (see Fig. 6-5).

$$G = \frac{1}{C}\sum_{i=1}^{C}\sqrt{(u_{ix})^2 + (u_{iy})^2} \tag{6.17}$$

$$O = \frac{1}{C}\sum_{i=1}^{C}\sqrt{(\hat{u}_{ix} - u_{ix})^2 + (\hat{u}_{iy} - u_{iy})^2} \tag{6.18}$$

where C is the number of macroblocks with reliable motion vectors, $\hat{\mathbf{u}}_i = [\hat{u}_{ix}\ \hat{u}_{iy}]$ is the MPEG motion vector of macroblock $MB_i$ and $\mathbf{u}_i = [u_{ix}\ u_{iy}]$ is the global motion vector. Like the activity power flow, the distribution of motion intensity across a shot is represented with the second order statistics, the mean $\mu_G$ and $\mu_O$, and variance $\sigma_G^2$ and $\sigma_O^2$ of the motion intensity of the P-frames in a sequence.

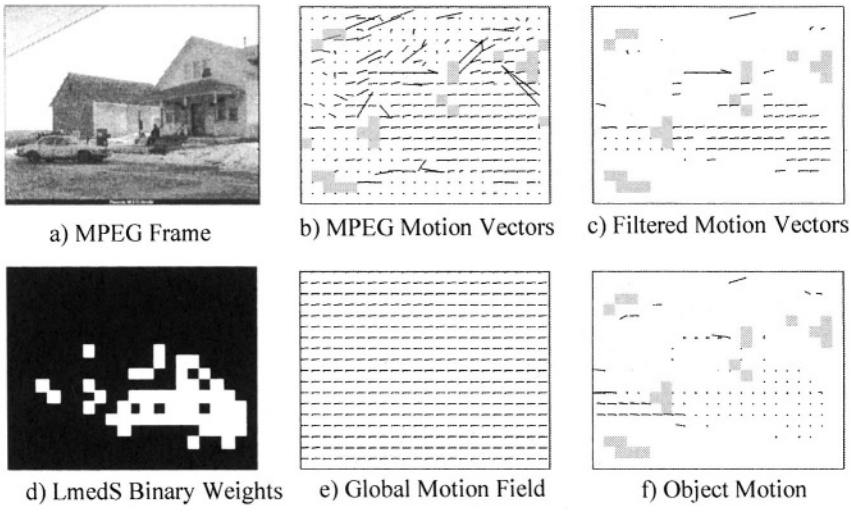



*Figure 6-5.* Motion vector field processing results for a frame with a pan to the right tracking a car (bottom left of frame). N.B. Shaded blocks in motion vectors fields are I-blocks.

## 6.2 RBF Network Classification

A simple clustering network such as a radial basis function (RBF) network can be trained from a large sample of shots from MPEG sequences (news, sport, drama, scenery, etc.) to classify the shot content [9]. We propose that a video shot can be suitably characterised by an 10-dimensional feature vector

$$\boldsymbol{x} = (\mu_I, \sigma_I^2, \mu_C, \sigma_C^2, \mu_W, \sigma_W^2, \mu_G, \sigma_G^2, \mu_O, \sigma_O^2) \qquad (6.19)$$

and this vector is used as the input to the RBF (see Fig. 6-6) for classification into one of the four categories.

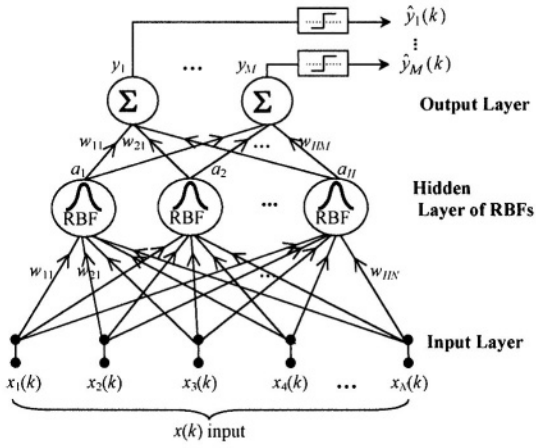



*Figure 6-6.* Diagram of RBF network structure with 10-dimensional input vector, H=4-node hidden layer, and M=4-node output layer.

We generalise the activation of the RBF nodes to a non-symmetrical Gaussian receptive field by using the square of *Mahalanobis* distance in the Gaussian function. The output of the *h*th hidden node ($1 \le h \le H$, and $H$=4 video classes) due to input sample $\boldsymbol{x}(k)$ ($k$=1,2,...K large) is therefore

$$a_h(k) = \exp[-\mathbf{w}_h (\hat{\mathbf{x}}_h - \mathbf{x}(\mathbf{k}))^T \Sigma_h^{-1} (\hat{\mathbf{x}}_h - \mathbf{x}(\mathbf{k}))] \qquad (6.20)$$

where $\hat{\mathbf{x}}_h$ and $\mathbf{x}(\mathbf{k})$ are, respectively, the *10*-dimensional position vector of the centre of the $h^{th}$ hidden node and the input test sample to the node, and $\Sigma_h$ is the distance scaling matrix of the node's receptive field. $\Sigma_h$ gives the width or the spread of influence of the node and is simply the covariance

matrix of the training samples assigned to, or captured by the $h^{th}$ node cluster. Furthermore, if the components of the feature vector $\boldsymbol{x}$ are uncorrelated to one another (this assumption is not quite true with $\boldsymbol{x}$ in Eq. (6.19) but is used to save computing time), then

$$\Sigma_h^{-1} = \begin{bmatrix} \sigma_{11}^{-2} & 0 & \ldots\ldots & 0 \\ 0 & \sigma_{22}^{-2} & \ldots & 0 \\ \ldots & & & \\ 0 & 0 & & \sigma_{1010}^{-2} \end{bmatrix} \tag{6.21}$$

in which $\sigma_{ii}^2$ are the variances of the ten components of the characteristic vector $\boldsymbol{x}$ in Eq. (6.19). The vector $\mathbf{w_h}$ contains weights connecting the $i^{th}$ input to the $h^{th}$ hidden node. The squared error of the hidden nodes in response to the $k^{th}$ training sample is minimized to train $\mathbf{w_h}$ using a gradient descent approach [10], which iteratively adjusts the weights using Eq. (6.22), in which $\eta$ is a learning rate which is set to 0.02, and $T_h(k)$ is the desired activation of the $h^{th}$ hidden node in response to the $k^{th}$ training sample.

$$\Delta w_{hi}(k) = -\eta \frac{a_h(k)}{a_h(k)_{\max}} \left( T_h(k) - \frac{a_h(k)}{a_h(k)_{\max}} \right) \frac{\left| x_i(k) - \hat{x}_{hi} \right|^2}{\sigma_{hi}^2} \tag{6.22}$$

The output layer of the RBF network consists simply of *linear summation* units with linear activation in contrast to the sigmoid activation of the output nodes in a multilayer perceptron. The network output from the $m^{th}$ node due to the data input vector $\boldsymbol{x}(k)$ is therefore

$$y_m(k) = \sum_{h=1}^{H} w_{hm} a_h(k) \tag{6.23}$$

where $w_{hm}$ is the coefficient or the weight from the $h^{th}$ hidden node to the $m^{th}$ output node ($1 \le m \le M$, and $M=H$). The $M$x$H$ weight matrix $\boldsymbol{w}$ can be simply *computed* so that the norm-2 of the output error is minimised. The well known pseudo-inverse solution is

$$\mathbf{w} = \mathbf{T}\mathbf{a}^T \left( \mathbf{a}\mathbf{a}^T \right)^{-1} \tag{6.24}$$

where $\boldsymbol{T}$ is the desired (target) $M$x$K$ output matrix as the result of applying $K$ training samples to the network, and $\boldsymbol{a}$ is the $H$x$K$ activation matrix (output) from the hidden layer in which activations below a small threshold are excluded to avoid instabilities of the pseudo-inverse solution in Eq. (6.24).

This supervised training process is done using the input feature vector from 30 video shots per hidden node, satisfying the requirements regarding the size of the training set [11].

## 7. RESULTS AND CONCLUSIONS

A database consisting of 240 video shots was built. This database consisted of 60 video shots from each category (S, N, D, O). The sports clips came from a wide range of sports including Aussie Rules Football, Rugby, Field Hockey, Cycling, Surfing, Snow-Skiing, Netball, and others. The shots from the news category were all of newsreaders in the studio, as it is believed that any classification system would need to look for shots of this nature to identify news sequences. The shots were gathered from a variety of different news programs. The scenery category contains clips varying from cityscapes to natural scenes, but with no object motion and smooth camera motion, such as long pans, as is typical for this genre of program. The drama category has the broadest range of shots. All shots in this category contain human interaction, but the spatial content other than this can vary greatly. The camera and object motion can also vary greatly from shot to shot, although generally within a shot the variations are smoother than in the sports category.

The RBF was trained with 30 shots from each category randomly selected from the database. After training the network, all 240 shots were presented to the network for classification. This process was undertaken ten times and the average success of the classification is presented in Table 6-1.

*Table 6-1.* RBF Network Classification Results

| | | RBF Classification | | | |
|---|---|---|---|---|---|
| | | S | N | O | D |
| Category | S | 0.93 | 0.00 | 0.03 | 0.04 |
| | N | 0.01 | 0.85 | 0.09 | 0.05 |
| | O | 0.09 | 0.06 | 0.74 | 0.11 |
| | D | 0.17 | 0.16 | 0.18 | 0.49 |

Sport is the best performed category as this domain is the most specific, containing large variations across a video shot, with some periods of high object and camera motion and other periods which are quiet. The other well defined category is News, which contains almost no motion and no variation from frame to frame. The classification errors present in this category are due to the broad nature of the two other categories. While they both contain less variation within a shot than those in the sport category, the amount of motion from shot to shot is not as well defined, especially for drama.

Consequently, the clustering of the broader categories is not as precise and clusters will overlap, resulting in errors in classification. Consequently, other visual features such as face detection, and audio features may be needed to enable more robust clustering. These results are promising enough to indicate the possibility of using these techniques as the first step in an indexing process in order to minimise the search space by partitioning a video a database into smaller clusters based on the genre of the video sequences.

## REFERENCES


1. W.A.C. Fernando et al., Scene change detection algorithms for content-based video indexing and retrieval, *IEE Electronics and Communication Engineering Journal,* 117-125 (June 2001).
2. D.T. Nguyen and W. Gillespie, A Video Retrieval System Based on Compressed Data from MPEG Files, *Proceedings of IEEE TENCON 2003,* Bangalore, India (October 2003)
3. MPEG-1, ISO/IEC 11172-2, 'Information Technology – Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s', Part 2: Video, (1993).
4. W. Gillespie and D.T. Nguyen, Filtering of MPEG Motion Vector Fields for use in Motion-Based Video Indexing and Retrieval, *Proceedings 7th International Symposium on Digital Signal Processing for Communication Systems,* Gold Coast, Australia, 8-11 (December 2003).
5. J.M. Odobez and P. Bouthemy, Robust Multiresolution Estimation of Parametric Motion Models, *Journal of Visual Communication and Image Representation,* **6**(4), 348-365 (December 1995).
6. K. Jinzenji, S. Ishibashi, and H. Kotera, Algorithm for automatically producing layered sprites by detecting camera movement, *Proceedings IEEE International Conference on Image Processing ICIP1997,* 767-770 (November 1997)
7. P. J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection,* (John Wiley, 1987).
8. P. Meer, D. Mintz, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review", *Int. Journ. of Computer Vision,* **6**(1), 59-70 (1991).
9. W. J. Gillespie and D.T. Nguyen, Classification of Video Shots Using Activity Power Flow, *IEEE Consumer Communications and Networking Conference CCNC2004,* Las Vegas, USA, (Jannuary 2004).
10. T. A. Hoang, Wavelet-Based Techniques for Classification of Power Quality Disturbances, PhD Thesis, School of Engineering, University of Tasmania, Oct 2002
11. Sascha Spengenberg, The RBF Network Receiver, http://www.ee.ed.ac.uk/~ssp/project/html/, site last viewed Oct 2003.

# PART 2:

# ERROR-CONTROL CODING, CHANNEL ACCESS, AND DETECTION ALGORITHMS

*This page intentionally left blank*

Chapter 7

# UNEQUAL TWO-FOLD TURBO CODES

*Application to Digital Image Transmission*

Cagri Tanriover[1] and Bahram Honary[2]
[1] *HW Communications Limited, St George's Works, St George's Quay, Lancaster, LA1 5QJ, UK;* [2]*Department of Communication Systems, Lancaster University, LA1 4YR, UK*

**Abstract:** Two-fold turbo codes have been recently introduced as the least complex member of the multifold turbo code family, and they have been shown to outperform their conventional symmetric and asymmetric cousins in the waterfall region. This article introduces a simple method that allows two-fold turbo codes to offer bi-level unequal error protection. Also, the application of the unequal error protection to a still image coding technique, i.e. hue-saturation-luminance coding, is explained and the pixel error performance results under Gaussian conditions are presented. According to the simulation results, the proposed scheme outperforms its conventional counterpart by up to 0.7 dB. Furthermore, with the new unequal error protection coding scheme, under certain channel conditions, significant savings in decoder iterations are achieved, without compromising the received image quality. Although the unequal error protection is limited to bi-level within the scope of this article, same technique can be used to provide multilevel error protection.

**Key words:** Turbo codes, iterative coding, image transmission, multifold turbo codes, two-fold turbo codes, unequal error protection, multimedia data transmission

## 1. INTRODUCTION

Recently, a new approach to turbo coding technique, i.e. multifold turbo coding [1, 2, 3], has been introduced that significantly improves the error performance of conventional turbo codes in the waterfall region. The key concept behind multifold turbo coding is to enhance the randomness of the conventional turbo codes [4], which is mainly injected by the pseudo-random interleaving, and to bring the weight spectrum closer to binomial weight distribution of a random code [5, 6]. This is achieved simply by

dividing the information bit sequence into equally sized segments, and taking different combinations of those segments such that each segment appears in at least two groups. Each combination is consequently interleaved and encoded as in the conventional turbo coding scheme. At the decoder, multiple estimates per segment are generated during iterative decoding. It should be noted that the number of estimates per segment (also referred to as the 'fold' of the code) is equal to the number of groups in which a segment appears. Therefore, higher the fold of the code, the more reliable the decoding becomes.

So far the multifold turbo coding technique has been applied in its two-fold form, where each segment appears in exactly two groups. Two-fold technique has been applied to both symmetric and asymmetric turbo codes with well known component codes such as Berrou's g[37,21], g[23,35] (i.e. primitive 16 code), CCSDS' g[23,33], and Costello's Big Numerator Accumulator (BNAC) codes. In all simulations two-fold turbo codes have outperformed their conventional counterparts with moderate information blocks such as 4096 and 8192 bits. For the details of this study the reader is referred to Tanriover [2] et al.

This chapter describes the two-fold encoding in detail and introduces a simple technique that allows two-fold turbo codes to offer bi-level unequal error protection (UEP). To demonstrate the practicality of the UEP scheme, transmission of Hue-Saturation-Luminance (HSL) coded still digital images is also presented. The scalable structure of the two-fold based UEP technique, and its suitability to different types of data transmission are also discussed in the later sections of the article.

## 2. TWO-FOLD TURBO CODING

Two-fold turbo codes are the least complex members of the multifold turbo code family. We start by describing the coding procedure for the two-fold codes using Fig. 7-1 as a reference. Each information block at the input of the two-fold encoder needs to be split into three segments of equal length. Therefore, we can represent an information block $I$ as the concatenation of three segments as in equation (7.1), where the special operator <> represents serial concatenation.

$$I\{123\}=I\{1\}\diamond I\{2\}\diamond I\{3\} \qquad (7.1)$$

Prior to interleaving, by taking two segments at a time, all possible segment couples are concatenated as indicated in equations (7.2) through (7.4). All the concatenated segment couples, except one, are then interleaved

using different interleaving sequences. In our example in Fig. 7-1, I{12} and I{13} are interleaved whereas I{23} is left as is.

$$I\{12\}=I\{1\}\langle\rangle I\{2\} \tag{7.2}$$

$$I\{13\}=I\{1\}\langle\rangle I\{3\} \tag{7.3}$$

$$I\{23\}=I\{2\}\langle\rangle I\{3\} \tag{7.4}$$

Using the same recursive systematic convolutional (RSC) encoder, all the segment couples (both uninterleaved and interleaved) are encoded and the corresponding parity sequences are stored. In Fig. 7-1, the parity sequences generated for I{12}, I{13}, and I(23} are P{12}, P{13}, and P{23}, in respective order. Note that the information block, I{123}, appears as part of the codeword since the code is systematic. The mother code rate of a two-fold turbo encoder is always 1/3, which can be increased by puncturing codeword. However, in the scope of this chapter, only the unpunctured two-fold turbo codes will be considered.

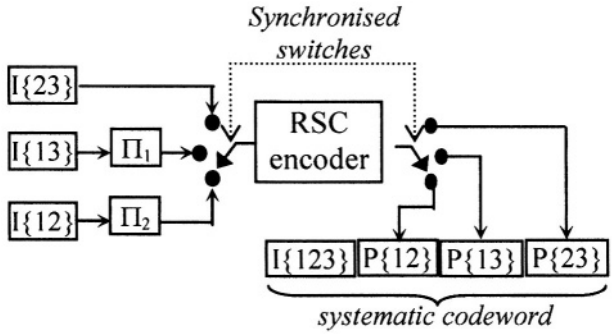



*Figure 7-1.* RSC two-fold encoder

Due to the structure of the two-fold turbo coding, each information segment is encoded twice as part of two different segment couples. At the decoder, for each segment two different estimates from two different constituent decoders are obtained, whereby the reliability of a posteriori information is improved in comparison to conventional turbo decoding.

Another advantage of two-fold turbo codes is that they introduce a remarkably low decoding complexity overhead compared to their conventional cousins while achieving better error performance. Although an

extra component decoder is required for decoding two-fold codes, as each component decoder has to deal with an information block that is 1/3 times shorter than a conventional turbo component decoder's, the overall constituent decoder complexity stays the same. However, two-fold decoding does introduce a small complexity overhead compared to conventional turbo decoding, which is due to the demultiplexing and multiplexing operations between consecutive iterations, which are shown in Fig. 7-2. The dashed horizontal lines represent the constituent decoder pipelines. $L_k\{ab\}$ denotes the extrinsic output of decoder k, associated with the combination of segments a and b. L` stands for the extrinsic input to a constituent decoder. As can be seen from Fig. 7-2, after each iteration, the output of each decoder is demultiplexed to its constituent segments, which are subsequently multiplexed with the associated segments from other decoders, before the next iteration. Note that during multiplexing, segments from a decoder are never fed back to it prior to the following iteration. Multiplexing and demultiplexing operations for the two-fold turbo decoders discussed here increase the complexity by 0.01% in comparison to their conventional counterparts [3].

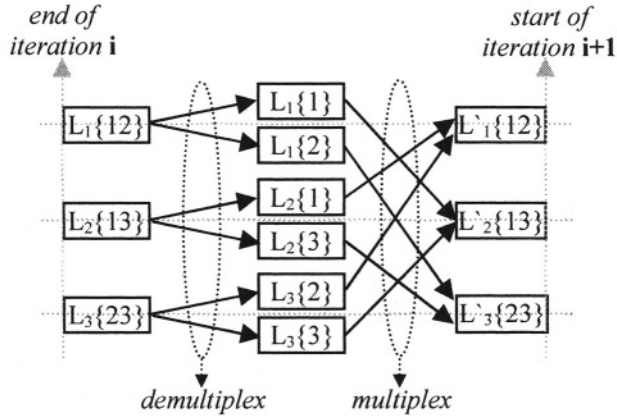



*Figure 7-2.* Demultiplexing and multiplexing between iterations.

## 3. TWO-FOLD BASED UEP

In the two-fold coding scheme described until now, as each segment is encoded twice, every information bit is protected equally. However, by a simple modification, the same scheme can be used for providing two-level

unequal error protection as well. More importantly, the UEP scheme described for two-fold turbo codes can also be extended to higher fold turbo codes to offer multilevel UEP.

Fig. 7-3 shows the modified two-fold encoder in Fig. 7-1, which introduces an additional protection level to segment 1. This is achieved simply by concatenating the first information segment with the information segment couple I{23}, and leaving I{13} and I{12} unchanged. Note that this simple modification slightly reduces the mother code rate from 1/3 to 3/10, as P{23} is transformed to P{123}, which is 50% longer. The encoding procedure remains exactly the same as described in section 2.

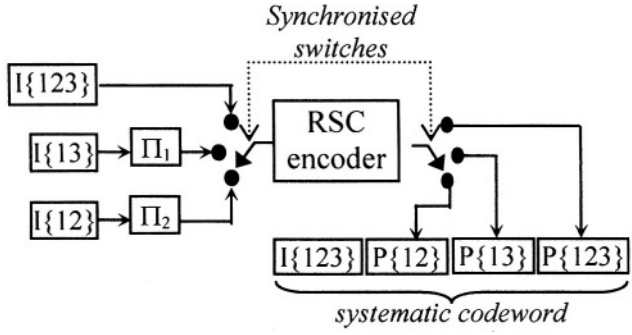



*Figure 7-3.* RSC two-fold encoder used for UEP

At the decoder, as segment 1 is encoded three times, three different likelihoods are generated for that segment. Similarly, if segment 2 was to be protected more heavily than segments 1 and 3, in this case I{13} in Fig. 7-1 would be extended to I{123}. A similar modification could also be made to better protect segment 3 compared to others. Increasing the length of I{12} and I{13} would cause the length of interleavers $\Pi_1$ and $\Pi_2$ and the corresponding deinterleavers at the decoder to increase as well, and hence higher the overall processing overhead. Therefore, from an implementation point of view, treating segment 1 as the better protected segment at all times is beneficial as no interleaving/deinterleaving overhead is introduced.

## 4. UEP APPLICATION TO STILL IMAGE TRANSMISSION

The two-level UEP offered by the modified two-fold turbo codes is well-suited for the transmission of hue-saturation-luminance (HSL) coded digital

images. Description of colour using hue, saturation and luminance has been introduced by the International Commission on Illumination (CIE) [7], and relates very closely to human perception of colour.

As the human eye is known to be more sensitive to light intensity changes than to colour changes, among the three components, the luminance is the most significant. Hue and saturation components add the colour vector to the light intensity and they carry equally significant perceptive information, to which the human eye is less sensitive. In order to illustrate the information content of HSL images, consider Fig. 7-4, where the source image 'truck' and its HSL decomposition are presented.

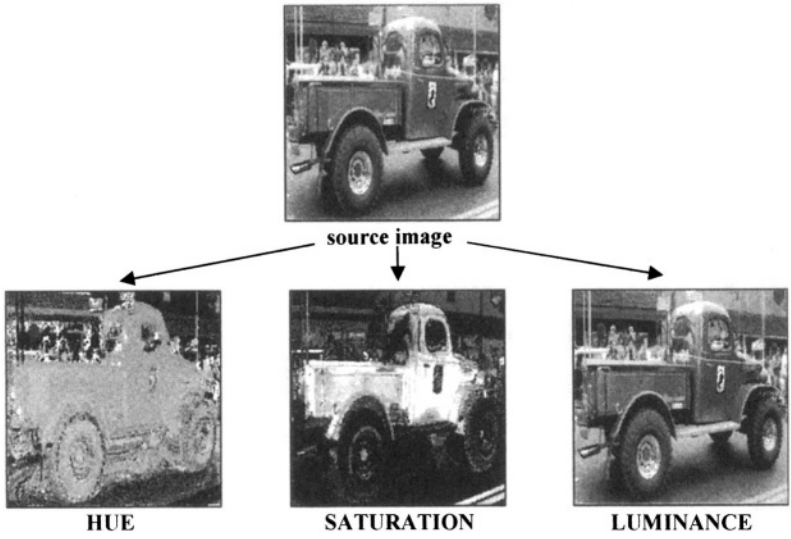



*Figure 7-4.* Hue, saturation and luminance decomposition

Each pixel in the source image is represented by three bytes (one byte per HSL component). In order to decompose the source image, bytes corresponding to each colour component are mapped to a new plane, whereby a grey-scale image for that component is constructed (Fig. 7-4).

Notice how limited the image detail is on the hue and saturation planes. This is because the hue and the saturation of the source image do not change as fast as the luminance, and therefore the H and S planes have visually static areas on them. However, by only looking at the luminance plane, one can easily perceive almost all image details, such as the contours, shapes, and light changes, in the source image.

Due to the sensitivity of our eyes to the light variations, pixel errors on the L plane are more easily noticed that the ones on the H and S planes, and therefore during transmission over a noisy channel, the L component needs to be better protected than the other two. Equal error protection can be applied to the H and S planes, as there is no significant difference in the information they carry. Therefore, in the proposed two-fold based UEP scheme, segment 1 is dedicated to the transmission of the L plane, whereas segments 2 and 3 are allocated to the H and S planes, respectively.

## 5. PIXEL ERROR RATE (PER)

While assessing the error performance of still image transmissions, considering pixel errors rather than bit or frame errors closely represents how the human eye actually receives channel errors. Depending on the location of bit errors, visually, two still images with identical bit error rates may be perceived quite differently [3, 8]. Therefore, the error performance results presented in this chapter are in terms of PER, which is described in the rest of this section.

In order to quantify the visual disturbance ($\Delta$) introduced by the channel noise, the colour difference between the received and the transmitted pixels needs to be calculated. For an n-colour digital image, the maximum possible colour displacement between any two received pixels is (n-1). If, for a pixel i, the transmitted and the decoded pixel values are denoted as, $t_i$ and $r_i$, respectively, $\Delta$ for i can be calculated as in equation (7.5).

$$\Delta_i = \frac{|r_i - t_i|}{n-1} \tag{7.5}$$

For a transmitted source image with K pixels, the PER can be calculated as in equation (7.6). Note that the PER can be interpreted as the average colour corruption on a digital image.

$$PER = \frac{1}{K}\sum_{i=1}^{K} \Delta_i \tag{7.6}$$

## 6. RESULTS

The PER performance of HSL coded image truck has been evaluated using a conventional turbo code (CL) and its modified two-fold (MTF) equivalent. The mother code rate for the CL is 1/3, whereas for the MTF this is 3/10. A moderate information frame size of 4608 bits was chosen for both schemes. As the component code, the 4-state g[7,5] RSC code was used and the decoding was performed in parallel using the max-log MAP algorithm.

Fig. 7-5 presents the comparative PER performance of the CL and the MTF for 2, 4, 8 and 16 iterations. It can be seen that at PER=$10^{-5}$, 4 MTF iterations (4i-MTF) provides roughly 0.7 dB gain over the CL scheme (4i-CL). More importantly, at Eb/No ratios higher than 0.75 dB, the pixel corruption of the decoded images with 4 MTF iterations is far less than that of the 8 and 16 CL iterations.

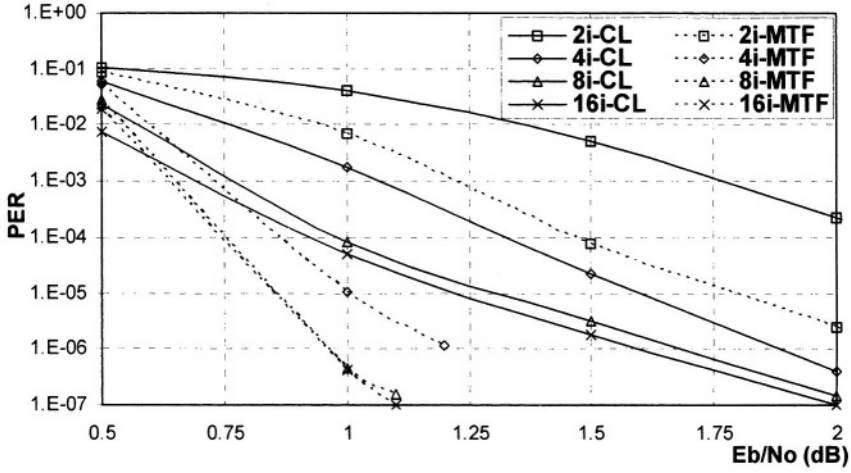



*Figure 7-5.* Pixel error performance.

To illustrate the PER of each plane, Fig. 7-6 has been included, which clearly shows the superior pixel error performance of the MTF over the CL scheme. For the H and S planes, the 1.0 dB gain at PER=$10^{-5}$, which is achieved by the MTF over the CL, is due to the error performance improvement of the two-fold turbo codes over the conventional turbo codes. However, the improved error performance for the L plane is a combined gain of the two-fold turbo coding and the UEP offered by the MTF scheme. Comparison of the L-MTF and the H-MTF and S-MTF curves shows that the unequal error protection scheme introduces an additional 0.2 dB gain to

the two-fold turbo coding at PER=$10^{-5}$. Most of all, the L-MTF curve achieves about 1.2 dB gain over the L-CL curve at the same PER.

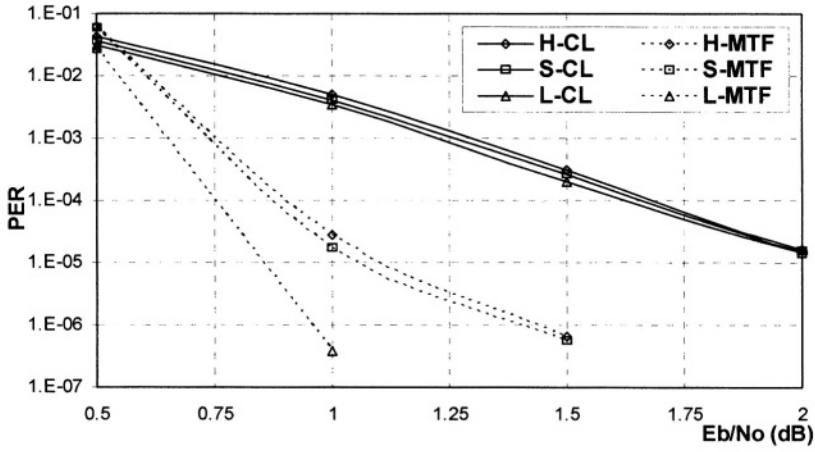



*Figure 7-6.* Pixel error performance of HSL, planes

## 7. CONCLUSIONS

In this chapter, an unequal error protection scheme using modified two-fold turbo codes has been introduced for reliable transmission of HSL coded still images in noisy channels. Using the PER to quantify the visual disturbance on a pixel basis, a conventional turbo code performance has been compared to its modified two-fold counterpart. It has been shown that for 4 iterations, 0.7 dB gain over the CL scheme could be achieved using the MTF turbo code at PER=$10^{-5}$. Moreover, for $E_b/N_o$ higher than 0.75 dB, with only 4 MTF iterations, the decoded HSL images can have better visual quality than those decoded with 16 CL iterations.

Besides the two-fold turbo codes discussed here, alternative unequal error protection schemes can also be developed by modifying higher fold turbo codes, and the number of protection levels can be increased as required. Such systems could be extremely versatile for future wireless multimedia applications. One example to this might be the transmission of compressed audio, still image and textual information within the same multifold turbo code information frame, which can offer different protection levels for each data type allowing efficient utilization of bandwidth.

## ACKNOWLEDGEMENTS

Authors wish to thank Dr Shu Lin for his technical feedback on two-fold turbo codes.

## REFERENCES

1. C. Tanriover, J. Xu, S. Lin, and B. Honary, "Multifold turbo codes," *Proc. ISIT'01,* Washington D.C., USA, June 2001, p 145.
2. C. Tanriover, J. Xu, S. Lin, and B. Honary, "Improving turbo code error performance by multifold coding," *IEEE Communications Letters,* Vol. 6, No. 5, May 2002, pp 193-195.
3. C. Tanriover, *Improved turbo codes for data transmission,* PhD thesis, Lancaster University, UK, April 2002.
4. C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error correcting coding and decoding: turbo codes," *Proc. Int. Conf. on Communications,* Geneva, Switzerland, May 1993, pp. 1064-1070.
5. G. Battail, "A conceptual framework for understanding turbo codes", *IEEE J. Select. Areas Commun.,* Vol. 16, No. 2, Feb. 1998, pp. 245-254.
6. E. Biglieri and V. Volski, "The weight distribution of the iterated product of single-parity-check codes is approximately Gaussian," *IEE Electronics Letters,* Vol. 30, June 1994, pp. 923-924.
7. General information on the International Commission on Illumination, http://members.eunet.at/cie/frameaboutcie.html.
8. P. Chippendale, C. Tanriover, and B. Honary, "Enhanced image coding for noisy channels," *Proc. 7th IMA International Conference,* Cirencester, UK, December 1999, pp 94-103.

Chapter 8

# CODE-AIDED ML JOINT DELAY ESTIMATION AND FRAME SYNCHRONIZATION

Henk Wymeersch and Marc Moeneclaey*
*Digital Communications Research Group, Dept. of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, 9000 GENT, BELGIUM*
{hwymeers,mm}@telin.UGent.be

**Abstract:** We present a novel maximum-likelihood (ML) algorithm for joint delay estimation and frame synchronization. The algorithm operates on coded signals and exploits the code properties by accepting soft information from the MAP decoder. Issues of convergence are addressed and we show how computational complexity may be reduced without any performance degradation. Simulation results are presented for convolutional and turbo codes, and are compared to performance results of conventional algorithms both in terms of mean square estimation error (MSEE) and BER. We show that code-aided delay estimation always improves the MSEE, but not necessarily the BER. On the other hand, code-aided frame synchronization is mandatory, in order to avoid either significant BER degradations or the need for very long pilot sequences.

**Key words:** turbo synchronization, delay estimation, frame synchronization, EM algorithm

## 1. INTRODUCTION

In packet-based communications, frames arrive at the receiver with an unknown propagation delay that varies from packet to packet. When delay estimation (DE) is performed by means of a conventional non-data aided (NDA) algorithm [1], the resulting estimate exhibits an ambiguity, due to the cyclostationary nature of the transmitted signal. Resolution

*This work has been supported by the Interuniversity Attraction Poles Program P5/11 - Belgian Science Policy.

of this ambiguity is known as frame synchronization (FS). Frame synchronization can be accomplished by exploiting the presence of a known pilot sequence in the transmitted data stream [2]. Since a frame synchronization failure gives rise to the loss of an entire packet, its probability of occurrence should be made sufficiently small. At the same time the pilot sequence must not be too long as it reduces the spectral efficiency of the system.

Although conventional estimation algorithms perform well for uncoded systems, a different approach needs to be taken when powerful error-correcting codes are used. These operate typically at low SNR, making the estimation process more difficult. By exploiting the knowledge of certain code properties, a more accurate estimate may be obtained. Iterative delay estimation and detection is performed in [3] but the problems of convergence and frame synchronization were not addressed, nor were comparisons made with conventional NDA estimation schemes. In [4] soft bits from the decoder are fed to a Mueller and Muller timing error detector in an iterative fashion. Yet another approach was taken in [5]: delay estimation and frame synchronization are performed through soft-bit combining. A theoretical framework for code-aided estimation was proposed in [6] but only applied to phase estimation.

In order to combine frame synchronization and decoding, various approaches have been proposed: [7] uses a list-based synchronizer and makes the pilot sequence part of the codeword, thus forcing the coder into a sequence of known states. The decoder verifies this sequence to determine whether or not frame synchronization is achieved. In [8] the so-called *path surface metric,* based on on the forward and backward metrics in the BCJR decoding algorithm [9], is used for frame synchronization. The properties of this metric change when the decoder is not synchronized. Recently, a frame synchronizer based on the distribution of the log-likelihood ratios at the output of the decoder was investigated [10]. Finally, in [11], we investigated an EM-based frame synchronization algorithm. However, the impact of imperfect delay estimation has not been considered in the above papers.

This chapter addresses the problem of *joint* delay estimation and frame synchronization for coded systems. Based on [6] we make use of the EM-algorithm [12] to derive a method to perform joint DE&FS. Convergence issues of the EM-based algorithm are discussed in detail. Through computer simulations, we make comparisons, in terms of mean square estimation error (MSEE) and BER, with known schemes from literature. We show that the EM DE algorithm is especially well suited to situations where conventional NDA algorithms fail to provide reliable estimates. Finally, we demonstrate that as far as FS is concerned, ap-

plication of the EM FS algorithm leads to a very significant reduction in the required number of pilot symbols, as compared to conventional FS algorithms.

## 2. SYSTEM DESCRIPTION

We assume data symbols transmitted in frames of $N$ symbols. Frames consist of a pilot sequence (**p**) of length $L$ and a data sequence (**d**) of length $N - L$. Some form of energy detection is assumed to roughly determine the arrival of a burst within $M_\tau$ symbol intervals. In the presence of an unknown timing delay, $\tau$, the received signal is given by

$$r(t) = \sum_{n=0}^{N-1} a_n p(t - \tau - nT) + n(t) \tag{8.1}$$

where $\mathbf{a} = [\mathbf{pd}]$ is the vector of transmitted M-PSK[1] symbols with $|a_k| = 1$ and $T$ is the symbol duration. The transmit pulse, $p(t)$, is a square-root cosine-roll-off unit energy pulse with roll-off $\alpha$ and one-sided bandwidth $B = (1+\alpha)/(2T)$, and $n(t)$ is a complex AWGN process with spectral density $N_0/E_s$, with $E_s$ the energy per symbol. The incoming signal is bandlimited and filtered at a rate $1/T_s$. When $T_s$ is sufficiently small, any subsequent processing can be performed in the digital domain. We further break up $\tau$ as $\tau = k_\tau T + \varepsilon_\tau$, where $k_\tau \in \{0, \ldots, M_\tau - 1\}$ and $|\varepsilon_\tau| < T/2$ denote the integer part and the fractional part of the time delay, respectively. The DE algorithm involves the estimation of the continuous parameter $\varepsilon_\tau$ or $\tau$, whereas FS refers to the estimation of the discrete parameter $k_\tau$. Estimation of $\varepsilon_\tau$ and $\tau$ will be denoted 'DE' and 'total DE', respectively, wherever it is appropriate to make such a distinction.

## 3. ML ESTIMATION THROUGH THE EM ALGORITHM

Assume we want to estimate a (discrete or continuous) parameter $b$ from an observation vector **r** in the presence of a so-called nuisance vector **a**. The maximum likelihood estimate of $b$ maximizes the log-likelihood function:

$$\hat{b}_{ML} = \arg\max_{\tilde{b}} \left\{ \ln p\left(\mathbf{r} \middle| \tilde{b}\right) \right\} \tag{8.2}$$

where

[1]generalization to other constellations is straightforward

$$p\left(\mathbf{r}\left|\tilde{b}\right.\right) = \int_{\mathbf{a}} p\left(\mathbf{r}\left|\mathbf{a},\tilde{b}\right.\right) p\left(\mathbf{a}\right) d\mathbf{a}.$$

Often $p\left(\mathbf{r}\left|\tilde{b}\right.\right)$ is difficult to calculate. The EM algorithm [12] is a method that iteratively solves (8.2). Defining the complete data $\mathbf{x} = [\mathbf{r},\mathbf{a}]$, the EM algorithm breaks up in two parts: the Expectation part (Eq. 8.3) and the Maximization part (Eq. 8.4):

$$Q\left(\tilde{b},\hat{b}^{(n)}\right) = \int_{\mathbf{x}} p\left(\mathbf{x}\left|\mathbf{r},\hat{b}^{(n)}\right.\right) \ln p(\mathbf{x}|\tilde{b}) d\mathbf{x} \tag{8.3}$$

$$\hat{b}^{(n+1)} = \arg\max_{\tilde{b}} \left\{ Q\left(\tilde{b},\hat{b}^{(n)}\right)\right\}. \tag{8.4}$$

It has been shown that $\hat{b}^{(n)}$ converges to a stationary point of the likelihood function under fairly general conditions [12]. However, when the initial estimate $(\hat{b}^{(0)})$ is not sufficiently close to the ML value, the EM algorithm may converge to a local maximum or a saddle point instead of the global maximum of the likelihood function. To avoid these convergence problems, we propose the following solution [11]. Assuming we have $K$ initial estimates $\left\{\hat{b}_1^{(0)},\ldots,\hat{b}_K^{(0)}\right\}$, we apply the EM algorithm ((8.3)-(8.4)) $K$ times, each with a different initial estimate; after convergence this will result in $K$ tentative estimates $\left\{\hat{b}_1,\ldots,\hat{b}_K\right\}$. The final estimate of $b$ is the tentative estimate with the largest likelihood:

$$\hat{b} = \arg\max_{\hat{b}_k} \left\{ \ln p\left(\mathbf{r}\left|\hat{b}_k\right.\right)\right\}. \tag{8.5}$$

As the computation of the likelihood function $p\left(\mathbf{r}\left|b\right.\right)$ is generally intractable, we resort to the following approximation:

$$\hat{b} = \arg\max_{\hat{b}_k} \left\{ Q\left(\hat{b}_k,\hat{b}_k\right)\right\}. \tag{8.6}$$

The EM algorithm can easily be extended to acquire the Maximum a Posteriori (MAP) estimate of $b$ by taking the a priori distribution $p(b)$ into account in (8.3).

## 4. CODE-AIDED DE AND FS

Denoting by $\mathbf{r}$ a random vector obtained by expanding $r\left(t\right)$ onto a suitable basis, we now make use of the EM algorithm for estimating the propagation delay, $\tau$. Let us define the complete data as $\mathbf{x} = [\mathbf{r}, \mathbf{d}]$. Taking (8.1) into account, we obtain, ommiting terms that do not depend on $\tilde{\tau}$ and $\mathbf{d}$

$$\ln p(\mathbf{r}|\tilde{\tau},\mathbf{d}) \propto \sum_{k=0}^{N-1} \Re\left\{\int_{-\infty}^{+\infty} r^*(t)\, a_k p(t-kT-\tilde{\tau})\, dt\right\}.$$

Since $\mathbf{d}$ and $\tilde{\tau}$ are independent, (8.3) then becomes

$$\begin{aligned} Q(\tilde{\tau},\hat{\tau}) &= \mathrm{E}_{\mathbf{d}}\left[\ln p(\mathbf{r}|\tilde{\tau},\mathbf{d})|\,\hat{\tau},\mathbf{r}\right] \\ &= \sum_k \Re\left\{\mathrm{E}_{\mathbf{d}}\left[a_k|\,\hat{\tau},\mathbf{r}\right] y^*(kT+\tilde{\tau})\right\} \end{aligned}$$

where $y(t) = r(t) \star p^*(-t)$. Hence, $y(t)$ is the signal obtained by filtering the received signal $r(t)$ by a filter, $p^*(-t)$, matched to the transmit pulse. We can break up $Q(\tilde{\tau},\hat{\tau})$ as follows:

$$Q(\tilde{\tau},\hat{\tau}) = C_p(\tilde{\tau}) + C_d(\tilde{\tau},\hat{\tau}) \tag{8.7}$$

where

$$C_p(\tilde{\tau}) = \sum_{i=0}^{L-1} \Re\left\{y^*(iT+\tilde{\tau})\, p_i\right\} \tag{8.8}$$

and

$$C_d(\tilde{\tau},\hat{\tau}) = \sum_{i=0}^{N-1} \Re\left\{y^*((i+L)T+\tilde{\tau})\, \mu_i(\mathbf{r},\hat{\tau})\right\} \tag{8.9}$$

wherein

$$\begin{aligned} \mu_i(\mathbf{r},\hat{\tau}) &= \mathrm{E}_{d_i}\left[d_i|\mathbf{r},\hat{\tau}\right] \\ &= \sum_{\{\alpha_l\}} P\left[d_i = \alpha_l|\mathbf{r},\hat{\tau}\right]\alpha_l \end{aligned} \tag{8.10}$$

denotes the a posteriori average of the data symbol $d_i$. Here $\{\alpha_l\}$ is the set of constellation points. The quantity $\mu_i(\mathbf{r},\hat{\tau})$ can be interpreted as a soft symbol decision: it is a weighted average of all possible constellation points. The a posteriori symbol probabilities in (8.10) can be provided by the MAP decoder [9]. This implies that the algorithm can be applied to a wide variety of codes such as convolutional codes, turbo codes, LDPC codes, repeat-accumulate codes etc. Application of (8.4) yields the following iterative algorithm for total DE:

$$\hat{\tau}^{(n+1)} = \arg\max_{\tilde{\tau}} \left\{C_p(\tilde{\tau}) + C_d\left(\tilde{\tau},\hat{\tau}^{(n)}\right)\right\}. \tag{8.11}$$

The algorithm starts with $n = 0$ from some initial estimate $\hat{\tau}^{(0)}$. How such an initial estimate may be obtained, will be discussed in section 5.

## 5. CONVENTIONAL DE AND FS

A data-aided (DA) estimate of the delay can be obtained from the pilot symbols as

$$\hat{\tau} = \arg \max_{\tilde{\tau}} C_p(\tilde{\tau}) \tag{8.12}$$

where $C_p(\tilde{\tau})$ is defined in (8.8). Note that in (8.12) the portion of the signal corresponding to the data symbols is not exploited.

A NDA delay estimator, such as an Oerder&Meyr (O&M) estimator [1], does make use of the signal portion containing the unknown data symbols:

$$\hat{\varepsilon}_\tau = -\frac{T_s}{2\pi} \arg \sum_k |y(kT_s)|^2 \exp\left(-j2\pi \frac{kT_s}{T}\right). \tag{8.13}$$

Observe that the estimator (8.13) only provides an estimate of the fractional part of the delay, rather than the total delay $\tau$. This is due to the cyclostationary nature of the transmitted signal. Hence, the NDA fractional DE must be combined with a frame synchronization algorithm that estimates the integer part of the delay. A well known DA FS algorithm is [13]:

$$\hat{k}_\tau = \arg \max_{\tilde{k} \in \{0,\ldots,M_\tau - 1\}} C_p\left(\tilde{k}T + \hat{\varepsilon}_\tau\right). \tag{8.14}$$

The performance of this correlation technique is close to that of the ML frame synchronization rule for uncoded transmission [2].

## 6. CONVERGENCE PROPERTIES

When the initial delay estimate, to be used in (8.11), is provided by a NDA estimator, it may or may not be close to the true value $\tau$ (depending on the value of $k_\tau$). In order to illustrate the convergence properties of the EM total DE algorithm (8.11) in the case of (rate 1/3) turbo encoded BPSK, we consider Fig. 8-1.

The top part of the figure shows, for different SNR, $\mathrm{E}\left[e^{(n+1)}\right] - e^{(n)}$ as a function of $e^{(n)}$, with $e^{(n)} = \left(\hat{\tau}^{(n)} - \tau\right)/T$ denoting the normalized delay estimation error at the $n$-the iteration[2]. The negative and positive zero-crossings of $\mathrm{E}\left[e^{(n+1)}\right] - e^{(n)}$ correspond to the stable and unstable equilibrium points of the EM algorithm. The stable equilibrium points are at $e^{(n)} = k$, with $k \in \mathbb{Z}$ whereas the unstable equilibrium points are

[2] note that Fig. 8-1 is independent of the iteration index 'n'

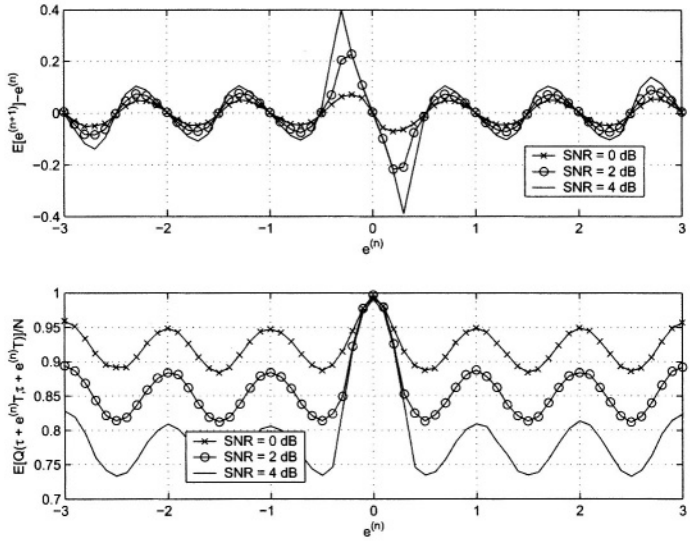



*Figure 8-1.* Convergence behavior for EM delay estimation for 50% roll-off as a function of the estimation error for rate 1/3 turbo coded BSPK transmission. These results are obtained through computer simulation.

at $e^{(n)} = k + 1/2$, with $k \in \mathbb{Z}$. Hence, the acquisition range of the EM algorithm for BPSK is $|e^{(0)}| < 0.5$, corresponding to a maximum allowable initial delay error magnitude of $T/2$. We note that for convolutional codes (results not shown), the periodic trellis will result in a periodicity in $\mathrm{E}\left[e^{(n+1)}\right] - e^{(n)}$; no periodicity is present for turbo-coding, because the corresponding trellis is time-varying.

The bottom part of the figure shows, as a function of $e^{(n)}$ and for different SNR, $\mathrm{E}\left[Q\left(\hat{\tau}^{(n)}, \hat{\tau}^{(n)}\right)\right]$. The stable (unstable) equilibrium points correspond to local maxima (minima) of $\mathrm{E}\left[Q\left(\hat{\tau}, \hat{\tau}\right)\right]$. These curves reflect the earlier observed non-periodicity associated with the turbo code. For a convolutional code (results not shown) $\mathrm{E}\left[Q\left(\hat{\tau}^{(n)}, \hat{\tau}^{(n)}\right)\right]$ will be periodic, making the FS process more error-prone.

In general, an obvious way around any inherent periodicity in the code, is placing a pseudo-random bit-interleaver between the encoder and the mapper[3], as happens in BICM schemes [14]. As the interleaver breaks all periodicity, the corresponding FS algorithm will have convergence properties similar to those of the FS algorithm for a random-like

[3] the mapper converts groups of bits to symbols in the constellation

code (such as the turbo code from Fig. 8-1). We therefore expect the EM algorithm to be able to detect any delay shift in any coded system, even when no pilot symbols are present.

From the above it follows that using a NDA delay estimate from (8.13) as initial estimate for the EM algorithm (8.11) might cause convergence to an erroneous equilibrium point when $k_\tau$ is nonzero. If we prefer a DA initial estimate (8.12), a long pilot sequence is needed to keep the variance of the estimate within acceptable limits. Instead, we propose to apply the EM algorithm with NDA initialization, but with $KM_\tau$, rather than one initial estimate:

$$\hat{\tau}_k^{(0)} = kT/K + \hat{\varepsilon}_\tau \quad \text{for } k \in \{0, 1, \ldots, KM_\tau - 1\} \tag{8.15}$$

where $\hat{\varepsilon}_\tau$ is obtained from the NDA algorithm (8.13), and the integer $K \geq 1$ is a design parameter. Applying the EM algorithm will result in $M_\tau K$ tentative estimates. The final delay estimate is then obtained according to (8.6) with $b = \tau$. This way, we can be sure that $K$ initial estimates yield a corresponding initial normalized error $e^{(0)}$ within the acquisition range of the EM algorithm. Strictly speaking $K = 1$ is sufficient, but we will point out in the next section the advantage of taking $K > 1$. In the remainder of this chapter we will denote the EM algorithm with $KM_\tau$ initial values by 'EM-K'.

In the case of perfect FS (i.e., $k_\tau$ is known) this EM algorithm can easily be specialized into a purely DE algorithm by retaining from (8.15) only the $K$ initial estimates closest to $k_\tau T$ and applying algorithm (8.11). Similarly, the EM algorithm can be modified to a FS algorithm by fixing $\hat{\varepsilon}_\tau$ and then applying (8.6) with $b = kT/K + \hat{\varepsilon}_\tau$.

## 7. PERFORMANCE RESULTS

We evaluate the performance of the EM algorithm for DE and FS when applied to a convolutionally coded and a turbo coded system with BPSK mapping, $T_s = T/4$ and a roll-off of 0.1.

The convolutional code is systematic and recursive with rate 1/2, generator polynomials $(21, 37)_8$ and constraint length 5; codewords consist of 512 BPSK symbols (not counting pilot symbols).

The turbo code consists of the parallel concatenation of two convolutional encoders, separated by an interleaver; codewords consist of 999 BPSK symbols. At each iteration of the EM algorithm, we have reinitialized the (iterative) turbo-decoding algorithm with uniform a priori information.

Maximization in (8.11) is performed through the well-known Newton-Raphson algorithm. For signal reconstruction (i.e., computing $y(nT + \tilde{\tau})$

in (8.8) and (8.9) from the samples $y(kT_s)$ we have used a 6-tap polynomial Lagrange interpolator.

The performance of the DE algorithm is evaluated in term of the MSEE. The MSEE is compared with the Modified Cramer-Rao bound (MCRB). The MCRB is a lower bound for the MSEE of any unbiased estimator [15]. The DE algorithms and the FS algorithms are further evaluated in terms of their BER performance.

**Computational complexity** Fig. 8-1 gives the impression that applying the EM-1 algorithm (8.11) with $M_\tau$ initial estimates should be sufficient to avoid any convergence problems and thus to always obtain the ML estimate. However, occasional convergence to an incorrect fixed point may still occur when the initial estimate of $\varepsilon_\tau$ is too far away from the ML estimate, i.e., when the O&M algorithm gives rise to an outlier. In order to cope with those situations, we propose to use the EM-2 algorithm, rather than the EM-1 algorithm. Although this approach appears to double the computational complexity, this is not necessarily so: denoting the decoding time per frame by $D$, and the number of iterations of (8.11) by $I$ (i.e., the number of EM iterations), the computational complexities of EM-K is of the order $DIKM_\tau$. However, by increasing $K$, the number of iterations $I$ can be reduced: from Fig. 8-1 we see that smaller estimation errors lead to faster convergence (i.e., less EM iterations). So, not only is EM-K able to solve convergence problems caused by outliers, it is able to do so at fewer iterations.

For iterative decoding algorithms, complexity can be further reduced by merging the decoding iterations with the estimation iterations [6]: at each EM iteration, only one decoding iteration is performed; at the start of a given decoding iteration, the decoder is initialized with the a priori probabilities obtained during the previous (EM) iteration. With this technique, the number of EM iterations needed to achieve convergence will be higher, but each EM iteration will have low computational complexity; the net effect is a substantial reduction in computational complexity.

**Delay estimation** Let us consider in Fig. 8-2 the performance related to the convolutional code. On the left part of Fig. 8-2 we show the MSEE of the O&M estimator and the EM-1 estimator, assuming perfect FS. We observe that application of the EM algorithm yields the smaller MSEE. The EM-1 estimator is very close to the MCRB for SNR above 2.0 dB. To see how this translates in BER performance, we observe the right part of Fig. 8-2. The O&M estimator results in BER degradations of around 0.5 dB, as compared to the perfect synchronization case. Application

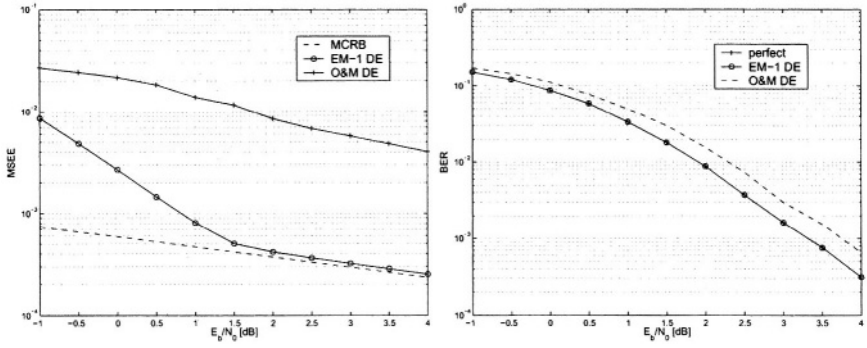



*Figure 8-2.* Fractional delay estimation performance for the convolutional code in terms of MSEE (left) and BER (right). Perfect frame synchronization is assumed.

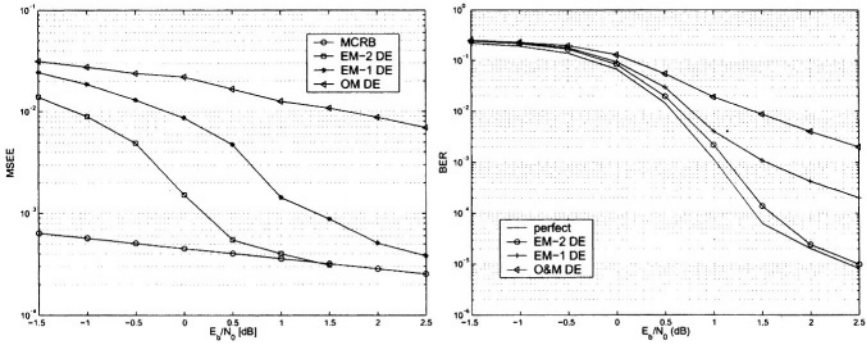



*Figure 8-3.* Fractional delay estimation performance for the turbo code in terms of MSEE (left) and BER (right). Perfect frame synchronization is assumed.

of the EM-1 estimator results in a negligible BER degradation for all considered SNR. We have verified (results not shown) that when the roll-off $\alpha$ is increased, the BER degradation of the O&M estimator compared to perfect synchronization performance is reduced. Hence, is such cases the EM-1 estimator will reduce the MSEE, but will not always noticeably improve the BER performance.

For the turbo code, we show results in Fig. 8-3. Again the EM-1 estimator reduces the MSEE as compared to the O&M estimator. However, the MCRB is not reached for any of the considered SNR values. Application of the EM-2 algorithm, as described in section 6, reduces the MSEE even more. The MCRB is now reached for an SNR above 1.5 dB. It is interesting to investigate the corresponding BER performance (in the right part of Fig. 8-3): the O&M estimator results in significant

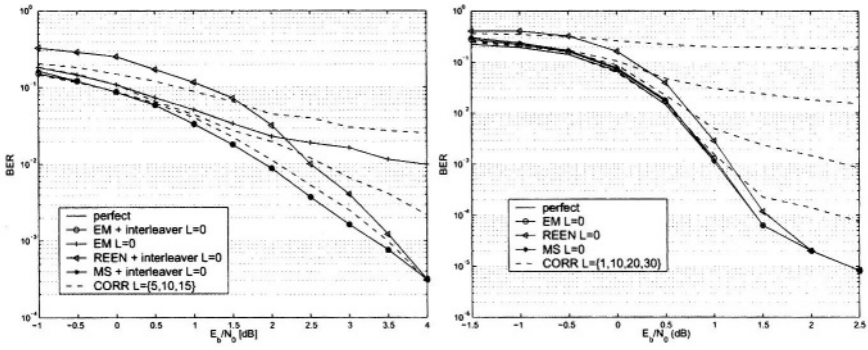



*Figure 8-4.* Frame synchronization performance for the convolutional code (left) and the turbo code (right). Perfect fractional delay estimation is assumed.

BER degradations (over 1.5 dB at a BER of $10^{-3}$). The EM-1 estimator can somewhat reduce the degradation to less than 1.0 dB. The EM-2 estimator on the other hand, is able to reduce the BER degradation to less than 0.1 dB for all SNR.

We conclude that although the EM-K estimator always reduces the MSEE degradation (expressed in dB) as compared to the O&M estimator, this does not necessarily translate in a comparable gain in terms of BER degradation. The choice to use either O&M, EM-1 or EM-2 has to be made on a case by case basis.

**Frame synchronization** Under the assumption of perfect fractional DE (i.e., $\hat{\varepsilon}_\tau = \varepsilon_\tau$) and $M_\tau = 3$, Fig. 8-4 shows the BER performance resulting from various FS algorithms (we remind that $L$ denotes the length of the pilot sequence, expressed in symbols):

- CORR-L: the conventional correlation rule (8.14).
- EM-L: the EM-based FS algorithm (i.e., (8.6) with $b = k_\tau$) with $M_\tau$ initial estimates.
- REEN-L: a code-aided algorithm from [16]. It is similar to (8.11), only now correlation is not performed with soft data symbols but with the sequence of data symbols obtained by re-encoding and re-mapping the decoded information sequence.
- MS-L: a code-aided algorithm from [10]. This algorithm is based on Mode Separation: the log-likelihood ratios (LLR) at the output of the decoder have a bimodal distribution. The distance between these two modes is large (resp. small) when FS is achieved (resp. not achieved). The delay shift corresponding to the maximum distance is the estimated

delay shift. Note that this algorithm cannot easily be modified to perform joint FS and DE.

When an interleaver is present between the coder and the modulator, this is mentioned explicitly in the labels of the graphs.

As far as the convolutional code is concerned, we observe that for the correlation rule a pilot sequence of around 15 symbols is necessary if we wish to avoid high BER degradations. The EM-based algorithm without interleaver has fairly poor performance when $L = 0$. However, when we include an interleaver between the encoder and the mapper, the EM algorithm causes no noticeable performance degradation. The re-encoding rule achieves good BER performance only for SNR above 4 dB. Finally, the MS algorithm has a BER performance similar to the EM FS algorithm. However, we have verified (results not shown) that the MS FS algorithm has a higher FS error rate (i.e., a higher percentage of incorrect estimates of $k_\tau$).

Similar results are shown for the turbo code in the right part of Fig. 8-4. Without a pilot sequence, the re-encoding algorithm results in a small BER degradation for SNR below 2.0 dB. The other code-aided algorithms achieve almost perfect BER performance for all considered SNR. The data-aided correlation rule requires many pilot symbols and gives rise to an error floor.

We conclude that the EM algorithm is able to perform FS without any performance degradation even when no pilot sequence is present. Consequently, as compared to conventional algorithms that need pilot symbols, the EM algorithm may increase the overall spectral efficiency of the system.

**Joint delay estimation and frame synchronization** In Fig. 8-5, we perform joint fractional DE and FS. The conventional algorithms (i.e., NDA DE with DA FS) result in performance degradations as compared to the reference BER curves (e.g., at $\mathrm{BER}=10^{-3}$, around 0.4 dB for the turbo code and 1.5 dB for the conv. code). In both cases the BER degradations is mainly due to imperfect NDA DE. The EM algorithm is able to reduce the degradation of the delay estimator and at the same time perform frame synchronization without requiring a pilot sequence.

# 8. CONCLUSIONS

This contribution has considered the problem of joint delay estimation (DE) and frame synchronization (FS) in coded systems. Based on the EM algorithm, we have shown how joint ML DE & FS may be performed and how convergence issues may be approached. We have compared the performance of the proposed scheme with known algorithms from

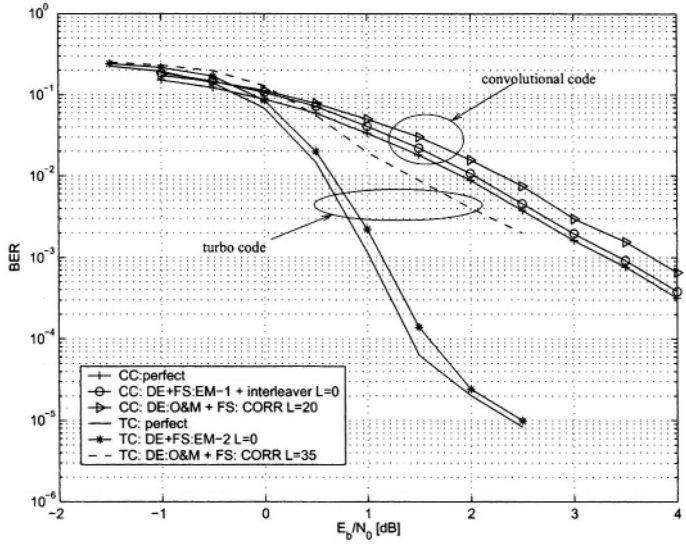



*Figure 8-5.* Joint FS and DE for the convolutional code (CC) and the turbo code (TC)

literature. The considered performance measures are the mean square estimation error and the BER. Through simulation we have shown that in those cases where conventional schemes fail to deliver reliable delay estimates, the proposed algorithm achieves excellent BER performance, albeit at an increased computational complexity. In any case, code-aided FS can be used to avoid the use of long pilot sequences, thus increasing the overall spectral efficiency.

# REFERENCES

1. M. Oerder and H. Meyr. "Digital filter and square timing recovery". *IEEE Trans. on Comm.,* 36:605–611, May 1988.

2. J.L. Massey. "Optimum frame synchronization". *IEEE Trans. on Comm.,* com-20(2):pp. 115–119, April 1972.

3. C. Herzet, V. Ramon, L. Vandendorpe and M. Moeneclaey. "EM algorithm-based timing synchronization in turbo receivers". In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Hong Kong, April 2003.

4. A.R. Nayak, J.R. Barry and S.W. McLaughlin. "Joint timing recovery and turbo equalization for coded partial response channels". *IEEE Trans. on Magnetics,* 38(5):2295–2297, Sept. 2002.

5. B. Mielczarek and A. Svensson. "Timing error recovery in turbo coded systems on AWGN channels". *IEEE Trans. on Comm.,* 50(10):pp. 1584–1592, Oct. 2002.

6. N. Noels, C. Herzet, A. Dejonghe, V. Lottici, H. Steendam, M. Moeneclaey, M. Luise and L. Vandendorpe. "Turbo-synchronization: an EM algorithm approach". In *Proc. IEEE International Conference on Communications (ICC),* Anchorage, May 2003.

7. M.K. Howlader and B.D. Woerner. "Decoder-assisted frame synchronization for packet transmission". *IEEE Journal on Selected Areas in Comm.,* 19(12):pp. 2331–2345, Dec. 2001.

8. J. Sodha. "Turbo code frame synchronization". *Signal Processing Journal, Elsevier,* 82:pp. 803–809, 2002.

9. L.R. Bahl, J. Cocke, F. Jelinek and J. Raviv. "Optimal decoding of linear codes for minimising symbol error rate". *IEEE Trans. on Information Theory,* 20:pp. 284–287, March 1974.

10. T.M. Cassaro and C.N. Georhiades. "Frame synchronization for coded systems over AWGN channel". *IEEE Trans. on Comm.,* 52(3):484–489, March 2004.

11. H. Wymeersch and M. Moeneclaey. "ML frame synchronization for turbo and LDPC codes". In *Proc. 7th Int. Symp. on DSP and Comm. Systems,* Coolangatta, Australia, December 2003.

12. A.P. Dempster, N.M. Laird and D.B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society,* 39(1):pp. 1–38, 1977. Series B.

13. P. Robertson. "A generalized frame synchronizer". In *Proc. GLOBECOM,* pages 365–369, Dec. 1992.

14. G. Caire, G. Taricco and E. Biglieri. "Bit-interleaved coded modulation". *IEEE Trans. on Information Theory,* 44:927–946, May 1998.

15. N.A. D'Andrea, U. Mengali and R. Reggiannini. "The modified Cramer-Rao bound and its applications to synchronization problems". *IEEE Trans. on Comm.,* 42:1391–1399, March 1994.

16. U. Mengali, R. Pellizzoni and A. Spalvieri. "Soft-decision-based node synchronization for Viterbi decoders". *IEEE Trans. on Comm.,* 43(9):2532–2539, Sept. 1995.

Chapter 9

# ADAPTIVE BLIND SEQUENCE DETECTION FOR TIME VARYING CHANNEL

Mohammad N. Patwary [1], Predrag Rapajic [1], Ian Oppermann [2]
[1] *School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia, e-mail: m.patwary@student.unsw.edu.au, p.rapajic@unsw.edu.au*
[2] *University of Oulu, Finland, P.O. Box 4500 FIN-90014, e-mail: Ian@ees2.oulu.fi*

**Abstract:** In this paper a new blind sequence detection algorithm has been proposed. It has a significant robustness against bit shift ambiguity in comparison to existing blind sequence detection algorithm with coded system. The Viterbi algorithm is extended to a blind form with new branch metric criterion. Significant gain (as much as 8dB) has been achieved in Signal to Noise Ratio (SNR) at the BER level of $10^{-3}$ due to the resistance against the error propagation.

**Key words:** Maximum Likelihood (ML), Per-Survivor Processing (PSP), Viterbi algorithm, BPSK, Bit shift ambiguity, Inter symbol interference (ISI).

## 1. INTRODUCTION

Recently different blind sequence detection algorithms have been proposed in [1-7]. The Viterbi algorithm has extended to the blind form where the branch metrics has been calculated from the short time average of the squared error and the instants of the Viterbi algorithm is defined from all possible candidate sequences within that specific short time average [1,4,6,7]. Test vector has to be generated to build some priory knowledge as proposed in [2]. Both of these blind sequence detection algorithms are without the knowledge of the channel response. These algorithms have delay and bit-shift ambiguity which are most common problem in any existing blind detection procedure. Using this algorithm there would be at least two most probable candidates, which is due to the bit-shift even though we are expecting a unique sequence. One of the most important weaknesses of [1,2]

is that for every set of the transmitted sequence there would be a pair of candidate sequences but we have to choose only one. However, there is no comment on this decision criterion. Recently, another simplified maximum likelihood (ML) sequence detection algorithm, namely Per-Survivor Processing (PSP) had been proposed in [8]. In PSP it has been considered to transmit a preamble to have the initial estimate of the channel. Also it has been assumed that the first and the last bits are known. In this paper we proposed an algorithm that clarifies how to obtain a unique sequence without sending any preambles or midambles. New branch metric parameter has been proposed to implement Viterbi algorithm for ML detection. Using the proposed algorithm we found the followings:

1. One of the most common problems with blind detection namely bit shift ambiguity has been eliminated under realistic scenarios.

2. Computational complexity is exactly the same as in [1,4] but SNR gain is 8dB against baud rate sampling and 1.5dB over double sampling.

3. For the same level of BER ( e.g $10^{-3}$) there is 8dB SNR gain over the algorithms in [8] where preambles are used for the initial estimate of the channel and the first bit and last bit at the truncation are assumed to be known.

4. The proposed algorithm is applicable to multiple inputs multiple output (MIMO) system with optimal ordering scheme.

Rest of the chapter has been organized in the following order. In section 2 the system model has been discussed that has been used for the algorithm performance analysis. Proposed algorithm has been described in section 3. The algorithm has been described in step-by-step manner at the end of the section. Simulation results and some comparisons with existing result presented in section 4 and finally the conclusion in section 5.

## 2. SYSTEM MODEL

A fast fading wireless communication channel with Inter-Symbol Interference (ISI) has been considered. The transmitter-receiver Structure of the proposed detection algorithm is shown in Fig. 9-1. Blocks of information sent to the convolution encoder. Encoded information sequence block mapped into base-band modulation scheme and transmitted to the Inter-

Symbol Interference channel, which includes additive white Gaussian noise (AWGN) to the original transmitted signal.

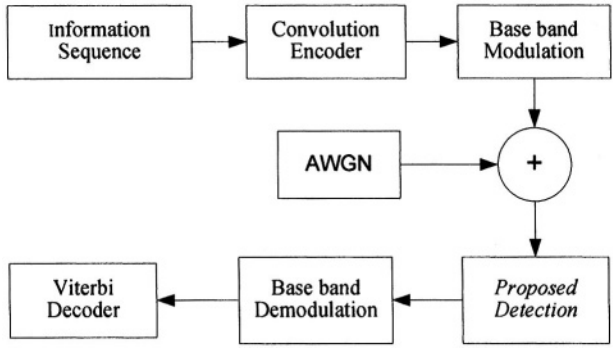



*Figure 9-1.* Proposed Transmitter-Receiver Structure.

Then the noise-contaminated signal processed using the proposed blind detection algorithm, followed by passing through the base band demodulator and decoder. Let us assume the transmitted sequence with a block length L is $\mathbf{S} = [a_0, a_1, a_2 \ldots\ldots a_{k-1}, a_k \ldots\ldots a_{l,-1}]$.

$$y_k = \sum_{k=1}^{N-1} a_{k-n} h_n + n_k \tag{9.1}$$

$$\hat{\mathbf{y}}_k = \begin{bmatrix} \hat{a}_{k-N} & \hat{a}_{k-N-1} & \cdot & \cdot & \cdot & \hat{a}_{k-2N+1} \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \hat{a}_{k-1} & \hat{a}_{k-2} & \cdot & \cdot & \cdot & \hat{a}_{k-N} \\ \hat{a}_k & \hat{a}_{k-1} & \cdot & \cdot & \cdot & \hat{a}_{k-N+1} \end{bmatrix} \times \begin{bmatrix} \hat{h}_0^k \\ \hat{h}_0^k \\ \cdot \\ \cdot \\ \hat{h}_{N-1}^k \end{bmatrix} \tag{9.2}$$

where, $n_k$ is additive white Gaussian noise and mark ^ denotes the estimate and N defines the number of channel co-efficient (rays) of the assumed channel. It should be mention that the length of memory should be greater than that of the channel length. For simplest case, let N=2 and memory length 4.

$$\hat{\mathbf{y}}_k = \hat{\mathbf{A}}_k \hat{\mathbf{h}}_k \tag{9.3}$$

$$\hat{\mathbf{A}}_k = \begin{bmatrix} \hat{a}_{k-2} & \hat{a}_{k-3} \\ \hat{a}_{k-1} & \hat{a}_{k-2} \\ \hat{a}_k & \hat{a}_{k-1} \end{bmatrix}, \hat{\mathbf{h}}_k = \begin{bmatrix} \hat{h}_0^k \\ \hat{h}_1^k \end{bmatrix}, \hat{\mathbf{y}}_k = \begin{bmatrix} \hat{y}_{k-2} \\ \hat{y}_{k-1} \\ \hat{y}_k \end{bmatrix}$$

It is assumed that M is the segment size, that is, in each step of processing there are M number of samples taken into account. Hence, on the detection process the candidate vector size at initial acquisition (synchronization) and onward is (M+N-1) where N is the length of the ISI channel. For each instant, it would be required to calculate $2^{(M+N-1)}$ branch metrics. In [1,4] branch metric whose cost factor derived from squared error criterion. The condition used in [1] is that the total no of sample processed at some instance, i. e, the length of the segment (M) should be greater than the length of the ISI channel. For time varying channel the length of the segment determines a trade-off between accuracy of the response estimation and the tracking ability of the channel. In practical implementation, however, a sorter segment is more advantageous [1].

## 3. DESCRIPTION OF THE ALGORITHM

Let us assume that the ISI channel $\mathbf{h} = [h_0, h_1]^T$ and the segment size M has been considered. Hence the length of the candidate vector will be (M+N-1). For example if M=5 is assumed, the original information block transmitted within the segment of samples are $\mathbf{S} = [a_0 \quad a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5]$. Then the received sample matrix within the segment is

$$\mathbf{y} = \mathbf{A}\mathbf{h} + \mathbf{n} \tag{9.4}$$

where

$$\mathbf{A} = \begin{bmatrix} a_1 & a_0 \\ a_2 & a_1 \\ a_3 & a_2 \\ a_4 & a_3 \\ a_5 & a_4 \end{bmatrix}$$

and $\mathbf{n}$ is the AWGN. According to [1], at the receiving end the branch metric those calculated using squared error criterion given by the following equation:

$$J_k = (\hat{\mathbf{A}}_k \hat{\mathbf{h}} - \mathbf{y}_k)^T (\hat{\mathbf{A}}_k \hat{\mathbf{h}} - \mathbf{y}_k) \tag{9.5}$$

where $\hat{A}$ is an estimate of the candidate metric and the minimum value of $J_k$ gives the true block of sequence of length $m$ with higher confidences factor, At the same time there would be another block of sequence with same confidence factor, which is the mirror replica (with opposite sign) of the true candidate block. That is if $\mathbf{A}_k$ is the true sequence block with squared error $\min(J_k)$ then there would be another sequence block in the candidate state set with same value of $J_k$ which would be exactly $-[\mathbf{A}_k]$. The equality of the confidence factors of the true and replica sequence block make it difficult to construct a one to one correspondence with true transmitted sequence and a candidate sequence in any ML blind sequence detection scheme. Hence causes bit shift or phase ambiguity. To avoid this problem in this context it is proposed to calculate branch metric using the following relation.

$$\alpha_k = \sqrt{(\mathbf{h}_r - \hat{\mathbf{h}})^T (\mathbf{h}_r - \hat{\mathbf{h}})} \tag{9.6}$$

where $\hat{\mathbf{h}} = \hat{\mathbf{A}}^{-1} \cdot \mathbf{y}$ and $\mathbf{h}_r$ is the reference channel which is an ideal inter-symbol interference (ISI) free channel of same span length as the estimated channel. That is

$$\mathbf{h}_{r(SISO)} = [1 \quad 0 \quad \ldots \quad \ldots \quad 0_{n-1}]^T \tag{9.7}$$

$$\mathbf{H}_{r(MIMO)} = \mathbf{I} \tag{9.8}$$

when the channel coefficients are considered real and when the channel is considered complex then

$$\mathbf{h}^c_{r(SISO)} = \left[\left(\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}\right) \quad 0 \quad \ldots \quad \ldots \quad 0_{n-1}\right] \tag{9.9}$$

$$\mathbf{H}^c_{r(MIMO)} = \left(\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}\right) \cdot \mathbf{I} \tag{9.10}$$

when the channel is considered complex and where n is channel span size. Eqs. (9.7) and (9.9) refer to the channels with single input single output system and the eqs. (9.8) and (9.10) correspond to multiple input multiple output (MIMO) systems respectively. Eqs. (9.7) and (9.8) representing a reference channel with real coefficients and (9.9) and (9.10) for complex coefficients. It has to be mention that the matrix $\hat{\mathbf{A}}$ with candidate vector is

always a non-square matrix according to condition mentioned above and hence $\hat{\mathbf{A}}^{-1}$ is a Moore-Penrose inverse.

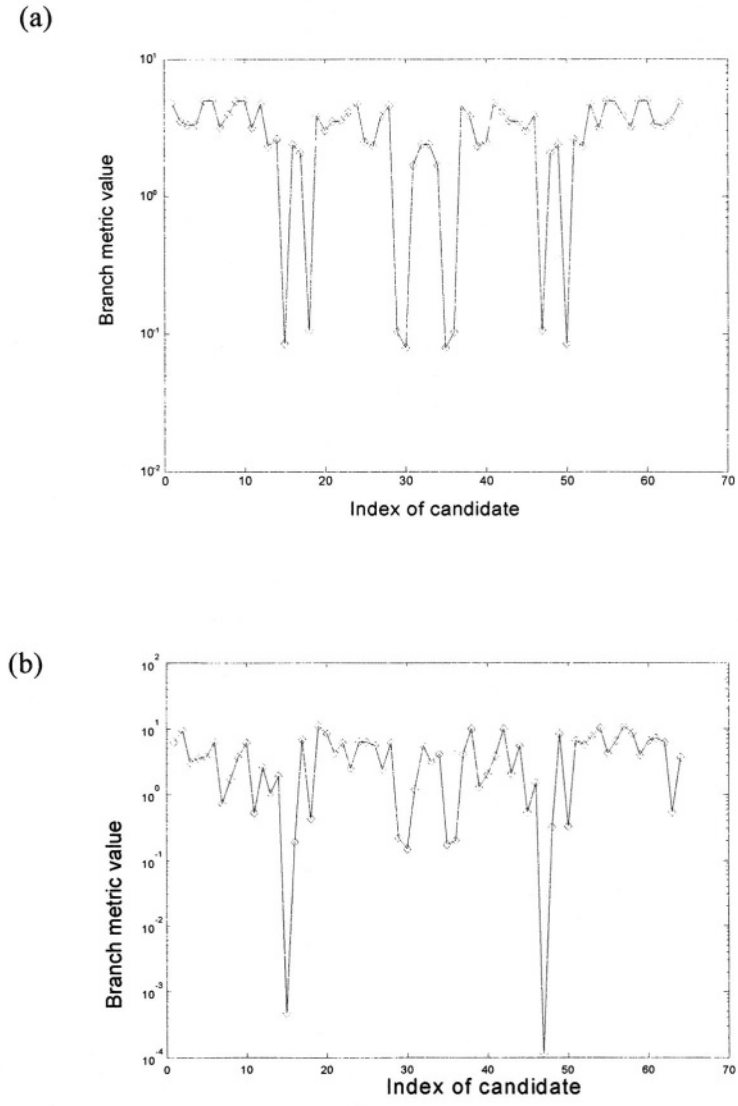



*Figure 9-2.* Calculated branch metric; (a) with the algorithm in [1], (b) with the proposed algorithm

Fig.9-2 (a) and (b) show the calculated branch metric at any instantaneous block of sequence with 6 bits. In Fig.9-2. (a), it has been observed that there are two branch metrics with minimum cost which are exact mirror replica to one other. In [1,2,4], no recommendation had but had been proposed about

this problem. Hence it is very difficult to make a one to one correspondence with any specific candidate vector. In contrary, using the proposed branch metric criteria there always exist a unique candidate vector with the minimum cost and its mirror replica, which causes bit shift ambiguity, will have totally different cost which is higher than the cost of the true block. A sample branch metric value with all possible 6-bit long candidate vectors with 10dB SNR are given in Fig.9-2 (b). As we considered (L+1) bits have been transmitted, through an ISI channel length 2, there are L samples to containing the desired block of information. From these L samples there are M samples that has been processed at each instant and sliding the segmentation window up to the dead end by moving one sample forward in each step. By doing so there would be K possible processing steps where K = (L – M +1) and in every step there are R ($= 2^{(M+N-1)}$) candidate vectors to consider to calculate branch metric. Hence the size of the matrix **C** that contains the calculated branch metric at any instant becomes (R ✕ K). Then the Viterbi algorithm has been used to track surviving states with minimum cost from the matrix **C** and eventually the possible block of transmitted sequence. This detected sequence block has been sent to the convolution decoder at the receiver to recover the original information that has been transmitted from the transmitting end.

**Proposed Algorithm**

***Step. 1:*** Initialize, segment size M, channel length N, the number of samples to process.

***Step. 2:*** Build the matrix C by calculating $\alpha_{i,j}$, by using the equation

$$\alpha_k = \sqrt{(\mathbf{h}_r - \hat{\mathbf{h}})^T (\mathbf{h}_r - \hat{\mathbf{h}})}$$

where i = 1, 2, …R, and j= 1, 2, …K.

***Step. 3:*** Calculate path metric $\beta_{i,L}$ for all possible paths from the values in matrix C where i = 1, 2, ...R and $\beta_{i,L}$ can be calculated from the equation below

$$\beta_{i,0} = \alpha_{i,0}$$

$$\beta_{i,j+1} = \beta_{i,j} + \min\left(\alpha_{j+1}^{(1)}, \alpha_{j+1}^{(-1)}\right)$$

where $\alpha_{j+1}^{(1)}$ and $\alpha_{j+1}^{(-1)}$ represents the branch metric of the state (j+1) proceeded from state j with 1 and -1 as the latest bit respectively.

***Step. 4:*** Choose the path with minimum path metric $\beta_{i,L}$ and track the bits through the path, which will be considered as the detected sequence. More improvement has been found while proposed branch metric has been used along with the short time averaged squared error. The modified branch metric defined as

$$\mu_k = J_k \cdot \alpha_k \tag{9.11}$$

Subsequently, the improved blind sequence detection algorithm steps are as follows including steps 1-2 as mentioned earlier.

***Step 5:*** Calculate the branch metric $\mu_k = J_k \cdot \alpha_k$ for each possible candidate vector.

***Step 6:*** For each branch leading to node $(a_k, a_{k-1}, a_{k-2})$ to $(a_{k-1}, a_{k-2}, a_{k-3})$ compare $\mu_k$ for all possible candidates vectors and select the candidate with minimum branch metric $\mu_k$. Using proposed detection algorithm there would be a unique set of candidate vector with minimum $\mu_k$.

***Step 7:*** Starting from initial state calculate the path metric $\gamma_k^m$ for each node up to the dead end of the train of sequence and select the sequence train with minimum path metric where *k* is the time instant, varies from 1 to the length of the train of sequence and *m* is the state (node) at any instant, varies from 1 to $L^M$ (L is the modulation constellation number, 2 for BPSK; M is the segmentation window). The values of $\gamma_k^m$ would be calculated by using the following equations:

$$\gamma_1^m = \mu_1^m \tag{9.12}$$

$$\gamma_{k+1}^m = \gamma_k^m + \min(\mu_{k+1}^j) \quad \text{where} \quad j \in \langle 1, \ldots L^M \rangle \tag{9.13}$$

# 4. SIMULATION RESULTS

## 4.1 BER Performance

Jakes channel of length 2 has been considered for BER performance analysis, while the proposed algorithm has been implemented in the receiver. For the system with coding, the generator polynomial $\left[1 + D + D^2, 1 + D^2\right]$

has been used. Binary phase shift keying (BPSK) has been used as base band modulation. Simulation results have been shown in Fig. 9-3. to Fig. 9-6. In Fig. 9-3., the BER performance of the proposed algorithm compared with the system with known pilot symbol at certain interval (one option for Per-Survivor Processing (PSP) [8]) has been shown. Fig. 9-4 shows the BER performance comparison among the proposed algorithm, PSP [8] and the work in [4]. From Fig. 9-3. and 9-4., it can be concluded that to reach at BER level $10^{-3}$ there is 8dB SNR gain has been achieved with the proposed algorithm over the recently proposed PSP ML detection technique [8]. Proposed algorithm requires higher computational complexity than that of PSP, where PSP requires the initial estimate of the channel. Fig 9-5(a) shows a comparison between the blind scheme in [1] and the proposed algorithm. A gain of 3dB has been achieved over the algorithm proposed in [1] for baud rate sampling. Fig. 9-5(b) shows a comparison of the BER performance with branch metric $\alpha_k$ (estimated using only the Euclidian distance) and the improved branch metric $\mu_k = J_k \cdot \alpha_k$ (estimated using both the Euclidian distance along with the squared error). For both of the cases a system with convolutional code has been considered. In Fig. 9-5(c), the BER performance of the proposed algorithm for the system with coding has been compared with exiting blind algorithms. As mentioned earlier, by increasing the segment length it is possible to improve the BER performance significantly with the cost of higher computational complexity. The proposed algorithm can be implemented on existing DSPs regardless of the increased computational complexity.

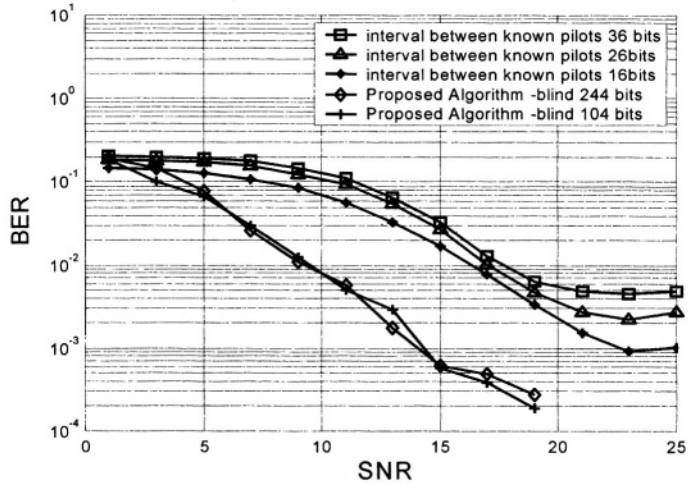



*Figure 9-3.* BER performance of the proposed algorithm.

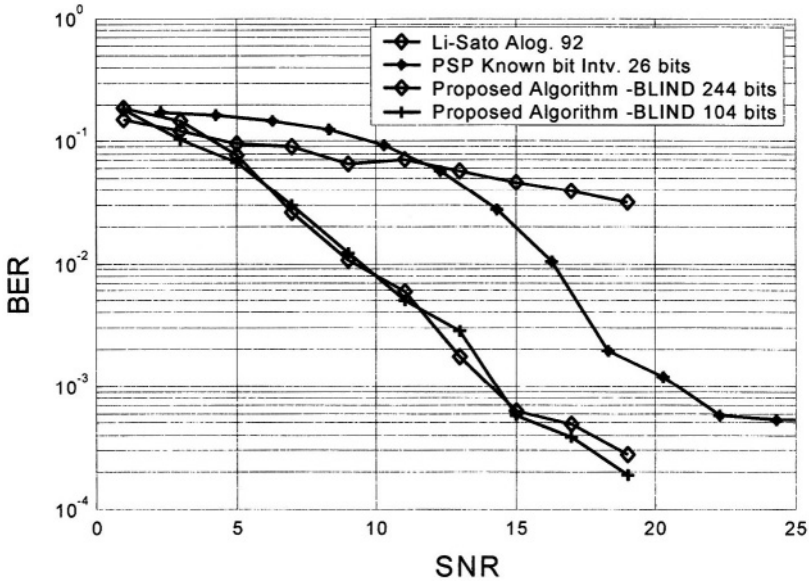
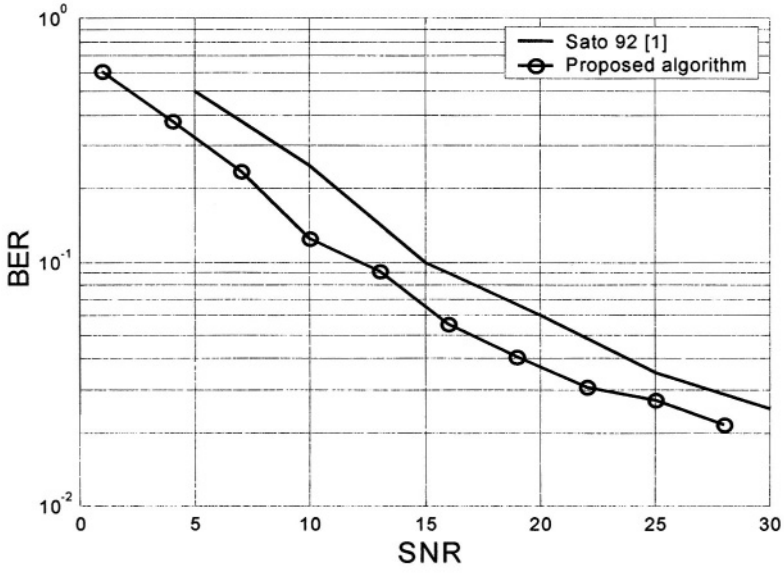
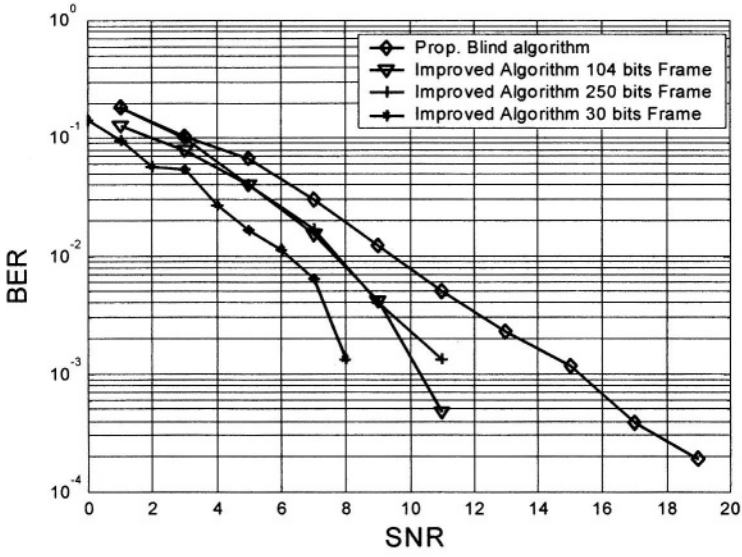


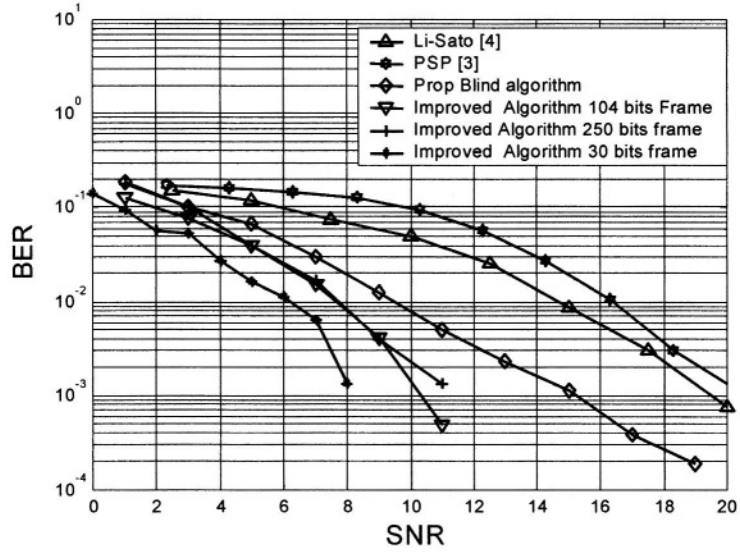

*Figure 9-4.* BER performance Comparison of the proposed algorithm with some existing ML detection techniques.




(*a*)




(*b*)




(*c*)

*Figure 9-5.* BER performance; (a) with the proposed algorithm, (b) with the proposed algorithm using channel length 2 and conv. code [7,5], (c) comparison with the existing algorithm.

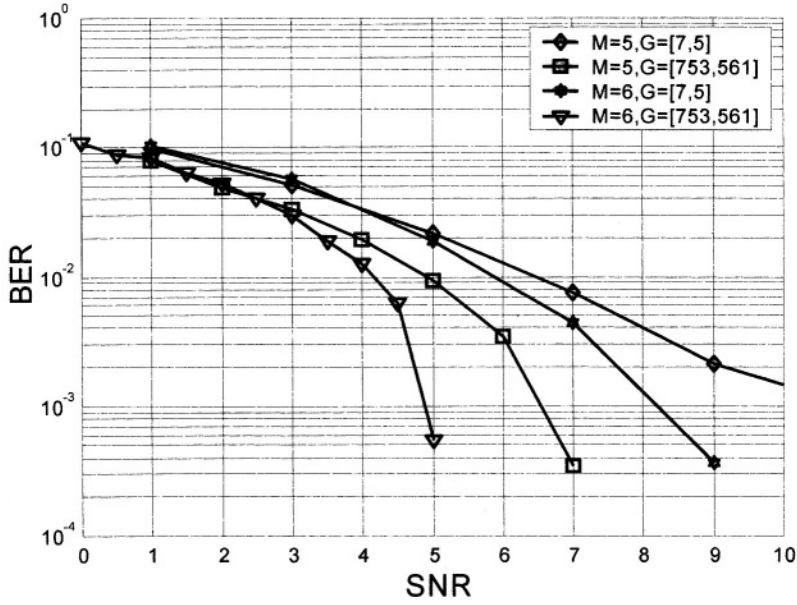



*Figure 9-6.* Influence of the segment length and the generator metric on BER performance

## 4.2 Receiver scalability

As mentioned in [1], the accuracy in the estimate of the channel increases by increasing the length of the processing or segmentation window. Consequently, the receiver BER performance improves. To achieve better BER performance the computational complexity will be exponentially higher. Fig. 9-6. shows the effects of the variation in the segmentation window for a different coded system. The higher the constrain length of the convolutional coded scheme the better the BER performance due the characteristics of the convolutional code. It is concluded that the proposed algorithm is highly scalable in terms of the computational complexity. Further BER improvements are achieved by additional computational complexity. Fig. 9-6 shows a comparative study of BER performance of the proposed receivers with various segmentation windows and constrain lengths of the used coded systems. From Fig. 9-6 it is concluded that a receiver can be designed with required BER (QoS) level at the expanse of computational complexity. The parameters involved with this trade-off are the segmentation window and the constrain length of the coded system.

## 5. CONCLUSION

A new blind sequence detection exploiting Viterbi algorithm has been proposed. Bit shift ambiguity due to the channel uncertainty can be completely reduced by using the proposed algorithm. One of the most common problems with blind detection namely bit shift ambiguity has been eliminated under realistic scenarios. Computational complexity is exactly the same as in [1,4], while SNR gain is 8dB at baud rate sampling and 1.5dB at double sampling. With an appropriate ordering scheme, the proposed algorithm can be implemented for MIMO receiver.

## REFERENCES

1. Y. Sato, " Blind Sequence Detection and Its Application to Digital Mobile Communication", IEEE Journal on Selected Area on Communications, Vol. 13, No. 1, January 1995, pp. 49–58.
2. Xiaohua Li, "Blind Sequence Detection Without Channel Estimation", IEEE Trans. on Signal Processing, vol. 50, No 7, July 2002, pp . 1735 – 1746.
3. H. Kubo, K. Murakami and T. Fujino, "Adaptive Maximum-Likelihood Sequence Estimation by means of Combined Equalization and Decoding in Fading Environments", ", IEEE Journal on Selected Area on Communications, Vol. 13, No. 1, January 1995, pp. 102–109.
4. Y. Li, B. Vucetic and Y. Sato, "Decoding of Convolutional Codes by Viterbi Blind Algorithm", Proc. of ICCS/ISITA , Singapore, Nov. 16-20, 1992, pp. 877-881.
5. K. Metzger, "A New Concept for Simplified Viterbi detection", Proc. of ICCS/ISITA, Singapore. Nov. 16-20, 1992, pp. 1307-1311.
6. Y. Sato, "A Blind Sequence detection over Rapidly Time -Varying channel and its Algebraic Structure", Proc. of ICCS/ISITA, Singapore. Nov. 16-20, 1992, pp. 1317-1321.
7. H. Oda and Y. Sato, 'Viterbi algorithms as Blind Identification under Continuously distributed Data" , Proc. of IEEE International Symposium on Information Theory, p-120, Budapest, June 24-28, 1991.
8. R. Raheli, A. Polydoros and C. K. Tzou, "Per-Servivor Processing: A General Approch to MLSE in Uncertain Environments", IEEE Trans. on Communications, Vol. 43, No 2/3/4, Feb/Mar/Apr 1995, pp. 354 – 364.

*This page intentionally left blank*

Chapter 10

# OPTIMUM PSK SIGNAL MAPPING FOR MULTI-PHASE BINARY-CDMA SYSTEMS

Yeong-Jin Seo and Yong-Hwan Lee
*School of Electrical Engineering and INMC, Seoul National University*

**Abstract:** Although the CDMA system can efficiently support multiple users, may suffer from peak-to-average power ratio (PAPR) increases as the number of users increases. As a result, it needs highly linear power amplifiers with a large back off. Recently, a new CDMA scheme, called binary CDMA (B-CDMA), has been proposed to alleviate this problem by quantizing the envelope of multi-user CDMA signals into a small number of levels, while preserving the advantages of CDMA signaling [1]. The performance of B-CDMA system is mainly determined by the quantization and detection error. The quantization noise can be minimized using the Lloyd-max algorithm [2]. In this chapter, the optimum PSK signal is designed to minimize the detection error in multi-phase B-CDMA systems. Finally, the analytic results are verified by computer simulation.

**Key words:** Binary-CDMA, PSK, MP B-CDMA

## 1. INTRODUCTION

One of major drawbacks of multi-code CDMA systems is high peak-to-average power ratio (PAPR) due to the aggregation of multiple spreading codes. As a result, multi-code CDMA transmitters require the use of highly linear power amplifiers with a large back off. Binary CDMA (B-CDMA) is a new modulation method that quantizes the signal amplitude into a small number of levels and employs PSK-modulation for transmission with constant envelope [1]. Thus, the B-CDMA can alleviate the need of linear power amplifiers, while preserving the advantages of CDMA signaling such

as the soft capacity and robustness to interference. However, the performance can significantly be affected by the quantization process.

The B-CDMA signal can be generated by various methods including the pulse-width (PW), multi phase (MP) and code selection (CS) methods [1]. The PW B-CDMA signal is obtained by converting the magnitude of multi-level signal into a finite number of pulse width. Thus, the transmission bandwidth of the PW B-CDMA increases as the quantization level increases. In practice, the signal can be quantized into two levels to accommodate the increase of transmission bandwidth. The MP B-CDMA is generated by transforming the signal amplitude into a finite number of PSK signal constellation. The CS B-CDMA is generated by a two-step process. In the first step, the subset of spreading codes is selected to reduce the number of signal levels. In the second step the selected code is modulated using the MP B-CDMA scheme.

Optimum quantization of the signal amplitude can be achieved by using the Lloyd-max algorithm [2]. However the PSK signal constellation has not been optimized analytically for the MP B-CDMA. This is mainly due to the fact that the mean square error of the chip decision error and the symbol error in the PSK system cannot be represented in a simple form. The two signal points having the largest distance after the quantization are PSK-mapped so that they have the largest distance on the PSK signal constellation [3]. For a given PSK signal constellation, the decision region can be determined so as to minimize the Bayes cost criterion [4]. However it may not be optimum because the bit error rate (BER) performance can be more affected by the PSK signal mapping than the decision region. In this paper, we optimize the PSK mapping points of the MP B-CDMA signal in additive white Gaussian noise (AWGN) channel. Since the CS B-CDMA is the same as the MP B-CDMA except the code selection block [5], the analytic design can also be applied to optimum design of the CS B-CDMA signal.

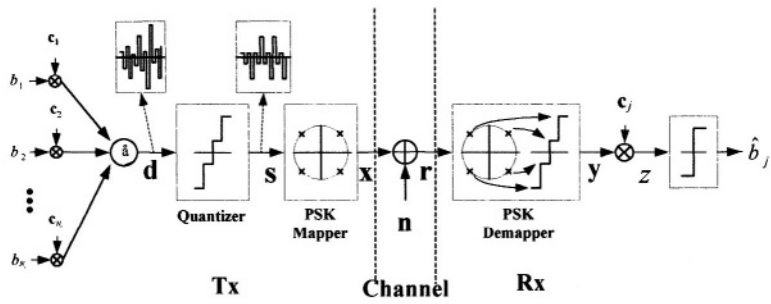



*Figure 10-1.* MP B-CDMA transceiver structure.

Section 2 describes the structure of the MP B-CDMA system. In Section 3, the noise due to the quantization and chip detection error is analyzed. The PSK mapping points are optimized to minimize the chip detection error using an iterative method. The proposed signal mapping scheme is evaluated using computer simulation in Section 4. Conclusions are summarized in Section 5.

## 2. SYSTEM MODEL

In the MP B-CDMA system, the sum of multiple users data is quantized into a finite number of levels and then modulated using a PSK modulation scheme. Fig. 10-1 depicts the transceiver structure of a baseband-equivalent MP B-CDMA system with quantization level 4, where $b_i$ and $\mathbf{c}_i$ respectively denote the bit and spreading code of the $i$ -th user, and the signal $\mathbf{d}$ denotes the sum of multiple users data, given by

$$\mathbf{d} = \sum_{i=1}^{N_c} b_i \mathbf{c}_i \tag{10.1}$$

Here, $N_c$ is the total number of users.

The aggregated user signal $\mathbf{d}$ is quantized at the chip-level. The output of the quantizer can be represented as

$$\mathbf{s} = f_q(\mathbf{d}) \tag{10.2}$$

where $f_q(\mathbf{d})$ denotes the quantization function that maps the signal $\mathbf{d}$ in the quantization region $\Phi_i$ onto the signal point $m_i$ at the chip-level. Then, the quantized signal $\mathbf{s}$ is PSK mapped as

$$\mathbf{x} = f_{map}(\mathbf{s}) \tag{10.3}$$

where $f_{map}(\mathbf{s})$ denotes the mapping function that maps the signal $\mathbf{s}$ onto the PSK constellation. Fig. 10-2. depicts the quantization and PSK-mapping region of the MP B-CDMA system, where $d_{ij}$ denotes the distance between the quantized signal point $i$ and $j$, $\Omega_i$ denotes the detection region of the PSK-modulated signal point $i$, $N_{pt}$ is the number of signal points after the quantization. Note that the guard phase is required in the MP B-CDMA system to reduce the decision errors between the signal points having the largest distance [3].

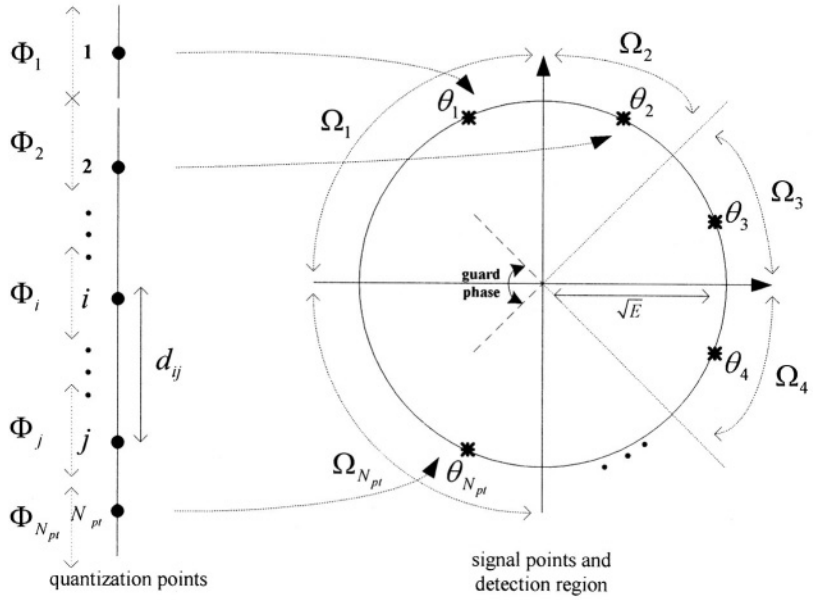



*Figure 10-2.* PSK modulation of MP B-CDMA

The received signal can be written as

$$\mathbf{r} = \mathbf{x} + \mathbf{n} \tag{10.4}$$

where $\mathbf{n}$ denotes the AWGN term. The PSK demodulator transforms the phase information of $\mathbf{r}$ into the magnitude

$$\mathbf{y} = f^{-1}{}_{map}(\mathbf{r}) \tag{10.5}$$

The demodulated signal $\mathbf{y}$ is despreaded using the spreading code $\mathbf{c}_j$ of the $j$-th user to detect the user data $\hat{b}_j$ as

$$\hat{b}_j = \begin{cases} 1 & , \text{ if } \mathbf{y} \cdot \mathbf{c}_j \geq 0 \\ 0 & , \text{ if } \mathbf{y} \cdot \mathbf{c}_j < 0 \end{cases} \tag{10.6}$$

Let $s_j$, $\delta_j$ and $e_j$ be the signal component, quantization noise and noise due to chip detection error at the $j$-th chip, respectively. Then, the demodulated signal $\mathbf{y}$ at the $j$ -th chip can be represented as

$$y_j = s_j + \delta_j + e_j \tag{10.7}$$

## 3. OPTIMUM PSK SIGNAL CONSTELLATION FOR MP B-CDMA

Since there is no correlation between $\delta_j$ and $e_j$, the variance of $y_j$ can be written as

$$\sigma_y^2 = \sigma_\delta^2 + \sigma_e^2 \tag{10.8}$$

where $\sigma_\delta^2$ and $\sigma_e^2$ are the variance of $\delta_j$ and $e_j$, respectively. Assuming that the spreading code of each user has unit power, the variance of $d_j$ corresponding to the $j$-th chip of **d** is equal to the number of multi-codes, $N_c$. The variance of the quantization noise per each chip, $\sigma_\delta^2$, can be represented as [6]

$$\sigma_\delta^2 = \sum_{i=1}^{N_{pt}} \int_{\Phi_i} (x - m_i)^2 \frac{1}{\sqrt{2\pi N_c}} \exp\left(-\frac{x^2}{2N_c}\right) dx \tag{10.9}$$

The optimum $m_i$ and $\Phi_i$ minimizing $\sigma_\delta^2$ can be obtained using the Lloyd-max algorithm [2]. The Lloyd-max algorithm can find the optimum quantization level in an iterative manner so as to minimize the quantization noise power. The variance of $e_j$ can be calculated as

$$\sigma_e^2 = \sum_{i=1}^{N_{pt}} \sum_{j=1}^{N_{pt}} p(j \mid i) p_s(i) d_{ij}^2 \tag{10.10}$$

where $p_s(i)$ is the probability density function (pdf) of signal point $i$ given by

$$p_s(i) = \int_{\Phi_i} \frac{1}{\sqrt{2\pi N_c}} \exp\left(-\frac{x^2}{2N_c}\right) dx \tag{10.11}$$

*Table 10-1.* Mapping phase

| $E_b/N_o$ | Conventional | Exhaustive search | Proposed |
|---|---|---|---|
| 10 dB | -135°, -90°, -45° | -120°, -85°, -46° | -118°, -75°, -30° |
| 11 dB | -135°, -90°, -45° | -125°, -91°, -45° | -124°, -80°, -34° |
| 12 dB | -135°, -90°, -45° | -127°, -85°, -46° | -129°, -84°, -38° |

and $p(j \mid i)$ is the probability that signal point $i$ is misdetected to $j$ [6]

$$p(j \mid i) = \int_{\Omega_j} \int_0^{\infty} \frac{1}{2\pi} e^{-\gamma_s \sin^2(\theta_r - \theta_i)} R e^{-(R - \sqrt{2\gamma_s}\cos(\theta_r - \theta_i))^2/2} dR d\theta_r \quad (10.12)$$

Here, $\gamma_s$ is the chip energy to noise power ratio ( i.e., $E_c / N_o$ ) and $\theta_i$ is the phase of signal point $i$ . Provided that $\gamma_s$ is high enough, most of chip detection errors are associated with the decision to the adjacent signal points. Thus the variance of the chip detection error can be approximated as

$$\sigma_e^{\,2} = \sum_{i=1}^{N_{pt}} \sum_{j=i_-, i_+} p(j \mid i) p_s(i) d_{ij}^{\,2} \quad (10.13)$$

where $i_-$ and $i_+$ denote the adjacent signal points of signal point $i$.

Since (10.12) involves nonlinear functions, we approximate $p(j \mid i)$ for ease of mathematical analysis. It can be shown that

$$\begin{aligned} p(j \mid i) &= \int_{\Omega_j} \left[ \frac{1}{2\pi} e^{-\gamma_s \sin^2(\theta_r)} \int_0^{\infty} R e^{-(R - \sqrt{2\gamma_s}\cos(\theta_r))^2/2} dR \right] d\theta_r \\ &= \int_{\Omega_j} \left[ \frac{1}{2\pi} \left\{ \begin{array}{l} e^{-\gamma_s} + e^{-\gamma_s \sin^2(\theta_r)} \times \\ \quad \sqrt{\gamma_s \pi}\cos(\theta_r) erfc(-\sqrt{\gamma_s}\cos(\theta_r)) \end{array} \right\} \right] d\theta_r \end{aligned} \quad (10.14)$$

Since the first term in (10.14) can be ignored at high $E_b / N_0$, (10.14) can be approximated as

$$\begin{aligned} p(j \mid i) &\approx \int_{\Omega_j} \frac{1}{2\pi} e^{-\gamma_s \sin^2(\theta_r)} \sqrt{\gamma_s \pi} \\ &\qquad \cos(\theta_r) erfc(-\sqrt{\gamma_s}\cos(\theta_r)) d\theta_r \\ &= \int_{\Omega_j} \frac{1}{2\pi} e^{-\gamma_s \theta_r^2} \left[ \begin{array}{l} e^{-\gamma_s(\sin^2(\theta_r) - \theta_r^2)} \times \\ \sqrt{\gamma_s \pi}\cos(\theta_r) erfc\left(-\sqrt{\gamma_s}\cos(\theta_r)\right) \end{array} \right] d\theta_r \end{aligned} \quad (10.15)$$

For a small $\theta_r$, $p(j \mid i)$ can further be approximated as

$$p(j \mid i) \approx \int_{\Omega_j} \frac{1}{2\pi} e^{-\gamma_s \theta_r^2} \sqrt{\gamma_s \pi} erfc\left(-\sqrt{\gamma_s}\right) d\theta_r \quad (10.16)$$

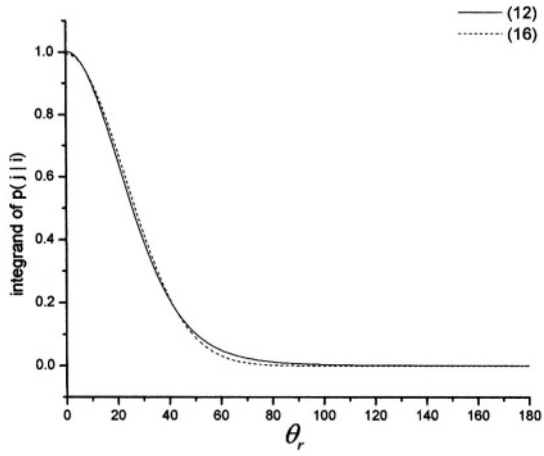



*Figure 10-3.* Approximation of $p(j \mid i)$

As $\theta_r$ increases, the difference between the integrand of (10.12) and (10.16) increases. However the difference between the two integrals is negligible. Fig. 10-3. compares the integrand of (10.12) and (10.16) when $E_b / N_0 = 11dB$, $N_c = 32$, $\theta_i = 0°$ and spreading factor $N_{SF} = 128$. Numerical results indicate that the approximated $p(j \mid i)$ is quite valid in nominal operating condition.

Let $\Omega_j$ be the misdetection region of signal point $i$ to $j$ and the corresponding range be from $\omega_j$ to $\omega_j'$, in (10.16) where, $\omega_j' = \omega_{j+1}$ and $\omega'_{N_{PM}+1} = \omega_1$. Because integrand in (10.16) decreases abruptly as $\theta_r$ increases and is not periodic, $\omega_j'$ can be set to be infinite. Thus, the variance of $e_j$ can be written as

$$\sigma_e^2 = \sum_{i=1}^{N_{PM}} \sum_{j=i_-,i_+} \left( \int_{\omega_j}^{\infty} \frac{1}{2\pi} e^{-\gamma_s \theta_r^2} \sqrt{\gamma_s \pi}\, erfc\left(-\sqrt{\gamma_s}\right) d\theta_r \right) \times \left( \int_{\Phi_i} \frac{1}{\sqrt{2\pi N_c}} \exp\left(-\frac{x^2}{2N_c}\right) dx \right) d_{ij}^2 \tag{10.17}$$

We can find the optimum PSK mapping points $(\theta_1, \theta_2, \cdots, \theta_{N_{PM}})$ that minimize $\sigma_e^2$ using an iterative method. The optimum phase can be found by

$$\frac{\partial \sigma_e^{\,2}}{\partial \theta_i} = 0 \tag{10.18}$$

It can be shown that $\theta_i$ is represented as a function of $\theta_j$

$$\theta_i = f_i(\theta_j) \quad , \quad \forall j \neq i \tag{10.19}$$

It can easily be shown that $\sigma_e^{\,2}$ is a convex function of $\theta_i$ and it has a unique global minimum. We can find $\theta_i$ using an iterative method with an arbitrary initial value. Note that the optimum $\theta_i$ is a the function of $E_b / N_0$. This implies that the optimum PSK mapping points are associated with the value of $E_b / N_0$.

## 4. PERFORMANCE EVALUATION

To verify the performance improvement, we evaluate the performance of the MP B-CDMA system with $N_{psk} = 7$ (i.e., 8-PSK) and one guard phase in AWGN channel using computer simulation. The MP B-CDMA uses an extended PN sequence with $N_{SF} = 128$ and $N_c = 32$ as the spreading code.

From (10.18), the optimum phase $\theta_i$ is determined by

$$\theta_1 = \frac{1}{3\gamma_s}\left(-\theta_2\gamma_s - 4\pi\gamma_s - 2\sqrt{\gamma_s} \times \sqrt{3\ln\left(\frac{4d_{17}^{\,2}p_s(1)}{2d_{12}^{\,2}p_s(1)+2d_{12}^{\,2}p_s(2)}\right)+\theta_2^{\,2}\gamma_s+2\theta_2\pi\gamma_s+\pi^2\gamma_s}\right)$$

$$\theta_2 = -\frac{4\log\left(\frac{2d_{23}^{\,2}p_s(2)+2d_{23}^{\,2}p_s(3)}{2d_{12}^{\,2}p_s(1)+2d_{12}^{\,2}p_s(2)}\right)-\theta_1^{\,2}\gamma_s+\theta_3^{\,2}\gamma_s}{2(\theta_1-\theta_3)\gamma_s} \tag{10.20}$$

$$\theta_3 = \frac{4\log\left(\frac{2d_{34}^{\,2}p_s(3)+2d_{34}^{\,2}p_s(4)}{2d_{32}^{\,2}p_s(2)+2d_{23}^{\,2}p_s(3)}\right)+\theta_2^{\,2}\gamma_s}{2\theta_2\gamma_s}$$

The proposed PSK mapping points $\{\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\}$ can be found by iteratively searching (10.20). Conventional PSK mapping points have an equal distance on the signal constellation with one guard phase, that is,

$\theta_1 = -135°, \theta_2 = -90°, \theta_3 = -45°$. Alternatively, the optimum PSK points can also be found by exhaustive search.

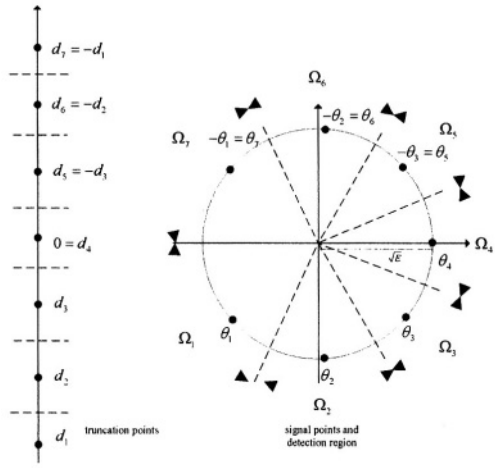



*Figure 10-4.* Proposed 8-PSK signal constellation of the MP B-CDMA

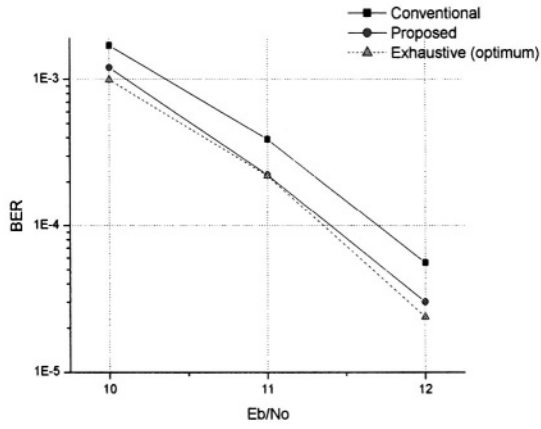



*Figure 10-5.* BER performance due to different phase mapping

Table 10-1. summarizes the conventional, exhaustively searched optimum and proposed PSK mapping points. The proposed point $\hat{\theta}_1$ is similar to exhaustively searched optimum point, but $\hat{\theta}_2$ and $\hat{\theta}_3$ are quite

different from the optimum points. Fig. 10-5. depicts the BER performance of MP B-CDMA with these mapping points. It can be seen that the proposed mapping points can provide the BER performance comparable to the exhaustively searched optimum ones, although $\hat{\theta}_2$ and $\hat{\theta}_3$ are different from the optimum points. This may imply that $\hat{\theta}_1$ with the largest quantization level is the dominant factor on the BER performance.

Fig. 10-6. depicts the performance of the MP B-CDMA system with the mapping points optimized for each $E_b / N_0$. It can be seen that the mapping points optimized under a nominal $E_b / N_0$ condition can be used for certain variation of $E_b / N_0$ without noticeable performance variation.

## 5. CONCLUSIONS

In this chapter, we have analytically determined the optimum PSK mapping points of the MP B-CDMA in AWGN channel. The PSK mapping point is optimized by minimizing the mean square error due to chip detection error. The optimum mapping points and obtained using an iterative method using an analytic expression. The simulation results show that the proposed signal constellation can provide near optimum performance, yielding a BER performance gain of about $0.4dB$ over the conventional one.

## REFERENCE

1. H. S. An, S. M. Ryu and S. W. Na, "Introduction to Binary CDMA," Proc. of *JCCI 2002,* Apr. 2002.
2. T. M. Cover, *Information Theory,* Wiley-Interscience, 1991.
3. E. K. Hong, M. G. Ahn, W. M. Lee and S. M. Ryu, "Design of Signal Constellation for MP/CDMA," Proc. of *JCCI 2001,* Apr. 2001.
4. W. M. Lee, E. K. Hong and Y. S. Park, "Design of Optimum Threshold for Chip-Level Multi Phased MC-CDMA System with Nonlinear Process," North-East Asia IT Symposium.
5. S. P. Kim, M. J. Kim and H. S. An and S. M. Ryu, "A Constant Amplitude Coding for CS-CDMA System," Proc. *of JCCI 2002,* Apr. 2002.
6. J. Y. Ko and Y. H. Lee, "Performance Analysis of Binary-CDMA Transceivers in AWGN Channel," *JCCI 2003. Ch. VI-C-1,* pp. 1-4. Apr. 2003.

Chapter 11

# A COMPLEX QUADRAPHASE CCMA APPROACH FOR MOBILE NETWORKED SYSTEMS

K.L. Brown, M. Darnell
*Institute of Integrated Information Systems, University of Leeds, Leeds, LS1 9JT, UK*

**Abstract:** A novel multiple access coding approach for mobile networked systems is presented based on complex quadraphase coding, matched error control coding and channel adaptation. This approach provides an adaptive multiple-access communications capability where residual spreading gain, coding gain, data rates and channel loading are variables. Based on a blind estimate of the channel state, these parameters can be optimised to satisfy user Quality of Service Requirements. With these techniques, the foundation is laid for a new class of Collaboratively Coded Multiple Access Systems that can operate stand-alone or as hybrid CDMA/CCMA communications systems. A principal advantage of this approach is that high user and service density can be accommodated while the overall system complexity can be managed to tractable levels.

**Key words:** CCMA, Complex Quadraphase, Adaptive CDMA Networks

## 1. INTRODUCTION

Mobile radio systems, in principal, may use several approaches to subdivide channel resources to support multiple user services. The most successful approaches in terms of user density rely on subdivision in time

over discrete frequency bands or variations of Code Division Multiple Access, CDMA. Of these approaches, CDMA has the greatest potential to provide increased user density if suitable decoding methods can be employed at the receiver. The engineering reasoning was substantially extended by the seminal work on Maximum Likelihood decoders by Verdú and others [1]. The replacement in CDMA of the correlative detector by estimating detectors that approach the performance of the Maximum Likelihood Detector then makes it appropriate to look back at the information theoretic principles of the T-adder channel as presented by Kasimi and Lin [2] and the developments that followed by Gallager [3] and Mathys[4]. Other novel encoding and decoding schemes were also described in several published sources [5-9]. It can be seen that the potential of CDMA is more closely aligned with Collaborative Coded Multiple Access, CCMA, than classical variants of CDMA. CCMA may generally be defined by its characteristics such that:

i. sequence sets are optimised for total system capacity
ii. sequence sets are larger than the average lengths of the codewords
iii. some mechanism is employed to select or exploit codeword characteristics as a function of channel characteristics or user/system requirements

Clearly, the advantage was seen by Khachatrian and Martirossian [9], whose high rate T-adder sequence sets were introduced as Synchronous CDMA sequences. T adder Channel sequences can be CCMA codes sets when exploited by channel adaptation and by maximum likelihood decoding. These sequences, though not Welch Bound Equivalent, WBE, have rate sums approaching $r = \left(i \times 2^{i-1} + 1\right)/2^{i}$ which is considerably greater than correlation recovered techniques and classical S-CDMA where the rate sum approaches $r = 1$ from below.

Even with the best known decoding procedures, CCMA has limits in terms of the maximum number of users. Therefore, it is necessary to use other techniques in conjunction with CCMA if very large numbers of users are required. P.Z. Fan and M. Darnell [10] introduced a hybrid scheme combining the strengths of CCMA and SSMA while managing computational complexity due to joint processing of the CCMA codes.

In this work, a CCMA system with complex quadraphase coding is introduced for wireless networked systems. Error control coding is matched in this system to achieve high performance. The system applies adaptive coding and multiplexing techniques for high utilization across a variety of data rates and user requirements. The sequence generation, rate matched error control coding, and adaptive control logic will be demonstrated. System performance will be characterised through simulation in Rayleigh

fading channels and the adaptation metric for the system will be derived through simulation.

## 2. CHANNEL DESCRIPTION

As part of a cell based mobile communications system, an efficient channel coding scheme is derived to support wireless network connectivity between a base station and the mobile receiver as shown in Fig. 11-1. The link between base station and Cellular Network Service Provider and Internet Service Provider is assumed to be outside of the channel description. To support Internet personal and multimedia like data, the subject link between the base station and the Mobile receivers is a conventional connection and the transport layer is serviced by Asynchronous Transfer Mode, ATM, carrying Internet Protocol. The base station to mobile receiver link is addressed in this chapter.

The radio channel is assumed to be Rayleigh fading. The channel within the cell is subdivided by applying high rate complex quadraphase codes and matched error control. The user data is recovered from the channel by synchronously demodulating and decoding with a sequential Maximum likelihood decoder and by using iterative error control decoding. The sequential decoding process provides metrics for channel estimation and data quality that may be used in conjunction with user requirements to apply channel adaptation.

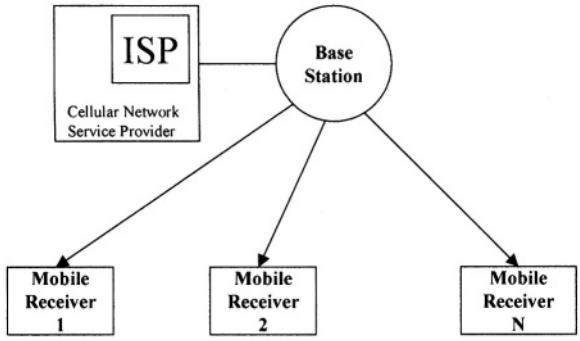



*Figure 11-1.* Simulated Milti-Access Channel Model.

## 3. COMBINING AND CODING

Advances in Digital Signal Processing and manufacturing will soon allow low power devices to exploit the advantages of complex alphabet coding and modulation. A new class of sequences that may operate as quadraphase CCMA sequence sets are derived and have some similarity in construction to the complex Type I CDMA sequences introduced by Brown *et.al.* [10,11]. These sequences are WBE sequences and are optimal under overloaded conditions. They may be generated using the following method: Let *A, B* and *C* represent the second order generator matrices for Type II sequences, where n is the order of the matrices, such that

$$A = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix}, \quad B = \begin{vmatrix} 1 & j \\ -j & -1 \end{vmatrix} \quad \text{and} \quad C = \begin{vmatrix} 1 & j \\ 1 & -j \end{vmatrix}.$$

*C* is a Hadamard generateor and B and C are Type I generators. The first generator, $D_k$, $k = n^2$, is given by

$$D_2 = A \otimes B = \begin{bmatrix} 1 & 1 & j & j \\ 1 & -j & j & -j \\ -j & -1 & -1 & -1 \\ -j & 1 & -1 & 1 \end{bmatrix}.$$

An optimal set of TYPE II codes is given by applying the following steps where $k \geq 3$ for all working sets:

i. Generate a working set by successively applying $D_k = D_1 \otimes D_{k-1}$ for $2^k \geq m$, interchanging *B* and *C*
ii. Sum all rows
iii. Remove rows where sums $\neq 0$
iv. Expurgate rows alternately about the centre of the matrix to obtain the desired length user sequences (columns)

The matrices are now balanced and orthogonal row correlation properties provide sufficient random properties to allow decodability of the columns assuming partial cooperation. Since the code and data alphabet lie on the complex roots of unity where $a_{j,k}$ and $b_{j,k} \in e^{-\pi nj/2}$, and $n \in \{0,1,2,3\}$, the elements of *D* also lies on the root of unity after modulation. These sequences then serve as directly modulatable sequence sets over the same alphabet and construct a $\pi/2$ QPSK signal for combining. The data is now encoded by binary Low Density Parity Check, LDPC, or Turbo codes. The selection of block and interleaver lengths, respectively, is chosen to mitigate fading while imposing acceptable delay. After the data is encoded, the encoded stream is mapped to $\pi/2$ phasing to modulate the code sets.

## 4. DATA RECOVERY AND ADAPTATION

Data recovery is performed after frequency translation by a complex sequential decoder over $R(4)$. The decoder search performs a minimum least squares match of the received code word with all possible true codewords. The solution with the minimum least square difference is also the most likely codeword. Since the codeword sums are uniquely decodable, the modulating character of the sequence with the minimum least square difference is also the most likely transmitted symbol. Since the decoder decision points occur at the $\pi/2$ QPSK reference, the decoder can pass a partitioned decision to the ECC decoder to perform soft decision binary ECC decoding. The binary decisions from the ECC decoder are mapped to the original complex data input alphabet for analysis.

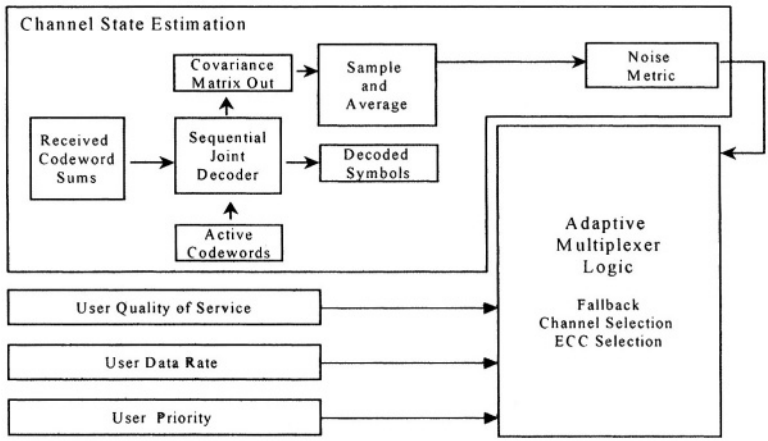



*Figure 11-2.* Adaptation state block diagram.

With the objective of exploiting the relationship between processing gain and the number of active users, it has been demonstrated previously[11,12] that blind channel adaptation can be achieved based on the observation of the signal detection statistics. Here, the minimum least square difference already determines the most likely transmitted codeword. This computation yields a time series of vectors whose magnitudes form a measure of channel performance. The log of the magnitude, if sampled and averaged with a period less than the channels coherence time, provides a useful estimate of the channel's performance without additional computation. This performance metric is strongly related to the received SNR and BER. Since the metric is computed with little additional computational complexity, it can be used as the dominant metric for channel adaptation.

The adaptation space is mapped over the spreading gain or equivalent user codeword distance that increases as the number of active users is decreased and coding gain which in some cases can be increased by the selection of lower rate error control codes and/or longer block lengths in the case of LDPC codes. The decision state vectors then contain the required user quality of service, user data rate, user priority, channel state variables and a channel adaptation metric. The decision flow is shown in Fig. 11-2.

## 5. COMPUTER SIMULATION

To construct a system model, a string within the MAC model is derived for computer modeling as shown in Figure 1. The user code sequences are reconstructed in a matrix containing all codewords and are modulated by a random vector over the complex alphabet. This data set represents independent data from each transmitting user. The set is then combined and the sum corrupted by a simulated Rayleigh slow fading channel with additive white Gaussian noise. All channels of the multiplexed signal are then recovered by a simulated receiver performing a sequential search of the codewords exhibiting the minimum square difference from the received codeword. An adaptation metric computation is included in the simulation to verify the correlation between error rate performance, SNR and the adaptation metric slope. All simulations were performed with the assumption of perfect bit synchronization and perfect phase estimation by the receiver.

## 6. SIMULATION RESULTS

The simulation was performed with the objective of verifying the performance of the coding sequences performance and the adaptive potential of the system. The adaptive properties given by the increase in coding and spreading gain property are particularly important when channel noise characteristics vary over a large range and when users are able to use multiple codewords to increase throughput.

The data presented to the channel is transported by the ATM protocol structure. Thus it is important to determine the contributions of bit errors to the ATM header and payload with respect to $E_b/N_o$. It is reasonable to use this as a reference in estimating the performance of higher-level protocols and the end application. A plot of the ATM header and payload error statistics with respect to $E_b/N_o$ is shown in Fig. 11-3.

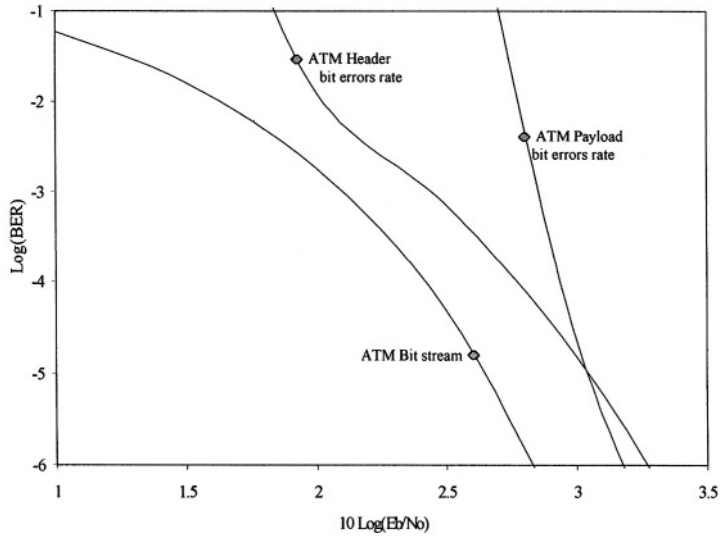



*Figure 11-3.* $E_b/N_o$ vs. BER ATM bit streams, header and payloads for turbo encoded fully loaded channel in AWGN.

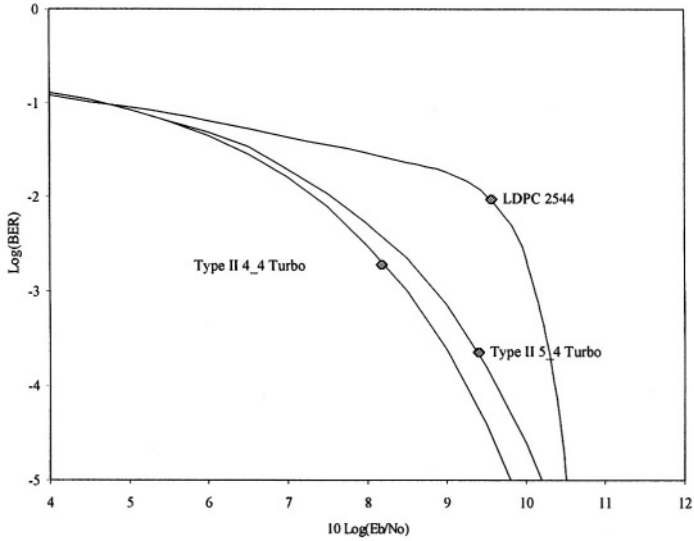



*Figure 11-4.* BER vs. $10\ \mathrm{Log}(E_b/N_o)$ for fading channel

Simulations for fading channels using turbo coding and LDPC coding were performed for fully loaded and overloaded systems. The turbo coder

used is the original 0.5 rate, 16 state coder with a 4000 bit length interleaver. The LDPC code is an irregular code with weight 3 columns and block size of 2742. The fade velocity is 100 km/h. The bit error rate versus $E_b/N_o$ performance of the system is shown in Fig. 11-4.

The performance of the adaptation metric was simulated across the full range of bit error rates. Simulation results show that the log of the error metric closely conforms to a straight-line relationship with Log of the bit error rate. In addition, the results were repeatable with small numbers of averages between $10^{-2}$ and the limit of the testing range of $10^{-6}$ bit error rate. The results of the error metric simulations are shown in Fig. 11-5. Over this range, the error metric is a reliable channel estimator.

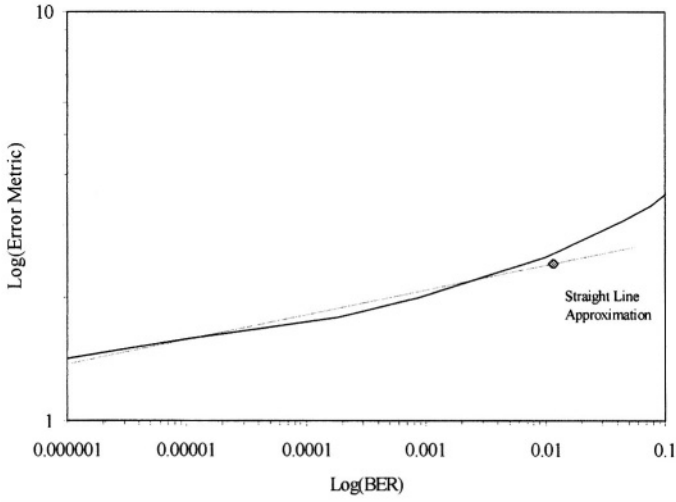



*Figure 11-5.* BER vs. Log(Error Metric) for fading channel.

## 7. SUMMARY

The construction techniques and simulations developed in this work demonstrate the potential for quadraphase coded adaptive CCMA and data recovery method for the overloaded MAC. For a fully loaded large system, the systems performance is identical to fully loaded systems using Type I complex quadraphase codes. For overloaded systems, the incremental user performance penalty meets with expectation for WBE encoded systems.

In comparison to other collaborative coded systems, this approach produces small alphabet size prior to quantising, which is twice that of real

valued T-adder channel codes proposed by Khachatrian and Martirossian [9]. The system has better performance because the sequences are WBE and use maximum likelihood decoding. In addition, the combined alphabet size is approximately the square root of the code sum alphabet produced by the Collaborative Codes as analyzed by Soysa *et.al.* [5]. This provides the potential benefits of much lower implementation complexity and better performance.

It has been shown in this work that the computed adaptation metric, with a small amount of reverse channel capacity, can provide the necessary input to control the adaptation process. In addition, if one assumes that channel noise characteristics vary slowly, relative to the symbol rate, it can be shown that the adaptation metric is a good estimator of the channel state with little additional computational complexity with increasing numbers of users. Thus, for most practical instances, the adaptive components add little complexity to the system's overall complexity, even as the number of users becomes large. If a parallelised sequential maximum likelihood decoder can be economically packaged, this approach may allow the capabilities of CCMA to replace classical CDMA in many applications.

## REFERENCES


1. A. Tulino, and S. Verdú, Asymptotic Analysis of Improved Linear Receivers for BPSK-CDMA Subject to Fading, *IEEE Transactions on Information Theory,* **IT-19**(8), 1544-1555 August (2001).
2. T. Kasami and S. Lin, Coding for a Multiple Access Channel, *IEEE Transactions on Information Theory,* **IT**(2), 129-137, March (1976).
3. R. G. Gallager, A Perspective on Multi-access Channels, *IEEE Transactions on Information Theory,* **IT-31(2),** 124-142, March (1985).
4. P. Mathys, A Class of Codes for a T Active User Out of N Multiple - Access Communications System, *IEEE Transactions on Information Theory,* **36**(6), 1206-1219, November, (1990).
5. S. Soysa, F. H. Ali, and S. A. G. Chandler, Comparison of T-User M-PSK CV_CCMA and Multi-level Modulation, *Proceedings of ISCTA,* July, (1997) pp. 419-422.
6. B. Honary, L. Kaya, G. S. Markarian, and M. Darnell, Maximum-likelihood decoding of array codes with trellis structure, *IEE Proceedings-I,* **140**(5), 340-345, October (1993).
7. J. A. Gordon and R. Barrett, Correlation-recovered Adaptive Majority Multiplexing, *Proceedings of IEE,* Vol.**118**(3/4), 417-422, March/April (1971).
8. P. Z. Fan, P. Z., and M. Darnell, Hybrid CCMA/SSMA coding scheme, *Electron. Letters,* Vol. **30**(25), 2105-2106, December, (1994).
9. G. H. Khachatrian and S. S. Martirossian, A new approach the design of codes for synchronous CDMA systems, *IEEE Transactions on Information Theory,* **IT-41,** 1503-1506, (1995).

*This page intentionally left blank*

Chapter 12

# SPATIAL CHARACTERIZATION OF MULTIPLE ANTENNA CHANNELS

Tony S. Pollock, Thushara D. Abhayapala, and Rodney A. Kennedy

*Wireless Signal Processing Program, National ICT Australia, and Department of Telecommunications Engineering, RSISE, The Australian National University, Locked Bag 8001, Canberra ACT 2601, Australia.*

(tony.pollock, thushara.abhayapala, rodney.kennedy)@nicta.com.au

**Abstract** In this chapter we present a realistic new model for wireless multiple-input multiple-output (MIMO) channels which is more general than previous models. A novel spatial decomposition of the channel is developed to provide insights into the spatial aspects of multiple antenna communication systems. By exploiting the underlying physics of free-space wave propagation we characterize the fundamental communication modes of a physical aperture and develop an intrinsic capacity which is independent of antenna array geometries and array signal processing. We show there exists a maximum achievable capacity for communication between spatial regions of space, which depends on the size of the regions and the statistics of the scattering environment.

**Keywords:** multiple antennas, capacity, antenna arrays, MIMO, channel modelling

## 1. INTRODUCTION

Multiple-Input Multiple-Output (MIMO) communications systems using multi-antenna arrays simultaneously during transmission and reception have generated significant interest in recent years. Theoretical work of [1,2] showed the potential for significant capacity increases in wireless channels via spatial multiplexing with sparse antenna arrays. With these developments comes the need for better understanding of the spatial properties of the wireless communications channel. The spatial properties of multiple antenna channels have significant impact on the capacity of MIMO systems, therefore, a good understanding of these properties

is required for effective design and implementation of wireless MIMO systems.

For randomly fading channels, much of the literature is limited to the idealistic situation of independent and identically distributed (i.i.d.) Gaussian channels, where the channel gains are modelled as independent Gaussian random variables (for example see [1,2]). The i.i.d. model corresponds to sufficiently spaced antennas such that there is no spatial correlation between antenna elements at the transmit and receive arrays, along with significant scattering between arrays. However, in practice, realistic scattering environments and limited antenna separation leads to channels which exhibit correlated fades.

For correlated fading, MIMO channel modelling can be approached via field measurements [3–6], and deterministic physical models such as ray tracing [7,8], where the significant characteristics of the channel are obtained and incorporated into the model. Such methods give an accurate characterization of the channel, however, they are computationally expensive and provide results for specific scenarios only. Finally, a statistical model can be postulated which attempts to capture the physical channel characteristics based on the basic principles of radio propagation [9–12]. These scattering models can often be used as simple analysis tools which illustrate the essential characteristics of the MIMO channel, provided the constructed scattering environment is reasonable.

With the notable exception of [10] and [12], the statistical models mentioned above have poor physical significance. In particular, the separate effects of the scatterers and the antenna correlation are not accounted for. As outlined in [10], the models assume that only the spatial fading correlation is responsible for the rank structure of the MIMO channel. In practice, however, high rank MIMO channels correspond not only to the low fading correlation, but also to the structure of scattering in the propagation environment.

The models presented in [10,12] allow for insight into the effects of spatial correlation and scattering, however, they are unfortunately limited to particular array geometries and model the scattering environment using a discrete representation. Therefore, although offering considerable insight into the scattering characteristics of the channel they are restricted spatially, in the sense that the antenna geometry is restricted to a particular array configuration and discrete scattering environments.

In contrast to previous models, the contribution of this chapter is a spatial channel model which includes the physical parameters of arbitrary antenna configurations and a tractable parameterization of the complex scattering environment. We approach the MIMO channel modelling problem from a physical wave field perspective. By using the

underlying physics of free-space wave propagation we explore the fundamental properties of the channel due to constraints imposed by the basic laws governing wave field behavior. Furthermore, we show that there exists a maximum achievable capacity for communication between spatial regions of space, which depends on the size of the regions and the statistics of the scattering environment. This bound on capacity gives the optimal MIMO capacity and thus provides a benchmark for future array and space-time coding developments.

## 2. CHANNEL MODEL

Consider the 2D MIMO system shown in Fig, 12-1, where the transmitter consists of $n_{\mathrm{T}}$ transmit antennas located within a circular aperture of radius $r_{\mathrm{T}}$. Similarly, at the receiver, there are $n_{\mathrm{R}}$ antennas within a circular aperture of radius $r_{\mathrm{R}}$. Denote the $n_{\mathrm{T}}$ transmit antenna positions by $\boldsymbol{x}_t = (\|\boldsymbol{x}_t\|, \theta_t)$, $t = 1, 2, \ldots, n_{\mathrm{T}}$, in polar coordinates, relative to the origin of the transmit aperture, and the $n_{\mathrm{R}}$ receive antenna positions by $\boldsymbol{y}_r = (\|\boldsymbol{y}_r\|, \varphi_r)$, $r = 1, 2, \ldots, n_{\mathrm{R}}$, relative to the origin of the receive aperture. Note that all transmit and receive antennas are constrained to within the transmit and receive apertures respectively, that is, $\|\boldsymbol{x}_t\| \leq r_{\mathrm{T}}, \forall t$, and $\|\boldsymbol{y}_r\| \leq r_{\mathrm{R}}, \forall r$. It is also assumed that the scatterers are distributed in the farfield from all transmit and receive antennas, therefore, define circular scatterer free regions of radius $r_{\mathrm{TS}} > r_{\mathrm{T}}$, and $r_{\mathrm{RS}} > r_{\mathrm{R}}$, such that any scatterers are in the farfield to any antenna within the transmit and receive apertures, respectively.

Finally, the random scattering environment is defined by the effective random complex scattering gain $g(\phi, \psi)$ for a signal leaving from the transmit aperture at an angle $\phi$, and entering the receive aperture at an angle $\psi$, via any number of paths through the scattering environment.

Consider the narrowband transmission of $n_{\mathrm{T}}$ baseband signals, $\{x_t\}$, $t = 1, \ldots, n_{\mathrm{T}}$, over a single signalling interval from the $n_{\mathrm{T}}$ transmit antennas located within the transmit aperture. From Fig. 12-1 the noiseless signal at $\boldsymbol{y}_r$ is given by

$$z_r = \sum_{t=1}^{n_{\mathrm{T}}} x_t \iint_{\mathbb{S}^1} g(\phi, \psi)\, e^{ik\|\boldsymbol{x}_t\| \cos(\theta_t - \phi)}\, e^{-ik\|\boldsymbol{y}_r\| \cos(\varphi_r - \psi)}\, d\phi d\psi. \quad (12.1)$$

where $\mathbb{S}^1$ denotes the unit circle.

Denote $\boldsymbol{x} = [x_1, x_2, \ldots, x_{n_{\mathrm{T}}}]'$ as the column vector of the transmitted signals, and $\boldsymbol{n} = [n_1, n_2, \ldots, n_{n_{\mathrm{R}}}]'$, as the noise vector where $n_r$ is the independent additive white Gaussian noise (AWGN) with variance $N_0 \in \mathcal{N}(0, 1)$ at the $r$-th receive antenna, then the vector of received signals

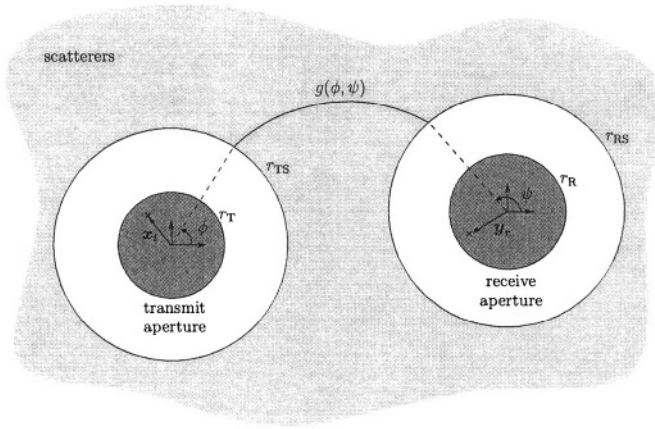



*Figure 12-1.* Scattering model for a 2D flat fading narrowband MIMO system. $r_T$ and $r_R$ are the radii of circular apertures which contain the transmit and receive antenna arrays, respectively. The radii $r_{TS}$ and $r_{RS}$ describe scatterer free circular regions surrounding the transmit and receive apertures, assumed large enough that any scatterer is farfield to all antennas. The scattering environment is described by $g(\phi,\psi)$ which gives the effective random complex gain for signals departing the transmit aperture from angle $\phi$ and arriving at the receive aperture from angle $\psi$, via any number of scattering paths.

$\boldsymbol{y} = [y_1, y_2, \ldots, y_{n_R}]'$ is given by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{12.2}$$

where $\boldsymbol{H}$ is the complex random channel matrix with $r,t$-th element

$$\boldsymbol{H}|_{r,t} = \iint_{\mathbb{S}^1} g(\phi,\psi) e^{ik\|\boldsymbol{x}_t\|\cos(\theta_t-\phi)}\, e^{-ik\|\boldsymbol{y}_r\|\cos(\varphi_r-\psi)}\, d\phi d\psi, \tag{12.3}$$

representing the channel gain between the $t$-th transmit antennna and the $r$-th receive antenna.

## 2.1 Channel Matrix Decomposition

Consider the modal expansion[1] of the plane wave [13]

$$e^{ik\|\boldsymbol{x}\|\cos(\theta_x-\phi)} = \sum_{n=-\infty}^{\infty} i^n J_n(k\|\boldsymbol{x}\|) e^{-in(\theta_x-\phi)}, \tag{12.4}$$

[1]Each mode, indexed by n, corresponds to a different solution of the governing electromagnetic equations (Maxwell's equations) for the given boundary conditions.

for vector $\boldsymbol{x} = (\|\boldsymbol{x}\|, \theta_x)$, and $J_n(\cdot)$ are the Bessel functions of the first kind.

Bessel functions $J_n(z)$, $|n| > 0$ exhibit spatially high pass behavior, that is, for fixed order $n$, $J_n(z)$ starts small and becomes significant for arguments $z \approx \mathcal{O}(n)$. Therefore, for a fixed argument $z$, the Bessel function $J_n(z) \approx 0$ for all but a finite set of low order modes $n \leq N$, hence (12.4) is well approximated by the finite sum

$$e^{ik\|\boldsymbol{x}\|\cos(\theta_x - \phi)} \simeq \sum_{n=-N}^{N} \overline{\mathcal{J}_n(\boldsymbol{x})} e^{in\phi}, \tag{12.5}$$

where $\overline{\mathcal{J}_n(\boldsymbol{x})}$ is the complex conjugate of function $\mathcal{J}_n(\boldsymbol{x})$, defined as the *spatial-to-mode* function

$$\mathcal{J}_n(\boldsymbol{x}) \triangleq J_n(k\|\boldsymbol{x}\|) e^{in(\phi_x - \pi/2)}, \tag{12.6}$$

which maps the sampling point $\boldsymbol{x}$ to the $n$th mode of the expansion (12.4). In [14] it was shown that $J_n(z) \approx 0$ for $n > \lceil ze/2 \rceil$, with $\lceil \cdot \rceil$ the ceiling operator. Therefore, we can define

$$N_{\mathrm{T}} \triangleq \lceil \pi e r_{\mathrm{T}} / \lambda \rceil, \tag{12.7}$$

$$N_{\mathrm{R}} \triangleq \lceil \pi e r_{\mathrm{R}} / \lambda \rceil, \tag{12.8}$$

such that the truncated expansions

$$e^{ik\|\boldsymbol{x}_t\|\cos(\theta_t - \phi)} \simeq \sum_{n=-N_{\mathrm{T}}}^{N_{\mathrm{T}}} \overline{\mathcal{J}_n(\boldsymbol{x}_t)} e^{in\phi}, \tag{12.9}$$

$$e^{-ik\|\boldsymbol{y}_r\|\cos(\varphi_r - \psi)} \simeq \sum_{m=-N_{\mathrm{R}}}^{N_{\mathrm{R}}} \mathcal{J}_m(\boldsymbol{y}_r) e^{-in\psi}, \tag{12.10}$$

hold for every antenna within the transmit and receive apertures of radius $r_{\mathrm{T}}$ and $r_{\mathrm{R}}$, respectively.

Substitution of (12.9) and (12.10) into (12.3), gives the closed-form expression for the channel gain between the $t$-th transmit antenna and $r$-th receive antenna as

$$\boldsymbol{H}|_{r,t} = \sum_{n=-N_{\mathrm{T}}}^{N_{\mathrm{T}}} \sum_{m=-N_{\mathrm{R}}}^{N_{\mathrm{R}}} \overline{\mathcal{J}_n(\boldsymbol{x}_t)} \mathcal{J}_m(\boldsymbol{y}_r) \iint_{\mathbb{S}^1} g(\phi, \psi) e^{in\phi} e^{-im\psi} d\phi d\psi. \tag{12.11}$$

From (12.11) the channel matrix $\boldsymbol{H}$ can be decomposed into a product of three matrices, which correspond to the three spatial regions of signal

propagation,

$$\boldsymbol{H} = \boldsymbol{J}_{\mathrm{R}} \boldsymbol{H}_{\mathrm{S}} \boldsymbol{J}_{\mathrm{T}}^{\dagger}, \tag{12.12}$$

where $\boldsymbol{J}_{\mathrm{T}}$ is the $n_{\mathrm{T}} \times (2N_{\mathrm{T}} + 1)$ *transmit aperture sampling* matrix,

$$\boldsymbol{J}_{\mathrm{T}} = \begin{bmatrix} \mathcal{J}_{-N_{\mathrm{T}}}(\boldsymbol{x}_1) & \cdots & \mathcal{J}_{N_{\mathrm{T}}}(\boldsymbol{x}_1) \\ \mathcal{J}_{-N_{\mathrm{T}}}(\boldsymbol{x}_2) & \cdots & \mathcal{J}_{N_{\mathrm{T}}}(\boldsymbol{x}_2) \\ \vdots & \ddots & \vdots \\ \mathcal{J}_{-N_{\mathrm{T}}}(\boldsymbol{x}_{n_{\mathrm{T}}}) & \cdots & \mathcal{J}_{N_{\mathrm{T}}}(\boldsymbol{x}_{n_{\mathrm{T}}}) \end{bmatrix}, \tag{12.13}$$

which describes the sampling of the transmit aperture, $\boldsymbol{J}_{\mathrm{R}}$ is the $n_{\mathrm{R}} \times (2N_{\mathrm{R}} + 1)$ *receive aperture sampling* matrix,

$$\boldsymbol{J}_{\mathrm{R}} = \begin{bmatrix} \mathcal{J}_{-N_{\mathrm{R}}}(\boldsymbol{y}_1) & \cdots & \mathcal{J}_{N_{\mathrm{R}}}(\boldsymbol{y}_1) \\ \mathcal{J}_{-N_{\mathrm{R}}}(\boldsymbol{y}_2) & \cdots & \mathcal{J}_{N_{\mathrm{R}}}(\boldsymbol{y}_2) \\ \vdots & \ddots & \vdots \\ \mathcal{J}_{-N_{\mathrm{R}}}(\boldsymbol{y}_{n_{\mathrm{R}}}) & \cdots & \mathcal{J}_{N_{\mathrm{R}}}(\boldsymbol{y}_{n_{\mathrm{R}}}) \end{bmatrix}, \tag{12.14}$$

which describes the sampling of the receive aperture, and $\boldsymbol{H}_{\mathrm{S}}$ is a $(2N_{\mathrm{R}} + 1) \times (2N_{\mathrm{T}} + 1)$ *scattering environment* matrix, with $p, q$-th element

$$\boldsymbol{H}_{\mathrm{S}}|_{p,q} = \iint_{\mathbb{S}^1} g(\phi, \psi) e^{i(q - N_{\mathrm{T}} - 1)\phi} e^{-i(p - N_{\mathrm{R}} - 1)\psi} d\phi d\psi, \tag{12.15}$$

representing the complex gain between the $(q - N_{\mathrm{T}} - 1)$-th mode of the transmit aperture and the $(p - N_{\mathrm{R}} - 1)$-th mode of the receive aperture[2].

The channel matrix decomposition (12.12) separates the channel into three distinct regions of signal propagation: free space transmitter region, scattering region, and free space receiver region, as shown in Fig. 12-1. The transmit aperture and receive aperture sampling matrices, $\boldsymbol{J}_{\mathrm{T}}$ and $\boldsymbol{J}_{\mathrm{R}}$, describe the mapping of the transmitted signals to the modes of the system, and the modes to received signals, given the respective positions of the antennas, and are constant for fixed antenna locations within the spatial apertures. Conversely, for a random scattering environment the scattering channel matrix $\boldsymbol{H}_{\mathrm{S}}$ will have random elements.

## 3. MODE-TO-MODE COMMUNICATION

It is well known that the rank of the channel matrix $\boldsymbol{H}$ gives the effective number of independent parallel channels between the transmit and receive antenna arrays, and thus determines the capacity of

[2]It is important to note the distinction between the *mode-to-mode* gains due to the scattering environment described by $\boldsymbol{H}_{\mathrm{S}}$, and the *antenna-to-antenna* channel gains described by $\boldsymbol{H}$.

the system. For the decomposition (12.12) the rank of $\boldsymbol{H}$ is given by $\min\{\text{rank}(\boldsymbol{J_T}), \text{rank}(\boldsymbol{J_R}), \text{rank}(\boldsymbol{H}_S)\}$, which for a large number of antennas and finite regions, becomes $\min\{2N_T + 1, 2N_R + 1, \text{rank}(\boldsymbol{H}_S)\}$. Therefore we see that the number of available modes for the transmit and receive apertures, determined by the size of the apertures, and any possible modal correlation or key-hole effects [15] (rank 1 $\boldsymbol{H}_S$) limit the capacity of the system, regardless of how many antennas are packed into the apertures.

Assume $n_T = 2N_T + 1$ and $n_R = 2N_R + 1$ antennas are optimally placed (perfect spatial-to-mode coupling) within the transmit and receive regions of radius $r_T$ and $r_R$, respectively, with total transmit power $P_T$. In this situation $\boldsymbol{J}_T\boldsymbol{J}_T^\dagger = \boldsymbol{I}$ and $\boldsymbol{J}_R^\dagger\boldsymbol{J}_R = \boldsymbol{I}$, hence the transmit and receive aperture sampling matrices are unitary and $\boldsymbol{H}_S$ is then unitarily equivalent to $\boldsymbol{H}$. The instantaneous channel capacity with no channel state information at the transmitter and full channel knowledge at the receiver [2] is then given by

$$C = \log\left|\boldsymbol{I}_{2N_R+1} + \frac{\eta}{2N_T + 1}\boldsymbol{H}_S\boldsymbol{H}_S^\dagger\right|, \tag{12.16}$$

where $\eta = P_T/N_0$ is the average SNR at any point within the receive aperture.

The ergodic capacity of uniform linear (ULA) and uniform circular (UCA) arrays are shown in Fig. 12-2 for an increasing number of antennas constrained within transmit and receive apertures of radius $0.8\lambda$, i.e. the physical size of the array remains fixed as the number of antennas is increased. Here we can see that by spatially constraining the antenna arrays the capacity growth saturates and, unlike the i.i.d. case, provides no further capacity improvement with increasing numbers of antennas. The mode-to-mode capacity (12.16) represents the intrinsic capacity for communication between two spatial apertures, giving the maximum capacity for all possible array configurations and array signal processing. We can see from (12.7) and (12.8) that the intrinsic capacity is limited by the size of the regions containing the antenna arrays (number of available modes), and the statistics of the scattering channel matrix (modal correlation).

Fig. 12-3 shows the radiation pattern of the first 6 modes of the circular and spherical apertures[3]. Each mode has a unique radiation pattern, therefore, mode-to-mode communication can be considered as a pattern diversity scheme, where the signals obtained by different modes may

[3]For extension of the model to the 3D spatial environment see [17].

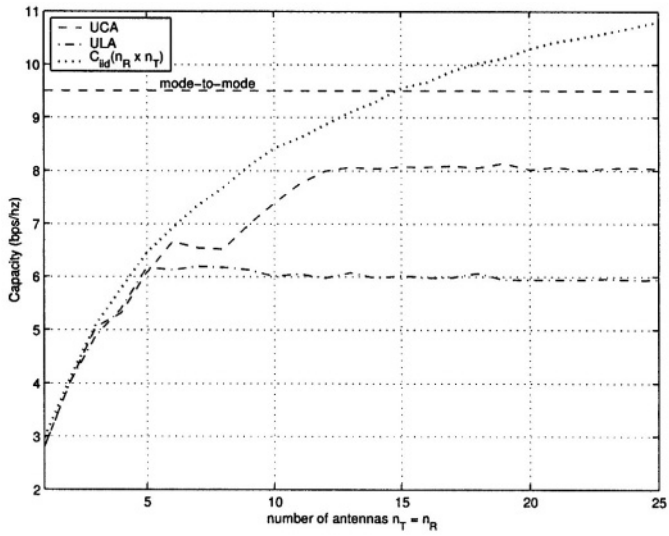



*Figure 12-2.* Ergodic capacity for increasing number of antennas of uniform linear (ULA) and uniform circular (UCA) arrays constrained within spatial regions of radius $0.8\lambda$ and isotropic scattering. The mode-to-mode capacity gives the maximum achievable capacity between the two apertures.

be combined to yield a diversity gain. However, the level of diversity achieved depends on the correlation between the modes, which strongly depends on the scattering environment as shown in the following section.

## 4. PROPERTIES AND STATISTICS OF SCATTERING CHANNEL MATRIX $H_S$

As the scattering gain function $g(\phi, \psi)$ is periodic with $\phi$ and $\psi$ it can be expressed using a Fourier expansion. For this 2D model with circular apertures a natural choice of basis functions are the orthogonal circular harmonics $e^{in\phi}$ which form a complete orthogonal function basis set on the unit circle[4], thus express

$$g(\phi, \psi) = \frac{1}{4\pi^2} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \beta_m^n e^{-in\phi} e^{im\psi}, \qquad (12.17)$$

[4] with respect to the natural inner product $\langle f, g \rangle = \int_0^{2\pi} f(\phi)\overline{g(\phi)}d\phi$

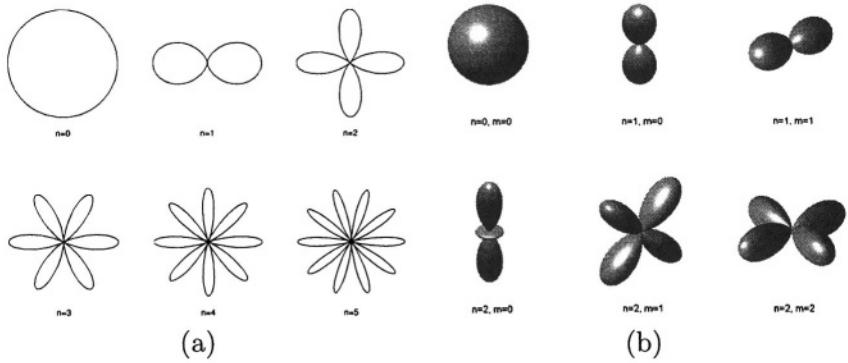



*Figure 12-3.* Radiation patterns of the first 6 modes of a (a) circular and (b) spherical aperture.

with coefficients

$$\beta_m^n = \iint_{\mathbb{S}^1} g(\phi,\psi) e^{in\phi} e^{-im\psi} d\phi d\psi. \tag{12.18}$$

Therefore, letting $n = q - N_{\mathrm{T}} - 1$, and $m = p - N_{\mathrm{R}} - 1$ denote the transmitter mode and receiver mode index, respectively, the scattering environment matrix coefficients are given by

$$\boldsymbol{H}_{\mathrm{S}}|_{p,q} = \beta_{p-N_{\mathrm{R}}-1}^{q-N_{\mathrm{T}}-1} = \beta_m^n. \tag{12.19}$$

Thus the random scattering environment can be parameterized by the complex random coefficients $\beta_m^n$, $n = -N_{\mathrm{T}}, \ldots, N_{\mathrm{T}}$, $m = -N_{\mathrm{R}}, \ldots, N_{\mathrm{R}}$, which gives the scattering gain between the $n$-th transmit mode and the $m$-th receive mode, and $\boldsymbol{H}_{\mathrm{S}}$ becomes

$$\boldsymbol{H}_{\mathrm{S}} = \begin{bmatrix} \beta_{-N_{\mathrm{R}}}^{-N_{\mathrm{T}}} & \cdots & \beta_{-N_{\mathrm{R}}}^{N_{\mathrm{T}}} \\ \beta_{-N_{\mathrm{R}}+1}^{-N_{\mathrm{T}}} & \cdots & \beta_{-N_{\mathrm{R}}+1}^{N_{\mathrm{T}}} \\ \vdots & \ddots & \vdots \\ \beta_{N_{\mathrm{R}}}^{-N_{\mathrm{T}}} & \cdots & \beta_{N_{\mathrm{R}}}^{N_{\mathrm{T}}} \end{bmatrix}. \tag{12.20}$$

Assuming a zero-mean uncorrelated scattering environment (Rayleigh), the scattering channel is characterized by the second-order statistics of the scattering gain function $g(\phi,\psi)$, given by,

$$E\left\{g(\phi,\psi)\overline{g(\phi',\psi')}\right\} = G(\phi,\psi)\delta(\phi-\phi')\delta(\psi-\psi'), \tag{12.21}$$

where $\delta(\cdot)$ is the Kronecker delta function, and $G(\phi,\psi) = E\left\{|g(\phi,\psi)|^2\right\}$ is the 2D power spectral density (PSD) of the modal correlation function,

$$\gamma_{n-n';m-m'} \triangleq E\left\{\beta_m^n \overline{\beta_{m'}^{n'}}\right\} = \iint_{\mathbb{S}^1} G(\phi,\psi) e^{i(n-n')\phi} e^{-i(m-m')\psi} d\phi d\psi, \tag{12.22}$$

and represents the scattering channel power over departure and arrival angles $\phi$ and $\psi$, normalized such that the total scattering channel power

$$\sigma^2_{\mathbf{H}_S} = \iint_{\mathbb{S}^1} G(\phi,\psi) d\phi d\psi = 1. \tag{12.23}$$

For the special case of uniform PSD, $G(\phi,\psi) = 1/4\pi^2$, the modal correlation becomes

$$\gamma_{n-n',m-m'} = \gamma_{0,0}\delta_{n-n'}\delta_{m-m'}, \tag{12.24}$$

corresponding to the i.i.d. $\{\beta_n^m\}$ case.

## 4.1 Modal Correlation in General Scattering Environments

Define $\mathcal{P}(\psi)$as the average power density of the scatterers surrounding the receiver, given by the marginalized PSD

$$\mathcal{P}(\psi) \triangleq \int_{\mathbb{S}^1} G(\phi,\psi)\, d\phi, \tag{12.25}$$

then, from (12.22) we see the modal correlation between the $m$ and $m'$ communication modes at the receiver is given by

$$\gamma_{m-m'} = \int_{\mathbb{S}^1} \mathcal{P}(\psi) e^{-i(m-m')\psi} d\psi, \tag{12.26}$$

which gives the modal correlation for all common power distributions $\mathcal{P}(\psi)$: von-Mises, gaussian, truncated gaussian, uniform, piecewise constant, polynomial, Laplacian, Fourier series expansion, etc. Similarly, defining $\mathcal{P}(\phi)$ as the power density of the scatterers surrounding the transmitter, we have the modal correlation at the transmitter

$$\gamma_{n-n'} = \int_{\mathbb{S}^1} \mathcal{P}(\phi) e^{i(n-n')\phi} d\phi. \tag{12.27}$$

As shown in [18] there is very little variation in the correlation due to the various non-isotropic distributions mentioned above, therefore

without loss of generality, we restrict our attention to the case of energy arriving uniformly over limited angular spread $\Delta$ around mean $\psi_0$, i.e., $(\psi_0 - \Delta, \psi_0 + \Delta)$. In this case the modal correlation is given by

$$\gamma_{m-m'} = \text{sinc}((m - m')\Delta)e^{-i(m-m')\psi_0}, \qquad (12.28)$$

which is shown in Fig. 12-4 for various modes and angular spread. As one would expect, for increasing angular spread we see a decrease in modal correlation, with more rapid reduction for well separated mode orders, e.g. large $|m - m'|$. For the special case of a uniform isotropic scattering environment, $\Delta = \pi$, we have zero correlation between all modes, e.g., $\gamma_{m-m'} = \delta_{m-m'}$.

Fig. 12-5 shows the impact of modal correlation on the ergodic mode-to-mode capacity for increasing angular spread at the transmitter and isotropic scattering at the receiver[5] for 10dB SNR. We consider transmit and receive apertures of radius $0.8\lambda$, corresponding to $2\lceil \pi e 0.8 \rceil + 1 = 15$ modes at each aperture. For comparison, the capacity for a 15 antenna ULA and UCA, contained within the same aperture size is presented. Also shown is the $15 \times 15$ antenna i.i.d. case, corresponding to the rich scattering environment with no restrictions on the antenna placement, i.e., $r_T, r_R \to \infty$.

The mode-to-mode capacity is the maximum achievable capacity between the two apertures, and represents the upper bound on capacity for any antenna array geometry or multi-mode antennas constrained within those apertures. All four cases show no capacity growth for angular spread greater than approximately 60°, which corresponds to low modal correlations ($\ll 0.5$) for the majority of modes, as seen in Fig. 12-4.

## 5. DISCUSSION

In this chapter we have presented a novel multiple antenna channel model which includes the spatial aspects of a MIMO system not previously considered. The spatial channel model developed includes the physical parameters of arbitrary antenna configurations (number of antenna and their location) and a tractable parameterization of the complex scattering environment.

Using the model we have developed a new upper bound on the capacity for communication between regions in space. Using the underlying physics of free space wave propagation we have shown that there is a

[5]This models a typical mobile communication scenario, where the receiver is usually surrounded by scatterers, and the base station is mounted high above the scattering environment.

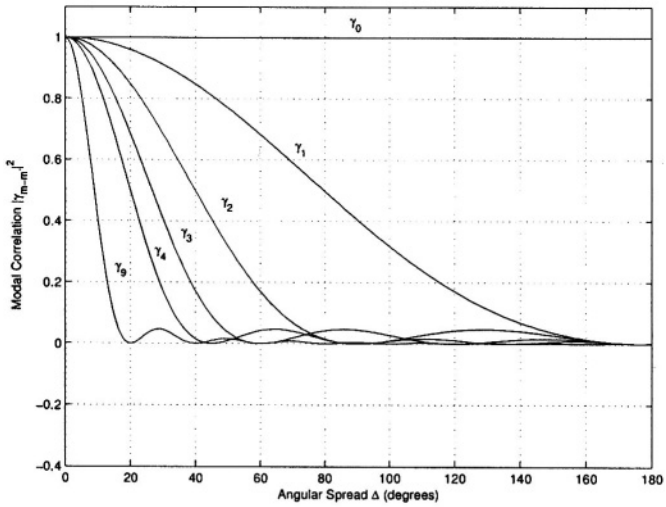



*Figure 12-4.* Modal correlation versus angular spread $\Delta$ of a uniform limited power density surrounding the aperture.

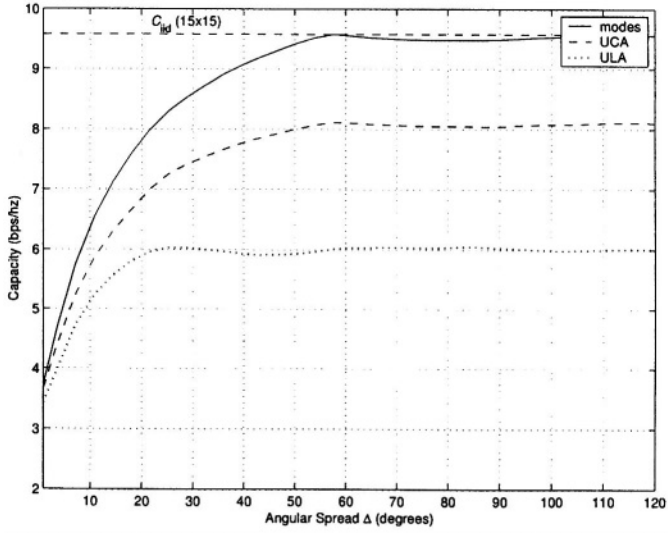



*Figure 12-5.* Capacity versus angular spread at the transmitter for mode-to-mode communication (modes), uniform linear array (ULA), and uniform circular array (UCA), within spatial regions of radius $0.8\lambda$ and isotropic receiver scattering. The mode-to-mode capacity gives the maximum achievable capacity between the two apertures.

fundamental limit to capacity for realistic scattering environments. By characterizing the behavior of possible communication modes for a given aperture, the upper bound on capacity is independent of antenna configurations and array signal processing, and provides a benchmark for future array and space-time coding designs.

In this chapter we have restricted the analysis to 2D circular apertures, however, extension to arbitrary shaped regions can be achieved by using a different choice of orthonormal basis functions (e.g. see [17, 19]), however, with the exception of spherical apertures [17], finding analytical solutions for more general volumes poses a much harder problem.

## REFERENCES

1. E. Telatar, "Capacity of multi-antenna gaussian channels," Tech. Rep., ATT Bell Labs, 1995.
2. G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications,* vol. 6, no. 3, 1998.
3. J.P. Kermoal, P.E. Mogensen, S.H. Jensen, J.B. Anderson, F. Frederiksen, T.B. Sorensen, and K.I. Pedersen, "Experimental investigation of multipath richness for multi-element transmit and receive antenna arrays," in *IEEE Vehicular Technology Conference,* Tokyo, Japan, 2000, pp. 2004–2008.
4. K.I. Pedersen, J.B. Anderson, J.P. Kermoal, and P.E. Mogensen, "A stochastic multiple-input multiple-output radio channel model for evaluation of space-time coding algorithms," in *IEEE Vehicular Technology Conference,* Boston, MA, 2000, pp. 893–897.
5. W. Yu, M. Bengtsson, B. Ottersten, D.P. McNamara, P. Karlsson, and M.A. Beach, "A wideband statisitical model for NLOS indoor wireless MIMO channels," in *IEEE Vehicular Technology Conference (Spring),* Birmingham, Al, 2002.
6. J.W. Wallace and M.A. Jensen, "Spatial characteristics of the mimo wireless channel: experimental data acquisition and analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* Salt Lake City, Utah, 2001, pp. 2497–2500.
7. G. Athanasiadou, A. Nix, and J. McGeehan, "A microcellular ray-tracing propagation model and evaluation of its narrow-band and wide-band predictions," *IEEE Journal on Selected Areas in Communications,* vol. 18, pp. 322–335, 2000.
8. G. German, Q. Spencer, A. Swindlehurst, and R. Valenzuela, "Wireless indoor channel modeling: statistical agreement of ray tracing simulations and channel sounding measurements," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Salt Lake City, UT, 2001, vol. 4, pp. 778–781.
9. Da-Shan Shiu, G.J. Foschini, M.J. Gans, and J.M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications,* vol. 48, no. 3, pp. 502–513, 2000.

10. D. Gesbert, H. Bolcskei, D. Gore, and A. Paulraj, "Outdoor mimo wireless channels: Models and performance prediction," *IEEE Transactions on Communications,* vol. 50, no. 12, pp. 1926–1934, 2002.
11. J.W. Wallace and M.A. Jensen, "Modeling the indoor MIMO wireless channel," *IEEE Transactions on Antennas and Propagation,* vol. 50, no. 2, pp. 591–599, 2002.
12. A.M. Sayeed, "Deconstructing multi-antenna fading channels," *IEEE Transactions on Signal Processing,* vol. 50, no. 10, pp. 2563–2579, 2002.
13. D. Colton and R. Kress, *Inverse acoustic and elctromagnetic scattering theory,* Springer-Verlag, Berlin, 1992.
14. H.M. Jones, R.A. Kennedy, and T.D. Abhayapala, "On the dimensonality of multipath fields: spatial extent and richness," in *ICASSP 2002,* Florida, 2002.
15. D. Chizhik, G.J. Foschini, M.J. Gans, and J.M. Kahn, "Capacities of multi-antenna transmit and receive antennas: correlation and keyholes," *Electronics Letters,* vol. 36, no. 13, pp. 1099–1100, 2000.
16. F. Demmerle and W. Wiesbeck, "A biconical multibeam antenna for space-division multiple access," *IEEE Transactions on Antennas and Propagation,* vol. 46, no. 6, pp. 782–787, 1998.
17. T.D. Abhayapala, T.S. Pollock, and R.A. Kennedy, "Novel 3D spatial wireless channel model," in *IEEE Vehicular Technology Conference (Fall),* Orlando, Florida, USA, 2003.
18. T.S. Pollock, T.D. Abhayapala, and R.A. Kennedy, "Introducing space into MIMO capacity calculations," *Journal on Telecommunications Systems,* vol. 24, no. 2-4, pp. 415–436, 2003.
19. D. A. B. Miller, "Spatial channels for communicating with waves between volumes," *Optics Letters,* vol. 23, pp. 1645–1647, 1998.

## Acknowledgments

National ICT Australia is funded through the Australian Government's *Backing Australia's Ability initiative,* in part through the Australian Research Council.

Chapter 13

# INCREASING PERFORMANCE OF SYMMETRIC LAYERED SPACE-TIME SYSTEMS

Phillip Conder, Tadeusz Wysocki
*Telecommunications and Information Technology Research Institute, University of Wollongong, Australia*

**Abstract** A number of decoding schemes have been proposed for Layered Space-Time systems, such as the Ordered Successive Interference Cancellation and the Sorted QR Decomposition. We describe here a new addition to that group, increasing the performance of Layered Space-Time decoding by using the Sorted QR Decomposition technique to construct a list of constellations to be passed to a Maximum Likelihood decoder.

This paper shows that significant performance improvement can be obtained for symmetric systems, where there is an equal number of transmit and receive antennas. It shows that the proposed scheme, has a roughly linear increase in complexity compared to SQRD. To overcome this increase in computational complexity, an adaptive system is described that has similar performance with reduced complexity.

**Key words:** Layered space-time systems, sorted QD decomposition, V-BLAST, adaptive decoding.

## 1. INTRODUCTION

In recent years, the demand for wireless communications has been increasing at a rapid pace, with more emphasis to provide higher rates, and improved quality in terms of reliability. It was shown, [1] [2], that employing multiple antennas both at the transmitter and receiver promises huge capacity increases in a multipath fading environment. Indeed, the capacity increases about linearly with the number of transmit and receive antennas.

The complexity of Maximum Likelihood (ML) decoding of such systems increases exponentially with transmit antenna numbers and constellation size. A number of sub-optimal decoding schemes with lower computational complexity have been proposed such as Zero Forcing (ZF), the Ordered Successive Interference Cancellation (OSIC) [3] [4] and the QR Decomposition [5].

These decoding systems perform best when the number of receive antennas is greater than the number of transmit antennas, while performance is less-optimal when antenna numbers are equal. This is due to the loss of diversity in the decoding process.

The paper is ordered as follows. Section 2 gives a brief system description of Layered Space-Time codes and Maximum Likelihood decoding. In Section 3 we review Layered Space Time system decoders such as ZF and OSIC, while Section 4 describes the Sorted QR Decomposition (SQRD). Section 5 introduces a method that uses the SQRD to produce a list of symbol combinations which is used by an ML decoder. Section 6 introduces an adaptive Reduced SQRD (RSQRD) and Section 7 compares the complexity of previous schemes to the fixed and adaptive RSQRD.

## 2. LAYERED SPACE-TIME SYSTEMS DESCRIPTION

The Layered Space-Time Processing approach was first introduced by Lucent's Bell Labs, with their BLAST family of Space Time Code structures [6]. An uncoded Vertical Bell Laboratories Layered Space-Time (VBLAST) scheme, where the input bit stream is de-multiplexed into $n_t$ substreams, is considered in this paper. Let $n_t$ be the number of transmit and $n_r$ be the number of receive antennas, where $n_r \geq n_t$, and $s = (s_1, s_2, ...s_{n_t})^T$ denote the vector of transmitted symbols in one symbol period. The received vector $Y = (Y_1, Y_2...Y_{n_t})^T$ is

$$Y = Hs + n \tag{13.1}$$

where $n = (n_1, n_2, ...n_{n_t})^T$ is the noise vector of additive white Gaussian noise of variance $\sigma^2$ equal to $\frac{1}{2}$ per dimension. The $n_r \times n_t$ channel matrix

$$H = \begin{pmatrix} h_{1,1} & . & . & h_{1,n_t} \\ . & . & & . \\ . & & . & . \\ h_{n_r,1} & . & . & h_{n_r,n_t} \end{pmatrix} \tag{13.2}$$

contains independent identical distribution (i.i.d.) complex fading gains $h_{i,j}$ from the $j^{th}$ transmit antenna to the $i^{th}$ receive antenna. We

assume quasi-static flat fading where $H$ is constant over $L$ symbol periods.

## 3. LAYERED SPACE-TIME SYSTEMS DESCRIPTION

There have so far been a number of decoding methods proposed Layered such as the optimal Maximum likelihood decoder, simple nulling methods such as Zero forcing and cancellation methods such as SQRD.

### 3.1 MAXIMUM LIKELIHOOD DECODING

Maximum Likelihood decoding is achieved by minimising

$$\| Hs - Y \|^2 \tag{13.3}$$

for all elements of $s$, which are symbols of constellation of size $C$. This would produce a search of length $C^{m_t}$, which for a system using 4 transmit antennas and 16 QAM gives 65536 possibilities, far beyond being practically decoded in real time. This leads to a search for methods of decoding with a reduced computational complexity.

### 3.2 ZERO FORCING AND MINIMUM MEAN SQUARED ERROR DECODING

The sub-optimal but less complex V-BLAST detector was proposed [3] [4] as a reduced complexity method to decode Layered Space-Time systems. A nulling (ZF) process was first introduced, which uses a pseudo inverse of $H$ to produce estimates, $\tilde{s}$, of the individual symbols, which are then passed to individual decoders. Conceptually, each transmitted symbol is considered in turn to be the desired symbol and the remaining symbols are treated as interferers.

$$\tilde{s} = H^{\dagger} Y \tag{13.4}$$

where †is the Moore-Penrose pseudo inverse [3]. Another method of nulling with better performance is to modify the receiver antenna preprocessing to carry out Minimum Mean Squared Error (MMSE) rather than ZF.

$$\tilde{s} = \left( \left( \left( H^H H + \left( \sigma^2 I \right) \right)^{-1} \right) H^H \right) Y \tag{13.5}$$

where $\sigma^2$ is the noise variance. MMSE and ZF nulling have the disadvantage that some of the diversity potential of the receiver antenna array is lost in the decoding process. To take advantage of the diversity

potential, nonlinear techniques, such as Ordered Successive Interference Cancellation (OSIC) have been introduced [6] and shown to have superior performance.

## 3.3 OSIC DETECTOR

The OSIC decoding (as called V-BLAST decoding) algorithm uses the detected symbol $\widetilde{s}_i$, obtained by the zero forcing, to produce a modified received vector with $\widetilde{s}_i$ canceled out. This modified received vector has fewer interferers and better performance due to a higher level of diversity. This process is continued until all $n_t$ symbols have been detected. Obviously an incorrect symbol selection in the early stages will create errors in the following stages. Therefore the order in which the components are detected becomes important to the overall system performance.

Fig. 13-1 shows the Symbol Error Rate (SER) of Zero Forcing, OSIC, and Maximum Likelihood decoding for a system using 16-QAM with 4 transmit and 6 receive antennas. At a SER of $10^{-4}$ the difference between ZF and OSIC is approximately $5dB$, while the difference between OSIC and ML decoding is $2dB$. This demonstrates that there is only a small difference between OSIC and ML when the number of receive antennas is 50% more than the number of transmit antennas.

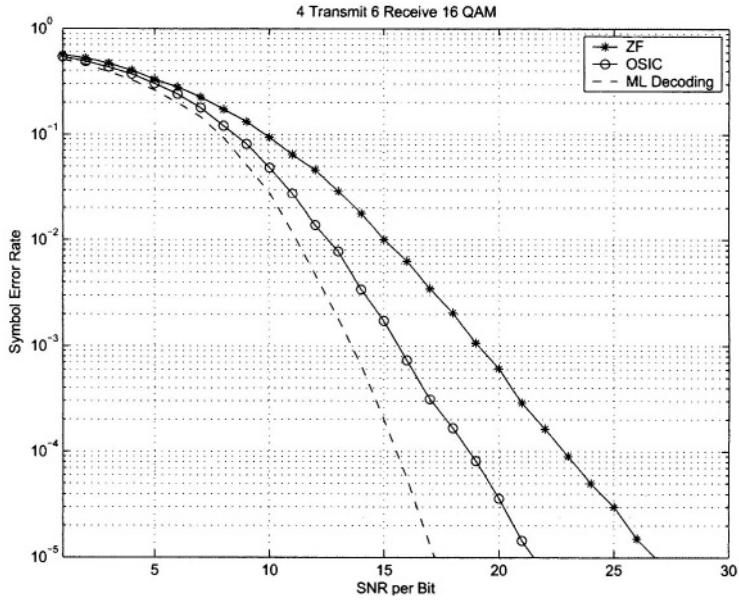



*Figure 13-1.* Performance of ZF, OSIC and ML decoding with 4 Transmit, 6 Receive antennas and 16 QAM.

## 3.4 SORTED QR DECOMPOSITION

The QR decomposition of the channel matrix $H$ was introduced in [5] as another method to decode Layered Space-Time systems. The $n_r \times n_t$ channel matrix $H$ is factorised into the unitary $n_r \times n_t$ matrix $Q$ and the upper triangular $n_t \times n_t$ matrix $R$.

$$H = Q.R \tag{13.6}$$

By denoting the column $i$ of $H$ by $h_i$ and column $i$ of $Q$ by $q_i$, the decomposition in equation (6) is described columnwise by

$$(h_1 ... h_{n_t}) = (q_1 ... q_{n_t}) \begin{pmatrix} r_{1,1} & \cdot & \cdot & r_{1,n_t} \\ & \cdot & & \cdot \\ & & \cdot & \cdot \\ 0 & & & r_{n_t,n_t} \end{pmatrix} \tag{13.7}$$

By multiplying the received vector $Y$ with the complex conjugate of matrix $Q$, an $n_t \times 1$ modified received vector

$$X = Q^H Y = Rs + \eta \tag{13.8}$$

is created from the $n_t \times 1$ received signal vector $Y$. The upper triangular matrix $R$ has the lowest layer (transmit signal $s_{n_t}$) described by

$$x_{n_t} = r_{n_t,n_t} s_{n_t} + \eta_{n_t} \tag{13.9}$$

The decision statistic $x_{n_t}$ is independent of the remaining transmit signals and can be used to estimate $\tilde{s}_{n_t}$

$$\tilde{s}_{n_t} = ML \left[ \frac{x_{n_t}}{r_{n_t,n_t}} \right] \tag{13.10}$$

where $ML$ is the Maximum Likelihood detector. This symbol is then used, by substitution, to detect $\tilde{s}_{n_t-1}$ from the equation

$$\tilde{s}_{n_t-1} = ML \left[ \frac{x_{n_t-1} - r_{n_t-1,n_t} . \tilde{s}_{n_t}}{r_{n_t-1,n_t-1}} \right] \tag{13.11}$$

This method of detection and substituting into upper layers is continued until all symbols are detected.

A number of techniques are based on using the QR decomposition [7]. One such, the Sorted QR decomposition is proposed in [8]. SQRD is based on the modified Gram-Schmidt algorithm [9]. The columns of $H$, $Q$, and $R$ are reordered in each orthogonalisation step to minimise

the magnitude of the diagonal elements of *R*. This method ensures that symbols with larger channel co-efficients $h_i$ are detected first while symbols with smaller $h_i$ are detected later to reduce error propagation. The SQRD algorithm has been shown in [8] to have similar performance to OSIC with lower computational complexity.

## 4. PERFORMANCE IN SYMMETRIC SYSTEMS

OSIC and SQRD have both been designed and shown to have performance similar that of ML decoding when the number of receive antennas is greater than than the number of transmit antennas. Fig. 13-2 shows the performance of SQRD for a symmetric system where the number of transmit and receive antennas are equal to 4.

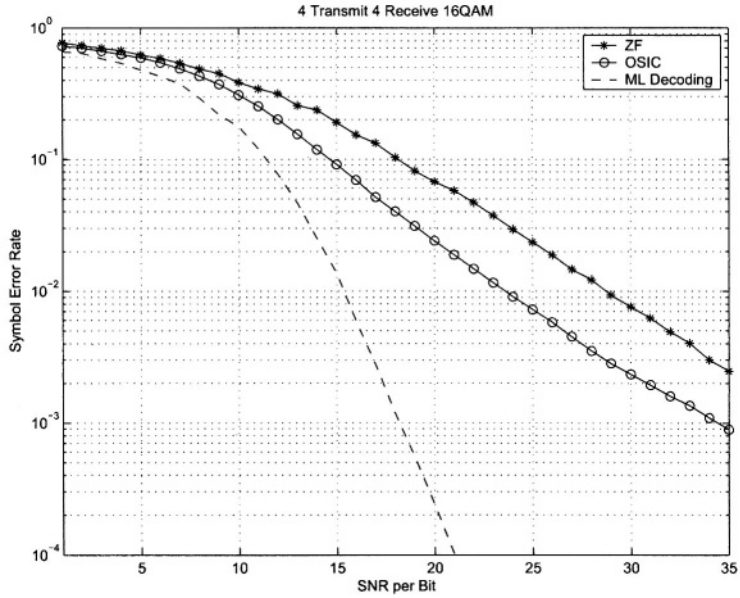



*Figure. 13-2.* Performance of ZF, OSIC and ML decoding with 4 Transmit, 4 Receive antennas and 16 QAM.

At a SER of $10^{-3}$ the difference between SQRD and ML decoding is over 15*dB*, compared to 2*dB* for the system of 4 transmit and 6 receive antennas shown in Fig. 1. This loss in performance for symmetric systems is due to the first decoding stage having a $G_{div} = 1$ and hence a higher SER which creates error propagation in other layers [9]. In each detection step $i = n_t \ldots 1$ a diversity of $G_{div} = n_r - i + 1$ is achieved.

This is demonstrated in Fig. 13-3, which shown the BER of each of the detection layers when there is no error propagating ie the previous detected symbols are correct.

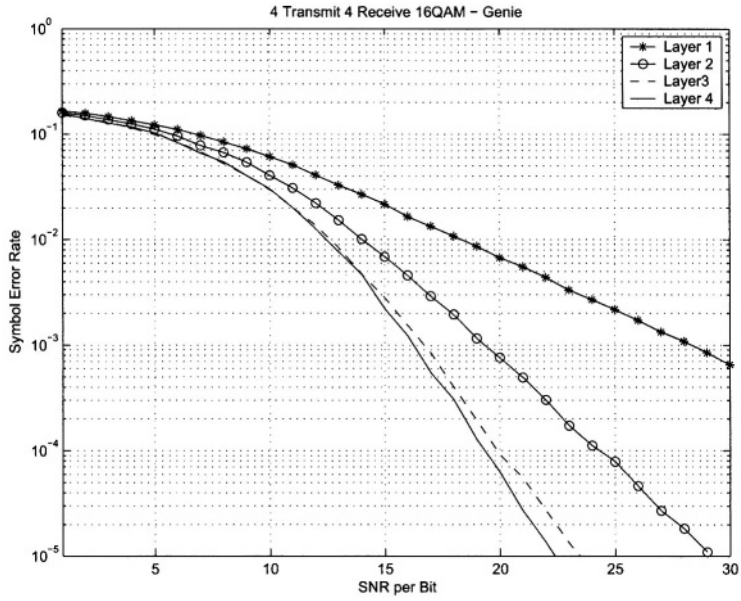



*Figure 13-3.* "Genie" decoding of system with 4 Transmit, 4 Receive antennas and 16 QAM.

## 5. REDUCED CONSTELLATION SEARCH FOR SYMMETRIC SYSTEMS

Both the OSIC and SQRD decoding approaches have the disadvantage that some of the diversity potential of the receiver antenna array is lost in the decoding process, particularly when $n_t = n_r$ and in early detection stages of decoding. To overcome this, a scheme called Reduced search using Sorted QR Decomposition (RSQRD) is introduced. The proposed scheme uses the SQRD method to produce a reduced constellation list, which is then used by the ML detector to determine which combination of symbols is the most likely.

The assumption made for RSQRD is that if $\widetilde{s}_i$ (the symbol estimate independent of other interfering symbols) is chosen incorrectly by the ML detector, then the correct solution will be close to $\widetilde{s}_i$. Therefore, a local search around $\widetilde{s}_i$ is performed to find the order of closest symbols.

In the lowest layer of SQRD, where the symbol being detected is independent of all other symbols the closest $k$ symbols to $\tilde{s}_i$ are found. By choosing $k$ symbols rather than just one symbol per layer, as with standard SQRD, the effect of wrong symbol selection producing error propagation is reduced. These $k$ symbols are then in turn, substituted into the next highest layer to find the nearest symbol for each $k$ symbol. Once the list is generated, the scheme performs the ML detection over all combinations in the list. In symmetric systems the majority of error propagation is caused by the first detected symbol because it has a $G_{div} = 1$. Therefore the largest gains obtained by RSQRD is when the constellation size is increased in the first detection stage.

## 6. ADAPTIVE RSQRD

In Section 5 it was described that the greatest improvements of the RSQRD were made by finding the $k$ most likely symbols in the first detected stage, where there was a reduced level of diversity, and then finding the combination for each value of $k$ . The size of $k$ was fixed to give a certain performance. This meant even when the correct combination of symbols was found the algorithm continued until the $k^{th}$ time.

Instead of finding $k$ combinations of symbols and then performing a Maximum Likelihood calculation, it would be far more efficient to perform the ML calculation for each combination of symbols after they have been detected and continue the search only if the ML solution is not found.

The Adaptive RSQRD algorithm works as follows: If the result of combination of symbols in (3) is less than a '*Threshold*' value the search is stopped, otherwise the search is continued to find the next combination of symbols. If after $k$ times no combination is selected, the combination with the smallest result from (3) is chosen. The important question being what is the optimal value of the '*Threshold*' variable?

The noise variance $(\sigma^2)$ and standard deviation $(\sigma)$ were trialled as the '*Threshold*' value and were both found to reduce the number of ML tests. Using the noise variance substantially reduced the number of ML tests for $SNR \leq 10db$, but increased the number of ML tests for higher $SNR's$. The standard deviation of noise was found to be optimal for $SNR > 10dB$. Since the greatest performance increase of RSQRD is when the $SNR > 15dB$, as shown in Fig. 13-4, using the noise variance $\sigma$ for the '*Threshold*' value is proposed for the adaptive scheme.

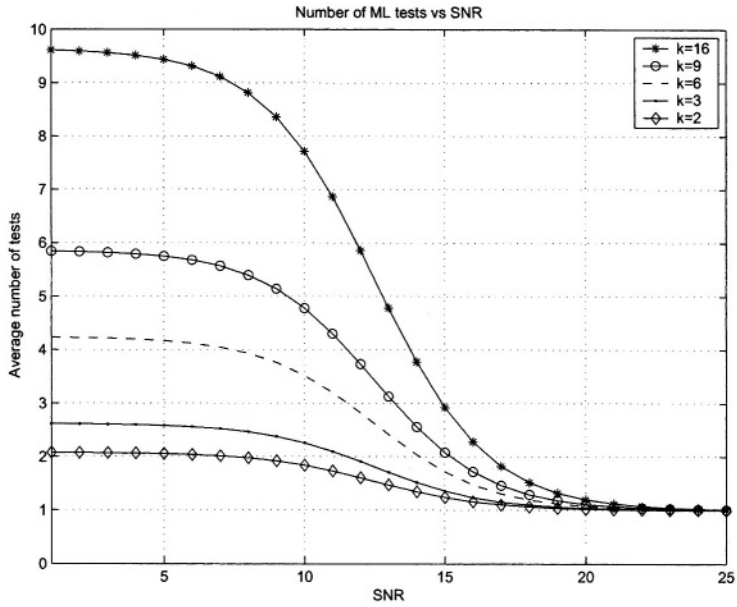



*Figure 13-4.* Average number of tests for various values of maximum $k$ vs $SNR$ for 16-QAM.

## 7. COMPARISON OF THE COMPLEXITY OF DIFFERENT SCHEMES

Fig. 13-4 shows average number of Maximum Likelihood tests versus $SNR$ for $k$ =16, 9, 6, 3 and 2. The system uses 4 Transmit 4 Receive antennas and 16-QAM. It can be seen from Fig. 13-4 that for $SNR > 20dB$ the number of ML tests approaches one for all values of $k$. We used of Monte Carlo simulation technique to find the number of ML tests for 16-QAM for various values of $k$ factor and $SNR$ ranging from 0 to $25dB$. After applying the non-linear least mean squares curve fitting method [10], the formula approximating the number of ML tests for 16-QAM is:

$$N_{MLtests} = \frac{k \times 0.54}{1 + e^{(SNR - 12.5)/2}} + 1 \qquad (13.12)$$

It was published in [11] that the decoding complexity of OSIC is approximately $\frac{27}{4}n_t^4$, while QR Decomposition based schemes have a decoding complexity of $\frac{29}{3}n_t^3$ [11]. These values are for systems with $n_t = n_r$ and do not take into account the assumption that the system has quasi-static flat fading, and $H$ is constant over $L$ symbol periods.

For this reason the computational complexity formulae of [7] were used as the basis for comparing RSQRD and OSIC. A single value was obtained for the computation complexity by counting real valued additions, multiplications and divisions as one floating point operation. The value of $k$ is an integer for the fixed scheme and is equal to $N_{MLtests}$ from (11) for the Adaptive scheme. The ratio of complexity of RSQRD and V-BLAST is given by:

$$\frac{C_{RSQRD}}{C_{OSIC}} = \frac{12n_t^3 + 18n_t^2 + Lk\left[(12n_t + 2)n_t\right]}{25n_t^4 + L\left[(18n_t)n_t\right]} \tag{13.13}$$

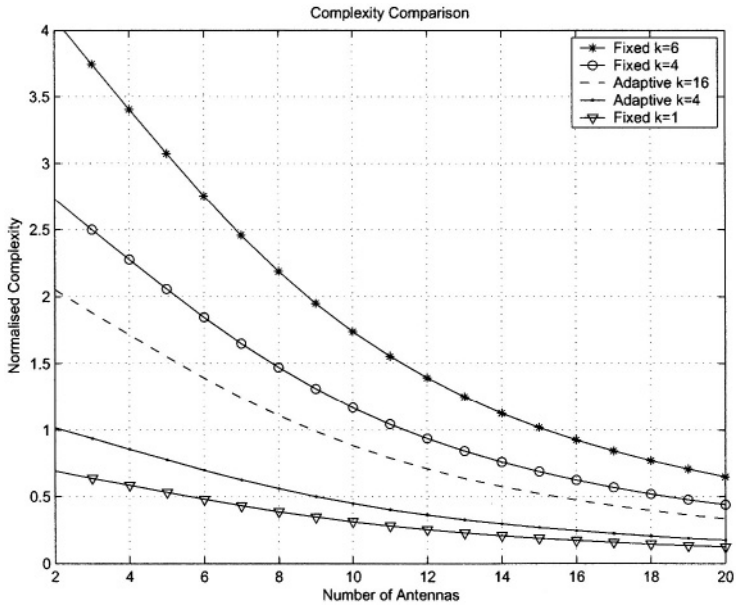



*Figure. 13-5.* Normalised Complexity comparison for fixed and Adaptive schemes, when L = 100 and SNR = 15dB.

Fig. 13-5 shows the relationship between the RSQRD and OSIC schemes for a number of different values of $k$, using (12) at $SNR = 15dB$. It can be seen that the complexity of the Adaptive RSQRD with $k = 4$ has a complexity lower than one (i.e. less than V-BLAST) for all antenna numbers, while the complexity of the Adaptive RSQRD with $k = 16$ has a complexity greater than one for antenna number lower than 11.

## 8. PERFORMANCE

Monte Carlo simulations were used to compare the performance of the proposed scheme and the standard Layered Space-Time detection algorithms.

The simulation results presented in this paper are as follows: Fig. 13-6 shows the comparison of the same decoders for a system using $n_t = 4$ and $n_r = 4$ antennas and 16-QAM. Fig. 13-7 shows the comparison of the same decoders for a system using $n_t = 4$ and $n_r = 6$ antennas. While Fig. 13-8 illustrates the comparison of the reduced constellation SQRD for a system with $n_t = 8$ and $n_r = 8$ antennas using QPSK, with various values of $k$.

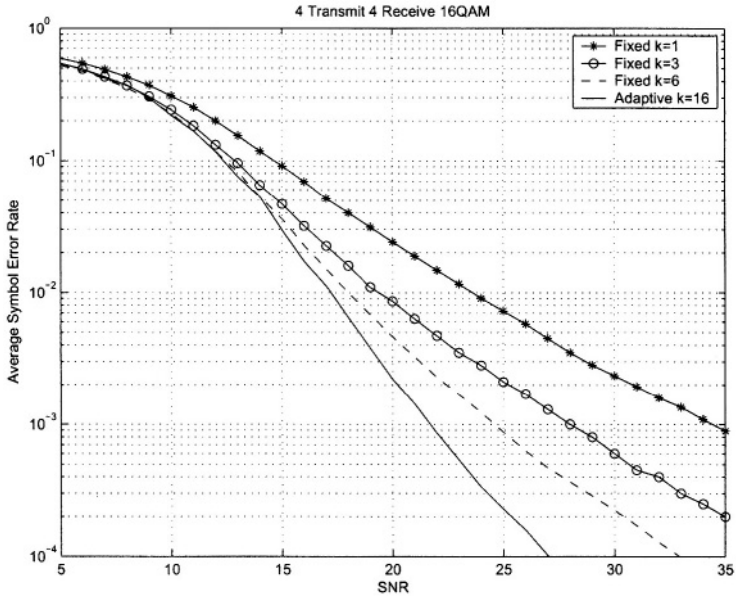



*Figure 13-6.* 4 Transmit, 4 Receive 16 QAM, RSQRD with constellation size 1, 3, 6 and Adaptive with k=16.

Fig. 13-6 shows the increase in performance between $k = 1$ and larger constellation size of 3 and 6 for a system using $n_t = 4$ and $n_r = 4$ antennas and 16-QAM. Approximately $5dB$ gain between SQRD and proposed scheme using $k = 3$ and $10dB$ for $k = 6$ at a SER of $10^{-3}$. The Adaptive system with $k = 16$ has the same performance as the fixed $k = 16$ system, with an increase of $14dB$ over the original SQRD system at a Symbol Error Rate of $10^{-3}$.

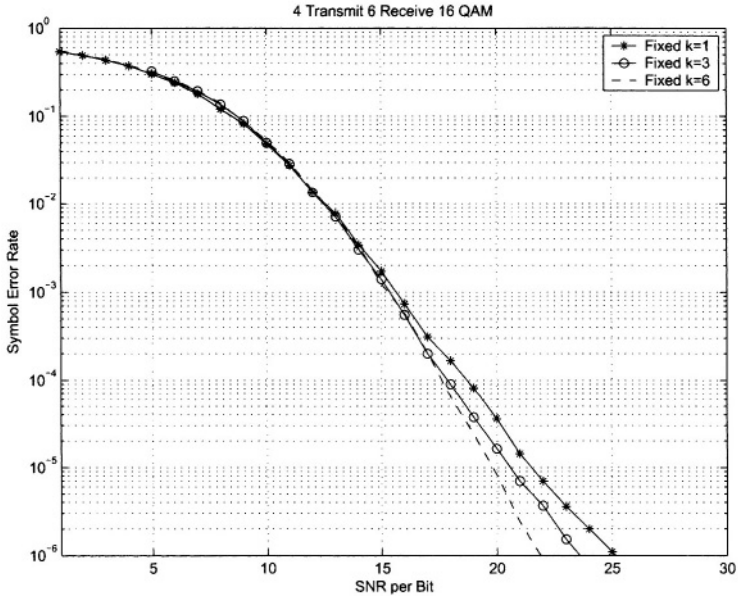



*Figure. 13-7.* 4 Transmit, 6 Receive 16QAM, RSQRD with constellation size 1, 3 and k=6.

Fig. 13-7 shows that there is only a small increase in performance between $k = 1$ and a larger constellation size of 3 and 6 for an asymmetric system. Approximately $1dB$ gain between SQRD and proposed scheme using $k = 3$ and $2dB$ for $k = 6$ at a SER of $10^{-5}$. Increasing the size of the lowest layer when $n_r > n_t$ brings only a small improvement because even the lowest layer, for the $n_t = 4$, $n_r = 6$ system, has a diversity level of 3.

From Fig. 13-8, it can be seen that there is a significant increase in performance between $k = 1$ (standard SQRD) and larger constellation sizes of 3 and 4 for symmetric systems. Approximately $8dB$ gain between SQRD and the proposed scheme using $k = 3$ and $10dB$ for $k = 4$ at a SER of $10^{-3}$, while there is only a small increase of $2dB$ for $k = 2$. Also of note is the result showing indistinguishable performance of the fixed and Adaptive systems with $k = 4$. This result was found to be the same for all $k$ using 16-QAM.

## 9. CONCLUSION

We have described a new improvement to increase the performance of Layered Space-Time systems, such as V-BLAST, by using the Sorted QR

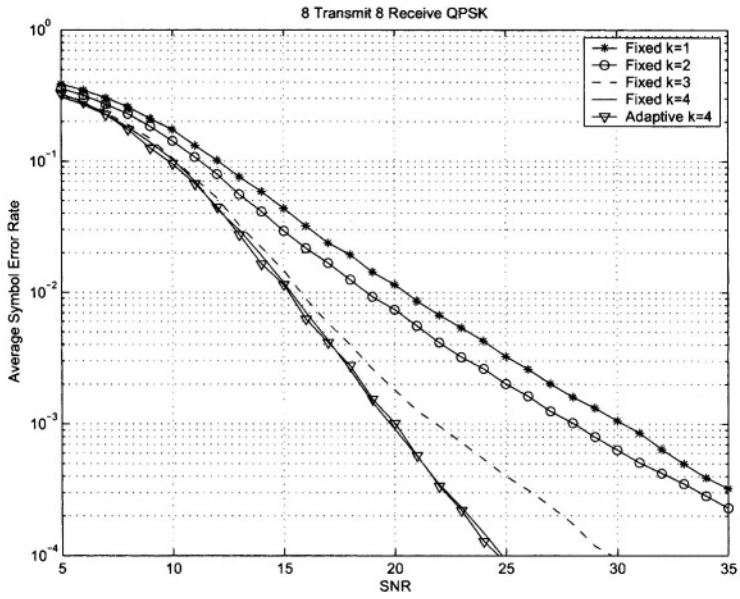



*Figure 13-8.* 8 Transmit, 8 Receive QPSK, RSQRD with constellation size 1, 2, 3, 4 and Adaptive with k=4.

Decomposition technique to construct a list of constellations to be passed to a Maximum Likelihood decoder. It was shown that a significant performance increase can be obtained by increasing the constellation size for the lowest layer. In addition, it was shown that while at high SNR's there is improvement when $n_r > n_t$, greatest improvement in performance is for symmetric systems, i.e. when $n_r = n_t$. This due to a unity diversity level for the first detected symbol which is then used to detect other symbols.

To overcome the increase in computational complexity an adaptive system was shown to have similar performance with a reduced complexity. By testing the combination of symbols after each detection step and varying the size of $k$ with the $SNR$, a computation complexity comparable to that of OSIC can be achieved with substantial performance increase.

The adaptive reduced constellation search scheme is not dependent on the Sorted QR Decompostion decoder and could be implemented on a OSIC using ZF or MMSE decoder described by earlier[3].

# REFERENCES


1. E. Telatar, "Capacity of Multi-antenna Gaussian Channels", AT & T Bell Labs, Murray Hill, NJ, Tech. Rep., 1995
2. G.J. Foschini and M.J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", Wireless Personal Communications, vol. 6, pp.311-335, 1998.
3. G.J. Foschini, "Layered Space-Time Architecture for Wireless Communications in a Fading Environment when using Multiple Antennas", Bell Labs Technical Journal, Autumn 1996.
4. P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel", Bell Labs., Lucent Technol., Crawford Hill Lab., Holmdel, NJ, Tech Rep., 1999.
5. D.Shiu and J.M. Kahn, "Layered Space-Time Codes for Wireless Communications using Multiple Transmit Antennas", in IEEE Proceedings of International Conference on Communication (ICC'99), British Columbia, June 1999.
6. G.D. Golden, G.J. Foschini, R.A Valenzuela, and P.W. Wolniansky, "Detection algorithm and initial laboratory results using the V-BLAST space-time communication architecture", Electronic Letters, vol. 35, no. 1, pp 14-15, January 1999.
7. B. Hassibi, "An efficient square-root algorithm of BLAST", Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference , 2000
8. D. Wubben, J. Rinas, R. Bohnke, V. Kuhn, K.D. Kammeyer, "Efficient Algorithm for Detecting Layered Space-Time Codes", 4th International ITG Conference on source and channel coding, Berlin, January 2002.
9. G. Strang, Linear Algebra and its Applications, Harcout Brace Jovanovich College Publishers, Orlando, Florida, third edition 1999.
10. R.A. Johnson, "Miller & Freund's Probability & Statistics for Engineers", Prontice-Hall, New Jersey, 1994
11. D.W. Waters and J.R. Barry, "Noise-Predictive Decision-Feedback Detection for Multiple-Input Multiple-Output Channels", IEEE International Symposium on Advances in Wireless Communications (ISWC02), British Columbia, September, 2002.

Chapter 14

# NEW COMPLEX ORTHOGONAL SPACE-TIME BLOCK CODES OF ORDER EIGHT

Jennifer Seberry[1], Le Chung Tran[2], Yejing Wang[1], Beata J. Wysocki[2], Tadeusz A. Wysocki[2], Tianbing Xia[1], Ying Zhao[1]

[1] *School of Computer Science, University of Wollongong, Australia,* [2] *School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Australia.*

**Abstract** In this chapter, three new AODs of order eight are given and they are used to construct new complex orthogonal space-time block codes. These new complex-valued codes are amenable to practical implementations as they allow for a more uniform spread of power among the transmit antennas while providing the same performance as other published codes of order eight.

**Key words:** Space-time codes, amicable orthogonal designs, complex orthogonal designs.

## 1. INTRODUCTION

Complex orthogonal space-time block codes (CO STBCs) based on Amicable Orthogonal Designs (AOD) [1] are known for the relatively simple receiver structure and minimum processing delay in case of complex signal constellations. The simplest CO STBC is an Alamouti code [3] for two transmit antennas, which is based on an amicable orthogonal pair of order two. Alamouti code has achieved the transmission rate one for two transmit antennas, while the CO STBCs for more than two transmit antennas cannot provide the rate one [2], but they can still achieve the full diversity for the given number of transmit antennas. It is noted that we consider here the *square* CO STBCs only. To date, only the CO STBCs based on full designs of order two and four have been proposed [3] and [4]. The known CO STBCs for higher number of transmit antennas, e.g. eight, have zeros in the code matrices that

results in the high peak-to-mean power ratio for the transmit antennas and impede their practical implementation.

The chapter is organized as follows. In Section 2, we introduce the new code designs. Section 3 provides expressions for Maximum Likelihood (ML) decoding of the transmitted symbols. Section 4 discusses the choice of signal constellation to provide optimal peak-to-mean power ratio, while Section 5 concludes the paper.

## 2. NEW CODS OF ORDER EIGHT

The construction of CO STBCs follows directly from CODs defined as follows.

DEFINITION 1 *A COD $Z = X + iY$ of order $n$ is an $n \times n$ matrix on the complex indeterminates $s_1, \cdots, s_t$, with entries chosen from $0, \pm s_1, \cdots, \pm s_t$, their conjugates $\pm s_1^*, \cdots, \pm s_t^*$, or their product with $i = \sqrt{-1}$, such that:*

$$Z^H Z = (\sum_{k=1}^{t} |s_k|^2) I_n \tag{14.1}$$

*where $Z^H$ denotes the Hermitian transpose of $Z$ and $I_n$ is the identity matrix of order $n$.*

For the matrix $Z$ to satisfy (14.1), the matrices $X$ and $Y$ must be a pair of AODs which implies that both $X$ and $Y$ are orthogonal designs themselves and $XY^T = YX^T$, where $(.)^T$ denotes matrix transposition. It has been shown in [1] that for order $n = 8$, the total number of different variables in the amicable pair $X$ and $Y$ cannot exceed eight. In Table 14-1, we record the number of variables in $X$ versus the number of variables in $Y$ with order eight. It has been shown in [5], that the

*Table 14-1.* Number of variables in an amicable pair with $n = 8$

| *Number of variables in Y* | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|---|---|---|---|
| *number of variables in X* | 0 | 0 | 0 | 1 | 4 | 4 | 4 | 5 | 8 |

construction of CODs can be facilitated by representing $Z$ as

$$Z = \sum_{j=1}^{t} A_j s_j^R + i \sum_{j=1}^{t} B_j s_j^I \tag{14.2}$$

where $s_j^R$ and $s_j^I$ denote the real and imaginary parts of the complex variables $s_j = s_j^R + i s_j^I$ and $A_j$ and $B_j$ are the coefficient matrices for

$$Z_1 = \begin{bmatrix} s_1 & s_2 & s_3 & 0 & s_4 & 0 & 0 & 0 \\ -s_2^* & s_1^* & 0 & -s_3 & 0 & -s_4 & 0 & 0 \\ -s_3^* & 0 & s_1^* & s_2 & 0 & 0 & -s_4 & 0 \\ 0 & s_3^* & -s_2^* & s_1 & 0 & 0 & 0 & s_4 \\ -s_4^* & 0 & 0 & 0 & s_1^* & s_2 & s_3 & 0 \\ 0 & s_4^* & 0 & 0 & -s_2^* & s_1 & 0 & -s_3 \\ 0 & 0 & s_4^* & 0 & -s_3^* & 0 & s_1 & s_2 \\ 0 & 0 & 0 & -s_4^* & 0 & s_3^* & -s_2^* & s_1^* \end{bmatrix}$$

*Figure 14-1.* A conventional COD of order eight.

$s_j^R$ and $s_j^I$, respectively. To satisfy (14.2), the matrices $\{A_j\}$ and $\{B_j\}$ of order $n$ must satisfy the following conditions:

$$\begin{aligned} A_j A_j^T &= I, \qquad B_j B_j^T = I, \ \forall j = 1, ..., t \\ A_k A_j^T &= -A_j A_k^T, \quad B_k B_j^T = -B_j B_k^T, \ k \neq j \\ A_k B_j^T &= B_j A_k^T, \ \forall k, j = 1, ..., t \end{aligned} \tag{14.3}$$

The conditions in (14.3) arc necessary and sufficient for the existence of AODs of order $n$. Thus, the problem of finding CODs is connected to the theory of AODs.

From the perspective of constructing CO STBCs, the most promising case is that in which both $X$ and $Y$ have four variables. This case has been considered as the known CO STBCs of order eight, corresponding to $COD(8; 1, 1, 1, 1)$ with all four variables appearing just once in each column of $Z$ as given in Fig. 14-1 [6]. In [7], L. C. Tran et. al. have introduced two new codes of order eight where some variables appear more often than other (more than once in each column), i.e., codes based on $COD(8; 1, 1, 2, 2)$ and $COD(8; 1, 1, 1, 4)$. These codes, namely $Z_2$ and $Z_3$, are given in Fig. 14-2 and 14-3, respectively. It is easy to check that these codes satisfy the condition 14.1. In [5], it has been envisaged that a $COD(8; 2, 2, 2, 2)$ exists. However, no construction of it has been derived in literature to date. Here, we propose the code of order eight based on the $COD(8; 2, 2, 2, 2)$ as given in Fig. 14-4.

$$Z_2 = \begin{bmatrix}
s_1 & s_2 & \frac{s_3}{\sqrt{2}} & \frac{s_3}{\sqrt{2}} & 0 & 0 & \frac{s_4}{\sqrt{2}} & \frac{s_4}{\sqrt{2}} \\
-s_2^* & s_1^* & \frac{s_3}{\sqrt{2}} & -\frac{s_3}{\sqrt{2}} & 0 & 0 & \frac{s_4}{\sqrt{2}} & -\frac{s_4}{\sqrt{2}} \\
\frac{s_3^*}{\sqrt{2}} & \frac{s_3^*}{\sqrt{2}} & -s_1^R + is_2^I & -s_2^R + is_1^I & \frac{s_4}{\sqrt{2}} & \frac{s_4}{\sqrt{2}} & 0 & 0 \\
\frac{s_3^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} & s_2^R + is_1^I & -s_1^R - is_2^I & \frac{s_4}{\sqrt{2}} & -\frac{s_4}{\sqrt{2}} & 0 & 0 \\
0 & 0 & \frac{s_4^*}{\sqrt{2}} & \frac{s_4^*}{\sqrt{2}} & s_1 & s_2 & -\frac{s_3^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} \\
0 & 0 & \frac{s_4^*}{\sqrt{2}} & -\frac{s_4^*}{\sqrt{2}} & -s_2^* & s_1^* & -\frac{s_3^*}{\sqrt{2}} & \frac{s_3^*}{\sqrt{2}} \\
\frac{s_4^*}{\sqrt{2}} & \frac{s_4^*}{\sqrt{2}} & 0 & 0 & -\frac{s_3^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} & -s_1^R + is_2^I & -s_2^R + is_1^I \\
\frac{s_4^*}{\sqrt{2}} & -\frac{s_4^*}{\sqrt{2}} & 0 & 0 & -\frac{s_3^*}{\sqrt{2}} & \frac{s_3^*}{\sqrt{2}} & s_2^R + is_1^I & -s_1^R - is_2^I
\end{bmatrix}$$

*Figure 14-2.* Code $Z_2$.

$$Z_3 = \begin{bmatrix}
s_1 & 0 & s_3^R + is_2^I & s_2^R + is_3^I & \frac{s_4}{2} & \frac{s_4}{2} & \frac{s_4}{2} & \frac{s_4}{2} \\
0 & s_1 & -s_2^R + is_3^I & s_3^R - is_2^I & \frac{s_4}{2} & -\frac{s_4}{2} & \frac{s_4}{2} & -\frac{s_4}{2} \\
-s_3^R + is_2^I & s_2^R + is_3^I & s_1^* & 0 & \frac{s_4}{2} & \frac{s_4}{2} & -\frac{s_4}{2} & -\frac{s_4}{2} \\
-s_2^R + is_3^I & -s_3^R - is_2^I & 0 & s_1^* & \frac{s_4}{2} & \frac{s_4}{2} & -\frac{s_4}{2} & \frac{s_4}{2} \\
-\frac{s_4^*}{2} & -\frac{s_4^*}{2} & -\frac{s_4^*}{2} & -\frac{s_4^*}{2} & s_1^R - is_3^I & s_2^* & s_3^R - is_1^I & 0 \\
-\frac{s_4^*}{2} & \frac{s_4^*}{2} & -\frac{s_4^*}{2} & \frac{s_4^*}{2} & -s_2 & s_1^R + is_3^I & 0 & s_3^R - is_1^I \\
-\frac{s_4^*}{2} & -\frac{s_4^*}{2} & \frac{s_4^*}{2} & \frac{s_4^*}{2} & -s_3^R - is_1^I & 0 & s_1^R + is_3^I & -s_2^* \\
-\frac{s_4^*}{2} & \frac{s_4^*}{2} & \frac{s_4^*}{2} & -\frac{s_4^*}{2} & 0 & -s_3^R - is_1^I & s_2 & s_1^R - is_3^I
\end{bmatrix}$$

*Figure 14-3.* Code $Z_3$.

$$Z_4 = \begin{bmatrix}
\frac{s_1}{\sqrt{2}} & \frac{s_1}{\sqrt{2}} & \frac{s_2}{\sqrt{2}} & \frac{s_2}{\sqrt{2}} & \frac{s_3}{\sqrt{2}} & \frac{s_4}{\sqrt{2}} & \frac{s_3}{\sqrt{2}} & \frac{s_4}{\sqrt{2}} \\
\frac{s_1}{\sqrt{2}} & -\frac{s_1}{\sqrt{2}} & \frac{s_2}{\sqrt{2}} & -\frac{s_2}{\sqrt{2}} & \frac{s_4^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} & \frac{s_4^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} \\
\frac{s_2^*}{\sqrt{2}} & \frac{s_2^*}{\sqrt{2}} & -\frac{s_1^*}{\sqrt{2}} & -\frac{s_1^*}{\sqrt{2}} & \frac{s_3}{\sqrt{2}} & \frac{s_4}{\sqrt{2}} & -\frac{s_3}{\sqrt{2}} & -\frac{s_4}{\sqrt{2}} \\
\frac{s_2^*}{\sqrt{2}} & -\frac{s_2^*}{\sqrt{2}} & -\frac{s_1^*}{\sqrt{2}} & \frac{s_1^*}{\sqrt{2}} & \frac{s_4^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} & -\frac{s_4^*}{\sqrt{2}} & \frac{s_3^*}{\sqrt{2}} \\
\frac{-s_4^R+is_3^I}{\sqrt{2}} & \frac{-s_3^R+is_4^I}{\sqrt{2}} & \frac{-s_4^R+is_3^I}{\sqrt{2}} & \frac{-s_3^R+is_4^I}{\sqrt{2}} & \frac{s_2^R-is_1^I}{\sqrt{2}} & \frac{s_2^R-is_1^I}{\sqrt{2}} & \frac{s_1^R-is_2^I}{\sqrt{2}} & \frac{s_1^R-is_2^I}{\sqrt{2}} \\
\frac{-s_3^R-is_4^I}{\sqrt{2}} & \frac{s_4^R+is_3^I}{\sqrt{2}} & \frac{-s_3^R-is_4^I}{\sqrt{2}} & \frac{s_4^R+is_3^I}{\sqrt{2}} & \frac{s_2^R-is_1^I}{\sqrt{2}} & \frac{-s_2^R+is_1^I}{\sqrt{2}} & \frac{s_1^R-is_2^I}{\sqrt{2}} & \frac{-s_1^R+is_2^I}{\sqrt{2}} \\
\frac{-s_4^R+is_3^I}{\sqrt{2}} & \frac{-s_3^R+is_4^I}{\sqrt{2}} & \frac{s_4^R-is_3^I}{\sqrt{2}} & \frac{s_3^R-is_4^I}{\sqrt{2}} & \frac{s_1^R+is_2^I}{\sqrt{2}} & \frac{s_1^R+is_2^I}{\sqrt{2}} & \frac{-s_2^R-is_1^I}{\sqrt{2}} & \frac{-s_2^R-is_1^I}{\sqrt{2}} \\
\frac{-s_3^R-is_4^I}{\sqrt{2}} & \frac{s_4^R+is_3^I}{\sqrt{2}} & \frac{s_3^R+is_4^I}{\sqrt{2}} & \frac{-s_4^R-is_3^I}{\sqrt{2}} & \frac{s_1^R+is_2^I}{\sqrt{2}} & \frac{-s_1^R-is_2^I}{\sqrt{2}} & \frac{-s_2^R-is_1^I}{\sqrt{2}} & \frac{s_2^R+is_1^I}{\sqrt{2}}
\end{bmatrix}$$

*Figure 14-4.* Code $Z_4$.

*Table 14-2.* Decision metrics for decoding code $Z_1$.

| *Variable* | *Decoding metric* |
|---|---|
| $s_1$ | $Arg\ \min_{s_1 \in S} \left[ \left\| \left(h_5 r_5^* + h_2 r_2^* + h_1^* r_1 + h_4^* r_4 + h_3 r_3^* + h_8 r_8^* + h_6^* r_6 + h_7^* r_7\right) - s_1 \right\|^2 + \left(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\right) \lvert s_1 \rvert^2 \right]$ |
| $s_2$ | $Arg\ \min_{s_2 \in S} \left[ \left\| \left(-h_8 r_7^* - h_2 r_1^* - h_6 r_5^* + h_5^* r_6 + h_7^* r_8 + h_1^* r_2 + h_3^* r_4 - h_4 r_3^*\right) - s_2 \right\|^2 + \left(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\right) \lvert s_2 \rvert^2 \right]$ |
| $s_3$ | $Arg\ \min_{s_3 \in S} \left[ \left\| \left(h_8 r_6^* - h_3 r_1^* + h_1^* r_3 - h_2^* r_4 + h_4 r_2^* - h_7 r_5^* + h_5^* r_7 - h_6^* r_8\right) - s_3 \right\|^2 + \left(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\right) \lvert s_3 \rvert^2 \right]$ |
| $s_4$ | $Arg\ \min_{s_4 \in S} \left[ \left\| \left(-h_8 r_4^* + h_7 r_3^* - h_5 r_1^* + h_6 r_2^* + h_1^* r_5 - h_2^* r_6 - h_3^* r_7 + h_4^* r_8\right) - s_4 \right\|^2 + \left(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\right) \lvert s_4 \rvert^2 \right]$ |

## 3. DECODING METRICS

In this section, a channel comprising eight transmit antennas and one receive antenna is examined. Let $R_{1\times 8}$, $H_{1\times 8}$ and $N_{1\times 8}$ be the matrices of received signals, of transmission coefficients and of noise, respectively. The transmit antennas are assumed to be sufficiently separately, so that the transmission coefficients between these transmit antennas and the receive antenna are independent of one another. The transmission equation is then as follows:

$$R = HZ + N$$

The transmitted symbols can be decoded following the Maximum Likelihood (ML) decoding scheme, which is expressed as:

$$\{\hat{s_k}\}_{k=1}^{8} = Arg \min_{\{s_k\}, s_k \in S} \|R - HZ\|_F \tag{14.4}$$

where $Arg(x)$ is the argument of $x$; $\|\Re\|_F$ is the Frobenius norm of the matrix $\Re$, i.e., the square root of the sum of all the magnitude squared elements of the matrix, and $S$ is the set of all possibilities of the transmitted symbols. We consider the conventional code $Z_1$ first. Since the transmitted block code $Z_1$ is orthogonal, then (14.4) can be extracted to four independent expressions for four corresponding symbols. The decision metrics for decoding the symbols of code $Z_1$ are given in Table 14-2. Similarly, the decision metrics for decoding codes $Z_2$, $Z_3$ and $Z_4$ are derived in Tables 14-3, 14-4 and 14-5, respectively.

*Table 14-3.* Decision metrics for decoding code $Z_2$.

| *Variable* | *Decoding metric* |
|---|---|
| $s_1$ | $Arg\ \min_{s_1 \in S} \Big[ \Big\| \frac{1}{2}\big( -h_4 r_4^* - h_3 r_4^* - h_4^* r_4 + h_3^* r_4 - h_8^* r_8 - h_8 r_7^* - h_7 r_7^* + h_8^* r_7 - h_7^* r_7 - h_3 r_3^* - h_7 r_8^* - h_8 r_8^* + h_7^* r_8 + 2h_5^* r_5 + 2h_6 r_6^* + 2h_2 r_2^* + 2h_1^* r_1 - h_3^* r_3 + h_4^* r_3 - h_4 r_3^*\big) - s_1 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_1 \rvert^2 \Big]$ |
| $s_2$ | $Arg\ \min_{s_2 \in S} \Big[ \Big\| \frac{1}{2}\big( h_4 r_4^* - h_3 r_4^* - h_4^* r_4 - h_3^* r_4 - 2h_6 r_5^* + 2h_5^* r_6 - h_7 r_7^* - h_7 r_8^* - h_7^* r_8 + h_8 r_8^* - 2h_2 r_1^* + h_8 r_7^* - h_3 r_3^* + 2h_1^* r_2 + h_3^* r_3 + h_4^* r_3 + h_4 r_3^* - h_8^* r_8 + h_7^* r_7 + h_8^* r_7\big) - s_2 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_2 \rvert^2 \Big]$ |
| $s_3$ | $Arg\ \min_{s_3 \in S} \Big[ \Big\| \frac{1}{\sqrt{2}}\big( h_1^* r_4 - h_2^* r_4 + h_8^* r_6 - h_7^* r_5 + h_1^* r_3 - h_7^* r_6 - h_5 r_7^* - h_6 r_7^* - h_5 r_8^* + h_6 r_8^* + h_3 r_1^* + h_4 r_1^* + h_3 r_2^* - h_4 r_2^* - h_8^* r_5 + h_2^* r_3\big) - s_3 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_3 \rvert^2 \Big]$ |
| $s_4$ | $Arg\ \min_{s_4 \in S} \Big[ \Big\| \frac{1}{\sqrt{2}}\big( -h_4^* r_6 + h_2^* r_7 + h_1^* r_7 - h_2^* r_8 + h_5 r_4^* + h_4^* r_5 + h_1^* r_8 + h_7 r_1^* + h_8 r_1^* + h_7 r_2^* - h_6 r_4^* + h_3^* r_5 + h_3^* r_6 - h_8 r_2^* + h_5 r_3^* + h_6 r_3^*\big) - s_4 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_4 \rvert^2 \Big]$ |

## 4. CHOICE OF SIGNAL CONSTELLATIONS

By examining the constituent matrices $\{A_j\}_{j=1}^4$, $\{B_j\}_{j=1}^4$, and the encoding matrix $Z_4$, for instance, it is easy to notice that the entries $z_{lk}$ ($l$=5...8, $k$=1...8) of $Z_4$ are composed of the real part of one indeterminate and the imaginary part of another indeterminate, e.g., $z_{51}=\frac{-s_4^R+is_3^I}{\sqrt{2}}$. This observation means that if the indeterminates $s_1,\ldots,s_4$ are chosen from the complex signal constellations where $s_j^R$ or $s_j^I$ ($j$=1...4) can be equal to zero, e.g., the QPSK constellation (1,-1,$i$,-$i$) then, some of the entries of the matrix $Z_4$ can be equal to zero depending on the transmitted data. Therefore, such constellations should be avoided. An example of a constellation where the power is evenly spread among the transmit antennas independently of the transmitted data is the QPSK constellation ($1+i$,1-$i$,-1+$i$,-1-$i$). This observation also holds for the other proposed codes $Z_2$ and $Z_3$.

## 5. CONCLUSION

In the chapter, we presented three new CO STBCs of order eight based on *COD*(8; 1, 1, 2, 2), *COD*(8; 1,1,1,4) and *COD*(8; 2, 2, 2, 2)

*Table 14-4.* Decision metrics for decoding code $Z_3$.

| *Variable* | *Decoding metric* |
|---|---|
| $s_1$ | $Arg\ \min_{s_1 \in S} \Big[ \Big\| \frac{1}{2}\big(h_5 r_5^* + h_5^* r_5 - h_7^* r_5 + 2h_2^* r_2 + 2h_1^* r_1 + 2h_4 r_4^* + 2h_3 r_3^* + h_5 r_7^* + h_6 r_6^* + h_8 r_6^* - h_8^* r_6 + h_6^* r_6 + h_7 r_5^* + h_6 r_8^* + h_8 r_8^* - h_6^* r_8 + h_8^* r_8 + h_7 r_7^* - h_5^* r_7 + h_7^* r_7\big) - s_1 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_1 \rvert^2 \Big]$ |
| $s_2$ | $Arg\ \min_{s_2 \in S} \Big[ \Big\| \frac{1}{2}\big(-2h_6^* r_5 + h_2 r_4^* + h_1 r_4^* - h_2^* r_4 + h_1 r_4^* + h_3^* r_2 - h_4^* r_2 - h_3 r_1^* - h_4 r_1^* + h_3^* r_1 - h_4^* r_1 + h_1^* r_3 - h_2^* r_3 - h_1 r_3^* - h_2 r_3^* + 2h_5 r_6^* + h_3 r_2^* + h_4 r_2^* - 2h_7 r_8^* + 2h_8^* r_7\big) - s_2 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_2 \rvert^2 \Big]$ |
| $s_3$ | $Arg\ \min_{s_3 \in S} \Big[ \Big\| \frac{1}{2}\big(h_1^* r_4 + h_2^* r_4 - h_7 r_5^* - h_5^* r_5 - h_7^* r_5 - h_1 r_4^* + h_2 r_4^* + h_3^* r_2 - h_4^* r_2 - h_3 r_1^* - h_4 r_1^* + h_4^* r_1 - h_3^* r_1 + h_1^* r_3 + h_2^* r_3 + h_5 r_7^* + h_1 r_3^* - h_2 r_3^* + h_5 r_5^* - h_3 r_2^* - h_4 r_2^* - h_6 r_6^* - h_8 r_6^* + h_6^* r_6 - h_8^* r_6 + h_8 r_8^* + h_6^* r_8 - h_8^* r_8 - h_7 r_7^* + h_6 r_8^* + h_5^* r_7 + h_7^* r_7\big) - s_3 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_3 \rvert^2 \Big]$ |
| $s_4$ | $Arg\ \min_{s_4 \in S} \Big[ \Big\| \frac{1}{2}\big(h_7 r_4^* + h_1^* r_5 + h_2^* r_5 + h_3^* r_5 + h_4^* r_5 + h_6 r_4^* - h_8 r_4^* - h_5 r_3^* - h_5 r_1^* - h_6 r_1^* - h_7 r_1^* - h_8 r_1^* - h_6 r_3^* + h_8 r_3^* + h_6 r_2^* - h_7 r_2^* + h_8 r_2^* - h_5 r_2^* - h_5 r_4^* + h_1^* r_6 - h_2^* r_6 + h_3^* r_6 - h_4^* r_6 - h_2^* r_8 - h_3^* r_8 + h_4^* r_8 + h_1^* r_7 + h_2^* r_7 - h_3^* r_7 - h_4^* r_7 + h_1^* r_8 + h_7 r_3^*\big) - s_4 \Big\|^2 + \big(-1 + \sum_{i=1}^{8} \lvert h_i \rvert^2\big) \lvert s_4 \rvert^2 \Big]$ |

where less time slots are wasted, compared to the conventional code. As a result, the new codes require lower peak-to-mean power ratios at transmit antennas to achieve the same bit error performance as the conventional code $Z_1$. Since, some of the variables in codes $Z_2$ and $Z_3$ appear more often than the others, these codes can be utilized for multi-modulation systems, where the variables appearing more frequently carry signals from higher order constellations than those appearing just once in each column. Moreover, because of the lack of zeroes in the design $Z_4$, this code is much easier to implement.

# REFERENCES


1. A. V. Geramita and J. Seberry, *Orthogonal designs: quadratic forms and Hadamard matrices,* Lecture notes in pure and applied mathematics, Vol. 43, Marcel Dekker, New York and Basel, 1979.
2. X.-B. Liang, "Orthogonal designs with maximal rates," *IEEE Trans. Inform. Theory,* Vol. 49, No. 10, pp. 2468–2503, Oct. 2003.

*Table 14-5.* Decision metrics for decoding code $Z_4$.

| *Variable* | *Decoding metric* |
|---|---|
| $s_1$ | $Arg\ \min_{s_1 \in S} \Big[\Big\| \frac{1}{2\sqrt{2}}\big( -2h_3r_4^* + 2h_4r_4^* + h_5r_5^* + h_6r_5^* + h_7r_5^* + h_8r_5^* + h_5r_6^* - h_6r_6^* + h_7r_6^* + h_6r_7^* + h_7r_7^* + h_8r_7^* + h_5r_7^* - h_8r_6^* + h_7r_8^* - h_8r_8^* + h_5r_8^* - h_6r_8^* - 2h_3r_3^* - 2h_4r_3^* + 2h_1^*r_2 - 2h_2^*r_2 - h_5^*r_5 - h_6^*r_5 + h_7^*r_5 + h_8^*r_5 - h_5^*r_6 + h_6^*r_6 + h_7^*r_6 - h_8^*r_6 + h_5^*r_7 + h_6^*r_7 - h_7^*r_7 - h_8^*r_7 + h_5^*r_8 - h_6^*r_8 - h_7^*r_8 + h_8^*r_8 + 2h_1^*r_1 + 2h_2^*r_1\big) - s_1 \Big\|^2 + \big( -1 + \sum_{i=1}^{8} |h_i|^2\big)|s_1|^2 \Big]$ |
| $s_2$ | $Arg\ \min_{s_2 \in S} \Big[\Big\| \frac{1}{2\sqrt{2}}\big( 2h_4r_1^* - 2h_4r_2^* + 2h_3r_2^* + h_6r_5^* - h_7r_5^* + h_5r_5^* + h_5r_6^* - h_6r_6^* - h_7r_6^* - h_8r_5^* + h_6r_7^* - h_7r_7^* - h_8r_7^* + h_5r_7^* + h_8r_6^* - h_7r_8^* + h_8r_8^* - h_6r_8^* + h_5r_8^* + 2h_3r_1^* + 2h_1^*r_4 - 2h_2^*r_4 + h_5^*r_5 + h_6^*r_5 + h_7^*r_5 + h_8^*r_5 + h_5^*r_6 - h_6^*r_6 + h_7^*r_6 - h_8^*r_6 - h_5^*r_7 - h_6^*r_7 - h_7^*r_7 - h_8^*r_7 - h_5^*r_8 + h_6^*r_8 - h_7^*r_8 + h_8^*r_8 + 2h_1^*r_3 + 2h_2^*r_3\big) - s_2 \Big\|^2 + \big( -1 + \sum_{i=1}^{8} |h_i|^2\big)|s_2|^2 \Big]$ |
| $s_3$ | $Arg\ \min_{s_3 \in S} \Big[\Big\| \frac{1}{2\sqrt{2}}\big( -2a_4r_6^* + a_7r_3^* + a_8r_3^* - h_5r_1^* - h_5r_4^* - h_6r_1^* - 2h_2r_6^* - h_7r_1^* - h_6r_4^* + h_7r_4^* - h_5r_2^* + h_8r_4^* - h_7r_2^* - h_6r_2^* - h_8r_2^* - h_8r_1^* + 2h_4r_8^* - 2h_2r_8^* - h_5r_3^* - h_6r_3^* - h_5^*r_2 - h_5^*r_4 + h_6^*r_4 + h_7^*r_4 - h_8^*r_4 + h_6^*r_2 - h_7^*r_2 + h_8^*r_2 + 2h_1^*r_5 + 2h_3^*r_5 + 2h_1^*r_7 - 2h_3^*r_7 + h_5^*r_3 - h_6^*r_3 - h_7^*r_3 + h_8^*r_3 + h_5^*r_1 - h_6^*r_1 + h_7^*r_1 - h_8^*r_1\big) - s_3 \Big\|^2 + \big( -1 + \sum_{i=1}^{8} |h_i|^2\big)|s_3|^2 \Big]$ |
| $s_4$ | $Arg\ \min_{s_4 \in S} \Big[\Big\| \frac{1}{2\sqrt{2}}\big( 2h_4r_5^* - h_5r_1^* + 2h_2r_5^* + h_8r_1^* - h_8r_3^* + h_6r_3^* + h_7r_3^* + h_6r_4^* - h_5r_4^* + h_6r_1^* + h_7r_4^* - h_8r_4^* - h_7r_2^* + h_6r_2^* - h_5r_2^* + h_8r_2^* - h_7r_1^* + 2h_2r_7^* - 2h_4r_7^* - h_5r_3^* + h_5^*r_2 + h_5^*r_4 + h_6^*r_4 - h_7^*r_4 - h_8^*r_4 + h_6^*r_2 + h_7^*r_2 + h_8^*r_2 + 2h_1^*r_6 + 2h_3^*r_6 + 2h_1^*r_8 - 2h_3^*r_8 - h_5^*r_3 - h_6^*r_3 + h_7^*r_3 + h_8^*r_3 - h_5^*r_1 - h_6^*r_1 - h_7^*r_1 - h_8^*r_1\big) - s_4 \Big\|^2 + \big( -1 + \sum_{i=1}^{8} |h_i|^2\big)|s_4|^2 \Big]$ |


3. S. M. Alamouti, A simple transmit diversity scheme for wireless communications, *IEEE J. Select. Areas Commun.,* Vol. 16, pp. 1451-1458, Oct. 1998.
4. V. Tarokh, H. Jafarkhani and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Trans. Inform. Theory,* Vol. 45, pp.1456–1467, July 1999.
5. D. J. Street, Amicable orthogonal designs of order eight, *Journal of Australian Mathematical Society* (A), Vol. 33, pp. 23–29, 1982.
6. O. Tirkkonen and A. Hottinen, Square-matrix embeddable space-time block codes for complex signal constellations, *IEEE Trans. Inform. Theory,* Vol.48, No.2, pp. 384–395, Feb. 2002.
7. L. C. Tran, J. Seberry, B. J. Wysocki, T. A. Wysocki, T. Xia and Y. Zhao,"Two new complex orthogonal space-time codes for 8 transmit antennas," *IEE Electronics Lett.,* Vol. 40, No. 1, pp. 55-56, Jan. 2004

# PART 3:

# HARDWARE IMPLEMENTATION

*This page intentionally left blank*

Chapter 15

# DESIGN OF ANTENNA ARRAY USING DUAL NESTED COMPLEX APPROXIMATION

Mattias Dahl[1], To Tran[1], Ingvar Claesson[1] and Sven Nordebo[2]
[1]*Blekinge Institute of Technology, School of Engineering, Department of Signal Processing, 372 25 Ronneby, Sweden;* [2]*Växjö University, School of Mathematics and Systems Engineering, 351 95 Växjö Sweden*

**Abstract:** This paper presents a new practical approach to complex Chebyshev approximation by semi-infinite linear programming. By the new front-end technique, the associated semi-infinite linear programming problem is solved exploiting the finiteness of the related Lagrange multipliers by adapting finite-dimensional linear programming to the dual semi-infinite problem, and thereby taking advantage of the numerical stability and efficiency of conventional linear programming software packages. Furthermore, the optimization procedure is simple to describe theoretically and straightforward to implement in computer coding. The new design technique is therefore highly accessible. The algorithm is formally introduced as the linear Dual Nested Complex Approximation (DNCA) algorithm. The DNCA algorithm is versatile and can be applied to a variety of applications such as narrow-band as well as broad-band beamformers with any geometry, conventional Finite Impulse Response (FIR) filters, analog and digital Laguerre networks, and digital FIR equalizers. The proposed optimization technique is applied to several numerical examples dealing with the design of a narrow-band base-station antenna array for mobile communication.

**Key words:** antenna array design, Dual Nested Complex Approximation, DNCA, optimization, semi-infinite linear program, real rotation theorem, Chebyshev approximation

## 1. INTRODUCTION

The array pattern synthesis of a non-uniformly spaced sensor array or beamformer [1,2] is closely related to the design of an FIR filter with arbitrary or non-linear phase response. The essential similarity is the finite-dimensional nature of the complex approximating functions [3-6].

For beamformers as for FIR filters the design problem is often cast as a finite-dimensional complex approximation problem [2,3]. It was shown that the (non-linear) complex Chebyshev approximation problem can be reformulated as an equivalent real semi-infinite linear program [2,7]. Classical least squares approximation methods can in many cases be used to obtain a desired solution [1]. However, when the design specification is given as a bound on the complex design error, the problem is naturally converted to a complex Chebyshev approximation problem.

Other possible methods to find the Chebyshev solution include quadratic programming [8]. It was shown that the semi-infinite linear program corresponding to the (real) Chebyshev approximation problem can be solved by using numerically efficient simplex extension algorithms [9]. In this context, it is also noted that some other recent approaches to complex FIR filter design such as a generalized Remez algorithm [10], and Tang's algorithm [5, 11] in fact also employ simplex extension algorithms. These results were later exploited for the design of digital FIR filters and digital Laguerre networks with complex Chebyshev error criteria [6,12].

Finitization can in principle give an arbitrarily accurate approximation of the complex Chebyshev solution but becomes exceedingly memory intensive as the grid spacing decreases [3]. The semi-infinite formulation deals directly with the true complex error and not an approximation thereof. The general complex approximation problems require an optimization formulation such as with Semi-Infinite Programming [6,8] (SIP), or Second Order Cone Programming [13,14] (SOCP). With SIP, a finite number of non-linear constraints are transformed to an infinite number of linear constraints, i.e. a linear combination of continuous functions. The problem can then be solved efficiently using algorithms which are extensions of the standard linear and quadratic programming algorithms such as the simplex algorithm [6,8,9]. With SOCP, the problem is solved without linearizing the constraints, and by using newly developed interior point methods [13-15].

An obstacle with the semi-infinite simplex extension algorithm [6,9,12] is the lack of commercially available software for efficient and reliable numerical solution of general complex approximation problems. In order to overcome the difficulties mentioned above, we present in this paper an applied semi-infinite front-end technique for complex Chebyshev approximation which is based on conventional finite-dimensional linear programming subroutines. The essence of the new technique, justified by the Caratheodory dimensionality theorem [16], is to exploit the finiteness of the related Lagrange multipliers by adapting conventional finite-dimensional linear programming to the semi-infinite linear programming problem.

By the proposed front-end technique, the complex Chebyshev approximation problem can be solved by taking advantage of the numerical

stability and efficiency of the given linear programming software package. Furthermore, the optimization procedure is simple to describe theoretically and straightforward to implement in computer coding. The new design technique should therefore be highly accessible for most design engineers.

In order to illustrate the flexibility and numerical efficiency of the proposed design technique we have included several design examples concerning the optimization of a narrow-band base-station antenna array for mobile communication in the 450 MHz band.

## 2. PROBLEM FORMULATION

To demonstrate the versatility of the complex approximation technique the optimization is performed from an application point of view. The design examples are taken from the mobile communication base-station area, or more precisely, design of an antenna array in the 450 MHz band. As a numerical example, consider the shielded planar hexagonal antenna array where the sensor elements are evenly distributed in sections on a hexagon, and with the incident wave front propagating in the same plane as the array, see Fig. 15-1 . The rather unusual configuration also illustrates the generality and the applicability for arbitrary basis functions (array response) for arbitrary geometries.

Consider the far-field and narrow-band case where the phase of the wave front is given by $e^{j(2\pi f_0 t - \mathbf{k}^T \mathbf{r})}$ where $f_0$ is the frequency, $t$ the time, $\mathbf{k} = \left(-2\pi f_0 / c\right) \cdot [\cos\varphi \quad \sin\varphi]$ the wave vector, $c$ the speed of wave propagation, $\varphi$ the angle of incidence and $\mathbf{r}$ the evaluated spatial point, see Fig. 15-1 . The hexagonal sectioned antenna array consists of antenna elements distributed along the ground-plane sections. One antenna element and the ground-plane together form a dipole. The dipoles have the spatial positions $(r_m, \varphi_m)$ where $m = 0, \ldots, M-1$ and $r_m$ and $\varphi_m$ are the distance and angle, respectively, between the middle of the hexagon to the dipole center. The phase center and origin of coordinates are located in the middle of the hexagon and the total array response $H(\varphi)$ is given as

$$H(\varphi) = \sum_{m=0}^{M-1} w_m a_m(\varphi) e^{j 2\pi f_0 \frac{r_m}{c} \cos(\varphi_m - \varphi)} \tag{15.1}$$

where $M$ is the total number of used/active dipoles in the antenna and

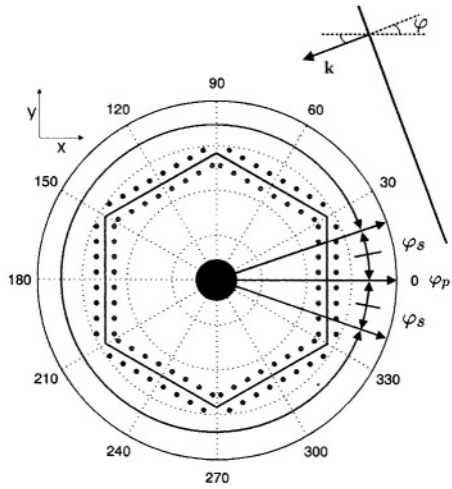



*Figure 15-1* A planar hexagonal antenna array with shield for the far-field where the outer '●' represents the sensor elements, the solid hexagon the ground-plane shield and the inner '●' represents the mirror source. The passband (main-lobe look direction) is defined by the angle $\varphi_p$ ($\varphi_p = 0$ in the figure) and the stopband (side-lobe region) by the interval $[\varphi_p+\varphi_s, \varphi_p-\varphi_s+360^\circ]$. The six ground-plane sections are positioned in directions $\varphi_g=[0^\circ,60^\circ,120^\circ,180^\circ,240^\circ,300^\circ]$.

$w_m$ is a complex weight. The radiation characteristics for a dipole located in $(r_m,\varphi_m)$ is denoted $a_m(\varphi)$. We only use three of the six array-sections at a time and consequently a subset of all dipoles are used. Depending on the angle of incidence $\varphi$ the three nearest heading sections with angle $\varphi_g$ are selected. The complex antenna array response for the angle $\varphi$ using vector notation is given by

$$\begin{aligned} H(\varphi) &= \mathbf{w}^H \mathbf{d}(\varphi) \\ &= \tilde{\mathbf{w}}^H \tilde{\mathbf{d}}(\varphi) \\ &= \begin{bmatrix} \Re\{w\} \\ \Im\{w\} \end{bmatrix}^T \begin{bmatrix} d(\varphi) \\ -jd(\varphi) \end{bmatrix}, \end{aligned} \tag{15.2}$$

where the complex array $\mathbf{w} = \Re\{\mathbf{w}\} + j\Im\{\mathbf{w}\}$ is an $M \times 1$ array vector of complex coefficients $w_m$ and $\mathbf{d}(\varphi)$ the corresponding array response vector of complex continuous and linearly independent transfer functions $d_m(\varphi) = a_m(\varphi) e^{j2\pi f_0 \frac{r_m}{c}\cos(\varphi_m-\varphi)}$, $m = 0,\ldots,M-1$. The $\tilde{\mathbf{w}}$ is an

$N \times 1$ real vector and $\tilde{\mathbf{d}}(\varphi)$ an $N \times 1$ complex vector where $N = 2M$. In order to make the passband extremely narrow, the passband in this formulation was restricted to a single point $\varphi_p$. The stopband is defined as $\Phi = [\varphi_p + \varphi_s, \varphi_p + 2\pi - \varphi_s]$ in radians.

## 2.1 The design specification

Consider the following design specification

$$\begin{cases} |H(\varphi)| \le \sigma(\varphi), \quad \varphi \in \Phi \\ H(\varphi_p) = 1 \end{cases} \tag{15.3}$$

where $\sigma(\varphi)$ is a prescribed strictly positive magnitude bound which leads to the minimax design formulation

$$\begin{cases} \min\limits_{\tilde{w} \in R^N} \max\limits_{\varphi \in \Phi} v(\varphi)|H(\varphi)| \\ H(\varphi_p) = 1 \end{cases} \tag{15.4}$$

where $v(\varphi) = 1/\sigma(\varphi)$. It is concluded that a solution to the specification in Eq. (15.3) exists if and only if the optimal objective value in Eq. (15.4) is less than or equal to one.

# 3. SEMI-INFINITE LINEAR PROGRAMMING

The optimal solution to the minimax formulation in Eq. (15.4) is given by the equivalent formulation

$$\begin{cases} \min \delta \\ v(\varphi)|H(\varphi)| - \delta \le 0, \quad \varphi \in \Phi \\ H(\varphi_p) = 1 \end{cases} \tag{15.5}$$

where $\delta$ is an additional real variable.

## 3.1 Semi-infinite linear programming formulation

Equation (15.5) corresponds to a non-linear optimization problem which is very difficult to treat as it stands due to the non-linear constraints. Equation (15.5) will therefore be converted to a semi-infinite linear programming problem. According to the *real rotation theorem* [17], a magnitude inequality in the complex plane can be expressed in the equivalent form

$$|z| \le \sigma \Leftrightarrow \Re\left\{ze^{j\theta}\right\} \le \sigma \quad \forall \theta \in [0,2\pi] \tag{15.6}$$

where $z$ is a complex number and $\sigma$ a real and positive number. By making use of Eq. (15.6), the design problem in Eq. (15.5) is reformulated as

$$\begin{cases} \min \delta \\ v(\varphi)\Re\left\{H(\varphi)e^{j\theta}\right\} - \delta \le 0, \quad (\varphi,\theta) \in \Phi \times \Theta \\ H(\varphi_p) = 1 \end{cases} \tag{15.7}$$

where $\Theta = [0,2\pi]$. In order to emphasize the linear structure of this formulation, Eq. (15.7) is finally rewritten as the following semi-infinite linear program

$$\begin{cases} \min \delta \\ \mathbf{a}^T(\varphi,\theta)\tilde{\mathbf{w}} - \delta \le 0, \quad (\varphi,\theta) \in \Phi \times \Theta \\ \mathbf{P}\tilde{\mathbf{w}} = \mathbf{p} \end{cases} \tag{15.8}$$

Where $\mathbf{a}(\varphi,\theta) = v(\varphi)\Re\left\{\tilde{\mathbf{d}}(\varphi)e^{j\theta}\right\}$, $\mathbf{P}$ is an $L \times N$ constraint matrix and $\mathbf{p}$ an $L \times 1$ constraint vector. The main-lobe constraint $H(\varphi_p) = 1$ in Eq. (15.7) is obtained by choosing $\mathbf{P}^T = \left[\Re\left\{\tilde{\mathbf{d}}(\varphi_p)\right\} \; \Im\left\{\tilde{\mathbf{d}}(\varphi_p)\right\}\right]$ and $\mathbf{p}^T = [1,0]$. The linear program in Eq. (15.8) is called semi-infinite since the number of variables (unknowns) are finite but the constraint set is infinite. For practical purposes in the implementation of the optimization algorithm, it is assumed that the set $\Phi$ is finite. Note that the total corresponding approximation problem is with respect to the true complex error since the phase parameter $\theta$ belongs to the infinite set $\theta \in [0,2\pi]$.

# 4. DUAL NESTED COMPLEX APPROXIMATION ALGORITHM

This section gives a description and an outlined convergence proof for the optimization algorithm. It is shown that this can be accomplished without explicit reference to the conceptually abstract dual formulation.

## 4.1 The complex approximation problem

The problem in Eq. (15.8) is referred to as Semi-Infinite Linear Programming (SILP) Problem and can be described using the general semi-infinite programming formulation

$$(P)\begin{cases}\min\limits_{\mathbf{w}} f(\tilde{\mathbf{w}}), \\ g_\alpha(\tilde{\mathbf{w}}) \le 0,\ \alpha \in \mathsf{A} \subset R^k \\ \tilde{\mathbf{w}} \in \tilde{W} \subset R^n\end{cases} \tag{15.9}$$

where $\tilde{\mathbf{w}}$ is an $N \times 1$ variable vector, $f(\tilde{\mathbf{w}})$ a convex continuous function $\tilde{W}$ a convex restriction set, $\mathcal{A}$ an infinite index set as a compact subset of Euclidean $k$-space, and $g_\alpha(\tilde{\mathbf{x}})$ a continuous constraint function which is convex for any fixed index $\alpha$.

## 4.2 Dual nested complex approximation

The Dual Nested Complex Approximation (DNCA) algorithm[1] to solve Eq. (15.9) is outlined below. Let $\mathcal{A}^{(k)}$ denote a sequence of finite subsets of the infinite index set $\mathcal{A}$ and initialize the algorithm with the subset $\mathcal{A}^{(0)}$.

1. Given $\mathcal{A}^{(k)} \subset \mathcal{A}$, solve the subproblem

$$\left(P^{(k)}\right)\begin{cases}\min\limits_{\mathbf{w}} f(\tilde{\mathbf{w}}),\ \tilde{\mathbf{w}} \in \tilde{W} \\ g_\alpha(\tilde{\mathbf{w}}) \le 0,\ \alpha \in \mathsf{A}^{(k)}\end{cases} \tag{15.10}$$

yielding the solution vector $\tilde{\mathbf{w}}_k$ and the Lagrange multiplier vector $\boldsymbol{\lambda}_k$.

[1] The DNCA optimization algorithm has been implemented in MATLAB™ and is available at http://www.bth.se/its/dnca/.

2. Reduce the subset by the inactive constraints

$$\mathcal{A}_R^{(k)} = \mathcal{A}^{(k)} \setminus \left\{ \alpha \in \mathcal{A}^{(k)} \mid (\lambda_k) = 0 \right\} \tag{15.11}$$

3. Define the entering index

$$\hat{\alpha}_e = \arg\max_{\alpha} g_\alpha \left( \tilde{\mathbf{w}}_k \right), \tag{15.12}$$

$$\mathcal{A}^{(k+1)} = \mathcal{A}_R^{(k)} \bigcup \left\{ \hat{\alpha}_e \right\} \tag{15.13}$$

and return to step 1 above.

The applicability of the algorithm above to complex approximation lies in the fact that Eq. (15.12) can easily be calculated when the approximation domain $\Phi = \Phi_1 \bigcup \Phi_2$ is finite. Let e.g. $g_\alpha(\mathbf{x})$ where $\mathbf{x} = \tilde{\mathbf{w}}$ be given by

$$g_\alpha(\mathbf{x}) = g_{\varphi,\theta}(\mathbf{x}) = \Re\left\{ \left( \mathbf{a}^T(\varphi)\mathbf{x} + b(\varphi) \right) e^{j\theta} \right\} + \mathbf{c}^T(\varphi)\mathbf{x} + d(\varphi) \tag{15.14}$$

where $\alpha = (\varphi, \theta) \in \mathcal{A} = \Phi \times \Theta$, $\mathbf{a}(\varphi)$ and $b(\varphi)$ are complex and $\mathbf{c}(\varphi)$ and $d(\varphi)$ are real, cf. Eq.(15.12). By employing Eq. (15.6), Eq. (15.12) can be calculated as $\hat{\alpha}_e = \left( \hat{\varphi}_e, \hat{\theta}_e \right)$ where

$$\hat{\varphi}_e = \arg\max_{\varphi} \left\{ \left| \mathbf{a}^T(\varphi)\mathbf{x}_k + b(\varphi) \right| + \mathbf{c}^T(\varphi)\mathbf{x}_k + d(\varphi) \right\} \tag{15.15}$$

$$\hat{\theta}_e = -\arg\left\{ \mathbf{a}^T\left( \hat{\varphi}_e \right)\mathbf{x}_k + b\left( \hat{\varphi}_e \right) \right\} \tag{15.16}$$

*Key observation 1:* Since the number of variables is $N+1$, note that an optimization software will usually give a total number of $N+1$ Lagrange multipliers greater than zero. The size of the so called *reference* set $\mathcal{A}^{(k)}$ is in fact $r \leq N+1$.

*Key observation 2:* The dual formulation suggests that the primal problem can be solved by considering a sequence of subproblems as in with increasing minimum cost and which is based only on finite subsets $\mathcal{A}^{(k)} = \{(\varphi_1, \theta_1), \ldots, (\varphi_r, \theta_r)\}$ consisting of no more than $N-1$ points of

$D = \Phi \times \Theta$. This observation constitutes the foundation for the development of the DNCA optimization algorithm.

*Key observation 3:* The number of variables is $N+1$ and the size of the *reference* set $A^{(k)}$ is only $r \leq N+1$. The constraint index $\hat{\alpha}_e = (\varphi_e, \theta_e)$ which is chosen to enter the basis $A^{(k)}$ is usually defined by the maximum constraint violation. The entering constraint $\hat{\alpha}_e$ is very likely to be independent of the small reference set $A^{(k)}$. This is the primary reason ensuring that the DNCA itself is a highly numerical stable procedure.

## 4.3 Convergence proof

Since the reduced subset $A_R^{(k)}$ yields the same solution $\tilde{\mathbf{w}}_k$ as the subset $A^{(k)}$ we have $f(\tilde{\mathbf{w}}_k) \leq f(\tilde{\mathbf{w}}_{k+1}) \leq f_{opt}$ and the sequence $f(\tilde{\mathbf{w}}_k)$ converges. However, it remains to be shown that the sequence $\tilde{\mathbf{w}}_k$ is not stuck in any state of cycling.

Convergence can be proved straightforwardly when the cost function $f(\tilde{\mathbf{w}})$ is strictly convex. Assume that $\tilde{\mathbf{w}}_k$ is not optimal, then $g_{\hat{\alpha}}(\tilde{\mathbf{w}}_k) > 0$. The claim is that there is a strict ascent, $f(\tilde{\mathbf{w}}_{k+1}) > f(\tilde{\mathbf{w}}_k)$, so that cycling cannot occur. Assume on the contrary that $f(\tilde{\mathbf{w}}_{k+1}) \leq f(\tilde{\mathbf{w}}_k)$. Define $\Omega_k = \left\{\tilde{\mathbf{w}} \mid g_\alpha(\tilde{\mathbf{w}}) \leq 0, \alpha \in A^{(k)}\right\}$ and $\Omega_{k+1} = \Omega_k \cap \left\{\tilde{\mathbf{w}} \mid g_{\hat{\alpha}}(\tilde{\mathbf{w}}) \leq 0\right\}$. Obviously $\tilde{\mathbf{w}}_k \in \Omega_k$, $\tilde{\mathbf{w}}_{k+1} \in \Omega_k$, $\tilde{\mathbf{w}}_{k+1} \in \Omega_{k+1}$ and $\tilde{\mathbf{w}}_k \notin \Omega_{k+1}$. Thus, $\tilde{\mathbf{w}}_k \neq \tilde{\mathbf{w}}_{k+1}$ and $f\left(t\tilde{\mathbf{w}}_k + (1-t)\tilde{\mathbf{w}}_{k+1}\right) < tf\left(\tilde{\mathbf{w}}_k\right) + (1-t)f\left(\tilde{\mathbf{w}}_{k+1}\right) \leq$ $tf\left(\tilde{\mathbf{w}}_k\right) + (1-t)f\left(\tilde{\mathbf{w}}_k\right) = f\left(\tilde{\mathbf{w}}_k\right)$ for any $0 < t < 1$ which contradicts the fact that $\tilde{\mathbf{w}}_k$ is optimum on $\Omega_k$.

If the cost function $f(\tilde{\mathbf{w}})$ is not strictly convex, strict ascent can still be obtained by requiring that the condition $f(\tilde{\mathbf{w}}_k) > f(\tilde{\mathbf{w}}_{k-1})$ be satisfied in order to execute Eq. (15.11) in Step 2 of the DNCA algorithm.

Since the reduced subset $A_R^{(k)}$ yields the same solution $\tilde{\mathbf{w}}_k$ as the subset $A^{(k)}$ we have $f(\tilde{\mathbf{w}}_k) \leq f(\tilde{\mathbf{w}}_{k+1}) \leq f_{opt}$ and the sequence $f(\tilde{\mathbf{w}}_k)$ converges. However, convergence to the optimum value $f_{opt}$ is non-trivial to show. A

theoretical framework to prove the global convergence of the DNCA algorithm is shown in [18].

If $g_{\hat{\alpha}}(\widetilde{\mathbf{w}}_k) \le 0$ then $\widetilde{\mathbf{w}}_k$ is the optimum solution since it is necessary and sufficient for an optimum solution that there exists Lagrange multipliers defined on a finite subset of $\mathcal{A}$. A practical stopping criteria is therefore $g_{\hat{\alpha}}(\widetilde{\mathbf{w}}_k) \le \varepsilon(\widetilde{\mathbf{w}}_k)$ where the tolerance parameter $\varepsilon(\widetilde{\mathbf{w}}_k) > 0$ and may depend on the current solution $\widetilde{\mathbf{w}}_k$.

## 5. DESIGN EXAMPLES

As an application example, consider the planar hexagonal antenna array as defined in Section 2. If the weighting function $v(\varphi)$ is uniformly distributed an equiripple solution is achieved, i.e. the lowest possible side-lobe level in the stopband with respect taken to one or more linear constraints. The array response used is given by Eq. (15.1) and the corresponding real variables $\widetilde{w}_m$, are defined as in Eq. (15.2). The specification in Eq. (15.3) is used to state the desired array design. The solution is obtained by using the DNCA algorithm as described in Section 3. The performance of using a hexagonal sectioned ground-plane shielded antenna array as in Fig. 15-1 is investigated.

The examples show the flexibility in design using the proposed semi-infinite front-end algorithm. The algorithm is capable of solving huge design problems with many antenna elements, see Fig. 15-2a and Fig. 15-3. Main-lobe steering, see Fig.l5-2b, and side-lobe control by incorporating arbitrary weighting functions, can at the same time be taken into consideration during the optimization process as well. All antenna array responses are designed for the frequency $f_0 = 450\,\mathrm{MHz}$ and the interspacing between antenna elements in the array is $d = \lambda/2 \approx 0.3$ meter. The distance between the source element and the ground-plane is $d_g = \lambda/4 \approx 0.15$ meter. Note that by choosing $f_0$ the design problem can easily be scaled or translated into different frequency bands. Fig. 15-2*a* illustrates a typical equiripple solution for an antenna with $M$ = 1024 elements. Each linear antenna section consists of 34 ($M = 3 \cdot 34 = 102$) isotropic dipole elements. The antenna look direction for the main-lobe is $\varphi_p = 0^\circ$, that is the main-lobe (passband) consists only of one point and is defined by one single point constraint $H(\varphi_p) = H(0) = 1$. The radius of the antenna is approximately 9 meters. There are 102 complex weights $\mathbf{w}$, which implies that $N$ = 1024 real variables $\widetilde{\mathbf{w}}$ are involved in the optimization process. The corresponding

convergence behavior in Fig. 15-3 shows the advance of variable $\delta$ which is continuously increasing to the optimal value $\delta_o$. The maxnorm $\|H\|$ has a more irregular but still overall decreasing behavior also approaching the optimal value $\delta_o$. In this design example the side-lobe (stopband) region starts at $\pm 1.5°$, and in all the other examples at $\pm 5.5°$ from the desired main-lobe direction. The angular grid resolution is $0.25°$ in this example and $0.5°$ in all others.

In the examples concerning the main-lobe steering the look direction $\varphi_p$ for the main-lobe is gradually increased from $0°$ to $35°$ in steps of $5°$. The main-lobe look direction is defined by one single point constraint $H(\varphi_p) = 1$, see Fig. 15-2*a*. Each linear antenna section consists of 10 $(M = 3 \cdot 10 = 30)$ isotropic dipole elements. The unusual symmetry in the hexagonal antenna compared to a circular design is obvious. However, by using the proposed design method it is possible to apply lobe steering to direct the main-lobe in the angular domain $\varphi$. The drawback is that several sets of weights must be used. Fig. 15-2b illustrate the performance of the 3-sectioned hexagonal antenna using 30 antenna elements for main-lobe directions $\varphi_p \in [0°, 5°, 10°, 15°, 20°, 25°, 30°, 35°]$. The radius of the antenna is approximately 3 meters. The loss in stopband attenuation performance between the $0°$ and $30°$ design is $\sim 3$ dB. Note that lobe steering in the direction $35°$ degrees can be obtained by counterclockwise switching antenna sections and reuse a mirrored weight setup of the $25°$ design. In this way, the number of weight sets will be kept reasonably low.

The ability to incorporate a weighting function $v(\varphi) = 1/\sigma(\varphi)$ is a useful option when designing antenna arrays. The array response in the angular domain can in that way be shaped in an arbitrary sense. The antenna look direction for the main-lobe and weighting function can be chosen arbitrarily. The final value of variable $\delta$ yields information about the amplitude margin with respect to the function $\sigma(\varphi)$ and is plotted as $\delta_o$ in the convergence behavior plots. This value of variable $\delta$ is an important design parameter. If the uniform weighting function $\sigma(\varphi) = 1$ (0 dB) is used we simply achieve the maximum side-lobe suppression by observing the final value of variable $\delta$. It is possible to add additional point constraints such as nulls in the stopband or constraints on the main-lobe.

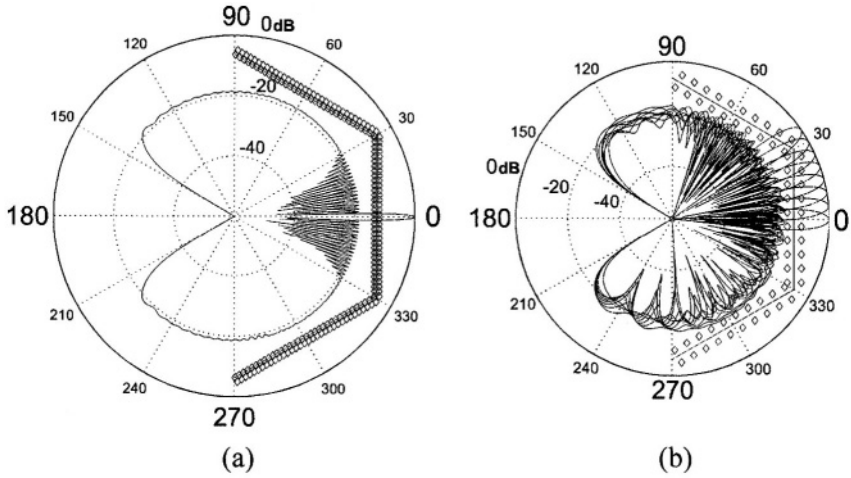



*Figure 15-2.* (a) Minimax design of a hexagonal antenna using 3 sections containing 34 antenna elements each. Side-lobe suppression ~18.5 dB. The design example consists of total *M=102* complex weights i.e. *N=204* real variables in the optimization. (b) Main-lobe steering $\varphi_p \in [0^\circ,5^\circ,10^\circ,\ldots,35^\circ]$. Minimax design of several lobes of an hexagonal antenna using 3 sections containing 10 antenna elements each.

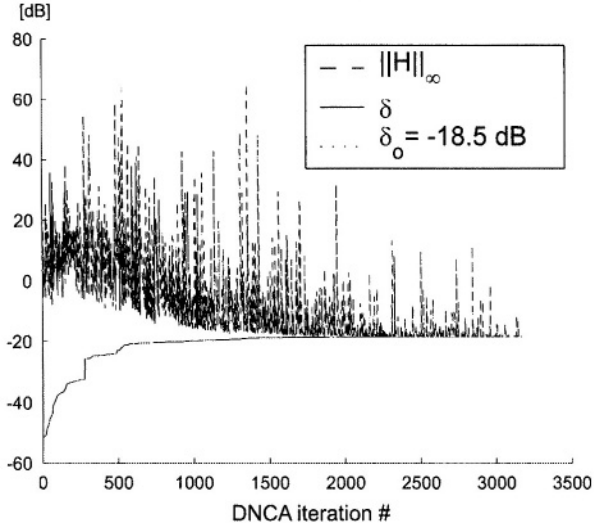



*Figure 15-3.* The monotonic convergence behaviour of $\delta$ corresponding to the minimax design in Fig. *15-2a.* The fluctuating maxnorm $||\boldsymbol{H}||$ and $\delta$ will converge to the same optimal value $\delta_o$.

## 6. SUMMARY

This paper solves antenna array Chebyshev approximation problems by exploiting the Caratheodory dimensionality theorem in a conventional linear programming software front-end. The semi-infinite linear programming theory is fairly recent [9] and there is lack of commercial software available for efficient numerical solution of the problem in Eq. (15.7). It is an extensive task to develop a specific software for semi-infinite linear programming to solve Eq. (15.7) if this software is also required to be stable and reliable with respect to numerical problems such as cycling phenomena, ill-conditioned matrix inversions, etc. On the other hand, the conventional finite-dimensional linear programming technique is well established and there is a lot of good, numerically reliable software available. The optimization technique proposed in this paper is therefore a convenient alternative which inherits the good numerical properties of the given linear programming subroutine. The computer code is significantly simplified in comparison with computer code which is tailored for semi-infinite linear programming. Moreover, the computational complexity is asymptotically equal. The proposed method is capable of solving large optimization problems such as huge antenna arrays and optimization variables. Extensive evaluations indicate the flexibility using the proposed front-end method.

## ACKNOWLEDGEMENT

This work has been supported by NUTEK, the National Swedish Board for Technical Development and KKS, The Knowledge Foundation, Sweden.

## REFERENCES

1. B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine,* vol.5, pp. 4-24, 1988.
2. R. L. Streit and A. H. Nuttall, "A general Chebyshev complex function approximation procedure and an application to beamforming," *Journal of the Acoustical Society of America,* vol. 72, pp. 181-190, 1982.
3. X. Chen, and T.W. Parks, "Design of FIR filters in the complex domain," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 35, pp. 144-153, 1987.
4. K. Preuss, "On the design of FIR filters by complex Chebyshev approximation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, pp. 702-712, 1989.

5. A. S. Alkhairy, K. G. Christian and J. S. Lim, "Design and characterization of optimal FIR filters with arbitrary phase," *IEEE Transactions on Signal Processing,* vol. 41, pp. 559-572, 1993.
6. D. Burnside and T. W. Parks, "Optimal design of FIR filters with the complex Chebyshev error criteria," *IEEE Transactions on Signal Processing,* vol. 43, pp. 605-616, 1995.
7. R. L. Streit and A. H. Nuttall, "A note on the Semi-Infinite Programming approach to complex approximation," *Mathematics of Computation,* vol. 40, pp. 599-605, 1983.
8. S. Nordebo, I. Claesson and S. Nordholm, "Weighted Chebyshev approximation for the design of broadband beamformers using quadratic programming," *IEEE Signal Processing Letters,* vol. 1, pp. 103-105, 1994.
9. E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces,* Wiley, New York, 1987.
10. M. Z. Komodromos, S. F. Russel and P. T. P. Tang, "Design of FIR filters with complex desired frequency response using a generalized Remez algorithm," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* vol. 42, pp. 274-278, 1995.
11. P. T. P. Tang, "A fast algorithm for linear complex Chebyshev approximations," *Mathematics of Computation,* vol. 51, pp. 721-739, 1988.
12. S. Nordebo and Z. Zang, "Semi-Infinite Linear Programming: A unified approach to digital filter design with time and frequency domain specifications," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing,* vol. 46, pp. 765-775, 1999.
13. H. Lebret and S. Boyd, "Antenna array pattern synthesis via convex optimization," *IEEE Transactions on Signal Processing,* vol. 45, pp. 526-532, 1997.
14. M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra and its Applications, Special Issue on Linear Algebra in Control, Signals and Image Processing,* 284:193-228, 1998.
15. D. Hertog, *Interior point approach to linear, quadratic and convex programming,* Kluwer Academic Publishers, 1994.
16. R. T. Rockafellar, *Convex analysis,* Princeton University Press, 1969.
17. T. W. Parks and C. S. Burrus, *Digital filter design,* John Wiley & Sons, Inc., 1987.
18. S. Nordebo, M. Dahl and I. Claesson, "Complex approximation with applications to antenna array pattern synthesis," in *Proc. of Electromagnetic Computations EMB01*, 2001.

Chapter 16

# LOW-COST CIRCULARLY POLARIZED RADIAL LINE SLOT ARRAY ANTENNA FOR IEEE 802.11 B/G WLAN APPLICATIONS

Serguei Zagriatski, and Marek E Bialkowski
*School of Inormation Technology and Electrical Engineering, The University of Queensland*

**Abstract:** Wireless Local Area Networks (WLAN) are new, fast growing telecommunication protocols. In order to attract the commercial market, antennas for WLAN applications have to feature low manufacturing cost and pleasant aesthetic appearance. Radial Line Slot Array (RLSA) antennas seem to provide such features. This article focuses on the design and development of Radial Line Slot Array antenna for WLAN applications in indoor environment. The presented theoretical results meet good return loss and specified radiation pattern requirements. A parameter study followed by an optimization procedure is presented. Discussion on level of influence of various parameters on return loss characteristic is described.

**Key words:** Radial Line Slot Array antenna, planar antennas for WLAN, access point antennas, antenna modelling

## 1. INTRODUCTION

Recent years have shown a growing demand for broadband communication services such as web-browsing, data transfer, voice communications and video streaming. Accessing these services in office buildings and homes is usually accomplished by forming wired connections to a communication network. An undesired effect of this action is an unnecessary decoration of buildings with many cables and access points. In particular, any changes involving repositioning of old or creating new access points require rewiring of the existing communication infrastructure, which

is usually time and labor consuming. These adverse attributes of wired connections are the main motivation for creating wireless access points.

There are several existing wireless networking standards which enable the wireless connectivity. These are shown in Table 16-1. The listed standards employ two Industrial, Scientific and Medical (ISM) 2.4GHz and 5.2 GHz frequency bands, which are dedicated for commercial, industrial and personal use. The key advantage of these frequency bands is that they are free of license requirement. This free-of-charge availability offers a significant reduction of deployment cost of the network, which is very attractive from the point of view of the network provider and the user.

*Table 16-1.* Summary of wireless network standards [1]

| Technology | Band | Signaling Rate/ Max Data Payload | Range | Modulation |
|---|---|---|---|---|
| 802.11 a | 5.2GHz | 54Mbps/32 Mbps | 20m | OFDM |
| 802.11 b | 2.4GHz | 11Mbps/5 Mbps | 50m | DSSS |
| 802.11 g | 2.4GHz | 54Mbps/32 Mbps | 50m | OFDM |
| Home RF | 2.4GHz | 10Mbps/5 Mbps | 50m | FHSS |
| Bluetooth | 2.4GHz | 1Mbps/750 kbps | 10m | FHSS |

Among the above quoted standards, the most established in the market is the IEEE 802.11b standard. Utilizing the low microwave frequency band of 2.4 GHz, it is capable of high penetration inside buildings and provides users with an adequate operational distance. Further boost of the 2.4GHz communication market is expected due to the introduction of IEEE 802.11g standard which also operates at the 2.4 GHz ISM band. This new standard offers the combined advantages of previously established standards IEEE 802.11b and g. They include high penetration rate because of the use of low microwave frequency band of 2.4 GHz (similarly as for IEEE 802.11b) and the high bit-rate transmission of 54Mbps, similar to that as offered by IEEE 802.11a. The multiple user access is accomplished through an Orthogonal Frequency Division Multiplexing (OFDM) transmission technology similar to that of IEEE 802.11a. Because of mixed mode operation capability, IEEE 802.11g offers the old 802.11b devices to operate at 11 Mbps and the new 802.11g devices to operate at 54Mbps over the same network. As the result, the IEEE 802b/g standard compatibility gives consumers perfect opportunity to an upgraded performance without having to be tethered to the 802.11b performance when in a mixed network.

## 1.1 Antennas for WLAN

The swift growth of WLAN standards and the primary focus on the radio transceiver design has led the antenna issues lagging behind the main stream of activities of WLAN communication system development. From the

technical point of view, the antenna operates as a transition between the transceiver and its surrounding environment. In order to attract a wide commercial market, it should be aesthetically pleasing (low profile) and of low cost. Usually two kinds of antennas are considered for use in WLAN. One is associated with a mobile user (or a peripheral device), while the other one concerns the network side (Access Point - AP). The first type of antenna needs to be small to minimize the overall size of the mobile or portable unit. A number of antennas capable to fulfill such requirement include printed planar monopoles, such as the Planar Inverted F antenna, and reduced-size microstrip patches usually involving a shorting pin as the size reduction mechanism.

The access point antennas do not introduce stringent limitations on the antenna size and its appearance. The primary performance objective for these antennas is to spread the signal to mobile or stationary devices within an enclosed space, e.g. a room. Perhaps due to these non-specific objectives, the area of designing antennas for WLAN AP has been largely overlooked. This is in contrast to the mobile units, for which antennas have been vigorously researched in recent years.

The simplest and most frequently utilized antenna for WLAN AP is a quarter-wavelength wire monopole positioned on a finite ground plane, as shown in Fig.16-1. This antenna, usually located at the ceiling of a room, exhibits an omni directional radiation pattern. Although very simple in construction, which is advantageous from the development point of view, it suffers from a considerable exposure to an accidental damage. This possible environmental hazard leads to necessity of employing a radome, which significantly increases profile of the antenna, and may be unacceptable in some applications. Another problem, which creates concerns among network users, is because of a possible intrusion into the network, as this antenna features a widely spread radiation pattern. This is unwelcome in applications requiring high level of security.

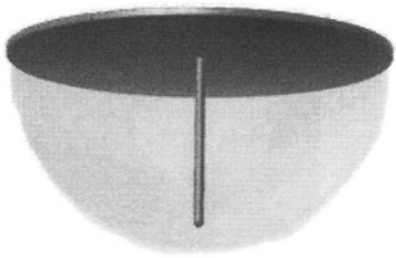


*Figure 16-1.* Quater-wavelength monopole

One alternative to a wire monopole is a microstrip patch antenna. It features low profile and offers a directional, conical radiation pattern when operating in a higher order $TM_{mn}$ mode [2]. In contrast to the monopole the patch antenna is robust against environmental mishaps, as the feeding point is located on its unexposed (rear) side. However it requires a suitable size ground plane to achieve proper operation. The reason is that the

electromagnetic energy is radiated through the perimeter of the patch, which should not be obstructed by the ceiling or wall structure. The size (diameter) of this antenna is close to one wavelength even without a ground plane.

An alternative to the wire monopole and the microstrip patch is a Radial Line Slot Array (RLSA) antenna [3]. Similarly to the patch operating in the higher order mode, this antenna is capable of producing a conical radiation pattern. Being of low profile, similarly as the patch, it can easily be hidden in the ceiling or in the wall, which is a very attractive property from the aesthetic point of view. With an appropriate slot distribution, the RLSA antenna can produce linearly or circularly polarized waves. The latter type of polarization is an important advantage of the antenna working in WLAN environment, where multi-path fading due to reflections from various scatterers can occur. Circular polarization is an effective way to tackle these problems. The reason is that during reflections the sense of circular polarization changes to the opposite type (for example from right-hand CP to left-hand CP) so that the waves undergoing single or an uneven number of reflections are rejected by a receiving antenna. In addition, the use of circular polarization permits freedom of user-end antenna orientation [4] when the transmitting and the receiving antennas are positioned horizontally. Another key factor for this antenna in terms of its competitiveness in a rapidly developing WLAN market is a fairly simple manufacturing process involving an inexpensive material, which leads to a low manufacturing cost.

Historically, the RLSA antenna has been well-known for its excellent performance at Ku-and Ka-bands in satellite communication applications [3]. Here we present the design of RLSA for 2.4GHz applications [5], which represents a lower end of the microwave spectrum.

## 1.2 RLSA antenna

The RLSA antenna was initially considered and investigated as a planar substitute to a well-known parabolic reflector. In the early sixties Goebels and Kelly [6] were first to demonstrate the use of a radial guide as a feeding structure of a moderate gain antenna. However an increased attention to this type of antenna was attracted only in the eighties when Goto et. al. proposed the RLSA antenna use for 12-GHz Direct Broadcast Satellite services in Japan [7-8]. Along with microstrip patch array (nicknamed a *squarial* antenna) the RLSA was considered as a main planar competitor to a parabolic reflector. In comparison with the microstrip patch array, the RLSA offers a higher level of radiation efficiency when the required antenna gain is in the order of 30dB or more [9]. For this value of gain the radiation efficiency of a microstrip patch array reduces to unacceptable level of 50% [8]. This is because of conduction losses in the feeding network, which

consists of a large length of a conducting transmission line from the input port to individual patches [10]. In the RLSA, a feeding network is accomplished using a low-loss radial cavity (filled with air or foam), which feeds slots in one of its circular shaped plates. Another advantage of RLSA over a microstrip patch array is the manufacturing cost, which is significantly lower.

Two different configurations of RLSA antenna have been proposed: double-layer and single-layer. The double-layer configuration feeding mechanism exploits a radial inward traveling wave in the parallel plate waveguide, whereas in the single layer a radial outward traveling wave is exploited. After overcoming some of its disadvantages [11], the single layer structure has been found as more attractive due to its simple design and manufacturing procedure. In the presented design procedure we consider the single-layer RLSA antenna variety.

A typical configuration of a single-layered RLSA antenna for DBS applications is shown in Fig. 16-2. It is formed by a cylindrical metal cavity, where top and bottom plates are separated by some fixed distance d (usually being smaller than a quarter-wavelength). The sidewall of the cavity is left either open or short-circuited.

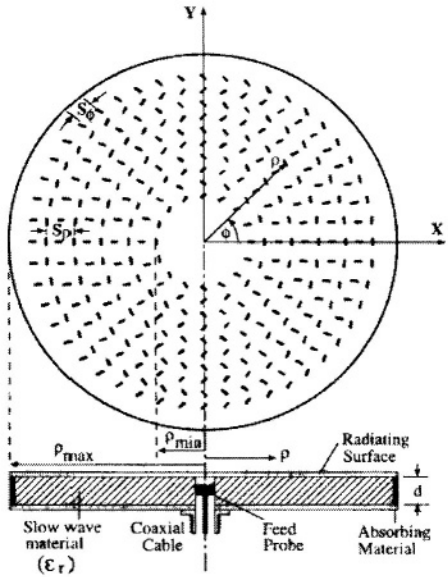



*Figure 16-2.* Typical structure of RLSA for Ku-, Ka-bands

In order to suppress grating lobes in the radiation pattern, the space inside the waveguide is filled with a low constant dielectric material (with relative dielectric constant being in the range of 1.5 to 2.0). The bottom plate includes a feed structure, which launches outwardly traveling wave in the

radial waveguide. The feed can be implemented using various coaxial to radial guide transitions. In order to provide radiation from the cavity into free space, the top plate bears number of slots through which the radial wave coupling and radiation occurs. By varying slots position and orientation it is possible to achieve different radiation patterns and polarizations.

## 2. FEED AND SLOT PATTERN DESIGN

The configuration of a circularly polarized RLSA antenna for WLAN applications is shown in Fig. 16-3. Here only one ring of slots is used to produce radiation pattern. The reasons for using only a single ring of slots are due to the requirement for small gain and to avoid an excessively large size of the antenna. Basic radiators are slot pairs arranged in a concentric ring. In practice, the slot pattern can be accomplished either using laser cutting in metal or by pressing method, if mass produced. The side walls are short circuited to avoid an undesired leakage of power. The bottom surface bears a disk-ended coaxial feed probe which launches an outward traveling radial TEM mode in the radial cavity. Due to the presence of the side wall a standing wave in the radial cavity is formed. This wave is coupled to slots, which in turn radiate power in free space. The first requirement for the feed and slot design is such that they should produce high return loss at the feed point.

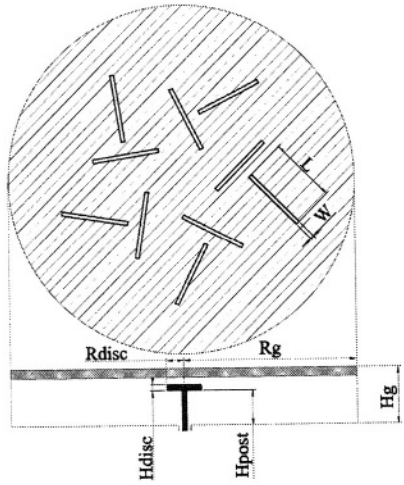



*Figure 16-3.* Top and side view of RLSA for WLAN

In order to accomplish the feed and slot design, a mixed theoretical approach is applied here. Initially, the feed is designed assuming that it

operates in an infinite parallel plate radial guide. This design and analysis stage neglects the presence of slots. In the next step, the slots size and position (with respect to the side shorting wall) are taken into account and the feed structure parameters are adjusted to produce good return loss with regard to the entire RLSA antenna structure. At this point, the slots orientation has to satisfy the requirement of high quality circular polarization. The analysis and design tasks are accomplished using a finite-element based electromagnetic field simulator High Frequency Structure Simulator (HFSS) of Ansoft. The flowchart in Fig. 16-4 shows the simulation procedure used in the design of RLSA for WLAN applications [12].

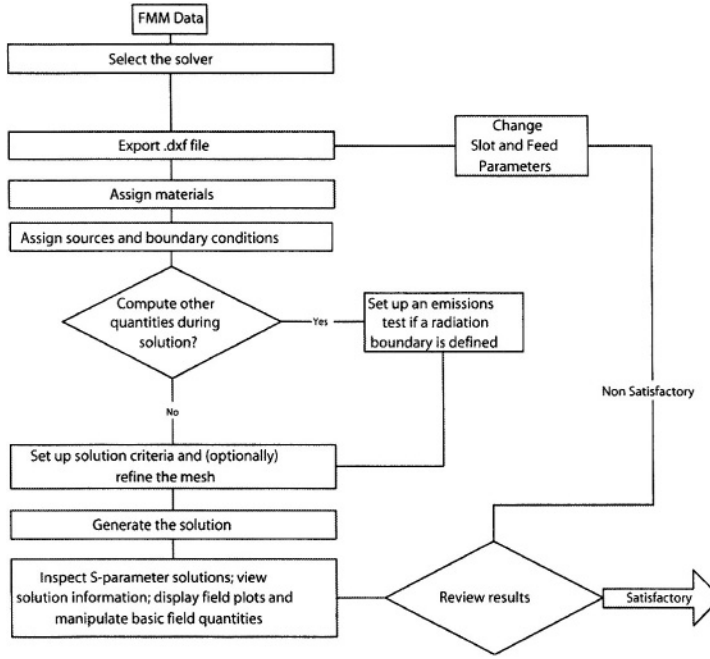



*Figure 16-4.* Simulation procedure adopted in this study

The design process is followed by the parameter study, which is also performed using HFSS. The investigated parameters concern the feed return loss, the shape and axial ratio of radiation pattern as a function of various parameters of the RLSA antenna. This study is important from the point of view of identifying the sensitive and non-sensitive antenna design parameters.

With regard to the RLSA antenna feed, a coaxially-fed, disk-ended probe is chosen. Its initial analysis and design is based on a Field Matching Method and resulting computer software described in [13,14,15]. In this

method, the structure of the feed in an infinite radial guide is divided into a number of cylindrical regions – I, II, and III (Fig. 16-5) – each of which may have a different relative permittivity: $\varepsilon_{rI}$, $\varepsilon_{rII}$, $\varepsilon_{rIII}$. The fields in each of regions I and II is expanded in terms of axially symmetric modes, with the only excitation source being the coaxial entry at the base of region I. Fields in region III are then expanded in terms of radial waveguide modes, and the requirement is to satisfy the boundary conditions at each of the region boundaries. The field expressions are formed by infinite series of modes, which in computations are truncated. Once a full set of simultaneous equation is formed, the field expansion coefficients are determined by applying a Galerkin procedure and standard algebraic equation solvers. Once the field coefficients are found, the driving point impedance and the associated return loss for the feed in an infinite radial guide is determined [3].

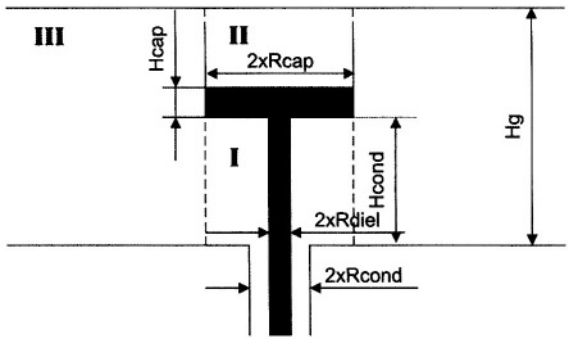



*Figure 16-5.* Disc-ended coaxial to waveguide transition

In the present design three parameters are varied: $R_{cap}$, $H_{cap}$, $H_{cond}$, while $\varepsilon_r$, $H_g$, $R_g$, $R_{cond}$, $R_{diel}$ are kept constant. Using 50Ω SMA receptacle as a base for the feed construction $R_{cond}$ and $R_{diel}$ are taken to be equal to 0.635mm and 2.03mm respectively, according to the standard data. The remaining parameters are chosen as follows: $\varepsilon_r$=2.7, $H_g$=14.9mm, $R_g$=125mm. Based on the infinite radial guide results the initial dimensions of the feed are obtained as: $R_{cap}$~7mm, $H_{cap}$~4mm, $H_{cond}$~8mm.

The next step in the design procedure is to analyze the entire RLSA antenna structure when the slots are present. This task is tackled with the help of HFSS. In order to achieve good return loss characteristics as well as to obtain circular polarization the following considerations are taken into account [16]. In order to satisfy circular polarization requirement, two slots have to be excited with the same amplitude and 90 degrees phase difference. This can be obtained if the slots are tilted 45 degrees with respect to radial direction (from the radial cavity centre towards to the wall) and separated a

quarter-wavelength apart along the radial direction. Using a pair slot radiator, the overall return loss due to a single slot is improved due to the fact that the reflected wave from one slot is cancelled by the reflected wave from the other when the circular polarization is used. As for the coupling, it can be minimized if the extension of one slot cuts the second at its center [8]. Details of slot pattern design are shown in Fig. 16-5. The lengths of the slots are chosen to be slightly smaller than half wavelength [4].

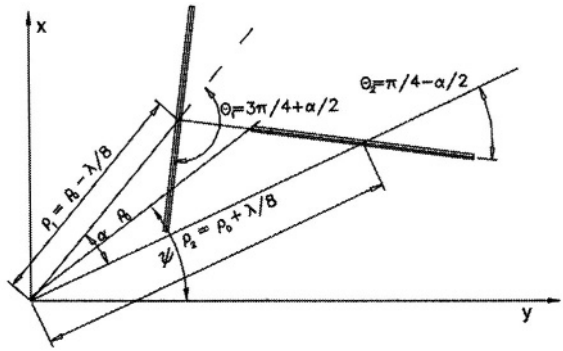



*Figure 16-5.* Slot pair geometry for theoretical design

The following formulas can be used to define position and tilt of each slot in the slot pair Eq. (16.1) [16]:

$$\rho_1 = \rho_0 - \frac{\lambda}{8} \qquad \psi_1 = \Psi + \frac{\alpha}{2} \qquad \theta_1 = 3\frac{\pi}{4} + \frac{\alpha}{2}$$
$$\rho_2 = \rho_0 + \frac{\lambda}{8} \qquad \psi_2 = \Psi - \frac{\alpha}{2} \qquad \theta_2 = \frac{\pi}{4} - \frac{\alpha}{2} \tag{16.1}$$

Table 16-2 shows the values of the antenna parameters, which satisfy conditions for achieving good performance in terms of Return Loss, Axial Ratio and Radiation Pattern.

*Table 16-2.* Final Dimensions of Antenna after Simulation

| | | | |
|---|---|---|---|
| Radial Waveguide Radius, $R_g$ | 121mm | Slot Width, W | 3mm |
| Radial Waveguide Height, $H_g$ | 15mm | Slot Height, L | 53mm |
| Dielectric Thickness | 2mm | Feed Cap Radius, $R_{cap}$ | 5mm |
| Dielectric Constant, $\varepsilon_r$ | 2.7 | Feed Cap Height, $H_{cap}$ | 2mm |
| Slot Tilt Angle, Θ | $45^0$ | Feed Conductor Height, $H_{cond}$ | 10mm |

## 3. RESULTS

Several antenna prototypes were manufactured in order to validate theoretical designs presented in the earlier section. The side view of one such antenna is shown in Fig. 16-7. The production of the antenna was divided into two parts. For the purpose of simplicity of testing process the antenna was formed by two parts. The first part included the bottom plate and the side wall. It was manufactured using an inexpensive metal spinning process. The coaxial feed was manufactured and attached to the bottom (base) part using standard SMA receptacle with a circular disk at the top of the tip. The second part was formed by the top conductive layer of antenna with slots. It was manufactured using a combination of metal spinning and laser cutting processes.

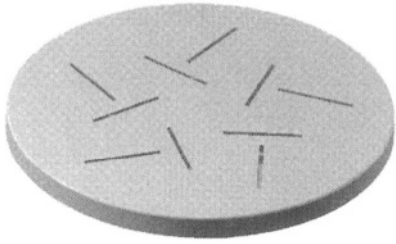


*Figure 16-7.* Manufactured CP RLSA

Fig. 16-8 shows the tolerance of return loss at 10dB level due to variations of the permittivity of the dielectric substrate from its nominal value by ±10%. The results shown in this figure reveal that the designer has an extra flexibility, which allows interchanging different layers with different parameters. In order to meet the specified tolerance levels, the dielectric layer was produced using a standard polycarbonate sheet with a relative dielectric constant $\varepsilon_r \sim 2.7$.

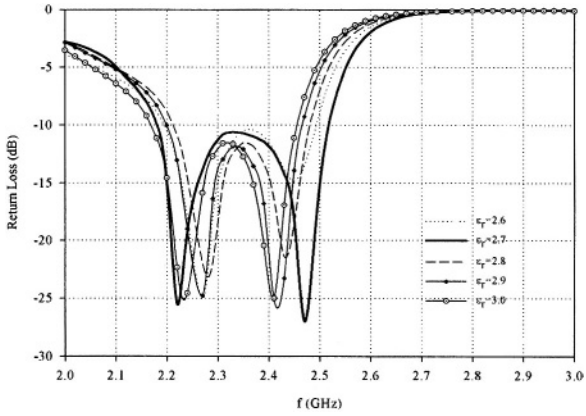



*Figure 16-8.* RL of RLSA with different dielectrics

Fig. 16-9 shows the comparison between theoretical and experimental dependences of return loss versus frequency for the antenna parameters presented in Table 16-2. In general, good agreement can be observed. It can be noticed that both simulation and measurement results indicate the 10dB return loss bandwidth of about 15%-17%. The discrepancy between the two graphs can be attributed to the mechanical mismatch and manufacturing tolerances of the top and bottom layers of the antenna as well as conduction losses which were not included in the HFSS simulations.

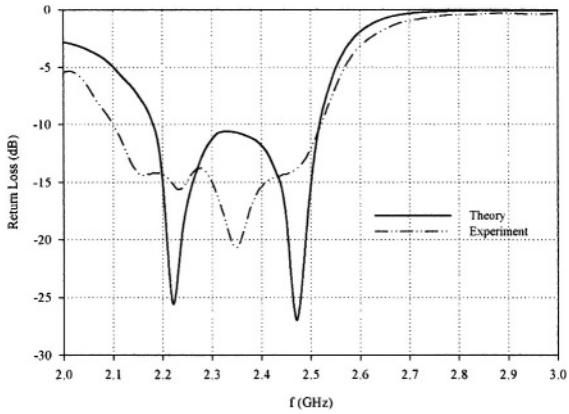



*Figure 16-9.* Theoretical vs. experimental results in terms of impedance match

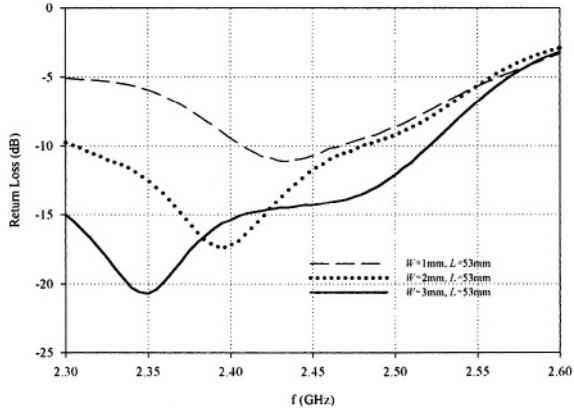



*Figure 16-10.* RL(f) for various slot widths

Next, the slot parameters were investigated in relationship to return loss. For this purpose several top layers of the antenna with different slot patterns

were produced while keeping the bottom layer the same. In Fig. 16-10, the return loss dependence over frequency is shown for the case when slot length L is kept constant at L=53mm and width W is changed from 1mm to 3mm. As can be seen from the graphs, increasing the slot's width W improves an overall return loss characteristic of the antenna.

Fig. 16-11 shows the return loss as a function of frequency when the slot's width W is kept fixed at 3mm and the slot's length is changed from 50mm to 53mm. Here, reduction of the slot's length leads to slight degradation of antenna performance in terms of return loss in the 2.4GHz WLAN frequency band.

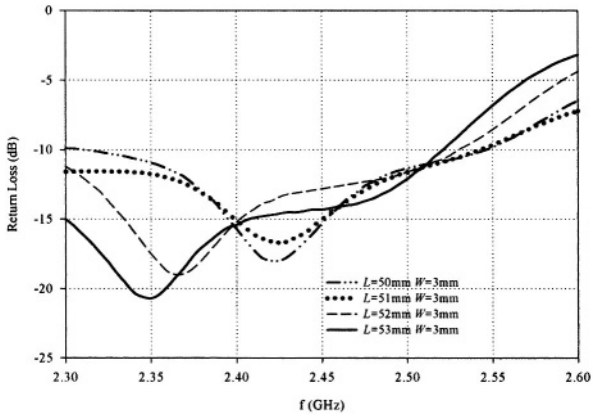



*Figure 16-11.* RL(f) for various slot heights

After verifying the antenna design with regard to impedance match, further attention was paid to radiation pattern shape and axial ratio characteristics (defining circular polarization performance). Fig. 16-12 represents two radiation patterns, as measured in two orthogonal planes normal to the top surface of antenna. It can be clearly seen that the antenna produces a radiation pattern with distinctive null in the broadside. Front-to-back ratio is 15dB and the main lobes are positioned at angle $\Psi \sim 30^0$ from the broadside. Further measurements revealed similar radiation patterns in any plane orthogonal to the slot surface, which means that antenna produced the required conical radiation pattern.

The results for axial ratio of the antenna as obtained at an angle of 30° from the broadside direction are shown in Fig. 16-13. Here the solid line represents the calculated values of axial ratio and the dashed line represents the measured values in the 2.4GHz WLAN frequency band. Both graphs well match each other. The two axial ratio curves are positioned below the 3dB level, which proves that the theoretical design and the manufactured antenna exhibit good quality circular polarization.

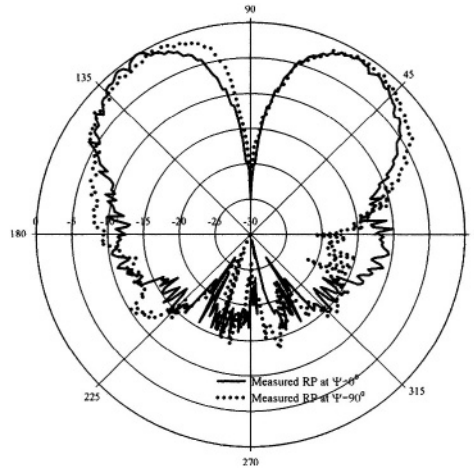



*Figure 16-12. Experimental radiation patterns in two orthogonal planes*

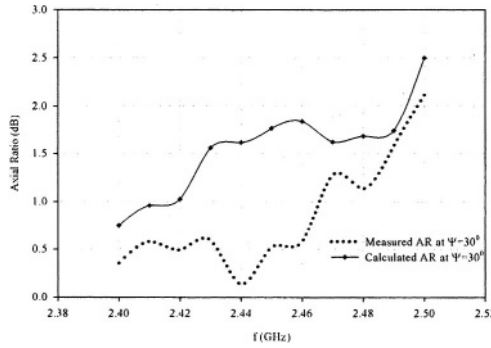



*Figure 16-13. Simulated and measured axial ratios at $\Psi=30^0$.*

## 4. CONCLUSIONS

There is a growing need for low cost and aesthetically pleasing Access Point antennas for rapidly developing 2.4GHz WLAN standards such as IEEE 802.11 b/g. Here we have proposed a Radial Line Slot Array (RLSA) antenna as a suitable candidate for such application. This antenna features a planar format with a pleasant aesthetic appearance, which is a welcoming attribute in the home and enterprise environments. In order to accomplish its design, in-house developed software and commercially available Ansoft HFSS have been used as design tools. A number of prototypes of this antenna have been manufactured and tested. All of them have shown good return loss performance and a conical radiation pattern with high quality

circular polarization across the entire 2.4GHz ISM frequency band. The obtained experimental results match well with the ones obtained in simulations. The proposed manufacturing technique is inexpensive and further significant reductions in cost can be expected if the antenna is mass produced.

## REFERENCES


1. P. Fowler, "5GHz Goes the Distance for Home Networking," *IEEE Microwave Magazine,* vol.3, no3 pp 49-55, September 2002.
2. K.L. Wong, F.S. Chang and T.W. Chiou, "Low-cost broadband circularly polarized probe-fed patch antenna for WLAN base station," *in Proc IEEE Antennas Propagat. Soc. Int. Symp. Dig.,* San-Antonio , TX, vol.2, pp. 526-529, 16-21 June 2002.
3. P. Davis and M. Bialkowski, "Experimental Investigations into a Linearly Polarized Radial Slot Antenna for DBS TV in Australia," *IEEE Trans. Antennas Propagat.,* vol. 45, pp. 1123-1129, July 1997.
4. K. Fujimoto, J.R. James, *Mobile Antenna Systems Handbook,* Norwood, MA: Artech House, 1994.
5. J. Takada et al., "Circularly Polarised Conical Beam Radial Line Slot Antenna", *Electronics Letters,* vol. 30, no. 21, pp. 1729-1703, Oct. 1994.
6. F. J. Goebels Jr. and K. C. Kelly, "Arbitrary Polarization From Annular Slot Planar Antennas," *IRE Trans. Antennas Propagat.,* vol. AP-9, pp. 342-349, July 1961.
7. N. Goto and M. Yamamoto, "Circularly Polaraized Radial Line Slot Antennas," *IECE Tech. Rep.,* vol. AP80-57, p. 43, Aug 1980 (in Japanese).
8. M. Ando, K. Sakurai, N. Goto, K. Arimura and Y. Ito, "A Radial Line Slot Antenna for 12 GHz Satellite TV Reception", *IEEE Trans. Antennas Propagat.,* vol. AP-33, pp 1347-1353, Dec. 1985.
9. E. Rammos, "New Wideband High-Gain Stripline Planar Array for 12 GHz Satellite TV," *Electron. Let.,* vol. 18, pp. 252-253, Mar. 1982.
10. M. E. Bialkowski, "Modelling of Planar Radial-Guide Antennas", *Proceedings of the International Conference on Microwaves, Radar and Wireless Communications,* Poland, Gdansk, May 20-22, 2002, pp. 205-217.
11. M. Takahashi, J. Takada, M. Ando and N. Goto, "A Slot Design For Uniform Aperture Field Distribution in Single-Layered Radial Line Slot Antennas," *IEEE Trans. Antennas Propagat.,* vol. 39, no. 7, pp 954-959, July 1991.
12. Ansoft® HFS™ Version 8.5 Full Manual
13. M. E. Bialkowski, "Analysis of a Coaxial-To-Waveguide Adaptor Including Disc-Ended Prove and a Tuning Post," *IEEE Trans. Microwave Theory Tech.,* vol. 43, pp. 344-349, Feb. 1995.
14. M.E. Bialkowski, "Analysis of Disc-Type Resonator Mount in Parallel Plate and Rectangular Waveguides," *Archiv fur Electronik und Ubertragungstechnik, AEÜ,* vol. 38, no. 5, pp. 306-311, 1984.
15. M. E. Bialkowski and S. T. Jellett, "Investigations into a Coaxial-To-Waveguide Transition Incorporating a Disc-Ended Probe," *1993 Asia-Pacific Microwave Conference Proceedings,* vol. 2, pp.13-16, Oct. 1993.
16. M. S.-Castaner, "Low-Cost Monopulse Radial Line Slot Antenna," *IEEE Trans Antennas Propagat.,* vol. 51, no. 2, pp. 256-263, February 2003.

Chapter 17

# SOFTWARE CONTROLLED GENERATOR FOR ELECTROMAGNETIC COMPATIBILITY EVALUATION

Piotr Gajewski, Jerzy Lopatka
*Telecommunications Institute, Military University of Technology, Warsaw, Poland, pgajewski@wel.wat.edu.pl, jlopatka@wel.wat.edu.pl*

**Abstract:** This paper presents a concept of a testbed designed for laboratory and field tests of wireless communication equipment and systems. The paper presents also a proposal of its use for evaluation of Quality of Service for defined particular scenarios. One of the testbed elements is a software defined arbitrary generator, that enables generation of both predefined signals and interference.

**Key words:** arbitrary signal generator, testbed, EMC

## 1. INTRODUCTION

Electromagnetic compatibility issues are significant problems for civilian, government and military wireless systems. This is caused by increasing number of various systems and continuously higher demands for bandwidth and data rates. The limited amount of available bandwidths causes a need for co-existence of different systems, assuming controlled degradation of each system performance by interference generated by another users of the system or another systems.

The use of COTS technology is an effective way to improve efficiency of military communications and information systems (CIS) development. In particular, some commercial wireless technologies have been recently tested for military applications. Many vendors offer military versions of such widely known wireless systems like GSM, TETRA, HIPERLAN (wireless

LAN) etc. The newest and perspective commercial technologies for wireless access should be also examined for military applications. This refers also to tactical systems, including peer-to-peer combat network radio (CNR) as well as radio access to the mobile tactical network by radio access points (RAPs). The necessity of coexistence of many radio networks based on different modulation methods (e.g. HF, DSSS) and different multiple access schemes (e.g. TDMA, CDMA) within limited area of military forces deployment creates specific demands for fulfilling electromagnetic compatibility criteria.

Another problem is coexistence of both civilian and military systems in the same bandwidth. In this case a precise legal solutions are required to describe rules of operations. In these cases law must decide which systems has priority, e.g. civilian landing system is more important than military data transmission system. Because both systems cannot work in the same time and in the same place, the less important system must be able to monitor if the superior one is working and cease its own operation for some period of time.

The very important problem for military wireless systems is their immunity against accidental interference caused by rapid changes of systems configuration and intentional jamming generated by enemy. The survivability of the system must be verified on all communications layers, and system structure and management procedures must be optimised to provide maximum quality of service (QoS). Generally, bit-error-rate (BER), signal-to-noise-ratio (SNR) or signal-to-interference-ratio (SIR) are the most frequently used measures to define QoS.

The radio resources allocation (RRA) is one of the main questions of wireless systems planning and management. The mobility of user, specific conditions of electromagnetic waves propagation, and heterogeneous phenomena of multimedia traffic create significantly difficult environment for developing effective radio resources allocation procedures, especially channel assignment, power control as well as data traffic control. The RRA algorithms should maximize the number of satisfied users within the available radio bandwidth. A user is satisfied if its session quality is below the acceptable level for an insignificant amount of time.

All these problems are essential for system users. Some of them can be solved using law regulations, some by proper system configuration during the design stage and some by analytic simulations. The most accurate information can be collected during system operation, but introduction of modifications at this stage is usually the most expensive. It seems that the most effective way to achieve, reliable and recurrent results are extensive laboratory tests, verified during field tests before system implementation [1].

## 2. TESTBED FOR ASSESSING QOS AND RRA METHODS

Future wireless personal communication systems should provide mobile users with a broad range of multimedia services including voice, data, and video with guaranteed quality of service (QoS).

The most often used QoS measures are the blocking probability and outage probability caused by significant increase of interference level in the service area. We can denote outage probability in interference environment as [2]:

$$P_{out}(t) = \Pr[BER > BER_{th}],$$

where $BER_{th}$ is the maximal acceptable value of BER for particular service, modulation scheme, access method, etc. Current value of BER is a random variable that is dependent on many factors like temporary SIR and SNR, number of users, services, etc.

Maximal utilization of system's radio resources can be achieved with RRA being an optimisation process. The optimisation should be realized both during system planning and QoS management when system operates. For this reason, we use some system data, which include system characteristics especially system disposed resources as well as traffic parameters. For example, analysis of CDMA systems shows that transmitters' powers, transmission rates, correlation characteristics of the codes, number of users and theirs activities, are the most important factors limiting the system capacity. A fast and accurate power control can compensate for a too large decrease in signal-to-noise ratio at the receiver input and keep it above the threshold level. However, in bad-link conditions, transmitter level can become too high, causing increase of interference in other ongoing calls. To limit such a phenomenon, two alternatives are proposed for CDMA systems in [3]:

- power and rate adaptation; in the case of bad conditions the transmission power is limited to $P_{max}$ while transmission rate is reduced to meet needed QoS,
- truncated rate adaptation; in bad conditions the data rate is adapted with a fixed transmission power $S < S_{min}$ or transmission is suspended, causing transmission delay.

To assign channels (codes or/and time slots), the channels segregation algorithm is proposed. The channel segregation algorithm reduces the processing power demand as well as avoids the particular planning of radio network. For codes segregation, the Self Organizing Feature Map algorithm,

based on neural network method has been developed and tested [2]. In this self-adaptive learning method, the codes are assigned in the order of the so-called priority list of codes. This list is made and updated based on the relationships between signal level and interference level estimated according to the spatial distribution of users. Here, the BER values are also used as a criterion for optimisation.

As the first step of system verification, the simplified analytical models are used to assess system performance by means of the specialized software tools. The field tests are the next step of technical solutions validation before the system implementation. To achieve the acceptable tests accuracy, an adequate laboratory testbed should be deployed.

The laboratory tests should be precisely organized and the following must be established:

- typical scenarios of operation,
- measured parameters and quality measures of the system,
- minimal requirements for each result,
- methodology of conducting each test.

Besides the typical situation tests, the testbed should enable examination of system performance in abnormal situations. These can be caused by very heavy traffic, rapid changes of system configuration, malfunctions of equipment, intentional jamming etc. [4].

## 3. TESTBED CONFIGURATION

Fig. 17-1 presents architecture of the testbed, used for wireless systems tests. It enables connection of devices under test (DUT) to the RF test unit that contains passive, star-shape network with 10 identical passive bi-directional branches, where each branch consists of a set of attenuators, and the fixed attenuators attenuate output power of DUTs. Their values are preset before the test, according to the assumed scenario. Variable attenuators are used for simulation of subscribers' movements, slow fading, changes of the system configuration, and the equipment malfunctions. Subscriber's simulators are PCs that are sources and sinks of sound and data. The software uses traffic models according to the defined scenario. The main server is a PC based server for a remote control of all units.

The testbed contains also a traffic simulator, modular software receiver (MSR) and arbitrary generator (ARB). They act as a source of interference. Traffic simulator can control the ARB and emulate the background RF traffic imitating other subscribers of the same system, subscribers of other systems and other sources of interference. The MSR [5] is a software defined receiver that acts as a radio monitoring station which can control the

arbitrary generator to imitate intentional jamming systems. The testbed can be also equipped with RF channel simulator that can simulate influence of real propagation circumstances on the system performance. It is also possible to establish a connection to other systems, i.e. ISDN wired system through radio access point. Measurements of quality of service between wireless system subscriber and ISDN subscriber are also possible.

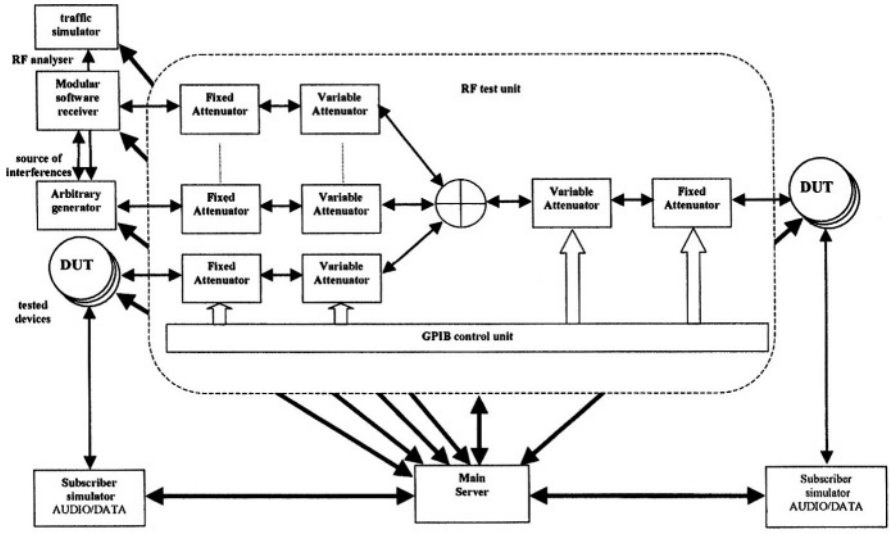



*Figure 17-1.* Block diagram of the testbed

## 4. ARBITRARY GENERATOR

The ARB can generate up to 5 different signals in the same time. It consists of three main parts: PCI board, generating baseband IQ signals, IQ generator and dedicated PC software. The ARB board contains 7 functional blocks (Fig. 17-2):

- DSP based real time IQ modulator for narrowband modulations;
- SRAM block for narrowband waveforms generation;
- DRAM block for wideband waveforms generation;
- digital up-converter (DUC) FIR signal filtration and interpolation,
- two DACs generating IQ signals and working with 100 MHz sampling frequency;
- external interfaces;
- control logic.

The DSP based modulator can perform a real-time narrowband modulation. Its input can accept both analog and digital signals in the acoustics band. The DSP is working in 34 kHz interrupt mode and enables generation of signals with bandwidth of up to 25 kHz. The DSP performs AM, FM, SSB, FSK, PSK and QAM modulations with software controlled modulation type and parameters. The modulator is realized on the basis of Texas Instruments TMS 320c50 processor.

The SRAM block consists of 2 independent modules, with capacity of 4 MB each. Hence the DUG accepts IQ pairs of 16 bit samples it gives 1M of IQ pairs. The modules can work alternatively in generation and load modes. RAM can be loaded from PC or from external interface e.g. DSP system.

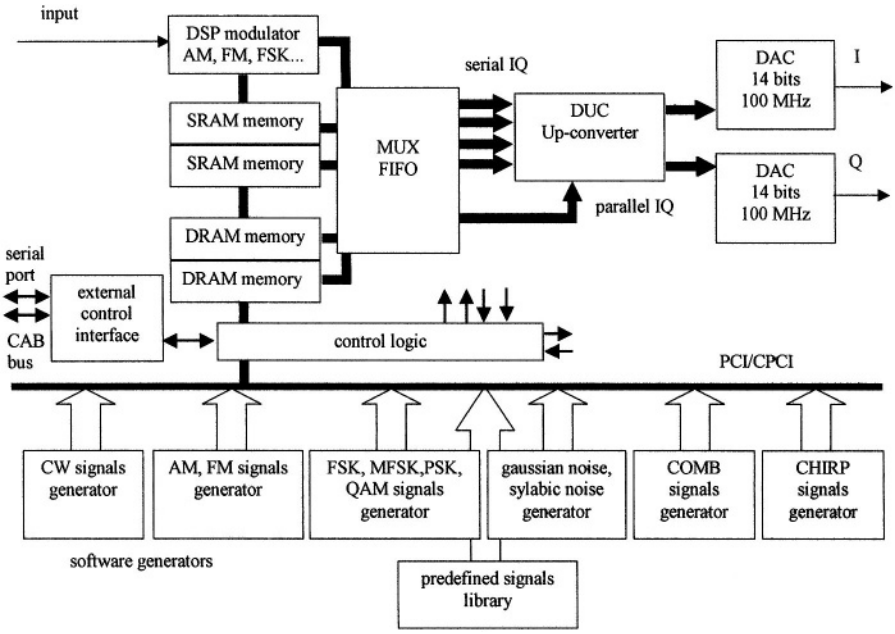



*Figure 17-2.* Block diagram of the generator

Memory space can be divided into parts, to store different waveforms. Specified sections of the waveforms can be played back once, in a sequence, or in a loop. The length of each waveform and DUG channel number, used for signal generation, is also programmed. Data rate of the waveform generation is variable and depends on the interpolation factor programmed in the DUC. During loading of one SRAM module, the other module can generate waveforms. Switching between modules is seamless for the DUC, so it is possible to change the waveform “on the run”.

DRAM block has a capacity of 2x128MB. It operates in a similar way to SRAM block, but because of the larger capacity, it can generate longer sequences that are necessary for signals played back with higher data rates. Data from DRAM can be also directly transmitted to the parallel input of DUC, with the 100MHz clock that enables generation of up to 70 MHz wide signal and its superimposition on up to 4 channels generated by DUC.

Digital up-converter (DUC), see Fig. 17-3., is a quad, multi-standard transmit chip GC 4116 from Gray Chip [6]. It contains four identical up-conversion channels. The input signal is filtered by 63 tap programmable FIR filter. It can be used to shape spectrum of the transmitted signal to meet particular standard requirements or to be used as a Nyquist filter, or for compensation of spectral distortion introduced by analogue filters, the IQ generator, and the amplifier or coupling devices.

The filtered signal can be interpolated in the FIR and CIC filters by a factor of 32-5792 that gives the minimum input sampling frequency about 17 kHz for the 100 MHz output frequency. The maximum input sampling frequency is about 3.1 MHz, and enables generation of signals with bandwidth up to 2,5 MHz. By combining two channels together it is also possible do increase the bandwidth twice. In IQ mode, the interpolation factor is divided by 2.

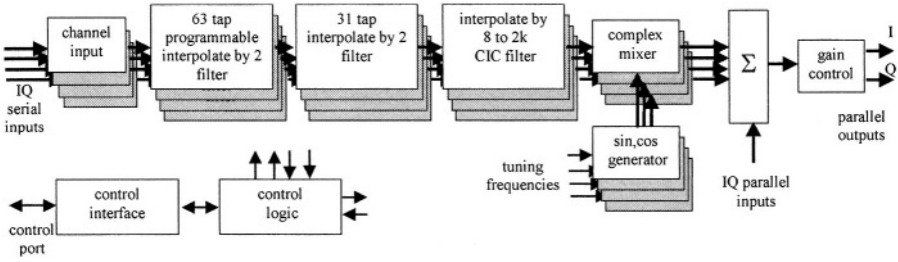



*Figure 17-3.* Digital up-converter block diagram

Interpolated signals are modulated, with resolution 0,02Hz, by programmable sine/cosine NCO, and the output signal can be a real or complex one. Output signals from four channels, can be added together with the programmed levels. The combined signal can be also added to a signal from the parallel 22 bit-wide DUC input. It enables introduction of a larger number of channels, but also make possible combining of the signal with wideband signals transmitted directly from the DRAM block.

The up-converter provides over 116 dB of spur free dynamic range that exceeds requirements of most of the standards.

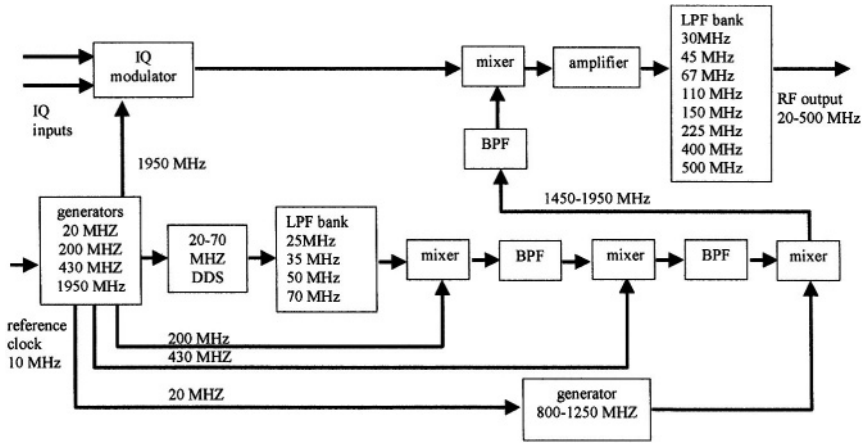



*Figure 17-4.* Frequency shifter block diagram

Applied digital-to-analog converters AD9772A from Analog Devices [7] are oversampling, 14-bit chips, optimised for baseband or IF waveform reconstruction applications requiring a high dynamic range. They integrate a complete, low distortion 14-bit DAC with a 2× digital interpolation filter and clock multiplier. The interpolation filter provides a low-pass response, useful for IQ applications, providing reduction in the complexity of the analogue reconstruction filter by suppressing the original upper in-band image by more than 73 dB. For direct IF applications, the filter response can be reconfigured to select the upper in-band image while suppressing the original baseband image.

The ARB also contains external interfaces to perform control in real-time:

- Coreco's Auxiliary BUS (CAB) – 32-bit bus working at 50 MHz [8]; it is a bus for communications between DSP board and ARB board. It enables remote programming and control of ARB, but also can be used for direct transmission of waveforms from the DSP board to DUC and signal generation "on line".
- Serial bus, the 50 MHz dedicated link between TI DSP-s and other boards; it can be used for ARB programming and control.
- PCI or CPCI bus; mainly used for board programming purposes and waveform download. CPCI version is dedicated for ruggedized solutions.

ARB contains also a set of software generators. They can generate both narrowband and wideband signals like CW, AM, FM, FSK, MFSK, PSK, QAM, OFDM etc. with selected parameters. For the interference generation, ARB also contains generators of Gaussian and syllabic noise, as well as COMB and CHIRP signals. All types of signals can be mixed together to create more complex signals. The generated signals are stored as files in a

predefined signals library on the HDD. Elaborated software allows also the use of recorded real signals and those artificially generated using Matlab, Excel or other external software.

Any IQ generator can be used as the frequency shifter. For tests of HF, VHF equipment, we proposed a fast synthesizer, working in the range of 20-500 MHz (see Fig. 17-4.). According to a stored sequence, it can perform up to 300 hops per second that can be used for imitation of frequency hoping radios.

## 5. CONCLUSIONS

The designed testbed is a flexible platform for multilevel analysis and tests of wireless equipment and systems reception of radio signals. One of its elements is an arbitrary generator that can generate a variety of interfering signals. It can work according to the predefined scenario, but also enables adaptive changes of the generated signals that imitate presence of the intentional jamming. The solution can be also used in automated radio-monitoring systems, EW applications etc.

## REFERENCES

1. Gajewski P., Lopatka J.: *New Wireless Technologies Challenges for Military CIS Development,* 4th NATO Symposium with Partners "Making C3 Policies a Reality", Brussels, 2002.
2. Gajewski P., Ziolkowski C.: *Resource Management for 3g and Beyond,* Workshop in XIV International Conference on Microwaves, Radars and Wireless Communications MICON-2002, Gdansk,2002.
3. Dixit S. et al: *Resource Management and Quality of Service in Third-Generation Wireless Networks,* IEEE Communications Magazine, February 2001, Vol. 39, No 2, pp.125-133.
4. Lopatka J.: *Modular Software Receiver for Radio Signal Analysis,* 6th International Symposium on Digital Signal Processing for Communication, Sydney 2001.
5. Mitola J. III: *Software Radio Architecture ,* Wiley-Interscience, 2000.
6. Texas Instruments: *GC4116– Multistandard quad DUC chip,* June 2002.
7. Analog Devices: *14-Bit, 160 MSPS TxDAC+ with 2. Interpolation Filter,* 2003.
8. Coreco Inc. :PYTHON/C6 – datasheet, 1998.

*This page intentionally left blank*

Chapter 18

# UNIFIED RETIMING OPERATIONS ON MULTIDIMENSIONAL MULTI-RATE DIGITAL SIGNAL PROCESSING SYSTEMS

Dongming Peng, [1] Hamid Sharif, [1] and Song Ci,[2]

[1]*Department of CEEN, University of Nebraska-Lincoln, Omaha, NE 68182, USA*

[2]*Department of CSESP, University of Michigan-Flint, Flint, MI 48502, USA*

**Abstract** The intense requirements of high-speed implementations of MultiDimensional (MD) Digital Signal Processing (DSP) systems justify the Application Specific Integrated Circuits (ASIC) designs and/or multiprocessor implementations. MD retiming has been recently proposed to improve the circuitry performance in high-level synthesis of single-rate MD DSP systems. This paper has conducted new theoretical analysis of MD multirate DSP systems modeled in data-flow graphs, and proposes intercalation of MD multirate systems so that unified MD retiming operations can be applied on multidimensional multirate DSP systems. By retiming and intercalation, full intra-iteration parallelism is achieved and functional elements can be executed simultaneously on circuits for the generic class of MD multirate DSP systems.

**Keywords:** Multirate Signal Processing, MD Data Flow Graph, Multidimenional Retiming.

## 1. INTRODUCTION

The intense requirements of high-speed implementations of MultiDimensional (MD) Digital Signal Processing (DSP) systems in practical applications justify the Application Specific Integrated Circuits (ASIC) designs and parallel processing implementations. These MD DSP applications include multimedia processing, computer vision, high-definition

television, medical imaging, remote sensing, and computation tasks in fluid dynamics. Due to the features of hierarchical signal analysis and multiresolution analysis, many of these applications are *multirate* in nature[2], meaning that the sample rates are not constant throughout the algorithm description.

There are many famous multirate MD DSP applications including Discrete Wavelet Transform [2, 20], Full Wavelet Transform [20], Multi-Wavelet Transform [1], M-ary Wavelet Transform [9, 20], Wavelet Packet Transform [10], Embedded Zerotree Wavelet coding [17], Set Partitioning In Hierarchical Trees [16], Spatial-Frequency Quantization [23, 24], and etc. The theory of multirate DSP systems has matured over the past decade [2, 20], but there have been only a few research papers, e.g., [7, 8, 25, 3, 5] reported in literature on a generalized theory for high level synthesis of multirate MD systems.

As one of important theories in high level synthesis of single-rate MD DSP algorithms, the recently proposed *MD* retiming [4, 13–15] improves the circuitry performance by guaranteeing that all functional elements can be executed simultaneously on circuits.

Most researches on retiming operations are focused on single-rate DSP algorithms only. There have been several papers ([5, 7, 8, 25, 26]) published in literature with research on applying retiming operations to multirate systems. For example, [26] has explored the use of retiming for attempting to reduce the execution time of one-dimensional Synchronous Data Flow Graphs (SDFG). It develops the basic definitions and results necessary to express and study SDFGs and review the problems faced when attempting to retime an SDFG in order to minimize clock period, and then present algorithms for doing this. [8] has reported valuable research on retiming graph construction to reduce interface register cost for building blocks based on 1-D multirate dataflow graphs, [5] has derived retiming for folding constraints which indicate how a multirate dataflow graph must be retimed for a given schedule to be feasible, and [25] has provided a new valid retiming of multirate graphs using the non-ordinary marked graph model and the reachability theory.

However, there are still many problems open regarding the application of retiming operations onto multirate DSP systems. For example, what are the necessary and sufficient conditions for the applicability of the retiming on the multirate dataflow graph [26]? Is the technique of retiming operation applicable to an ARBITRARY multirate DSP dataflow graph? Is there a unified methodology for retiming operations on both single-rate and multirate systems? The last problem is addressed in this paper based on a comprehensive modeling and analysis of multirate MD dataflow graphs. We have constructed in the paper a complete theo-

retical analysis of modeling MD multirate DSP algorithms in dataflow graphs. Based on the analysis, the paper proposes a novel technique of MD intercalation in an iteration space that is used for retiming.

## 2. RESEARCH BACKGROUND

An MD Data Flow Graph (MDFG) [4, 13–15], G=(V, E, d, t), is a node-weighted and edge-weighted directed graph modeling an MD DSP algorithm. V is the set of computation nodes. $E \subset V \times V$ is the set of edges representing the data flows and dependencies between nodes, d is the set of delay-weights (n-component vectors) on E (each edge is associated with an n-component vector as its delay-weight) and represents the MD delays of data flowing between two nodes, with $n$ being the number of dimensions of the algorithm, t is the set of computation times (in clock cycles) for the computation nodes.

An *iteration* is the execution of a loop body exactly once, i.e., executing the task corresponding to each node in V exactly once. By replicating an MDFG at multi-dimensionally indexed positions, we expand an *iteration space,* where each MDFG, excluding the edges with delay vectors different from (0, 0, ⋯, 0), is taken as a *cell* indexed by Cartesian coordinates. Those non-zero delay weighted edges within the MDFG give specifications of the dependencies between these cells in the iteration space. Due to the causality, a legal MDFG must have no zero-delay cycle, i.e., the summation of the delay vectors along any cycle path in the MDFG can not be (0, 0, ⋯, 0).

A multi-dimensional retiming operation on a node $u \in V$ redistributes the nodes in the cells in iteration space. The retiming vector r(u) of a node $u \in V$ represents the offset vector between the original cell containing u, and the one after retiming. To preserve dependencies in iteration space after retiming operations, delay-weights of edges change accordingly. Formally, for edge $e : u \rightarrow v$, we have the *retiming equation* [4]

$$d_r(e) = d(e) + r(u) - r(v) \tag{18.1}$$

where d(e) or $d_r$(e) is the delay-weight of edge e after or before retiming respectively. After retiming, an instance of node u in the cell indexed by **i** in iteration space is moved to cell **i-r(u)**.

Obtaining full inter-operation parallelism is equivalent to obtaining non-zero delays on all edges of the MDFG by retiming techniques such that the computation tasks corresponding to all nodes in the retimed MDFG can be executed simultaneously. The delay-weights on the edges outgoing from a computation node in the MDFG corresponds to the necessary storage of the data outgoing from this computation node to its following computation nodes. Therefore, another key purpose of the

retiming technique is to minimize the system storage requirement by modifying the delay-weights on edges in the MDFG.

## 3. THEORETICAL ANALYSIS OF MODELING MR MD DSP ALGORITHMS

Multirate DSP algorithm descriptions contain decimators and/or expanders to represent multirate data flows [5, 12, 18, 19]. Those decimators and expanders can be modeled by *multirate-weighted* $M(e)$ on edges. Consider an edge e: $u \rightarrow v$ connecting two computation nodes u and v in the MR-MDFG. The dependence relationship between data streams at both sides of edge e in the MR-MDFG is represented by:

$$P(m) \leftarrow Q(M(e) \times m - d(e)) \tag{18.2}$$

where $P$ is the data stream flowing out from node v and $Q$ is the data stream flowing out from node u. Applying traditional retiming vectors r(U) and r(V) on the nodes of U and V, we have the new offset-weight of e: U $\rightarrow$ V as

$$d_r(e) = d(e) + M(e)r(U) - r(v) \tag{18.3}$$

*Definition 3.1:* In an MR-MDFG, the *rate of a path P,* Z(p), is defined as Z(p)=$\prod_i M(e_i)$, where $\{e_i\}$ represents all edges along path $p$. *Definition 3.2:* The rate of an edge $e : u \rightarrow v$, $R(e)$, is defined as $R(e) = MAX[Z(p)]$ where the maximum is taken over all paths from the sources of the MR-MDFG to $v$ and going through edge e. *Definition 3.3:* We define an MR-MDFG as a *rate-balanced* MR-MDFG if, for any internal computation node in the MR-MDFG, the rates of this node's two input edges are equal. Otherwise, the MR-MDFG is defined as *rate-conflict. Definition 3.4: The delay of a path p,* $d(p)$, is defined as $\sum_{i=1}^{K}[(\prod_{j=0}^{i-1} M(e_j))d(e_i)]$, where $M(e_0)$ is assumed to be 1, and $e_1, e_2, \cdots$, and $e_K$ are all K edges along p from the beginning to the end.

*Lemma 3.1* If and only if an MR-MDFG is rate-balanced, for any edge in the MR-MDFG, e.g., e: $u \rightarrow v$, we have $R(e) = Z(p)$, where $p$ is ANY path from a source of the MR-MDFG to node v and goes through edge e, i.e., all $Z(p)$ are equal for ANY path $p$ that starts from a source of the MR-MDFG to $v$ and goes through e.

*Lemma 3.2:* Such relation always exists for the retiming operations along any path p in the multirate-MDFG: $\Delta d(p) = r(u) - R(p)r(v)$, where u and v are respectively the starting node and the destination node of path p.

*Theorem 3.1: 1) All rate-balanced MR-MDFGs are stable for retiming; 2) If in a branch-type rate-conflict MR-MDFG there exists a rate-conflict*

*edge whose primary paths have more than one common node, the MR-MDFG is not stable for retiming; 3) None of the cycle-type rate-conflict MR-MDFG is stable for retiming.*

*proof:*

1) Suppose u and v are any two nodes in a rate-balanced MR-MDFG. Assume two different paths $P_A$ and $P_B$ from u to v. Let $P_0$ be the path from one of the sources of the MR-MDFG to u. Since $P_A$ and $P_B$ have the same destination node (i.e., v), we can find a node $U_0$ such that $P_A$ and $P_B$ respectively go through edges $e_A$ and $e_B$ (the two input edges of node $U_0$), and path $U_0 \rightarrow v$ (called $P_C$) is the common part of $P_A$ and $P_B$. Thus we have $P_A = P_{A1} + P_C$ and $P_B = P_{B1} + P_C$ where paths $P_{A1}$ and $P_{B1}$ are respectively another part of $P_A$ and $P_B$. Finally, we conclude that $R(e_A) = R(e_B)$ (per *Definition 3.3)* $\Longrightarrow Z(P_0+P_{A1}) = Z(P_0+P_{B1})$ (per *Lemma 3.1)* $\Longrightarrow Z(P_0) \times Z(P_{A1}) = Z(P_0) \times Z(P_{B1})$ (per *Definition 3.1)* $\Longrightarrow Z(P_{A1}) = Z(P_{B1}) \Longrightarrow Z(P_{A1}) \times Z(P_C) = Z(P_{B1}) \times Z(P_C)$ $\Longrightarrow Z(P_A) = Z(P_B)$. This leads to the conclusion that all rate-balanced MR-MDFGs are stable for retiming if *Lemma 3.2* is considered.

2) In a branch-type rate-conflict MR-MDFG where there exists a rate-conflict edge e: $v \rightarrow v_0$ whose primary paths $P_A$ and $P_B$ have more than one common nodes, we can assume a node u (which is different from v) as one of the common nodes of $P_A$ and $P_B$. $P_A$ goes through $e_A$ and $P_B$ goes through $e_B$, where $e_A$ and $e_B$ are the two input edges of node v. If u is a source of the MR-MDFG, based on *Lemma 3.2,* we simply draw the conclusion that the MR-MDFG is not stable for retiming because $Z(P_A) \neq Z(P_B)$. If u is not a source of the MR-MDFG, we suppose that $P_{A1} : u \rightarrow v$ and $P_{B1} : u \rightarrow v$ are respectively a part of $P_A$ and $P_B$, and $P_{A2}$ and $P_{B2}$ are respectively another part of $P_A$ and $P_B$. Since $R(e_A) = Z(P_A)$, we have $Z(P_A) \geq Z(P_{A1+B2})$ based on *Definition 3.2,* where $P_{A1}$ is the first part and $P_{B2}$ is the second part of path $P_{A1+B2}$. Thus we have $Z(P_{A1}) \times Z(P_{A2}) \geq Z(P_{A1}) \times Z(P_{B2})$ $\Longrightarrow Z(P_{A2}) \geq Z(P_{B2})$. Similarly considering edge $e_B$, we also have $Z(P_{B2}) \geq Z(P_{A2})$. So $Z(P_{A2}) = Z(P_{B2})$, implying $Z(P_{A1}) \neq Z(P_{B1})$. Thus the MR-MDFG is not stable for retiming based on *Lemma 3.2.*

3) Suppose there is a rate-conflict edge e: $u \rightarrow v$ whose primary paths are $P_A$ and $P_B$ in a cycle-type rate-conflict MR-MDFG, and $P_A$ goes through e. Let the starting node of $P_B$ be s, a source of the MR-MDFG. Because $P_A$ is different from $P_B$, we have two paths from s to u: $P_B$ and $(P_A + P_B)$. Apparently $Z(P_B) \neq Z(P_B + P_A)$. Thus the MR-MDFG is not stable for retiming based on *Lemma 3.2. End of Proof.*

*Theorem 3.2:* If and only if an MR-MDFG is rate-balanced, such property exists: when we cut off all multirate-weighted edges (i.e., those edges whose multirate-weights are not unit matrices), and obtain sub-

graphs each of which is a single-rate MDFG, all those multirate-weighted edges in the original MR-MDFG that are pointed directly into the same subgraph belong to a rate-identical cut.

*proof:*

1) Suppose that an MR-MDFG is rate-balanced, and there are two edges $e_A$ and $e_B$ with different rates directly pointed into the same subgraph $G_1$. Let u and v be the nodes in $G_1$ that are directly connected to $e_A$ and $e_B$ respectively. Because $G_1$ is a connected graph, we can find a node t in it such that there are two paths $P_A$ and $P_B$ from the sources of the original MR-MDFG going through $e_A$ and $e_B$ respectively and having t as the same destination. Since all edges in $G_1$ are of single-rate, $Z(P_A)$ and $Z(P_B)$ are equal to $R(e_A)$ and $R(e_B)$ respectively. So we have $Z(P_A) \neq Z(P_B)$, which is contradictory to *Lemma 3.1.*

2) If an MR-MDFG is rate-conflict, we can find a rate-conflict edge e: $u \rightarrow v$ and let its primary paths be $P_A$ and $P_B$. Suppose u is in subgraph $G_1$ if all multirate-weighted edges in the MR-MDFG are cut off. Now we claim that there exist two multirate-weighted edges with different rates in the original MR-MDFG pointed directly into $G_1$. Otherwise, any two different paths from the sources of MR-MDFG to node u would have the same rate because all edges in $G_1$ are of single-rate. *End of Proof.*

# 4. THE TECHNIQUE OF MD INTERCALATION

*Lemma 4.1*: The cell dependence vectors in the iteration space that correspond to the multirate-weighted edges in an MR-MDFG have lengths of *O*(*N*) in the iteration space, where N is the input size of the DSP algorithm represented by the MR-MDFG.

*Proof:*

Assume an edge e: $U \rightarrow V$ with the offset-weight d(e) and the multirate-weight M(e) in the MR-MDFG. Suppose that a cell dependence vector in the iteration space corresponding to e is from the cell including U (indexed by S) to the cell including V (indexed by T). According to 18.2 and the concept of the cell dependence vector, the length of this cell dependence vector should be $|S-T| = |S-M(e)S+d(e)| = |(1-m_1)s_1+d_1,(1-m_2)s_2+d_2,(1-m_3)s_3+d_3,\cdots,(1-m_n)s_n+d_n| = O(Max(s_1,s_2,\cdots,s_n)) = O(N)$, where $m_1, m_2, m_3, \cdots, m_n$ are diagonal elements in matrix M(e), $d_1, d_2, d_3, \cdots, d_n$ are elements in offset-vector d(e), both M(e) and d(e) are taken as constant values independent of the DSP algorithm's input size N, and $s_1, s_2, s_3, \cdots, s_n$ are Cartesian coordinates of S in the iteration space ranged by the algorithm's input size N. *End of Proof.*

*Theorem 4.1*: The traditional MD retiming techniques cannot asymptotically reduce the minimum storage requirement in the hardware mapping of multirate MD DSP algorithms which are represented by MR-MDFGs.

*Proof:*

Consider an edge e: $U \rightarrow V$ in an MR-MDFG, and two retiming vectors r(U) and r(V). From 18.3 we have e's retimed offset weight $d_r(e) = d(e) + M(e)r(U) - r(v)$. A retiming vector is applied to the same computation node in all cells (MDFGs). In other words, the same computation node in all cells will be moved by the same distance in the iteration space according to the retiming vector, thus the length of the retiming vector (or the moving distance) should be a small value independent of N (the algorithm's input size). Otherwise the epilogue[14] or prologue [14] will be so big that the retiming operations are meaningless.

Because only r(U) and r(V) affect the changing of the length of the cell dependence vector between these two nodes in retiming operations, after retiming operations, the length of the cell dependence vector is still $O(N)$. Since to reduce the lengths of cell dependence vectors is the only benefit that can be exploited by retiming operations to reduce the minimum storage requirement, we reach the conclusion described in this theorem. *End of Proof.*

*Procedure 4.1*: *MD intercalation*

*Step 1. Partitioning the rate-balanced MR-MDFG into single-rate subgraphs:* Cut off all multirate-weighted edges from the MR-MDFG (i.e., those edges whose multirate-weights are not unit matrices) to obtain subgraphs each of which is a single-rate MDFG.

*Step 2. Creating the normalization matrix:* Suppose that the original MR-MDFG has been partitioned into K subgraphs: $G_1, G_2, \cdots, G_K$ in Step 1, and the rates of these subgraphs are represented by matrices $R_1, R_2, \cdots, R_K$. Assume $r_{i,1}, r_{i,2}, \cdots, r_{i,n}$ are the diagonal elements of the matrix $R_i$ $(1 \leq i \leq K)$. Create a special diagonal matrix $R_0$ called *normalization matrix* whose diagonal elements $r_{0,1}, r_{0,2}, \cdots, r_{0,n}$ are evaluated in the following way: $r_{0,j}$ is equal to the least common multiple of $r_{1,j}, r_{2,j}, \cdots, r_{K,j}$ $(1 \leq j \leq n)$.

*Step 3. MD expansion and intercalation in the iteration space:* Consider any a cell indexed by an n-component vector t in iteration space, and a subgraph $G_i$ $(1 \leq i \leq K)$ which is partitioned from the original MR-MDFG in the first step and located in this cell. The operation of MD intercalation on this subgraph in this cell is to move $G_i$ (including all the nodes and edges within $G_i$) to a new position in iteration space indexed by vector $\frac{R_0}{R_i} \times t$. Apply such operation of MD intercalation to

all partitioned subgraphs (in the first step) in all cells in the iteration space. *End of Proof.*

We call *Procedure 4.1* as "MD intercalation" because the MR-MDFG is partitioned into single-rate subgraphs and these subgraphs are "intercalated" in iteration space according to their rates.

*Theorem 4.2:* After the MD intercalation for rate-balanced MR-MDFG's, the lengths of the dependence vectors in n-D iteration space are independent of the DSP algorithm's input size.

*proof:*

Assume an edge e: $U \rightarrow V$ with the offset-weight d(e) and the multirate-weight M(e).

1) If its multirate-weight M(e) is a unit matrix, e is in a subgraph (assumed to be $G_i$) partitioned in Step 1. The dependence vector in the iteration space corresponding to e is from the cell including U (indexed by S) to the cell including V (indexed by T). The length of the dependence vector before MD intercalation is $|S - T| = |d(e)|$. Suppose that $G_i$'s rate is $R_i$. The instance of $G_i$ in the cell at S is moved to the position S'$=\frac{R_0}{R_i} \times S$. The instance of $G_i$ at T is moved to the position T'$=\frac{R_0}{R_i} \times T$ according to Step 3 in the procedure of MD intercalation. The length of the dependence vector is $|S'-T'| = |\frac{R_0}{R_i} \times S - \frac{R_0}{R_i} \times T| = |\frac{R_0}{R_i} \times d(e)|$, which is independent of the algorithm's input size (the range of the iteration space).

2) If e's multirate-weight M(e) is not a unit matrix, e must be between two subgraphs (assumed to be $G_i$ and $G_j$) partitioned in Step 1. Suppose U is in $G_i$ and V is in $G_j$, $G_i$ and $G_j$'s rates are $R_i$ and $R_j$ respectively. The dependence vector in the iteration space corresponding to e is from the cell including U (indexed by S) to the cell including V (indexed by T). The length of the dependence vector before MD intercalation is $|S - T| = |S - M(e)S + d(e)|$. After MD intercalation, the instance of $G_i$ (including U) in cell S is moved to the position S'$=\frac{R_0}{R_i} \times S$, and the instance of $G_j$ (including V) in cell T is moved to the position T'$=\frac{R_0}{R_j} \times T$. Thus the length of the dependence vector after MD intercalation is $|S' - T'| = |\frac{R_0}{R_i} \times S - \frac{R_0}{R_j} \times T| = |\frac{R_0}{R_j} \times M(e) \times S - \frac{R_0}{R_j} \times T| = |\frac{R_0}{R_j} \times (M(e)S - T)| = |\frac{R_0}{R_j} \times d(e)|$, which is independent of the algorithm's input size, or the range of the iteration space. *End of Proof.*

# 5. RETIMING FORMULATIONS FOR MR-MDFG'S

## RETIMING EQUATIONS FOR INTERCALATED ITERATION SPACE

Assume an edge e: $U \rightarrow V$ with offset-weight d(e) and multirate-weight M(e) in a rate-balanced MR-MDFG. Suppose in the first step of the MD intercalation the MR-MDFG is partitioned, and nodes U and V are in the subgraphs $G_i$ and $G_j$ respectively, with $R_i$ as the rate of $G_i$, and $R_j$ as the rate of $G_j$. $R_0$ is the normalization matrix.

1) If M(e) is a unit matrix, or e is a single-rate edge, $G_i$ is the same as $G_j$ and $R_i = R_j$. *Before MD intercalation:* retiming vector r(U) (or r(V)) is defined as the difference vector between the original position and the retimed position of node U (or V) in the iteration space in terms of Cartesian coordinates. Similar to [4, 13–15], the dependence vector before retiming is d(e), and the dependence vector after retiming is $d_r(e)$=d(e)+r(U)-r(V). *After MD intercalation:* nodes U and V in any cell in the iteration space are moved according to $R_i$, thus U and V are only possibly located in the grid whose positions are indexed by $\frac{R_0}{R_i} \times X$ ranged in the iteration space, where X is any n-component vector whose elements are integers. We redefine the retiming vector r(U) (or r(V)) after MD intercalation as *the moving distance within the grid* of node U (or V). As in the proof of *Theorem 4.2,* the dependence vector in the iteration space after MD intercalation yet before retiming is $\frac{R_0}{R_i} \times d(e)$. Furthermore, the dependence vector in the iteration space after MD intercalation and after retiming is $\frac{R_0}{R_i} \times d_r(e) = \frac{R_0}{R_i} \times (d(e) + r(U) - r(V)) = \frac{R_0}{R_j} \times (d(e) + r(U) - r(V))$.

2) If M(e) is not a unit matrix, in other words, if e is not a single-rate edge, $G_i$ is different from $G_j$ and $R_i$ is not equal to $R_j$. *Before MD intercalation:* retiming vector r(U) (or r(V)) is defined as the difference vector between the original position and the retimed position of node U (or V) in the iteration space in terms of Cartesian coordinates. Suppose in a cell indexed by S in the iteration space is found a copy of the MR-MDFG in which a node U is located. The dependence vector in the iteration space starting from this node U is $d(e) + S - M(e)S$. The dependence vector after retiming (yet before MD intercalation) is $(d(e)+S-M(e)S)+M(e)r(U)-r(V)$, based on 18.2 and 18.3 and the restriction that the dependence relationships should not be changed after retiming. *After MD intercalation:* as in the above paragraph, U (or V) is only possibly located in the grid whose positions are indexed by $\frac{R_0}{R_i} \times X$ (or $\frac{R_0}{R_j} \times X$) ranged in the iteration space, where X is any n-component

vector whose elements are integers. We also redefine the retiming vector r(U) (or r(V)) after MD intercalation as *the moving distance within the grid* for node U (or V). The dependence vector in the iteration space after MD intercalation yet before retiming is $\frac{R_0}{R_j} \times d(e)$ according to *Theorem 4.2.* Considering the dependence vector in the iteration space after retiming yet without MD intercalation, and the different moving distances of r(U) and r(V) because of MD intercalation, we have that the dependence vector in the iteration space after MD intercalation and after retiming is $\frac{R_0}{R_j} \times (d(e) + r(U) - r(V))$, with the same equation as for single-rate edge.

Based on the above analysis, by MD intercalation, we have *unified* MD retiming equations for single-rate edges and for multirate edges in the MR-MDFG. Moreover, the lengths of dependence vectors after MD intercalation and retiming are independent of the data positions in iteration space.

## SERIAL PROCESSING ORDER ON MD DATA SET

Suppose the system is n-D. A group of n n-dimensional unitary orthogonal vectors $\{S_1, S_2, \cdots, S_n\}$ is used to describe the processing order. Let (A, B) be the inner product of any two vectors A and B.

Consider any two data samples indexed by n-component vectors X and Y respectively in the MD data set. The processing order of X and Y is decided as the following. 1) If $(X, S_1) \neq (Y, S_1)$, the sample corresponding to the smaller one of the two inner products will be processed earlier. 2) Else if $(X, S_2) \neq (Y, S_2)$, the sample corresponding to the smaller one of the two inner products will be processed earlier. 3)···. n) Else if $(X, S_n) \neq (Y, S_n)$, the sample corresponding to the smaller one of the two inner products will be processed earlier.

Let $H_1$ be the maximum number of samples from the MD data set on a hyperplane indicated by equation $(X, S_1) = C_1$, where X is the n-component vector to index the sample, and $C_1$ is any constant value. Let $H_2$ be the maximum number of samples on a hyperplane indicated by equations $(X, S_1) = C_1$ as well as $(X, S_2) = C_2$, where $C_1$ and $C_2$ are any constant values. Let $H_3$ be ···. ···. Let $H_{n-1}$ be the maximum number of samples on a hyperplane indicated by equations $(X, S_1) = C_1$, $(X, S_2) = C_2$, ···, and $(X, S_{n-1}) = C_{n-1}$, where $C_1$, $C_2$, ···, and $C_{n-1}$ are any constant values.

*Definition 5.1*: The Eigen-function *F*(*X*) of a linear processing order on an MD data set represented by $\{S_1, S_2, \cdots, S_n\}$ is $F(X) = \sum_{i=1}^{n-1} H_i \times (X, S_i)$, where X is any n-component vector.

## THE COMPLETE RETIMING FORMULATION

Let $R_v$ be the storage cost for a computation node v in the MR-MDFG in hardware mapping. $R_v$ is the minimum storage necessary to store the data outgoing from node v and later consumed by other computation nodes in the MR-MDFG. Define W(u,v) as min $\{F(d(p_x))\}$ where the minimum is taken over $\{p_x | p_x$ is any path from node u to node v$\}$. Let t(u) be the time for the computation node u to perform a calculation. Define t(p) as $\sum_{i=0}^{i=K} t(v_i)$ where path p consists of nodes $\{v_i | i \in [1, K]\}$. Define t(u, v) as Max{t(p)}where the maximum is taken over {p | p is any path from u to v such that F(d(p))=W(u,v)}.

*Theorem 5.1*: The complete formulation of retiming for MD intercalated iteration space should be under the following constraints: 1) cost constraint: $R_u = Max(F(d(e) + r(u) - r(v))$ for any edge e: $u \rightarrow v$ outgoing from u in the MR-MDFG; 2) causality constraint: $F(r(v) - r(u)) \leq F(d(e))$ for any edge $u \rightarrow v$ in the MR-MDFG; 3) clock period constraint: $F(r(v) - r(u)) \leq W(u, v) - 1$ for all nodes u and v in the MR-MDFG such that $t(u, v) > c$, where c is the requirement of the system's minimum clock period.

*Proof.*

In the case of single-rate MR-MDFG, the MD intercalation will not lead to the partition of the MR-MDFG because the multirate-weight of any edge in the MR-MDFG is a unit matrix. Then the proof of this theorem is simply an extension from the 2-D case [4] to n-D case. The proof follows the similar proof for 2-D case. Note that we use unitary vectors in n-D space to designate the processing order.

In the case that not all edges in the MR-MDFG have unit matrices as the multirate-weights, the retiming vector r(u) has been redefined as the moving distance vector within the *grid* by intercalation (instead of within iteration space), but the retiming equations for an MR-MDFG are the same as the normal retiming equations for a single-rate MR-MDFG according to the analyses in this section. Based on the unified MD retiming equations for the single-rate edges and the multirate edges in the MR-MDFG given in this section, and considering that the three constraints for the case of single-rate MR-MDFG had been derived only from the same retiming equations in [4], we can conclude the same three constraints of the retiming formulations for the MR-MDFG, followed by the reasoning similar to the proof in [4]. *End of Proof.*

The complete formulation of the MD retiming for a multirate MD DSP algorithm represented by an MR-MDFG, which are preprocessed by the MD intercalation, is described as: *to find retiming vectors for the computation nodes in the MR-MDFG so as to minimize* $COST = \sum_{v \in V} R_v$ *un-*

*der the cost constraint, causality constraint and clock period constraint, where V is the set of all nodes in the MR-MDFG.*

There are many practical procedures that have been derived from this traditional single-rate formulation, some of which can be found in [4, 13–15, 6, 11]. Since we have proposed a unified formulation of retiming technology, the systematic procedures for determining retiming vectors in multirate systems according to the formulation can be established in the same way as those work in previous literature for single-rate systems.

## 6. CONCLUSIONS

An important new technique, MD intercalation, is proposed in the paper. Based on it, a complete UNIFIED formulation of retiming operations for MD multirate DSP algorithms represented by MR-MDFG's is proposed, where clock period constraint, cost minimization, causality constraint, and arbitrarily linear processing order are addressed.

## REFERENCES

1. M. Cotronei, L. B. Montefusco and L. Puccio, *"Multiwavelet analysis and signal processing,"* IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, Vol. 45, Aug. 1998, pp. 970 -987.

2. Seung-Jong Choi and J. W Woods, "Motion-compensated 3-D subband coding of video," *IEEE Transactions on Image Processing,* Vol. 8, Feb. 1999, pp. 155 -167.

3. T. C. Denk and K. K. Parhi, "Systematic design of architectures for M-ary tree-structured filter banks," *IEEE Signal Processing Society Workshop on VLSI Signal Processing,* Vol. 8, Sept. 1995, pp. 157 -166.

4. T. C. Denk and K. K. Parhi, "Two-dimensional retiming," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* Vol. 7, June 1999, pp. 198 -211.

5. T. C. Denk and K. K. Parhi, "Synthesis of folded pipelined architectures for multirate DSP algorithms," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* Vol. 6, Dec. 1998, pp. 595 -607.

6. F. Fernandez, A. Sanchez and A. Duarte, "An optimal software-pipelining method for instruction-level parallel processors based on scaled retiming," *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis,* June 2001, pp. 405 -410.

7. F. Fernandez and A. Sanchez, "Application of multidimensional retiming and matroid theory to DSP algorithm parallelization," *Proceedings of 25th EUROMICRO Conference,* Vol. 1, Sept. 1999, pp. 511 -518.

8. J. Horstmannshoff and H Meyr, "Optimized system synthesis of complex RT level building blocks from multirate dataflow graphs," *Proceedings of 12th International Symposium on System Synthesis,* Nov. 1999, pp. 38 -43.

9. H. Khalil, A. F. Atiya and S. Shaheen, "Three-dimensional video compression using subband/wavelet transform with lower buffering requirements," *IEEE Transactions on Image Processing,* Vol. 8, June 1999, pp. 762 -773.

10. F. G. Meyer, A. Z. Averbuch and J. O. Stromberg, *"Fast adaptive wavelet packet image compression,"* IEEE Transactions on Image Processing, Vol. 9, May 2000, pp. 792 -800.

11. Jun Ma, K. K. Parhi and E. F. Deprettere, "Derivation of parallel and pipelined orthogonal filter architectures via algorithm transformations," *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems,* Vol. 3, June 1999, pp. 347 -350.

12. D. Peng and M. Lu, "MD intercalation and retiming for a general class of MD multirate DSP systems," Proceedings of the Seventh Australia Conference on Parallel and Real-time Systems, Sydney, Australia, Nov. 2000, pp. 215-226.

13. N. L. Passos, E. H.-M. Sha and S. C. Bass, "Optimizing DSP flow graphs via schedule-based MD retiming," *IEEE Transactions on Signal Processing,* Vol. 44, Jan. 1996, pp. 150 -155.

14. N. L. Passos, E. H.-M. Sha and Liang-Fang Chao, "MD interleaving for synchronous circuit design optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Vol. 16, Feb. 1997, pp. 146-159.

15. N. L. Passos and E. H.-M Sha, "Achieving full parallelism using MD retiming," *IEEE Transactions on Parallel and Distributed Systems,* Vol. 7, Nov. 1996, pp. 1150-1163.

16. A. Said and W. A. Pearlman, *A new, fast, and efficient image codec based on set partitioning in hierarchical trees,* IEEE Transactions on Circuits and Systems for Video Technology, Vol. 6, June 1996, pp. 243 -250.

17. J. M. Shapiro, *Embedded image coding using zerotrees of wavelet coefficients,* IEEE Transactions on Signal Processing, Vol. 41, 1993, pp. 3445 -3462.

18. V. Sundararajan and K. K. Parhi, "Synthesis of folded, pipelined architectures for multi-dimensional multirate systems," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. 5, 1998, pp. 3089 -3092.

19. P. P. Vaidyanathan, "Multirate systems and filter banks," *Englewood Cliffs, NJ: Prentice-Hall,* 1993.

20. M. Vetterli, J. Kovacevic, em Wavelets and Subband Coding, Prentice Hall, 1995.

21. J. Wang and K Huang, "Medical image compression by using three-dimensional wavelet transformation," *IEEE Transactions on Medical Imaging,* Vol. 15, Aug. 1996, pp. 547 -554.

22. Michael Weeks and Magdy Bayoumi, "3-D discrete wavelet transform architectures," *IEEE International Symposium on Circuits and Systems (ISCAS '98),* June 1998, pp. IV-57 -60.

23. Zixiang Xiong, K. Ramchandran, M. T. Orchard, *Wavelet packet image coding using space-frequency quantization,* IEEE Transactions on Image Processing, Vol. 7, June 1998, pp. 892 -898.

24. Zixiang Xiong, K. Ramchandran, M. T. Orchard, *Space-frequency quantization for wavelet image coding,* IEEE Transactions on Image Processing, Vol. 6, May 1997, pp. 677 -693.

25. V. Zivojnovic and R. Schoenen, "On retiming of multirate DSP algorithms," *IEEE International Conference Proceedings on Acoustics, Speech, and Signal Processing,* Vol. 6, May 1996, pp. 3310 -3313.

26. T. W. O'Neil and E. H. -M. Sha, "Retiming synchronous data-flow graphs to reduce execution time," *IEEE Transactions on Signal Processing,* Vol. 49, Oct 2001, pp. 2397-2407.

Chapter 19

# EFFICIENT DECISION FEEDBACK EQUALISATION OF NONLINEAR VOLTERRA CHANNELS

Songsri Sirianunpiboon and John Tsimbinos
*Defence Science and Technology Organisation, PO BOX 1500, Edinburgh, South Australia 5111, Australia, Songsri.Sirianunpiboon@dsto.defence.gov.au, John. tsimbinos@dsto. defence.gov.au*

**Abstract** Intersymbol interference caused by nonlinear channel imperfections can be reduced by a Decision Feedback Equaliser (DFE). The desire to eliminate the need for a tentative decision that arises in the nonlinear channel DFE has led to the development of root finding methods for obtaining the correct symbols in such equalisers. This chapter offers several further developments to this approach. An a priori root selection method suitable for up to 3rd order nonlinear Volterra channels is given. An analysis of the conditions for monotonicity of Volterra channels is presented. Finally, an efficient symbol decision method suitable for general *N*th order nonlinear channels using a modified form of the bisection method is developed.

**Keywords:** Decision Feedback Equaliser, Channel Equalisation, Intersymbol Interference, Tentative Decision, Root Finding Method, Nonlinear Channel, Volterra Channel, Volterra Model

## 1. INTRODUCTION

Nonlinear intersymbol interference (ISI) caused by channel nonlinearities can lead to a significant increase in the error rate of communication and digital storage channels. A Decision Feedback Equaliser (DFE) incorporating nonlinear feedback filters as typically shown in Figure 19-1, is aimed at decreasing this effect [1-5]. Error recovery for such DFEs for *N*th order nonlinear channels has been analysed in [6].

Unlike linear channels, nonlinear intersymbol interference is caused by nonlinear components based on past as well as *present* symbols. In past work [2-5], this required the DFE to make a tentative decision estimate of the current symbol, followed by a final decision, as shown in Figure 19-2.

An approach that eliminates the need for a tentative decision, and improves the probability of error performance is a DFE using a *Root Method* [7], to make the symbol decisions, as shown in Figure 19-3. Figure 19-4 illustrates typical improvement in error probability given by the Root Method over the tentative decision method for a third order nonlinear channel. This paper expands on the root method approach in several ways. Section 3 gives an a priori correct root selection method, illustrated for 2nd and 3rd order nonlinear Volterra channels. Section 4 gives an analysis of the conditions for monotonicity of Volterra channels. Finally, Section 5 makes use of the monotonicity assumption to develop an efficient symbol decision method suitable for general *N*th order nonlinear channels using a modified form of the bisection algorithm.

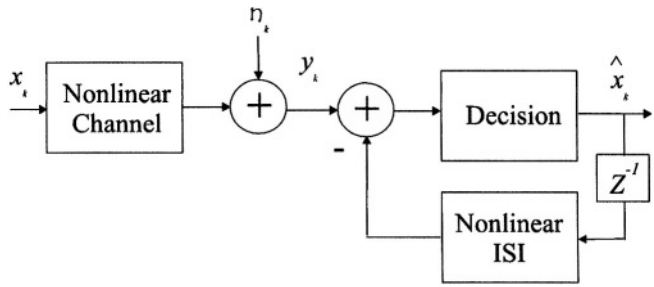



*Figure 19-1.* Nonlinear channel followed by conventional decision feedback equaliser

## 2. ROOT METHOD FOR THE NONLINEAR CHANNEL DFE

Assume an $N^{\text{th}}$ order nonlinear Volterra channel with input $x_k$ and output $y_k$ relationship given as:

$$y_k = \sum_{r_1=0}^{M_1} h_{r_1} x_{k-r_1} + \sum_{r_1=0}^{M_2} \sum_{r_2=0}^{M_2} h_{r_1 r_2} x_{k-r_1} x_{k-r_2} + \cdots + \sum_{r_1=0}^{M_N} \cdots \sum_{r_N=0}^{M_N} h_{r_1 \ldots r_N} x_{k-r_1} \cdots x_{k-r_N} + \eta_k, \tag{19.1}$$

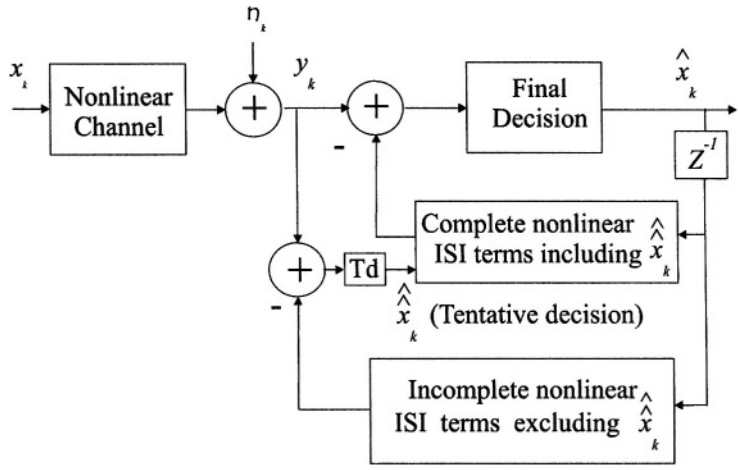



*Figure 19-2.* Nonlinear channel followed by decision feedback equaliser with Tentative Decision

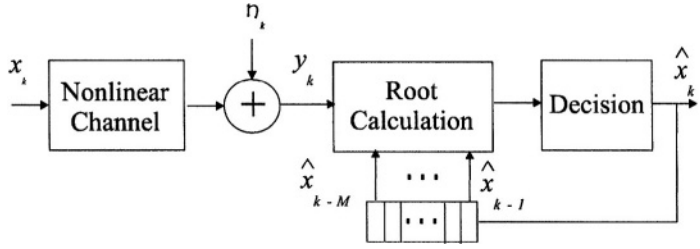



*Figure 19-3.* Nonlinear channel followed by Root Finding decision feedback equaliser

where $h_{r_1}, h_{r_1 r_2}, \ldots, h_{r_1 \ldots r_N}$ are the first, second, …, $N$th order Volterra kernels and we assume that $h$ is symmetric in its arguments. $\eta_k$ is additive white Gaussian noise with $\mathbf{E}(\eta_k) = 0$ and $\mathbf{E}(\eta_j \eta_k) = \sigma^2 \delta_{jk}$, for all $j$, $k$. In terms of the $k^{\text{th}}$ input symbol $x_k$, we can write (19.1) as

$$y_k = c_0(k) + c_1(k)x_k + \cdots + c_N(k)x_k^N + \eta_k \,, \qquad (19.2)$$

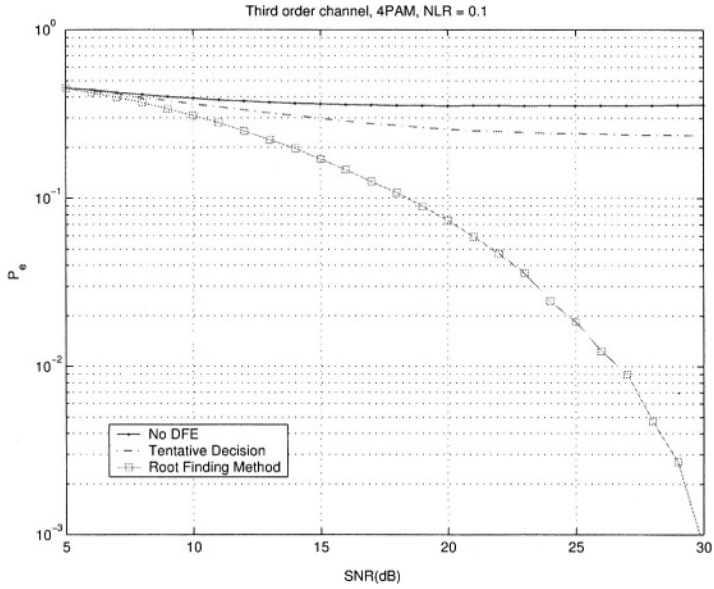



*Figure 19-4.* Typical improvement given by Root Finding method over the Tentative Decision method

where

$$c_0(k) = \sum_{r_1=1}^{M_1} h_{r_1} x_{k-r_1} + \sum_{r_1=1}^{M_2} \sum_{r_2=1}^{M_2} h_{r_1 r_2} x_{k-r_1} x_{k-r_2} + \cdots + \sum_{r_1=1}^{M_N} \cdots \sum_{r_N=1}^{M_N} h_{r_1 \ldots r_N} x_{k-r_1} \ldots x_{k-r_N},$$

and for $n = 1, 2, \ldots, N$,

$$c_n(k) = h_{0\ldots0} + \binom{n+1}{n} \sum_{r_{n+1}=1}^{M_{n+1}} h_{0\ldots0 r_{n+1}} x_{k-r_{n+1}} + \cdots + \binom{N}{n} \sum_{r_{n+1}=1}^{M_N} \cdots \sum_{r_N=1}^{M_N} h_{0\ldots0 r_{n+1}\ldots r_N} x_{k-r_{n+1}} \ldots x_{k-r_N}. \tag{19.3}$$

For each possible set of previous inputs given by $\{x_{k-1}, \ldots, x_{k-M}\}$, where $M = \max\{M_1, \ldots, M_N\}$, we will refer to the polynomial:

$$V(x_k|\, x_{k-1}, \ldots, x_{k-M}) = c_0(k) + c_1(k)x_k + \cdots + c_N(k)x_k^N \qquad (19.4)$$

as the *channel polynomial* for this set of previous inputs, where $c_0(k)$, $c_1(k)$, ..., $c_N(k)$ are defined in (2). Note that there is a channel polynomial for each set of previous inputs. The input-output relation (19.4) can be written as

$$y_k = V(x_k|\, x_{k-1}, \ldots, x_{k-M}) + \eta_k \,. \qquad (19.5)$$

Having received the channel output $y_k$, the problem is to decide (estimate) which symbol $x_k$ was sent. However, this involves knowing the correct values of the previous symbols that have been sent. For the DFE this problem is tackled by replacing $\{x_{k-1}, \ldots, x_{k-M}\}$ in (19.5), by the previous decisions $\{\hat{x}_{k-1}, \ldots, \hat{x}_{k-M}\}$. So the DFE decision problem becomes one of estimating $x_k$, given the model

$$y_k = V(x_k|\, \hat{x}_{k-1}, \ldots, \hat{x}_{k-M}) + \eta_k \,. \qquad (19.6)$$

One method for estimating $x_k$ is to solve

$$V(x_k|\, \hat{x}_{k-1}, \ldots, \hat{x}_{k-M}) - y_k = 0 \qquad (19.7)$$

and apply the standard quantiser $Q$ to the solution [1]. The *Root Method* of Redfern and Zhou [7], involves finding all the roots of (19.7). However, the problem is deciding which is the correct root and which are spurious. Redfern and Zhou propose that the root which is closet in Euclidean distance to an input symbol be selected. The DFE decision is then this closest input symbol. The main disadvantage of this method is that finding all of the roots of a polynomial can be very computationally intensive. The question addressed in this paper is whether there are circumstances under which we can obtain the single correct root of (19.7) without the need to expend effort finding the spurious roots.

Redfern and Zhou [7] briefly proposed a modification of this method to avoid the computational problem of finding all the roots of (19.7). The modification was to exploit the fact that there are a finite set of input symbols by choosing $\hat{x}_k$ to be the input symbol that minimises the magnitude of the left hand side of (19.7). Instead of finding all roots of (19.7), each symbol in the symbol set is substituted and the one that give the smallest value is chosen, as in

$$\hat{x}_k = \arg\min_{x \in \mathcal{S}} \left(y_k - V(x|\hat{x}_{k-1}, \ldots, \hat{x}_{k-M})\right)^2 \qquad (19.8)$$

where $\mathcal{S}$ is the symbol set. Our extensive analysis of the *symbol substitution method* is beyond the scope of this chapter, and will be published

elsewhere [10]. This chapter is more concerned with the development of an efficient root finding approach and the reduction of computational complexity.

Among nonlinear Volterra channels one can single out those which we will refer to as monotonic. The solution of (19.7) is equivalent to finding the mathematical inverse of the channel polynomial at $y_k$. To achieve this unambiguously in all cases the channel polynomial $V(x_k|x_{k-1},\ldots,x_{k-M})$ must be monotonically increasing or decreasing over the region occupied by the symbol set, for any given set of *previous inputs.* We will refer to the channel as being *monotonic* if the channel polynomial is *monotonically increasing* or *deceasing* over the symbol region, for *each* possible set of previous inputs. We will refer to the channel as being *monotonically increasing* (or *decreasing*) if the channel polynomial is *monotonically increasing* (or *decreasing*) over the symbol region for *all* of the *previous inputs.*

In this chapter we will show that under the condition that the channel is monotonic over the symbol region, the correct root of (19.7) can be chosen a priori, or located very efficiently. Although this chapter deals with general $N$th order channels, we demonstrate this for second order channels. We will also show that under this condition that we do not need to determine the roots of (19.7) at all, we only need to know in which decision interval the correct root lies. This leads to an efficient DFE algorithm based on a modified form of the bisection algorithm. Throughout this chapter we illustrate our analysis for the m-PAM symbol set (i.e., $x_k \in \{-(m-1),-(m-3),\ldots,-1,1,\ldots,(m-3),(m-1)\}$) however the analysis can be extended to the case of complex symbols.

Finally, we note that in the following we will often use the short hand notation $V(x_k)$ for the channel polynomial $V(x_k|x_{k-1},\ldots,x_{k-M})$ where no confusion arises.

## 3. A PRIORI ROOT SELECTION METHOD

This section will give details of how to select a priori the correct root of (19.7) for a second order channel. Consider the model of the second order channel polynomial:

$$y_k = c_2(k)x_k^2 + c_1(k)x_k + c_0(k) = V(x_k|x_{k-1},\ldots,x_{k-M}) \tag{19.9}$$

where $c_0(k), c_1(k)$ and $c_2(k)$ are defined in (19.3). Given $y_k$ we need to determine $x_k$ by solving the following:

$$V(x_k|x_{k-1},\ldots,x_{k-M}) - y_k = 0. \tag{19.10}$$

Let $\hat{x}_k$ be the estimate of $x_k$, then we have,

$$\hat{x}_k = -\frac{c_1(k)}{2h_{00}} \pm \frac{1}{2h_{00}}\sqrt{c_1^2(k) - 4h_{00}(c_0(k) - y_k)} \tag{19.11}$$

Substituting (19.9) for $y_k$ into (19.11) we have

$$\hat{x}_k = -\frac{c_1(k)}{2h_{00}} \pm \frac{|c_1(k) + 2h_{00}x_k|}{2h_{00}}. \tag{19.12}$$

Since we do not a priori know the value of $x_k$, we need to be able to choose the + or – sign in (19.12) so that it will give the correct value of $\hat{x}_k$ for all $x_k$ in the symbol set.

We can choose the + sign in (19.12) if $c_1(k) + 2h_{00}x_k \geq 0$, for all $x_k \in \mathcal{S}$, i.e., if $c_1(k) \geq 2(m-1)|h_{00}|$. Similarly, we can take the – sign in (19.12) for all $x_k \in \mathcal{S}$, if $c_1(k) \leq -2(m-1)|h_{00}|$. Therefore, the correct root is

$$\hat{x}_k = -\frac{c_1(k)}{2h_{00}} + \frac{\text{sign}(c_1(k))}{2h_{00}}\sqrt{c_1^2(k) - 4h_{00}(c_0(k) - y_k)} \tag{19.13}$$

provided that $|c_1(k)| \geq 2(m-1)|h_{00}|$. Finally, we note that if $h_0$ satisfies

$$h_0 \geq 2(m-1)\sum_{i=0}^{M}|h_{0i}|, \tag{19.14}$$

then we can make a single sign choice in (19.13), irrespective of the value of the previous input symbols. In this case the term $\text{sign}(c_1(k))$, in (19.13), can be replaced by $\text{sign}(h_0)$.

In practice, due to the presence of noise (19.12) may have no roots in the symbol region (or no roots at all). Thus, when no root lies inside the symbol range [-(m-1),(m-1)], i.e., when $V(-m+1)V(m-1) > 0$, then we choose $\hat{x}_k$ to be either -(m-1) or (m-1) depending on which symbol gives the value closer to zero.

For the third order channel the correct root can be obtained in a similar fashion. The derivation is more complicated, and space does not allow the inclusion of the details here. However, the results are as follows:

$$\hat{x}_k = s_1 + \exp\left(\frac{2\pi i}{3}\left[\frac{3}{2\pi}\arg\left(s_2^*\left(s_1 - \frac{c_2(k)}{3h_{000}}\right) - \pi\right)\right]\right)s_2 - \frac{c_2(k)}{3h_{000}} \tag{19.15}$$

where

$$s_1 = (r + (q^3 + r^2)^{\frac{1}{2}})^{\frac{1}{3}}, \tag{19.16}$$

$$s_2 = (r - (q^3 + r^2)^{\frac{1}{2}})^{\frac{1}{3}}, \tag{19.17}$$

$$q = \frac{c_1(k)}{3h_{000}} - \frac{c_2^2(k)}{9h_{000}^2}, \tag{19.18}$$

$$r = \frac{1}{6}\left(\frac{c_1(k)c_2(k)}{h_{000}^2} - \frac{3c_0(k)}{h_{000}}\right) - \frac{1}{27}\frac{c_2^3(k)}{h_{000}^3} \tag{19.19}$$

and $[x]$ denotes the closest integer to $x$. Also, the same type of analysis can be applied to complex m-QAM input symbols. This type of approach does not practically extend to higher than third order channels, so an alternative approach of root (symbol) selection for general $N$th order channel is proposed. This approach is based on a modified form of the bisection algorithm and given in section 5.

## 4. CONDITION FOR MONOTONIC CHANNEL POLYNOMIAL OVER THE SYMBOLS SET FOR ALL POSSIBLE PREVIOUS DECISIONS

In this section we will derive necessary and sufficient conditions for a quadratic channel to be monotonic. We will see that these conditions are precisely those given in the previous section for being able to determine the correct root for a quadratic channel. We will go on to briefly describe sufficient conditions for higher order channels to be monotonic. For the sake of convenience and simplicity of the notation, we write $s$ for $x_k$ and $s_1, s_2, \ldots, s_M$ for $x_{k-1}, x_{k-2}, \ldots, x_{k-M}$. Then for a second order Volterra channel the channel polynomial takes the form:

$$V(s|s_1, \ldots, s_M) = h_{00}s^2 + (h_0 + 2\sum_{j=1}^{M_2} h_{0j}s_j)s + (\sum_{i=1}^{M_1} h_i s_i + \sum_{i=1}^{M_2}\sum_{j=1}^{M_2} h_{ij}s_i s_j) \tag{19.20}$$

We now find the condition on this polynomial so that the turning point is outside the symbol range i.e., the channel polynomial to be monotonic in the symbol range. Here we use m-ary PAM input sequence $x_k$ as an example, so the symbol range is $[-(m-1), (m-1)]$

The channel polynomial will have no turning point on the interval [-(m-1), (m-1)] if and only if

$|\ V'(s|s_1, \cdots, s_M)\ | > 0$ on (-(m-1), (m-1)). Suppose that

$$V'(s|s_1, \cdots, s_M) \geq 0 \quad \text{on } [-(m-1), (m-1)] \tag{19.21}$$

with equality only possible at the end points of the interval. This implies that

$$h_0 \geq v(s|s_1, \cdots, s_M) \quad \text{on } [-(m-1), (m-1)] \tag{19.22}$$

where

$$v(s|s_1, \cdots, s_M) = -2h_{00}s - 2\sum_{j=1}^{M_2} h_{0j}s_j. \tag{19.23}$$

Since $v(s|s_1, \cdots, s_M)$ is linear in $s$, its maximum values on $[-(m-1), (m-1)]$ will occur at one of the end points. This implies that

$$\begin{aligned} h_0 &\geq \max_{s\in[-(m-1),(m-1)]} v(s|s_1, \cdots, s_M) \\ &= 2|h_{00}|(m-1) - 2\sum_{j=1}^{M_2} h_{0j}s_j. \end{aligned} \tag{19.24}$$

Similarly if

$$V'(s|s_1, \cdots, s_M) \leq 0 \quad \text{on } [-(m-1), (m-1)] \tag{19.25}$$

with equality only possible at the end points of the interval, we obtain

$$h_0 \leq -2|h_{00}|(m-1) - 2\sum_{j=1}^{M_2} h_{0j}s_j. \tag{19.26}$$

Thus, the channel polynomial (19.9), will be monotonic if and only if $h_0$ satisfies (19.24) or (19.26), for each possible set $\{s_1, \cdots, s_M\}$, i.e.,

$$\begin{aligned} h_0 \notin \bigcup_{s_1,\cdots,s_M\in\mathcal{S}} [&-2|h_{00}|(m-1) - 2\sum_{j=1}^{M_2} h_{0j}s_j, \\ &2|h_{00}|(m-1) - 2\sum_{j=1}^{M_2} h_{0j}s_j]. \end{aligned} \tag{19.27}$$

Checking for monotonicity of a quadratic channel from (19.27) can be difficult when M is large. This is because there may be values of $h_0$ which lie in possible gaps between the intervals in (19.27); finding these

will require an exhaustive search over all $\{s_1, \cdots, s_M\} \in \mathcal{S}$. A simpler sufficient condition for channel monotonicity is

$$h_0 \notin [-2|h_{00}|(m-1) - 2 \min_{s_1,\cdots,s_M \in \mathcal{S}} \sum_{j=1}^{M_2} h_{0j} s_j, \\ 2|h_{00}|(m-1) - 2 \max_{s_1,\cdots,s_M \in \mathcal{S}} \sum_{j=1}^{M_2} h_{0j} s_j] \tag{19.28}$$

which is equivalent to the condition

$$|h_0| \geq 2(m-1) \sum_{j=0}^{M_2} |h_{0j}|. \tag{19.29}$$

The condition for the third order channel polynomial to be monotonic over the symbols set for all possible set of previous decisions can be obtained in term of quadratic programming. The conditions are messy and will not be given here. However, a sufficient condition for monotonicity of general order channel is given below:

$$|h_0| > \max v(s|s_1, \ldots, s_M) \quad \forall s_i \in \mathcal{S} \tag{19.30}$$

where $v(s|s_1, \ldots, s_M)$ is defined in a similar way as in (19.23).

## 5. DECISION BY BISECTION METHOD

Assuming the channel is monotonic, it is not necessary to find the exact root of (19.7). Instead we need only find in which quantizer decision interval the root lies. The estimate $\hat{x}_k$ is simply the midpoint of that decision interval. The method is based on the bisection algorithm in which we determine if a function changes sign over some interval and, if so, deduce that the interval must contain the root. We evaluate the function at the interval's midpoint and examine its sign. Using the midpoint, we replace whichever limit has the same sign. After each iteration the bounds containing the root decrease by a factor of two. Therefore, the method only requires $\log_2(m) + 2$ where $m$ is the level of constellation evaluations of the polynomial and so $N(\log_2(m) + 2)$ multiplications for an $N$th order channel polynomial. So, given the order of channel and the constellation, this method offers the advantage of knowing the exact number of operation needed to make a decision. We use a modification of the standard bisection algorithm for which (19.7) is only ever evaluated at the decision boundaries. Here we apply the algorithm to m-PAM input symbol set. However, it can be extended to m-QAM input symbol

using the bisection method in the complex plane [9]. The decision by bisection algorithm is summarised below.

**Bisection algorithm:**

1 **if** $V(-m)V(m) > 0$ **then if** $V(-m) < V(m)$, $\hat{x}_k = -(m-1)$ **else** $\hat{x}_k = m - 1$; **quit**

2 Initialise $m_1 = -m$, $m_2 = m$,
$mid = 2 \times \text{ceil}((m_1 + m_2)/4)$.

3 **if** $V(m_1)V(mid) \geq 0$ **then** $m_1 = mid$, **else** $m_2 = mid$.

4 Set $mid = 2 \times \text{ceil}((m_1 + m_2)/4)$.

5 Repeat from step 3 until $m_2 - m_1 = 2$.

6 $\hat{x}_k = (m_1 + m_2)/2$; **quit.**

where $\text{ceil}(x)$ denotes the smallest integer larger that $x$.

The following illustrates the difference in CPU time usage between the root method by Redfern and Zhou and the proposed decision by bisection method. We use a third order Volterra system with the kernels as in [7], i.e., $h_0 = 1, h_1 = 0.5, h_2 = -0.3, h_{00} = 0.2M, h_{01} = 0.1M, h_{000} = 0.3M, h_{100} = -0.1M, h_{110} = -0.2M$ where $M$ is a scalar that alters the severity of the nonlinearity. We also define signal-to-noise ratio (SNR) in dB and the nonlinear-to-linear ratio (NLR), given by (19.31) and (19.32) respectively, as in [7].

$$\text{SNR} = 10 \log_{10} \left[ \frac{\text{var}(y_k - \eta_k)}{\text{var}(\eta_k)} \right] \tag{19.31}$$

and

$$\text{NLR} = \frac{\text{var}\left(\sum_{p=2}^{P} y_k^p\right)}{\text{var}(y_k^1)} \tag{19.32}$$

where $y_k^p$ is the output of the $p$th-order Volterra system.

The average CPU times associated with carrying out the root method and bisection method are shown in Figure 19-5. They were obtained by programming in Matlab in which the root finding algorithm is an internal function based on an eigenvalue method. In order to compare the root method's CPU time usage with the bisection method, the bisection method was also programmed in C and was called from within matlab. 100,000 data samples were used at each level of constellation to generate

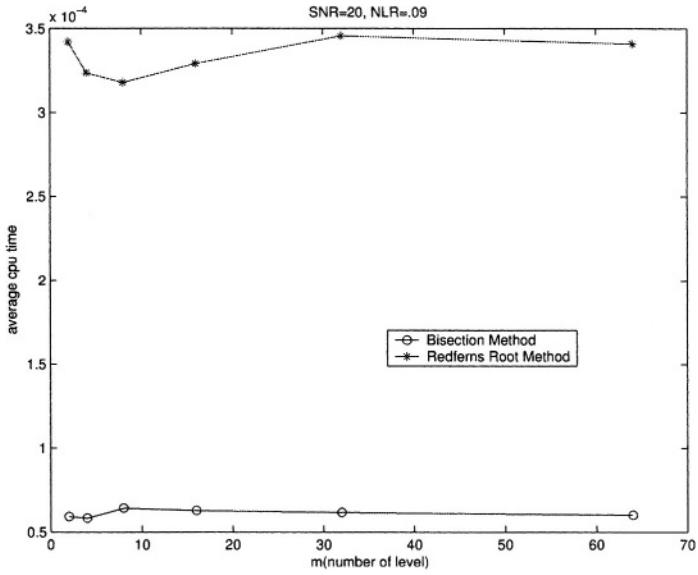



*Figure 19-5.* Average cpu time versus $m$(Constellation level) for the third order channel with NLR = .09 and SNR = 20.

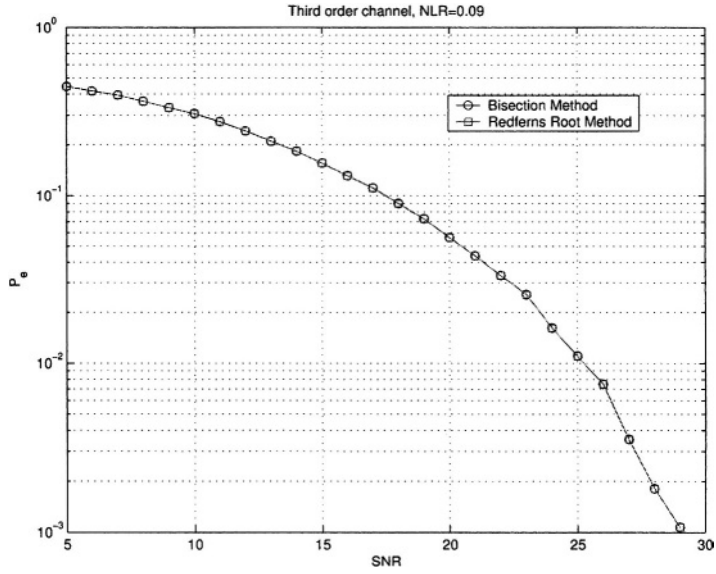



*Figure 19-6.* Probability of error $P_e$ versus SNR for the third order channel with NLR = 0.09

average CPU time usage. A NLR of 0.09 was chosen to ensure the channel polynomial satisfied monotonicity condition, allowing the bisection method to be applied. Figure 19-6 shows that the same probability of error $P_e$ was obtained for both methods at all SNR's. Figure 19-5 compares the time taken to extract a polynomial root by the root method with the time taken to compute the full decision of the bisection method. We see that the bisection method is computationally more efficient than the root method, by approximately a factor of 7, even without including the root method's time for searching through the set of $m$ symbols. The complexity of the decision by bisection is $O(N \log_2(m))$.

# 6. CONCLUSION

For decision feedback equalisation of nonlinear channels the root finding approach eliminates the need for a tentative decision and provides a significant improvement in the probability of error. The root finding method approach calculates all possible roots of the channel polynomial and then determines the correct root based on a minimum distance to a symbol criterion. In this paper we showed that under the condition that a nonlinear channel is monotonic over the symbol region, the output symbol of a Volterra based decision feedback equaliser can be chosen correctly and efficiently by prior selection of a root of the channel polynomial, rather than finding all roots. The conditions for channel monotonicity were given. We also proposed an efficient method of Volterra decision feedback equalisation based on a modified form of the bisection algorithm which follows from the channel monotonicity. It is shown that the bisection method gives a significant improvement in computational complexity over a range of constellation levels, while giving equivalent probability of error performance for all SNR levels. Further work, to be published in the future, involves the *symbol substitution method,* including a probability of error analysis, simulation results and comparisons.

# REFERENCES

1. D.D. Falconer, "Adaptive Equalization of Channel Nonlinearities in QAM Data Transmission Systems," *The Bell Systems Technical Journal,* Vol. 57, No. 7, Sept. 1978, pp. 2589-2611.
2. W. G. Jeon, J. S. Son, Y. S. Cho, Y. H. Lim, and D. H. Youn, "Nonlinear Equalization for Reduction of Nonlinear Distortion in High-Density Recording Channels," *IEEE Int. Conf. on Comms,* Seattle, Washington, USA, June 1995, pp. 503-507.
3. C. P. Callender, S. Theodoridis, and C. F. N. Cowan, "Adaptive non-linear equalisation of communications channels," *Elsevier Signal Processing,* Vol 40, 1994, pp. 325-333.

4. O. E. Agazzi and N. Seshadri, "On the Use of Tentative Decisions to Cancel Intersymbol Interference and Nonlinear Distortion (With Application to Magnetic Recording Channels)," *IEEE Transactions on Information Theory,* Vol 43, No. 2, March 1997, pp. 394-408.
5. W. E. Ryan, J. P. LeBlanc, and R. A. Kennedy, "Performance of RAM-Based Decision Feedback Equalizers With Application To Nonlinear Satellite Channels," *Fifth International Symposium on Signal Processing and its Applications,* Brisbane, Australia, Aug. 1999, pp. 407-410.
6. J. Tsimbinos, and L. B. White, "Error Propagation and Recovery in Decision-Feedback Equalizers for Nonlinear Channels," *IEEE Trans. on Comms.* 49(2), Feb. 2001, pp. 239-242.
7. A. J. Redfern, and G. T. Zhou,"A Root Method for Volterra System Equalization," *IEEE Signal Processing Letters,* 5(11), Nov.1998, pp. 285-288.
8. A. J. Redfern, and G. T. Zhou, "Decision Feedback Equalization for Volterra Systems-a Root Method," *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers, 1998,* Volume:1, 1998, pp.47-51.
9. H. S Wilf, "A Global Bisection Algorithm for Computing the Zeros of Polynomials in the Complex Plane" *Journal of ACM* 25(3), July 1978, pp. 415-420.
10. S. Sirianunpiboon and J. Tsimbinos, "Decision Feedback Equalisation of Nonlinear Channels by Symbol Substitution," in preparation.

Chapter 20

# A WIDEBAND FPGA-BASED DIGITAL DSSS MODEM

Kevin Harman, Adrian Caldow, Cindy Potter, Jon Arnold and Gareth Parker
*Defence Science and Technology Organisation, Australia*

**Abstract:** Direct sequence spread spectrum (DSSS) modulation has particular merit for channels subject to multipath propagation and narrowband interference. If the receiver is implemented in firmware on a field-programmable gate array (FPGA)-based platform, the high-speed parallel architecture of FPGAs can be exploited to realise sophisticated processing of wide bandwidth DSSS signals. This chapter discusses a wideband, burst-mode, 100 Mchip per second (100 Mcps) DSSS demodulator with an asynchronous feed-forward architecture that has been hosted on an FPGA-based digital receiver. The measured performance of this architecture is given and compared with that predicted via simulation and theory. Also discussed is the implementation of a frequency domain adaptive filter for interference suppression via narrowband excision and the FPGA design issues related to it.

**Key words:** Direct Sequence Spread Spectrum, Field Programmable Gate Array, asynchronous feed-forward, interpolation, noncoherent despreading, noncoherent demodulation, burst communications, interference suppression, DFT, FFT, filterbank.

## 1. INTRODUCTION

Direct sequence spread spectrum (DSSS) is a modulation technique in which a message signal is spread over a bandwidth that is typically much greater than that required for reliable communications. This frequency diversity provides resistance to channel defects such a multipath propagation and narrowband interference, while providing a relatively inconspicuous transmitted power spectrum which offers little degradation to co-channel

users. Additional robustness to narrow band interference, flat fading and fast fading can be achieved by incorporating interference mitigation, RAKE diversity combining and error control coding, respectively [1].

This chapter discusses a DSSS receiver that has been implemented in Field Programmable Gate Array (FPGA) logic. The primary driver for adopting an FPGA solution was to extend (in real time) the flexibility and sophistication of digital signal processing into the very wideband domain in which DSSS is most advantageous. It was also important to develop a modular, reconfigurable receiver platform, in which the complexity of the realised DSSS demodulator is scalable with the capabilities of the FPGA technology of the day. For example, DSSS lends itself to a modular solution in which a core demodulator is reusable; as one channel of a larger RAKE receiver, or used in conjunction with interference mitigation or error control coding. Other reasons for the FPGA approach include an ability to achieve a degree of platform independence and security of the firmware algorithms.

Consistent with FPGA capabilities at the conception of this project, the receiver discussed herein was initially specified to provide simplex communications at data rates ranging from 100 kbits/sec up to 4 Mbits/s. To ensure that the DSSS communications has minimal impact on other in-band communications, a minimum spreading factor of 10 (a processing gain of 10dB) was required. The resulting 100 Mchip/sec (Mcps) bandwidth provides some margin at 4Mbit/s, and a spreading factor of up to 1000 (30dB) at the lower data rate. These rates fully extended the FPGAs used to implement the first generation platform. The 4 Mbits/s data rate severely limits the spreading gains achievable, so to further reduce the impact on other communications, a burst mode of operation was required. This requires rapid acquisition for chip timing and as a consequence, an asynchronous feedforward architecture (AFF) forms the core of this receiver, avoiding the acquisition delays that would be present, particularly with very low chip SNR, in a closed-loop sampling architecture. Noncoherent demodulation was chosen for a similar reason.

At commencement of the receiver design, the available FPGA technology would only permit the implementation of this core functionality. However, as illustrated in Fig. 20-1, there is a steady growth in FPGA capacity, which accelerated rapidly in the period 1996 to 2004. It is also noteworthy that in addition to this increase, other advances are being made in FPGA technology, such as the incorporation of arithmetic and dedicated microprocessor functions. In order to best use any particular technology, a design needs to take into consideration such features, in addition to more general aspects of FPGA architectures such as a suitability to parallel processing and fixed point, finite precision arithmetic.

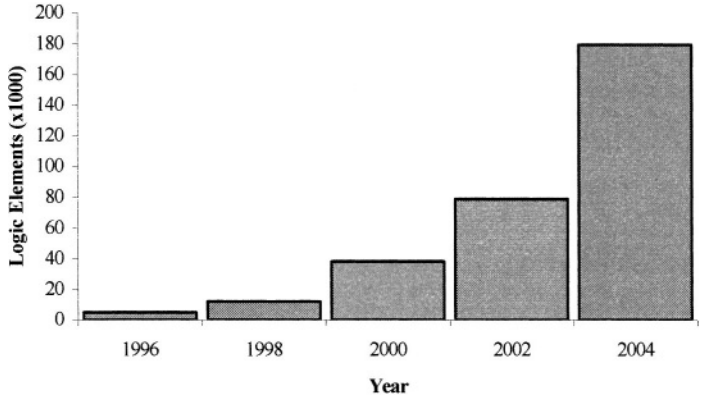



*Figure 20-1.* The growth trend in Altera FPGA capacity

Although it was realised that only core functionality could be achieved in the initial receiver design, it was also anticipated that additional features could be incorporated in the near future. Environmental surveys revealed that although multipath effects would degrade the DSSS communications, interference from multiple narrowband cochannel users could pose a greater threat to reliable DSSS performance. Consequently, the receiver has now incorporated a processing function to mitigate these affects.

In the remainder of this chapter, the discussion on the FPGA design of the receiver is expanded. In Section 2, the design issues associated with the core DSSS function are elaborated and the corresponding performance of the receiver under field trials is detailed. Section 3 discusses how the narrow band interference reduction function has been achieved using a discrete Fourier transform (DFT) filterbank, with interesting design aspects of the filterbank being the focus. The chapter is summarised in Section 4, which also outlines future work.

## 2. THE DSSS DEMODULATOR

The DSSS demodulator is the critical entity of the modular FPGA based receiver. In a stand-alone configuration it recovers the message signal from the received signal without any pre- or post-processing, and provides a certain bit error rate (BER) characteristic and maximum bandwidth (FPGA throughput) at a corresponding cost in FPGA resources. Thus it also establishes performance baselines and FPGA usage constraints for the

modular scenario that encompasses interference mitigation, RAKE combining and error control coding.

The signal processing elements required in a DSSS demodulator are well known [2], and include spreading code acquisition, tracking and despreading, followed by conventional demodulation of the underlying data. Less obvious is the appropriate architecture for each function, which depends on the target application.

As outlined in Section 1, this application demands real-time firmware based digital processing of a burst, wideband signal. From the conception of the project the transmitter was planned to be a low-cost, low-power, portable device, with the implication that frequency stability may be poor. These constraints make architecture selection challenging. FPGA throughput, which is limited by both the technology clock rate and the extent to which design parallelism is supportable, is challenged by the high sample rate implicit in wideband signal digitisation and by the need for sufficient oversampling (to support a particular performance) in the processing stages. Accurate carrier recovery is challenged by the low operating signal to noise ratio (SNR), poor carrier stability, severe multipath in the channel and low quality transmitter oscillators. Similarly, rapid chip timing recovery, required for burst signalling, is challenged by low SNR and poor clock stability.

## 2.1 The wideband AFF Architecture

An architecture which addresses these issues [3,4] is presented in Fig. 20-2. Here, the quadrature received signal is minimum (twice) oversampled to reduce throughput; the correlative detectors for spreading code acquisition, tracking and despreading are all noncoherent, permitting operation in the presence of frequency error; the chip timing is recovered with an asynchronous feed-forward (AFF) interpolation and tracking circuit, which adds only chip-scale latency to the timing recovery; and the underlying data modulation is differential QPSK (DQPSK), which permits noncoherent demodulation. To maximise the transmitted data rate, the modulation is unbalanced or dual-channel [2], whereby each of the in-phase and quadrature (I and Q) channels carry separate data, as well as different spreading codes.

Referring to Fig. 20-2, the signal processing commences with minimum oversampling of the chip-rate signal at baseband. To ease the throughput burden on the FPGAs, the sampled data may be taken in time-interleaved parallel streams, provided suitable parallel architectures exist for the subsequent processes.

Chip-matched filtering follows, and has the intention of maximising the SNR at the input to subsequent stages (any timing error from the AFF

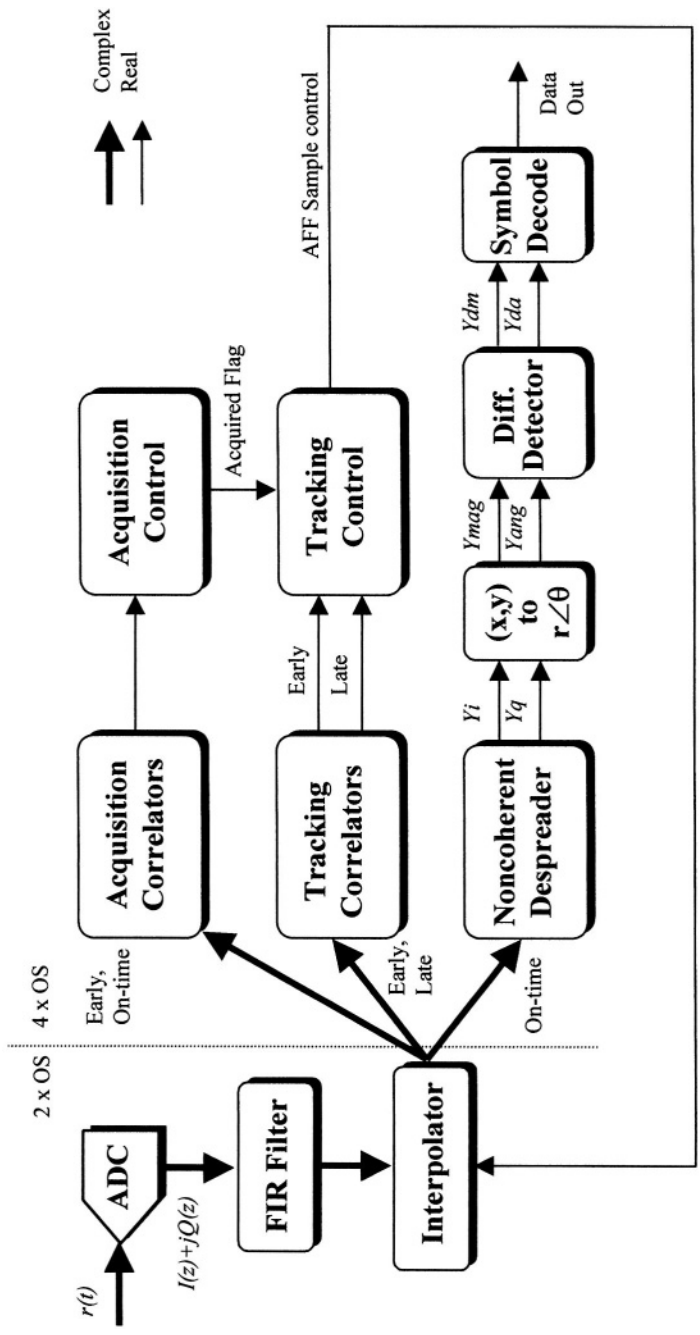



*Figure 20-2.* The wideband AFF architecture

approach will reduce the effectiveness of this matching, as noted in Section 2.3.1). FIR filters with a parallel, multiplierless design were used. This maximises throughput at the expense of utilisation, although the multiplierless approach is both rate and resource efficient.

Interpolators increase the effective sample rate from two to four samples per chip. The intention is to provide sampled data with sufficient timing resolution that no further timing adjustments are required for acceptable performance from the remaining signal processing. Detailed simulations indicated that four samples per chip linear interpolation was the optimal compromise between complexity and performance [5] (higher order interpolators were analysed but added little to tracking performance).

Spreading code acquisition is achieved using a noncoherent (envelope-detecting) parallel acquisition strategy [6,7]. To minimise acquisition time in support of burst signalling, the correlation length (which is restricted by FPGA capacity) is maximised via register-level design optimisation and placement, and two timing phases separated by ½-chip are tested simultaneously [6].

Tracking follows acquisition and performs both coarse (integral chip period) and fine (¼-chip period) timing adjustments to compensate for the asynchronous chip clocks on the link and maintain optimal sample selection at the output of the interpolator. An early-late gate delay-lock loop (ELDLL) approach is taken [2], but in the AFF architecture the loop decisions unconventionally drive sample (interpolator) state and code phase selection, rather than sample clock phase as in a conventional PLL. Again, the tracking loop uses noncoherent correlators, allowing tracking in the presence of frequency error.

Valid tracking ensures that the best sample instant to within ¼-chip resolution is known. With this synchronisation, the interpolator produces chip spaced complex samples $u(n)$ which are then despread to recover the underlying DQPSK symbols. To achieve this noncoherently a novel despreading scheme was proposed in which the $k^{\text{th}}$ symbol, $y^{(k)}$ is despread according to:

$$y^{(k)} = \frac{1}{N} \sum_{n=(k-1)N}^{kN-1} u(n)\left(p_I(n) + p_Q(n)\right), \tag{20.1}$$

where $p_I(n)$ and $p_Q(n)$ are the I and Q spreading sequences, $N = R_{ch}/R_{sym}$ is the correlation length (processing gain) and $R_{ch}$ and $R_{sym}$ are the chip and symbol rates respectively. The performance of this noncoherent despreading method is further discussed in Section 2.3.

Symbol demodulation is then achieved using conventional DQPSK techniques [8], which partially mitigate the uncompensated frequency error. A multiplierless CORDIC rectangular-to-polar converter [9] first converts the despread symbols to polar form. This simplifies the angle difference calculation performed by the differential detector to de-rotate the constellation. Then hard decisions are made on the de-rotated symbols. This processing chain adds only pipelining latency to the throughput and so has no ultimate effect on packet overhead during burst transmissions. However, it must be noted that this architecture supports the generic case of an arbitrary number of chips per symbol in the DSSS modulation, not the common case of one pseudo random binary sequence (PRBS) iteration per symbol with the implication that a long, noise-like PRBS can be used at any data rate, and the data rate can be adjusted independently of the spreading codes. Lacking synchronisation between the symbols and the spreading codes, symbol timing must be determined and tracked with a symbol timing recovery (STR) loop, and the lock time of this loop constitutes an overhead per burst packet (see Section 2.3.3).

## 2.2 Analysis of the noncoherent despreader

Attempted despreading by separately correlating the I and Q components of $u(n)$ with the corresponding spreading sequence, as would be done for coherent DSSS communications [2], is inappropriate for this noncoherent receiver. Moreover, the techniques of Viterbi [7], which consider noncoherent, but balanced QPSK. (where the same data modulates both I and Q channels) are also unsuitable.

In equation (20.1), $u(n)$ is correlated against both I and Q spreading sequences and it can be shown [10,11] that this gives rise to a signal which, after differential detection, has a mean that is close to the transmitted message symbol (although there is a rotation and slight attenuation that depend on the residual frequency error). The resultant additive noise component comprises both a random part related to the noise at the receiver input (this is double that which would be present in a coherent receiver, due to the correlation against both I and Q sequences) and also a deterministic part that is a consequence of the nonzero partial correlation between I and Q spreading sequences.

Fig. 20-3 illustrates this deterministic contribution to the noise on the despread, differentially detected symbols, for the case where there is no received random noise. The figure shows an overlay of the scatter plots for a residual frequency error increasing from 0 Hz to 100 kHz, all with 50 MHz chip rate and 500 kHz symbol rate. It can be observed that for zero frequency error, the partial correlations give rise to a purely radial

distribution, but as frequency error increases, an angular dispersion is introduced.

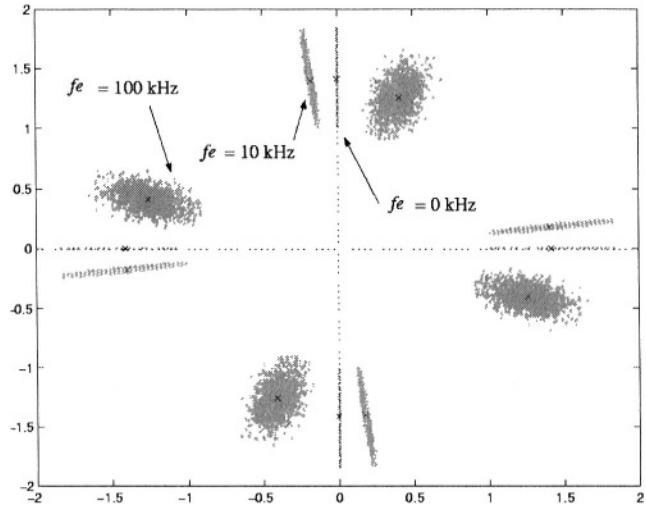



*Figure 20-3.* Scatter plot for differential detection with *Rch* = 50 MHz, *Rsym* = 500 kHz and frequency error equal to 0, 10 and 100 kHz

### 2.2.1 SNR of the noncoherent despreading

The losses associated with the noncoherent despreader are quantified in an earlier paper [10], which analyses the SNR of the demodulated signal. The differentially detected symbol SNR is plotted against the chip rate SNR in Fig. 20-4. The three curves show the effect of increasing the correlation length from $N = 50$ (lower curve) to $N = 2000$ (upper curve) and the performance of a coherent demodulator is shown by the dashed line.

The results in Fig. 20-4 clearly show that at low chip SNR there is a loss of approximately 5.3 dB compared to coherent demodulation. This loss is attributed to differential detection (2.3 dB) and the doubling of the random (receiver input-derived) noise variance (3dB). At low chip SNR, the partial correlations have negligible impact. At higher chip SNR (> 0dB), the partial cross-correlation terms become more dominant than the random noise resulting in the asymptotic behaviour shown. This suggests that the implementation penalty of the noncoherent despreader is potentially very large at high chip SNR.

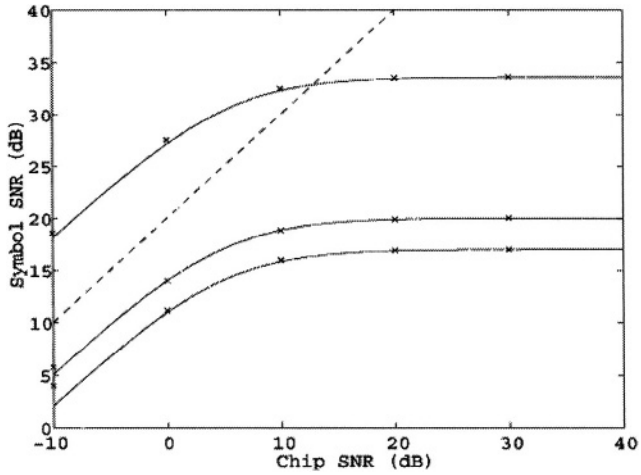



*Figure 20-4.* Differentially detected symbol SNR versus chip SNR for zero frequency error and with *N* increasing from 50 (lower) to 100 (middle) and 2000 chips/symbol (upper). Crosses indicate simulation results and the dashed trace is coherent demodulation for *N*=100

### 2.2.2 BER performance of the noncoherent despreading

Although SNR provides a convenient measure of signal degradation due to the noncoherent despreader, it does not readily translate into bit errors in this instance, as the noise is not circularly distributed and the component due to the non-zero partial-correlations is deterministic [10]. Whilst this deterministic component has an approximate Gaussian distribution, the tails are truncated, eliminating the infrequent but potentially large errors associated with a true Gaussian distribution.

The noise distribution for the noncoherent despreader has been analysed [11], leading to the BER model which has been used to generate Fig. 20-5. The solid trace shows the modelled performance for $N = 50$ (2Mbit/sec) and $N = 500$ (200kbit/sec) chips/symbol. (The discrepancy between the BER predicted using both techniques at 200 kbit/sec is due to contrasting methods [10,11] for modelling the effect of differential detection.) The dashed trace shows the BER performance that arises from a conventional mapping of symbol rate SNR using the Q function. It is clear that although noncoherent despreading using equation (20.1) has a limiting effect on the differentially detected symbol SNR, this does not proportionally limit the BER performance.

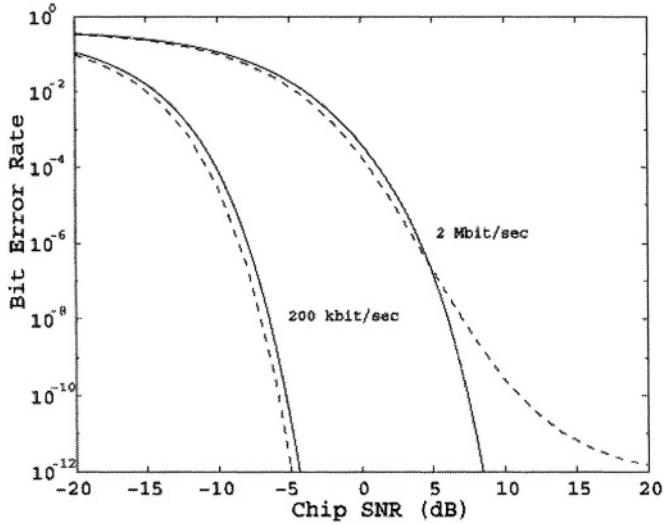



*Figure 20-5.* BER performance for the model (solid) and computed from the symbol rate SNR (dashed), with 50MHz chip rate.

## 2.3 Demodulator performance

In the test case, the chip rate was 50Mchip/s [1]; the underlying data rate was 200kbps or 2Mbps, DQPSK encoded; the spreading sequences were $14^{th}$ order m-sequences; 8-bit analogue-to-digital converters (ADCs) sampling at 100Msamples/s were used for digitisation; the FIR filters were 17-taps long, with 8-bit input, 18-bit internal, and 8-bit output precision; the complex acquisition correlators were 512-chips long; the CORDIC provided a six-stage rotation; and symbols were represented with 16-bit magnitude and 8-bit angle precision.

The signal processing card for the first generation platform was a 6U CompactPCI slave card with a 32-bit, 33MHz PCI interface, and a fixed FPGA array containing six Altera FLEX10K devices. Three of these were EPF10K100ARC240-1 devices, and the other three were EPF10K200SRC240-1 devices, providing a total of 45,000 logic elements. Average utilisation was 75%. In contrast, the second generation platform has a 3U extended-PCI motherboard that supports a 64-bit, 66MHz PCI interface, and has three high-density daughter-board sites for FPGA (or

[1] Firmware place and route for 100Mcps operation is feasible but non-trivial with the first generation platform. Accordingly, testing was completed at 50Mcps, but with the knowledge that the architecture is fully rate scalable.

other) modules. A single module with an Altera Stratix EP1S80 device provides 79,000 logic elements, which would be just 41% utilised with the DSSS demodulator functionality.

### 2.3.1 BER Performance in an AWGN Channel

The performance measure is BER versus symbol SNR, and losses are identified with respect to the BER curve of ideal DQPSK with no forward error correction. Chip SNR was the controlled parameter, with symbol SNR calculated using the expected processing gain at each data rate. Consequently measured losses incorporate any contributions due to the AFF architecture.

Fig. 20-6. shows the measured BER performance. The demodulator is seen to operate typically within a 6dB loss bound, diverging at higher SNR. As described in Section 2.2, a 3dB component of the loss is due to the noncoherent despreading approach. Other components [4] result from calibration (0.5dB), the interpolated AFF approach (0.5 to 1.5dB), and the effect of finite (¼ -chip) timing accuracy, which reduces the effectiveness of the matched filtering, and increasingly so with increasing SNR.

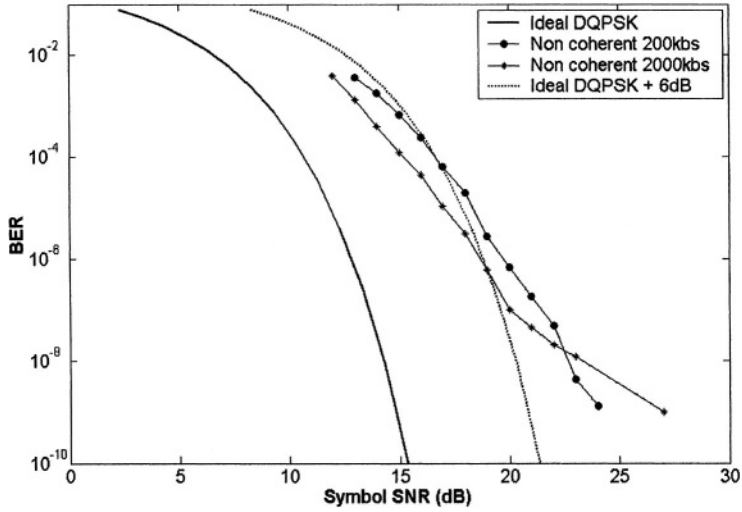



*Figure 20-6.* BER performance in AWGN

### 2.3.2 Robustness to multipath and interference

The wideband DSSS modulation method was intended to provide resilience to multipath and interferers in the urban channel scenario. Consider two examples encountered during wireless testing of the implemented test case. The first is illustrated in Fig. 20-7. This is a plot of the double-sided power spectral density of the received signal after downconversion to baseband. The vertical axis is in 'dB', and the horizontal axis spans the full DSSS main-lobe bandwidth of –50MHz to +50MHz. This snapshot was taken at an operating point of 0dB chip SNR over a 200m line-of-sight path. The salient features here are the main lobe of the desired DSSS signal, seen with relatively low power level, and the broad jammer (an undesired microwave data link) centred 13MHz below centre and some 6dB above the peak of the desired signal. In the presence of this jammer, the BER performance of the receiver at 2Mbps experienced a degradation of 2dB relative to the result of Fig. 20-6, which shows that the link remained robust despite the undesired signal. In contrast, a narrowband link, with similar power, at the jammed frequency would more likely be inoperable.

A second scenario is given in Fig. 20-8. This spectral plot (having the same format as Fig. 20-7) shows severe multipath fading in a difficult urban channel comprising both intra-building and inter-building paths over a 100m range. The link BER was 3.2x 10-3 at a 27dB (symbol) SNR operating point, which is quite consistent with the theoretical flat Rayleigh fading limit [2], taking into account the known demodulator losses.

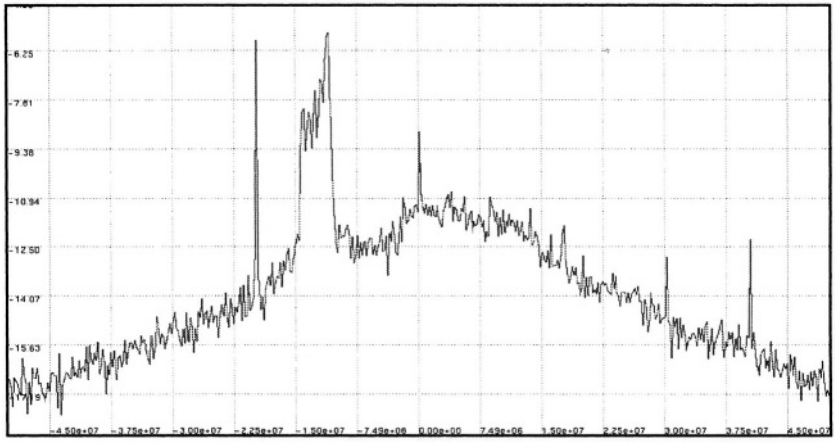



*Figure 20-7.* Received signal spectrum showing narrowband interference

### 2.3.3 Burst capability

The primary performance measure for the utility of the receiver operating in a burst mode is the time taken for valid demodulation after the start of each packet. With the given architecture this is a function of both chip acquisition time and symbol timing loop lock time. At the instant of acquisition the sampled data is valid within a ½-chip timing error, which is adequate for demodulation at a degraded BER. Within one or two chip tracking loop dwells (typically less than one symbol), this error will be

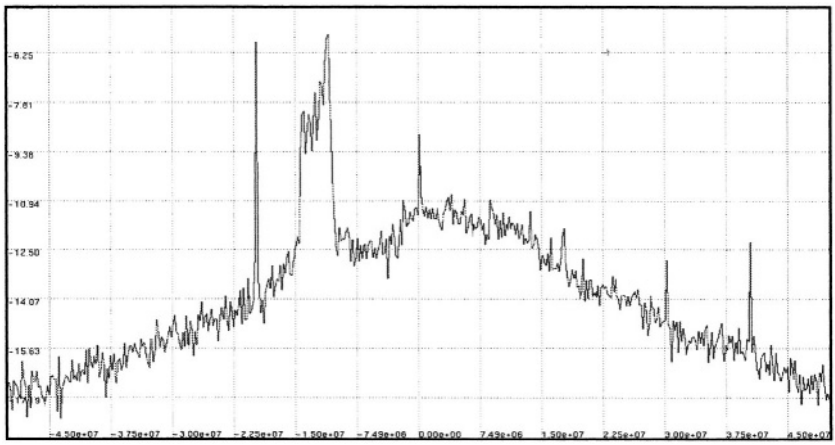



*Figure 20-8.* Received signal spectrum showing multipath in an urban channel

reduced to the expected ¼-chip margin and hence sample quality is effectively sufficient immediately after acquisition. Despreading and demodulation of these samples commences coincident with acquisition, however the recovered data will not be valid until symbol timing has been recovered. The STR loop operates in parallel with the despreading, and will provide a valid symbol clock within a lock time that is dependant on its loop settings and the current operating point. Thus the two significant contributors to output delay are acquisition time ($T_{acq}$) and STR lock time ($T_{lock}$).

Statistics of $T_{acq}$ for parallel correlator structures are well understood [6]. Measurement of mean $T_{acq}$ for the 2Mbps case over the normal SNR operating range established that $T_{acq} \leq 1\text{ms}$ is achievable [4]. Preliminary measurements of $T_{lock}$ above 17dB symbol SNR suggest that the lock period is of the order of 200us (i.e. hundreds of symbols) at 2Mbps. Thus a total delay of the order of 2ms would be a conservative estimate for the expected operating range. If packet overhead (content loss) of 1% was tolerable, this suggests that a packet duration as small as 200ms is feasible.

## 2.4 A coherent demodulation approach

Differential detection as a means of mitigating residual frequency error is appropriate for the burst context in which further frequency correction is likely to be too slow. The known penalty of this approach is 2.3dB for the DQPSK case [8].

An alternative with negligible firmware resource penalty is to employ a non-data-aided feedforward (NDAFF) carrier recovery scheme, such as those described by Classen *et al* [12] and Fitz [13]. Differential encoding and decoding can still be used, alleviating the need to establish an absolute phase reference, but the difference operation can be deferred until symbol decoding, allowing the 2.3dB loss to be recovered. In this approach, first frequency error and then residual phase error are separately estimated and compensated, with the quality of synchronisation dependant on the estimate variances. If averaging periods for acceptable variance are too long, this approach may not be suitable for the burst context.

An implementation with NDAFF synchronisation yielded a 2dB improvement when the estimate for the (slowly varying) frequency error was computed with an averaging period of the order of 1000 symbols, and that for the (rapidly varying) phase was computed over the order of 10 symbols. This solution, which required only 800 additional logic elements compared to the differentially coherent approach and fit easily on the first generation platform, highlights a key benefit of the FPGA-based receiver - the signal processing may be adapted for optimal performance without having to change any hardware.

# 3. INTERFERENCE SUPPRESSION

DSSS possesses a natural resistance to narrowband interference, since the correlation performed in the despreading process has the effect of spreading the energy of a narrow band interferer over the DSSS bandwidth. For large interferers, additional processing may be necessary to avoid significant degradation to the signal of interest (SOI). *Interference excision* [14,15] is a crude but effective technique that has the effect of notch filtering those frequency components that are contaminated by the interference. An effective method of implementing interference excision is to channelise the received signal into partially overlapping frequency bands using a DFT filterbank analyser [16]. Frequency channels identified as containing interference, usually via a comparison against a threshold, have their magnitude set to zero. Channel recombination using a filterbank synthesiser follows.

Computationally efficient algorithms for software implementation of a DFT filterbank exist [16] and rely heavily on multirate filtering concepts and fast Fourier transforms (FFTs). The same algorithms can be used as the basis for FPGA implementation, with modifications incorporated to best exploit the FPGA functionality; software DSP solutions are likely to be largely serial in nature and FPGA devices are predominantly logic, providing intrinsically parallel operations.

It should also be emphasised that incorporation of a DFT filterbank at the front end of the digital processing section has other benefits in addition to facilitating interference excision; more elaborate methods of mitigating interference are possible [17] and frequency domain equalisation can be incorporated to either complement or replace Rake processing.

## 3.1 Firmware Filterbank Development

A filterbank configured as a frequency domain adaptive filter is shown in Fig. 20-9. If the filterbank analyser channels are equally weighted, the synthesiser output will ideally be a time-delayed version of the input signal [16]. The filterbank is then said to exhibit 'perfect reconstruction' [18]. Excision of interference within the $k^{\text{th}}$ channel can be achieved by setting the weight $G_k = 0$. With an input sampling frequency $F_s$ Hz, at time $mM/F_s$, the $k^{\text{th}}$ analyser output channel, centred at $fk = kF_s/K$ Hz, is expressed as

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_k^{-kn}. \qquad (20.2)$$

The synthesiser output is given by

$$y(n) = \sum_{m=-\infty}^{\infty} f(n - mM)\frac{1}{K}\sum_{k=0}^{K-1} Y_k(m)W_k^{kn}, \qquad (20.3)$$

where $K$ is the total number of channels, $M$ is the decimation factor and $W_K = e^{j2\pi/K}$ is a 'twiddle factor'. $h(n)$ and $f(n)$ are the impulse responses of length $RK$ prototype analyzer and synthesiser filters, where $R$ is a real number that is chosen to achieve a specified prototype filter response. Using (20.2) and (20.3), it can be easily shown [16] that the analyser and synthesiser can be implemented using a combination of multirate processing and FFTs. Two common algorithms are the Weighted-Overlap Add and polyphase techniques [16]. The latter approach is well suited to hardware or firmware implementation.

The form of a $K$ channel polyphase filterbank analyser is shown in Figure 10. This structure includes a clockwise commutator to distribute the input data samples at rate $F_s$ Hz to the $K$ polyphase filters. There are $M$ such unique filters, each derived from a different decimation of the prototype filter $h(n)$ [19].

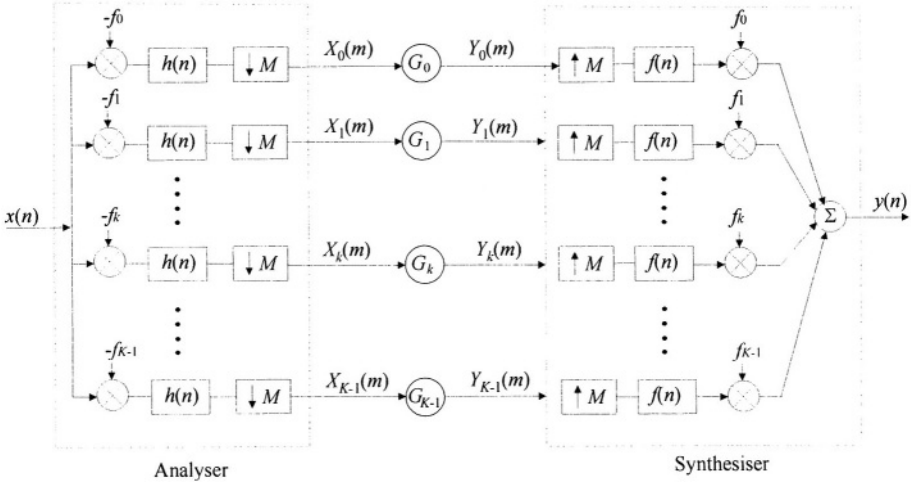



*Figure 20-9.* A filterbank configured as a frequency domain adaptive filter.

According to equation (20.2), a new vector of analyser output data, $[X_0(m), X_1(m), ..., X_{K-1}(m)]$ is calculated every $M$ input samples. Implied in this processing is an $I$ times upsampling, which can be achieved by the insertion of ($I$-1) zeros between consecutive channel samples. For every $M$th input sample a DFT (computed using an FFT for computational efficiency) is performed on the available data. Similar, but dual processing is performed in the synthesiser.

A direct implementation [19] of an 8 channel filterbank in FPGA would provide very limited capability and require approximately 63,000 logic elements and 256 real multiplier blocks. The target FPGA is an Altera EP1S80 Stratix device, with 79,000 logic elements and 176 (9bit) DSP blocks. A specification that would facilitate more meaningful interference mitigation would include $K$=128 channels and clearly, to achieve this with the target FPGA device, a more novel implementation is required. The main issues addressed in the FPGA implementation of the filterbank concerned trading off hardware resource usage, number of channels, maximum allowable sampling rate and fixed point numerical precision.

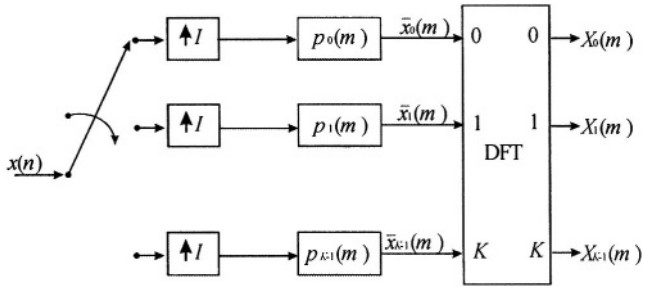



*Figure 20-10.* Polyphase filterbank analyser.

### 3.1.1 Numerical Precision

The numerical precision associated with each operation in the filterbank affects the degree to which the back-to-back analyser/synthesiser combination approximates perfect reconstruction. Polyphase filter, FFT twiddle factor weights and input, output and internal data all require truncation to a particular numerical precision. In order to achieve a target precision for the synthesiser output, a careful design of each stage needs to be undertaken. It was found through simulation that the precision of the polyphase filter coefficients has considerable impact on the reconstructed signal quality. For good accuracy, truncation should only be applied to the polyphase filter outputs, with full precision maintained in the internal arithmetic operations. In contrast, it was found that for the FFT process, *N* bit precision can be maintained by truncating the output of each butterfly stage to *N* bits. This results in a resource efficient design, with a modularity that also achieves design simplicity.

### 3.1.2 Vector arithmetic

Many vector operations can be efficiently implemented using either parallel processing or resource re-use. Parallel processing requires the presence of a dedicated resource for each vector element and can facilitate processing at a clock speed less than the input data rate. Resource reuse reduces the hardware resource requirements, but requires processing to be performed at greater than the input data rate. A combination of these two schemes can be used to yield an architecture which allows physical resource requirements to be traded off against processing speed. The target 128 channel filterbank was achieved using a combination of parallel processing and hardware re-use; parallel processing alone was found to be insufficient.

In the analyser, the input data is oversampled by a factor $I = K/M$ and a new vector of analyser output data is generated every $M$ input data samples. An explicit implementation would use zero padding of the input data, which would be computationally wasteful. Alternatively, this zero insertion can be achieved implicitly by merging the $I$ sets of $M$ unique polyphase filters into a single set of $M$, length $RI$ shift registers and $K$ accumulators. This approach achieves an efficient design by reducing both resource redundancy and latency. That is, the utilisation of each hardware resource is maximised. Similar techniques were used in the synthesiser.

### 3.1.3 FFT Implementation

In the $K$ point FFT and inverse FFT operations that are performed every $M$ input data samples, hardware re-use (for example, the techniques of Mirza [20]) enables the number of butterfly operations to be reduced from $K/2$ per FFT stage to $I$. Operations within each butterfly are executed at the maximum system sampling rate and data buffering is incorporated between consecutive FFT and inverse FFT stages. With this approach and utilizing unity stage one-twiddle factors, the resource requirements of the FFT are reduced by order $M$ compared with a brute force implementation.

### 3.1.4 Overall Complexity

To best implement the filterbank in FPGA logic, the specific architecture of the device needs to be considered and exploited. In this particular case, the Stratix EP1S80 contains 176 dedicated 9 bit DSP blocks. As much multiplication as possible should be targeted to these dedicated blocks, as an alternative logic implementation would consume considerable resources. Employing the design techniques discussed in sections 3.1.1 to 3.1.3 leads to a filterbank implementation with overall system resource requirements incorporating $4RI+8I(\log_2 K-1)$ real multiplier and $4RI-6I-2+12I(\log_2 K)$ real adder functional blocks. It can be shown [19] that it is possible to execute all multiplications using the available DSP blocks, provided a maximum 9 bit precision is used for all multiplier parameters. This results in a performance compromise that limits the reconstruction accuracy of a back-to-back analyser-synthesiser combination to a –34dB error. This should nevertheless prove satisfactory for the DSSS interference mitigation application. The final implementation of the 128 channel filterbank on the target FPGA device resulted in a resource usage of approximately 61,000 logic elements, 124 9bit DSP blocks, and approximately 38k RAM. Simulations confirm a sustained processing rate of 90 MHz; further optimisation is required to achieve the systems 100 MHz throughput.

## 4. SUMMARY

This chapter has provided a brief overview of the successful development of a reconfigurable FPGA-based DSSS receiver platform. As a result of this development, several key results were established.

An architecture for FPGA-based processing of very wideband, burst DSSS signals was presented and shown to offer performance within 6dB of ideal in an AWGN channel. The measured performance degradation can be attributed to limitations imposed on the system architecture for compatibility with burst signalling.

The feasibility and utility of FPGA-based receiver platforms was reinforced, with the given architecture being successfully implemented on an FPGA-based signal processing platform, and the signal processing shown to be versatile. Versatility was achievable by accommodating algorithmic changes without platform penalty, and through modular expansion consistent with FPGA technology trends.

One such modular expansion, which would offer robustness to a DSSS link in the presence of narrowband interference, was an interference excision front-end. A DFT filterbank approach to interference suppression was shown to be challenging but feasible for FPGA-based platforms in 2004, requiring careful design of the architecture and careful control of numeric precision.

In the future, this receiver system will exploit the growing density of FPGA devices to realise a highly robust link with the receiver comprised of a wideband RAKE DSSS demodulator with a high resolution filterbank front-end and the incorporation of powerful error-correcting codes. The FPGA-based signal processing platforms will continue to serve as tools for algorithm research and development.

## 5. REFERENCES

1. B. Sklar, *Rayleigh Fading Channels in Mobile Digital Communication Systems - Part I & II,* IEEE Communications Magazine, pp. 90-109, July 1997.
2. R. L. Peterson, R. E. Ziemer, and D.E. Borth, *Introduction to Spread Spectrum Communication,* Prentice Hall, New Jersey, 1995.
3. J. Arnold, A. Caldow, and K. Harman, *A Reconfigurable 100MChip/s Spread Spectrum Receiver,* Proc. IEEE International Conference on Acoustics Speech and Signal Processing, v.II, pp. 445-448, 2003.
4. K. Harman, A. Caldow and J. Arnold, *A Wideband FPGA-Based Digital Receiver and its Performance as a 50MChip/sec DSSS Demodulator,* Proc. 7th International Symposium on Digital Signal Processing and Communication Systems, pp.349-354, December 2003.

5. W.G. Cowley, and J. Choi, *B1.5 DSSS Modem Project Stage I Final Report,* Institute for Telecommunications Research, South Australia, 1999, (unpublished).
6. Polydoros, and C. L. Weber, *A Unified Approach to Serial Search Spread Code Acquisition-Part I & II,* IEEE Trans. Comm., Vol. 32, No. 5, May 1984, pp 542-560.
7. A.J. Viterbi, CDMA, *Principles of Spread Spectrum Communication,* Addsion Wesley, New York, 1995.
8. A. B. Carlson, *Communication Systems,* 3rd ed., McGraw-Hill, 1986.
9. R. Andraka, *A Survey of Cordic Algorithms for FPGAs,* Proceedings of 6th ACM/SIGDA International Symposium on FPGAs, pp 191-200, 1998.
10. G. Parker, K. Harman, J. Arnold and A. Caldow, *Noncoherent Detection of Unbalanced QPSK Direct Sequence Spread Spectrum Modulation,* Proc. 7th International Symposium on Digital Signal Processing and Communication Systems, pp. 355-360, December 2003.
11. A. Caldow, G. Parker, and K. Harman, *BER Analysis of a Noncoherent Demodulation of Unbalanced DQPSK DSSS,* to appear in Proc. IEEE International Symposium on Spread Spectrum Techniques and Applications, 2004.
12. F. Classen, H. Meyr, P. Sehier, *An All Feedforward Synchronization Unit For Digital Radio,* IEEE 43rd Vehicular Technology Conference, pp 738-741, 1993.
13. M. P. Fitz, *Equivocation in Nonlinear Digital Carrier Synchronisers,* IEEE Transactions on Communications, Vol.39, No.11, pp 1672-1678, 1991.
14. J. G. Proakis, *Digital Communications,* 2nd ed., McGraw-Hill, 1989.
15. L. B. Milstein, and P. K. Das, *An Analysis of a Real-Time Transform Domain Filtering Digital Communication System - Part I: Narrow-Band Interference Rejection,* IEEE Transactions on Communications, vol COM-28, pp.816-824, June 1980.
16. R. E. Crochiere, and L. R. Rabiner, *Multirate Digital Signal Processing,* Prentice Hall, 1983.
17. G. Parker, *Frequency Domain Restoration of Communications Signals,* Institute for Telecommunications Research, University of South Australia, 2001.
18. P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Prentice-Hall, 1993.
19. C. Potter, G. Parker, *An FPGA Frequency Domain Filter for DSSS Interference Mitigation,* Proc. 7th International Symposium on Digital Signal Processing and Communication Systems, pp. 355-360, December 2003.
20. M. Mirza, *B1.6 DSSS Modem Project; High Speed FPGA parallel-FFT multiple butterfly Preliminary Design Report, Task6 Part 2,* Institute for Telecommunications Research, University of South Australia, (unpublished), August 2001.

Chapter 21

# ANTENNAS FOR 5-6 GHZ WIRELESS COMMUNICATION SYSTEMS

Yuehe Ge[1], Karu P. Esselle[1] and Trevor S. Bird[2]
[1]*Department of Electronics, ICS, Macquarie University,* [2]*CSIRO ICT Centre*

Abstract: We review the development of thin, broadband, E-shaped microstrip patch antennas for high-speed (IEEE 802.11a) wireless communication systems operating in the 5.0 – 6.0 GHz frequency range. These antennas may be used in low-profile wireless communication devices such as wireless network adaptor cards in the PCMCIA (also known as PC or Cardbus) format. Importantly, most of these antennas have a height that can be accommodated inside a wireless device as thin as 5mm thickness. Two different twin antenna configurations are also been discussed. The first configuration has two closely spaced similar antennas that can be independently used for transmitting and receiving. It can also be used for space diversity. The second is a diversity antenna pair that provides both polarization and space diversity. In all cases, the reflection at the antenna input is < –10 dB within the two IEEE 802.11a bands (i.e. 5.15 – 5.35 GHz and 5.725 – 5.825 GHz) and the isolation between the two antennas is > 20 dB in the same frequency bands.

Key words: Microstrip antenna, WLAN, Wireless communication, Diversity, Array, PCMCIA, IEEE 802.11, Patch antenna, Broadband antenna, PC, Cardbus, Low-profile

## 1. INTRODUCTION

There is an increasing demand for wireless communication systems such as wireless local area networks (WLAN) and short-range wireless communications, supported by an expanding suite of standards, e.g. IEEE 802.11, HIPERLAN and Bluetooth. This demand has attracted significant interest in antenna designs that are preferable for such wireless communication systems. Many novel antenna structures have been proposed

for this purpose, and they operate in single or multiple bands [1-5]. Among them, microstrip patch and printed antennas have seen the natural favorites. They have inherent advantages of low profile, less weight, low cost, and ease of integration. Although early implementations of microstrip antennas suffered from small bandwidth, many techniques have been proposed recently [1,2] to improve their bandwidth. Two such techniques are the use of a thick substrate and slots cut strategically in the metallic patch. The probe-fed U-slot patch antenna [1] and its improved version, the E-shaped patch antenna [2], can be designed to provide a very good bandwidth.

Although the first generation wireless systems (e.g. IEEE 802.11b) were relatively slow compared with wired counterparts, latest wireless network standards provide a much faster bit rate with wireless convenience. In particular, wireless systems operating in the 5-6 GHz bands (e.g. IEEE 802.11a) have the ability to provide high-speed connectivity (> 50 Mb/s) in high-density environments. Although some current IEEE 802.11a systems operate only in the 5.15-5.35 GHz band, advanced implementations are emerging that make use of both the 5.725-5.825 GHz band and the 5.15-5.35 GHz band. Due to the availability of a wider frequency bandwidth and many more channels, such 802.11a systems are expected to provide superior robustness and bit error rate than 2.45 GHz systems (e.g. 802.11b, 802.11g) when used in busy environments where multiple wireless hot-spots may be operating concurrently.

In this chapter, E-shaped patch antennas are reviewed for use in low-profile wireless communication devices (e.g. PCMCIA wireless network adaptor cards) that operate in the 5-6 GHz frequency range. In particular, we consider devices as thin as 5mm. We outline the theory of the E-shaped patch antenna in the Section II. In the same section, we discuss a low-profile E-shaped patch antenna design on a small ground plane, which is suitable for compact, very thin communication devices. A design of a pair of E-shaped patches, for space diversity or for independent transmission and reception, is described in Section III. A more advanced diversity antenna, composed of two orthogonal E-shaped patch antennas for space and polarization diversity, is presented in Section IV and our conclusions are given in Section V.

## 2. SINGLE E-SHAPED PATCH ANTENNA

The general configuration of a probe-fed symmetrical E-shaped patch antenna is shown in Fig. 21-1. The patch is fed by a probe, attached to a coaxial line, through the ground plane. Ideally there is no physical substrate; the gap between the patch and the ground plane is filled by air. However, a low-loss foam material may be used to fill this gap, either fully or partly, for

mechanical stability. The parameters that characterize the antenna are the patch length and width (L, W), the height of the patch (H), the length of the middle wing ($L_S$), the widths of the three wings ($W_1$, $W_2$) and the position of the coaxial probe ($L_0$).

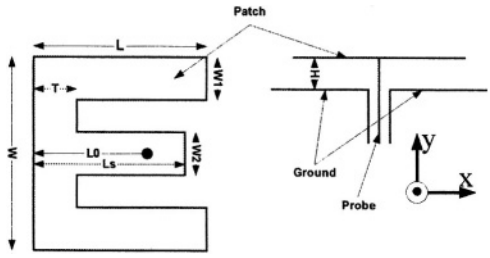



*Figure 21-1.* Configuration of a probe-fed symmetrical E-shaped patch antenna (from [6], © IEEE, reprinted with permission).

This antenna configuration has been previously investigated in [2] and [3]. The symmetrical E-shaped patch antenna has two resonant frequencies: the centre wing resonates at a higher frequency and the two side wings resonate at a lower frequency. Since the target bandwidth of the antenna is approximately 5-6 GHz (where the return loss should be < -10 dB), the two resonant frequencies should be selected to be around 5.25 GHz and 5.8 GHz as a starting point in the design process. The design process begins with the centre wing. Its resonance frequency, when it is connected to the two side wings, is much lower than when it is isolated. Hence, it is designed to resonate at about 6.7-6.9 GHz in isolation. Then the side wings, resonating around 5.8 GHz, are added. After adjusting the parameters of the antenna shown in Fig. 21-1, a reasonable design has been obtained with an antenna height of 5mm (Antenna 1). The predicted bandwidth of this design is much wider than the required bandwidth, indicating that thinner antennas can be designed to meet the requirements. Hence, this design process has been repeated for 4mm (Antenna 2) and 3.5mm (Antenna 3) height antennas. All designs were initially done using Ansoft Ensemble (SV) software that is very efficient, but they were tested or fine-tuned later using Ansoft HFSS 8.5 software, which can also model the effect of the finite ground plane. The parameters of the three antenna designs thus obtained are given in Table 21-1.

Fig. 21-2 shows the results of input reflection coefficient magnitude, $|S_{11}|$, from initial simulations (Ansoft Ensemble). All three designs, including the thin 3.5 mm antenna, cover the entire 5-6 GHz frequency band.

*Table 21-1.* The parameters of three E–shaped patch antennas (in mm).

| | L | W | W1 | W2 | L0 | Ls | T | H |
|---|---|---|---|---|---|---|---|---|
| Antenna 1 | 32.0 | 23.6 | 8.5 | 8.0 | 18.0 | 11.3 | 6.8 | 5.0 |
| Antenna 1 | 32.0 | 23.6 | 8.5 | 8.0 | 18.0 | 10.8 | 6.8 | 4.0 |
| Antenna 1 | 32.0 | 23.6 | 8.5 | 8.0 | 18.4 | 11.4 | 6.8 | 3.5 |

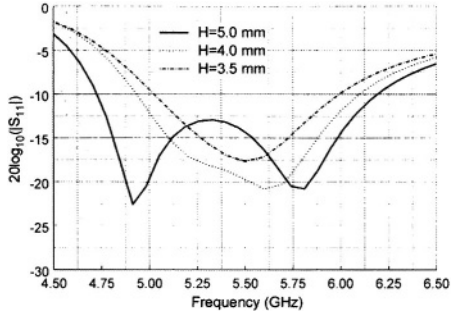



*Figure 21-2.* Computed input reflection of three E-patch antennas.

To verify the design process, the three antennas have been fabricated and the return loss has been measured with an HP 8720D network analyzer. The measured results of $|S_{11}|$ are given in Fig. 21-3. It can be seen that the experimental results agree quite well with the theoretical predictions, and the –10 dB bandwidth of the 3.5 mm antenna is wider than predicted.

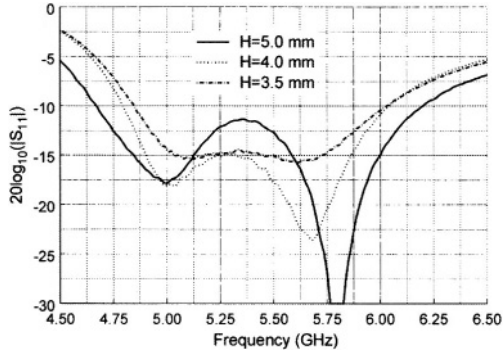



*Figure 21-3* Measured input reflection of three E-patch antennas.

The effect of the size of the ground plane on antenna bandwidth has also been studied. The ground plane sizes considered are $100\times100$ mm$^2$ (ground 1), $80\times80$ mm$^2$ (ground 2) and $60\times60$ mm$^2$ (ground 3). The measured $S_{11}$ with different ground planes are shown in Figs. 21-4 to 21-6. The bandwidths of the three antennas obtained from the measurements and

software simulations, respectively, are compared in Table 2 for various ground plane sizes and patch heights. It can be seen that the bandwidth increases with the height of the patch (H), while the size of the ground plane has only a little effect on the bandwidth, for the ground plane sizes considered [3].

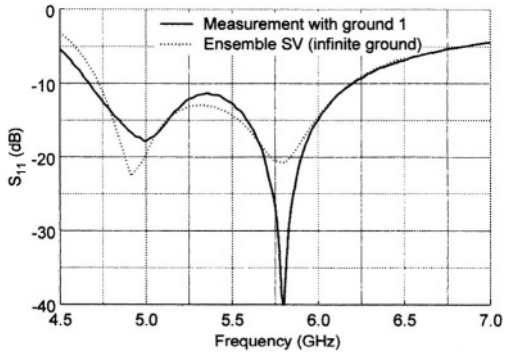



*Figure 21-4.* S11 of Antenna 1 on a 100×100 mm ground plane.

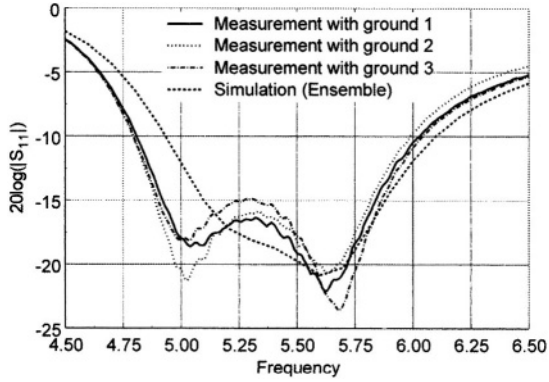



*Figure 21-5.* S11 of Antenna 2 on ground planes of size 100×100 mm (ground 1), 80×80 mm (ground 2) and 60×60 mm (ground 3).

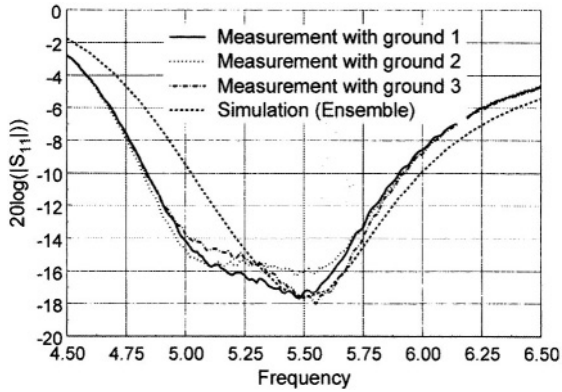



*Figure 21-6.* S11 of Antenna 3 on ground planes of size 100×100 mm (ground 1), 80×80 mm (ground 2) and 60×60 mm (ground 3).

*Table 21-2.* Comparison of the bandwidth of E-shaped patch antennas in Table 1.

| Bandwidth | Measurements | | | Theory (Ensemble, Infinite Ground) | | |
|---|---|---|---|---|---|---|
| | Ant. 1 | Ant. 2 | Ant. 3 | Ant. 1 | Ant. 2 | Ant. 3 |
| Ground 1 | 27.86% | 22.65% | 21.45% | | | |
| Ground 1 | | 21.7% | 21% | 26.38% | 20.72% | 17.8% |
| Ground 1 | | 22.2% | 19.9% | | | |

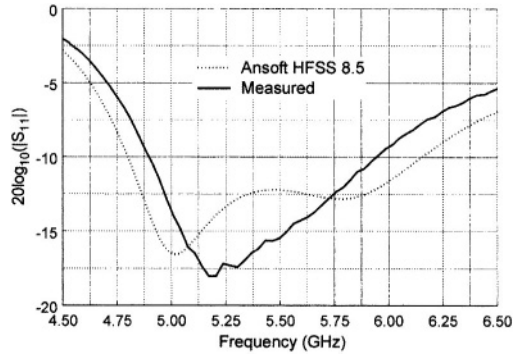



*Figure 7.* Measured and computed input refelection of a symmetrical E-shaped patch on a 54×35 mm ground plane (from [6], © IEEE, reprinted with permission).

Based on the above investigations, a single symmetrical E-shaped patch antenna suitable for compact wireless communication devices has been designed. The area of the ground is taken to be $54 \times 35$ mm$^2$, to correspond with a typical antenna extension of a PCMCIA card. The parameters of the antenna are: L=23.6 mm, W=32.0 mm, H=3.5 mm, $L_S$=18.4 mm, $L_0$=11.4

mm, $W_1$=8.5 mm, $W_2$=8.0 mm and T=6.8 mm. The inner and outer diameters of the 50Ω coaxial probe are 1.3 mm and 4.1 mm, respectively.

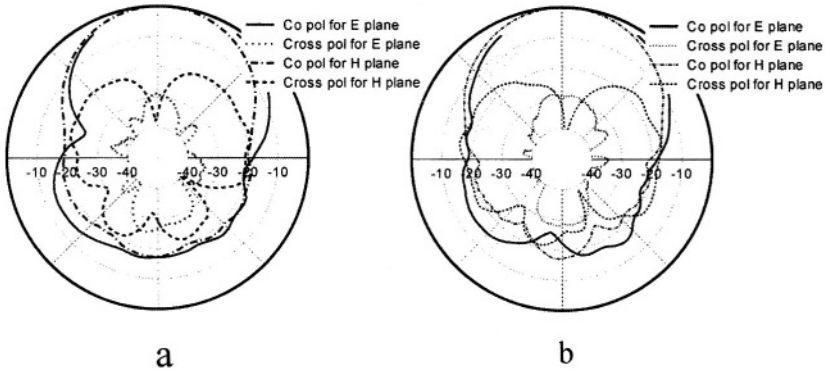



*Figure 21-8.* The measured radiation patterns of the E-shaped patch antenna: (a) at 5.25 GHz; (b) at 5.78 GHz (from [6], © IEEE, reprinted with permission).

This design has been achieved with Ansoft HFSS 8.5. The reflection $|S_{11}|$ from the simulation and the measurement is shown in Fig. 21-7. From the measured results, it is seen that $|S_{11}|$ is less than –10 dB in the frequency range of 4.8-5.9 GHz. The radiation patterns of the E-shaped patch antenna were measured and Fig. 21-8 shows the measured co-polarization and cross-polarization radiation patterns in the E and H planes at 5.25 GHz and 5.78 GHz. The half-power-beamwidth in the E plane is 65° at 5.25 GHz and 57.5° at 5.78 GHz. The H plane beamwidth is 55.5° at 5.25 GHz and 58.5° at 5.78 GHz. The measured gain in the frequency range 5-6 GHz is > 7 dBi and the measured cross-polarization levels are about 20 dB below the co-polarization levels.

## 3. DECOUPLED ANTENNA PAIR

Sometimes it is desirable to have two antennas on one wireless communication device. The two antennas may work independently - one for transmitting and the other for receiving – and avoids the need for a diplexer. In this case, good mutual coupling isolation is essential between the two antennas. In other cases, two antennas may used to achieve space and/or polarization diversity.

An antenna pair suitable for this purpose is shown in Fig. 21-9. It is composed of two E-shaped patches. They are aligned with the wings parallel

and each is fed by a coaxial probe. The antennas have the same dimensions as the antenna described in the previous section, i.e. L=23.1 mm, W=32.0 mm, H=3.5 mm, $L_S$=18.4 mm, $L_0$=11.0 mm, $W_1$=8.5 mm, $W_2$=8.0 mm and T=6.8 mm. The area of the ground plane is 37×70 $mm^2$. The separation between the two inputs (coaxial probes) is 35 mm. There is no material between the patches and ground plane other than the probe but this volume may be filled by a foam-type material for additional mechanical strength.

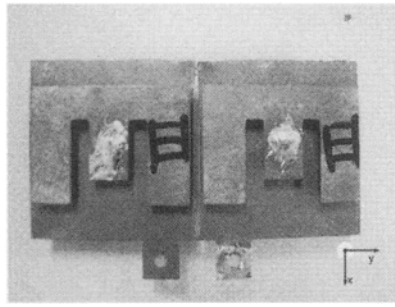


*Figure 21-9.* Decoupled pair of E-shaped patch antennas (from [6], © IEEE, reprinted with permission).

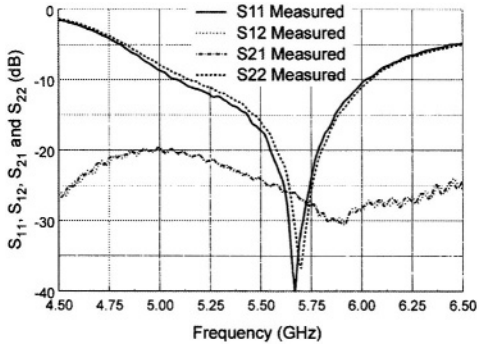



*Figure 21-10.* Scattering parameters of the decoupled E-shaped patch antenna pair (from [6], © IEEE, reprinted with permission).

This design has been tested using Ansoft HFSS 8.5. From the simulations it is found that good mutual coupling isolation (> 20 dB) could not be achieved with a standard continuous ground plane. To reduce mutual coupling, we introduced a slot in the ground, which is located between the two E-shaped patches, as can be seen in Fig. 21-9. With this slot, the mutual coupling is reduced significantly. Fig. 21-10 shows the two-port scattering

parameter magnitudes obtained from the measurements. $S_{11}$ and $S_{22}$ give the return loss at each antenna input and $S_{21}$ and $S_{12}$ indicate the mutual coupling between the two inputs. It can be seen that the input reflection is less than -10 dB over the frequency range of 5.15-5.95 GHz and the mutual coupling is also below –20 dB in the same bandwidth. The measured radiation patterns are plotted in Fig. 21-11. Fig. 21-11 (a) shows the patterns when one antenna is excited at 5.25 GHz (while the other antenna is simply match terminated) and Fig. 21-11(b) shows the patterns when the same antenna is excited but at 5.8 GHz. Due to symmetry, similar patterns are expected when the other antenna is excited. The measured gain is around 7 dBi throughout the operating bandwidth.

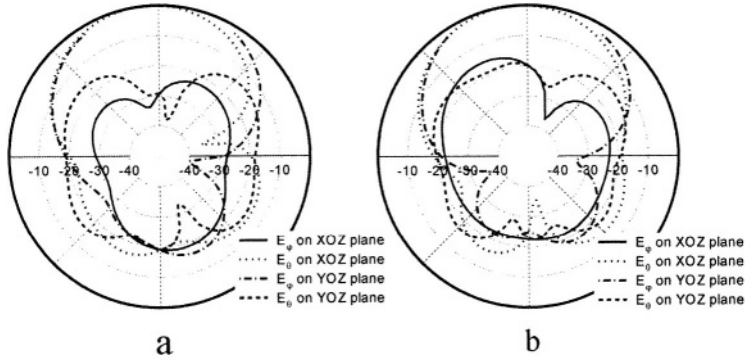



*Figure 21-11.* Radiation patterns of the decoupled E-shaped patch antenna pair. Antenna 1 is excited: (a) at 5.25 GHz; (b) at 5.78 GHz (from [6], © IEEE, reprinted with permission).

## 4. POLARISATION AND SPACE DIVERSITY ANTENNA

Signal fading due to multi-path interference is common in indoor environments and these effects can be reduced using antenna diversity techniques. The three common diversity techniques are space, pattern and polarization. In order to implement diversity in a wireless communication device, at least two antennas are required. A switching or signal combining circuit is also required between the antennas and the transceiver, to switch automatically between the two antennas or to combine the two signals to achieve optimum signal. The antenna pair described in the previous section can provide space diversity as they are separated in space by more than half

a wavelength apart. However, both antennas in that pair have identical radiation patterns and polarization characteristics, and hence, they do not provide any polarization or pattern diversity. Alternatively, an antenna pair that can provide both space and diversity is shown in Fig. 21-12. This antenna pair is composed of two E-shaped patch antennas, which are arranged orthogonal to each other in space. This makes the polarization of the two antennas different and complementary to each other. The ground plane area is $37\times64$ mm$^2$. The distance between the two coaxial probes is 35 mm. There is no substrate or other material between the patches and ground plane. The two antennas have different dimensions, as shown in Table 21-3.

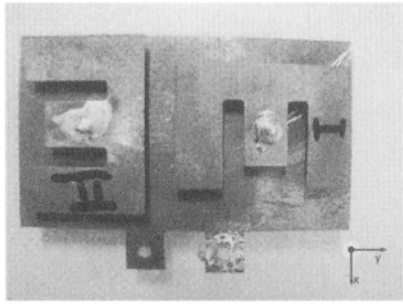


*Figure 21-12.* Diversity antenna composed of two orthogonal E-shaped patch antennas (from [6], © IEEE, reprinted with permission).

*Table 21-3.* The dimensions of the diversity antenna pair (in mm) (from [6], © IEEE, reprinted with permission).

| Ant. | L | W | H | Ls | $L_0$ | W1 | W2 | T |
|---|---|---|---|---|---|---|---|---|
| 1 | 22.6 | 32.0 | 3.5 | 18.4 | 11.5 | 8.5 | 8.0 | 6.8 |
| 2 | 22.1 | 32.0 | 3.5 | 18.4 | 11.1 | 8.5 | 9.0 | 6.8 |

The measured scattering parameters of the diversity antenna are shown in Fig. 21-13. It can be seen that the input reflection at each input ($S_{11}$ and $S_{22}$) is less than -10 dB over the frequency range of 5.15-5.95 GHz and the mutual coupling ($S_{21}$ and $S_{12}$) is below –20 dB in the same range. Figs. 14 shows the measured radiation patterns of the diversity antenna. Fig. 21-14(a) shows the patterns when only antenna 1 is excited at 5.25 GHz whereas Fig. 21-14(b) shows those when only antenna 2 is excited at 5.25 GHz. The far-field components $E_\varphi$ and $E_\theta$ on XOZ and YOZ planes are plotted in the two figures. It can be observed that, on XOZ plane, $E_\theta$ is the main (co-) polarization of antenna 1 but $E_\varphi$ is the main polarization of antenna 2. Similar orthogonality can be observed in the patterns on YOZ plane from the two antennas. In other words, the two antennas have complementary

polarizations. The measured gain in the 5-6 GHz frequency range is about 7 dBi.

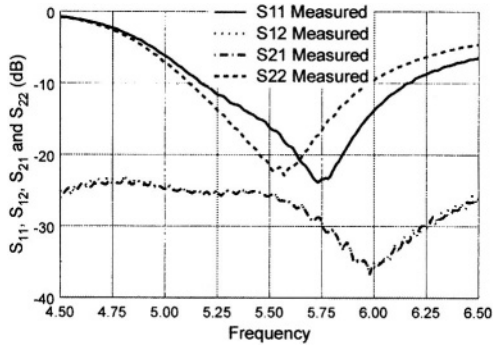



*Figure 21-13.* Scattering parameters of the diversity antenna pair (from [6], © IEEE, reprinted with permission).

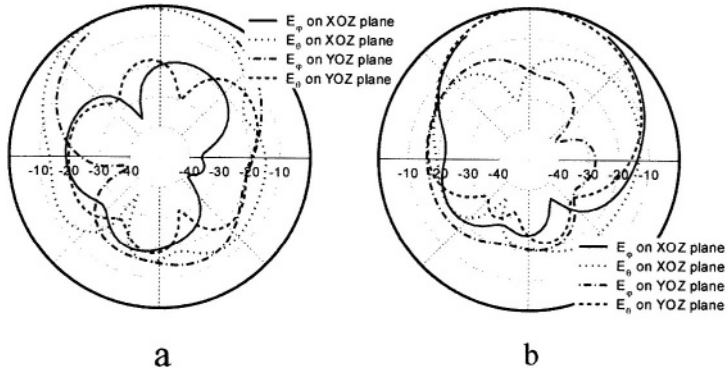



*Figure 21-14.* Radiation patterns when one antenna in the pair is excited at 5.25 GHz: (a) Port 1 is excited; (b) Port 2 is excited (from [6], © IEEE, reprinted with permission).

## 5. CONCLUSIONS

We have reviewed theoretical and experimental results of several broadband E-shaped patch antennas and diversity arrays, which are suitable for compact, thin wireless communication devices operating in the 5-6 GHz frequency range. Simulations and experiments indicate that the height of the

patch affects the bandwidth of an E-shaped patch antenna. However, our results clearly demonstrates that the antenna does not need to be thicker than 3.5 mm, in order to achieve good performance over both IEEE 802.11a band. In addition, two different arrays of E-shaped patch antennas were described, one providing space diversity and the other giving both space and polarization diversity. Experiments show that these antennas have a return loss > 10 dB in the two IEEE 802.11a wireless bands, i.e. 5.15-5.35 GHz and 5.725-5.825 GHz. In the case of two arrays, the isolation between the two antennas is > 20 dB in these frequency bands. In general this means that the correlation coefficient between the two antennas is very small. It also suggests that the two antennas may be used as independent transmitting and receiving antennas if necessary.

## 6. ACKNOWLEDGEMENT

The authors thank Dr Andrew Weily in Macquarie University and Mr Ken Smart from CSIRO for their help with the far field pattern measurements of the antennas. We also thank the CSIRO for providing access to their antenna range for pattern measurements. This research was supported by the Australian Research Council.

## REFERENCES

1. K. F. Lee, K. M. Luk, K. F. Tong, S. M. Shum, T. Huynh, and R. Q. Lee, "Experimental and Simulation Studies of the Coaxially Fed U-Slot Rectangular Patch Antenna", *Proc. Inst. Elec. Eng.,* pt. H, vol. 144, pp. 354-358, Oct. 1997
2. F. Yang, X.-X. Zhang, X. Ye, and Y. Rahmat-Samii, "Wide-Band E-Shaped Patch Antennas for Wireless Communications", *IEEE Trans. AP,* Vol. 49, No. 7, pp. 1094-1100, July 2001.
3. Y. Ge, K. P. Esselle, and T. S. Bird, "Broadband E-Shaped Patch Antennas for 5 - 6 GHz Wireless Computer Networks", *IEEE Antennas and Propagation Society (AP-S) International Symposium,* Columbus, OH, USA, June 2003.
4. T. Wu, S. Fang, and K. Wong, "A Printed Diversity Dual-Band Monopole Antenna for WLAN Operation in the 2.4- and 5.2-GHz Bands", *Microwave and Optical Technology Letters,* vol. 36, no. 6, pp. 436-439, March 20 2003.
5. H. Chuang, L. Kuo, C. Lin, and W. Chen, "A 2.4 GHz Polarization-Diversity Planar Printed Antenna for WLAN and Wireless Communication Systems", *2002 IEEE Antennas Propagat. Soc Int Symp Dig,* San Antonio, Texas, USA, pp. 76-79.
6. Y. Ge, K. P. Esselle, and T. S. Bird, "E-Shaped Patch Antennas for High-Speed Wireless Networks", accepted for publication in *IEEE Transactions on Antennas and Propagation,* Feb. 2005

# INDEX