USC–SIPI REPORT #313

A Precision Receiver for CDMA Communications Using Neural-Based Array Processors

by

Mithat C. Dogan and Jerry M. Mendel

January 1994

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 400
Los Angeles, CA 90089-2564 U.S.A.

Acknowledgment

I would like to express my sincerest gratitude to my research advisor, Professor Bing J. Sheu, Department of Electrical Engineering and Department of Biomedical Engineering, for his giudance and support throughout my Ph.D research work. I wish to extend my appreciation to Professor Theodore W. Berger, Department of Biological Engineering, Professor Murray Gershenzon, Department of Material Science, and Dr. Wai-Chi Fang, Jet Propulsion Laboratory for serving as my dissertation committee members. I would also like to thank Professor Edward K. Blum for also serving on my Qualifying Examination Committee.

I am very grateful to Professor Leonard M. Silverman, Dean of the Engineering School; Professor Hans H. Kuehl, Chairman of the Electrical Engineering - Electrophysics Department; Professor Robert A. Scholtz, Chairman of the Electrical Engineering - Systems Department; Professor Chrysostomos (Max) L. Nikias, Director of Integrated Media Systems Center, which is an Engineering Research Center of National Science Foundation; Ms. Ramona Gordon, Ms. Anna Fong, and Ms. Gloria Halfacre in the Electrical Engineering Program, for providing such a great research environment for my Ph.D. study at the University of Southern California. This research work conducted through connections with several research organizations including Integrated Media Systems Center (IMSC), Center for Neural Engineering (CNE) and Signal and Image Processing Institute (SIPI). Information exchange with Professor Leon O. Chua of Electrical Engineering and Computer Sciences Department at University of California, Berkeley, and Professor Tamas Roska of Computer and Automation Institute of the Hungarian Academy of Sciences, is very beneficial.

I would like to express my special appreciation for the 3-year fellowship support and encouragement from Ministry of Education, Taiwan. With these support and encouragement, I have the chance to come to USC for Ph.D study and obtain many precious experience and knowledge which change my life entirely.

Valuable discussions with graduate doctoral colleagues from VLSI Multimedia Laboratory were truly stimulating, including Dr. Sa Hyun Bang (now President of Standard Telecom Inc. in Santa Clara, CA), Dr. Robert C. Chang (now Associate Professor at National Chung Hsing University, Taiwan), Dr. Tony H. Wu (now project leader at Cirrus Logic Inc. Milpitas, CA), and Dr. Eric Y. Chou (now research scientist at HP ULSI Labs., Palo Alto, CA). Many Thanks to Steve H. Jen, Richard H. Tsai, Michelle Y. Wang, Wayne C. Young, Austin K. Cho, Alex Y. Park, Suat U. Ay, Vicky L. Zhang, and Andy C. Chu for friendly interaction.

Finally, I would like to express special thanks and appreciation to my parents, Guey-Chuen Chen and Yuh-Mei Her. I would like to thank my sisters Le-Lim Chan and Le-Chin Chen. I deeply appreciate my wife, Jessica Ming-Chieh Wu, for devoted support, encouragement, and love during my Ph.D. study, and my daughter, Angela I-Huey Chen, for her understanding and patience.

Contents

List Of Tables

List Of Figures

Abstract

With rapid advances of deep-submicron semiconductor technology and progress in communication systems, our lives enter a new Multimedia Era: communicating in any place and at any time as evidenced by the availability personal communication systems. Among all possible technologies, Code Division Multiple Access (CDMA) has been receiving more and more attention and the market of wireless communication with CDMA is booming since the year 1990.

CDMA is a spread spectrum technology. All of the CDMA users share the same bandwidth and their communication channels are separated by means of pseudorandom codes. The universal frequency reuse is crucial to the high spectral efficiency. To maintain high quality and high spectral efficiency in CDMA systems, the power difference of received signals should be as small as possible. In satellite communication, the high-power and low-power transmitters co-exist. In land communication, some users may be near the base-station and some may be far away. A 60 dB or more signal power difference is quite possible. This is called near-far problem. In 1986, S. Verdu illustrated in the theoretical derivation that the optimal near-far resistance detector could be achieved. However, no electronic implementation has been developed so far.

The one-dimensional compact neural network is very suitable for communication receivers in CDMA systems. Its architecture is based on a combination of the locally connected cellular neural network and the fully-connected Hopfield neural network. The compact neural network is a very efficient architecture for electronic implementation. It exhibits high degree of fault tolerance, high data throughput rate, and even low power consumption. By properly mapping the cost function of optimal near-far resistance detector onto the energy function of the compact neural network and applying the innovative hardware annealing technique, multiuser detector with optimized solution is achieved. Extensive computer simulation using MATLAB codes has been conducted. Satisfactory results are obtained. The neural network-based CDMA receiver design is a very convincing solution in future personal communication systems.

Chapter 1

Introduction

1.1 Personal Communication Ssytem (PCS)

With rapid advances in the technological development of wireless communications, the increasing significance of data and message communication and the regulatory and political climates over the past decade, wireless personal communication has become the fatest growing segemnt of telecommunication. From the viewpoints of portability and mobility, the communication coverage can be classified as: within a house or a building by using cordless phone or wireless local area networks (WLANs); within a small community or a city by extented cordles phone, WLANs and cellular phones; within a country or state by pagers, cellular phones, pagings, satellite-based wireless communication; and throughout the world by satellite-based wireless communication. Two-way voices, data transmission, messaging, paging, etc., are most popular applications in nowadays wireless personal communication. Two-way voice communication is real time. Wired telephones, cordless phones, and cellular mobile telephones are used. Paging alerts the paged party and transmits the number of a calling party or some alpha-numeric messages. It only provides only one-way communication. Messaging is not real time but can be used to transmit, store or retrieve messages. It includes electronic mail, voice mail, and facsimile. In general, technologies and systems providing wireless communications services can be categorized into seven distinct groups [1]: cordless telephones, cellular mobile radio systems, wide area wireless data systems, high-speed wireless local-area networks, paging/messaging systems, and satellite-based mobile systems and personal communication system (PCS).

Though the current technologies will develop continuously for enhancing their services, there are strong demands and factors for much more advanced technologies:

- rapid advances in semiconductor components;
- advanced developments in intelligent networks, network management, and service types;
- explosive growth of the number of wireless telephone subscribers;
- data privacy;
- requirement for hand-held communicators with multiple applications;
- pressure to integrate different technologies including messaging, paging, cordless phones, cellular phones, wired phones; and
- demand for wider scope and sophistication of multimedia sevices.

A vision of the ubiquitous telecommunication services known as the PCS emerged. In Europe, the Research into Advanced Communications in Europe (RACE) program was launched in 1988 to develop the third-generation mobile communication systems, Universal Mobile Communication System (UMTS) and

Mobile Broadband System (MBS). CDMA and TDMA are developed in RACE. The RACE program has two phases [2]. RACE I concentrated on system engineering, outline specifications, and key technologies. RACE II is concentrating on system integration and the prototyping of new services and applications. Table 1 lists the RACE II mobile projects [3].

Table 1 Race II mobile projects

Main Study Topic	**Project**
Radio access - TDMA	ATDMA
Radio access - CDMA	CODIT
Network principless	MONET
Low bit rate video coding	MAVT
Satellite integration	SAINT
Smart antennas	TSUNAMI
Broadband mobile	MBS

In USA, Department of Defense is making all out efforts to establish new communication systems for the 21th century battlefields. Robust information systems and rapid delopyment will be crucial factors for successful military operation in the highly mobile environment. Current systems have their limitations in supporting mobile operation in the presence of sporadic connectivity and variable bandwidth. Current commercial and military systems are primarily to provide voice communication in mobile environment. Most of the networks are with large immobile infrastructure. They are not suitable for rapid delopyment. The generic object of

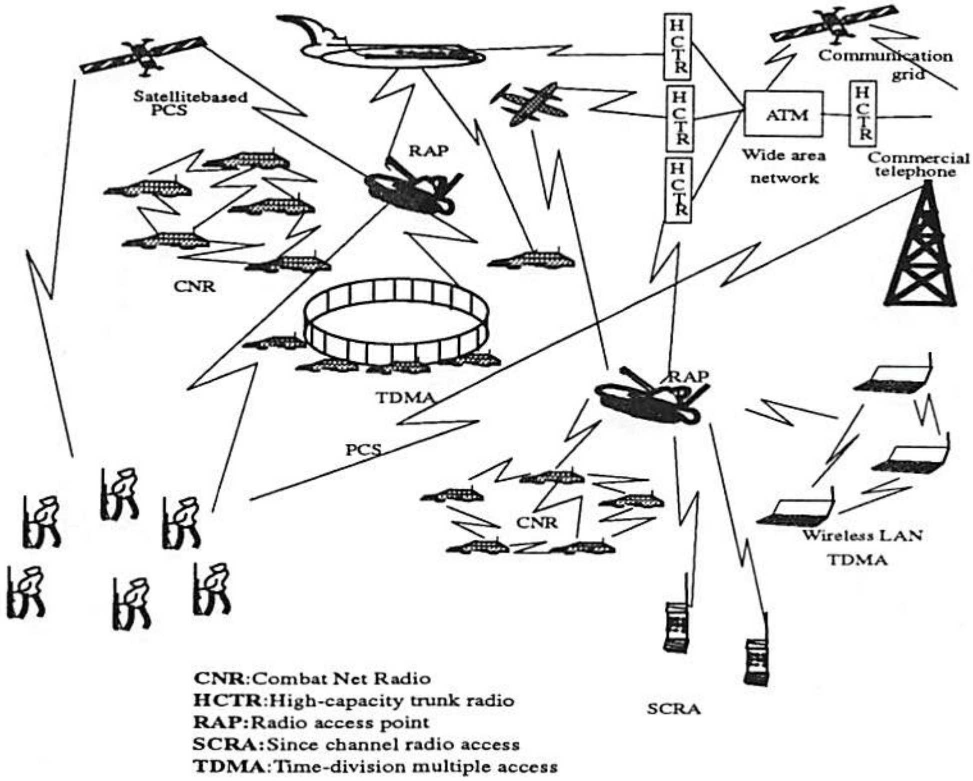


Figure 1.1: Future digital battlefield communication elements.

the efforts is to progress toward secure, seamless, theater-wide, multimedia communication for tactical users. Satellite-based PCS, directed broadcast systems (DBSs), terrestrial PCS, etc, are major systems. Future digital battlefield will have a mix of terrestrial and space-based communication to handle voice, data, and imagery. Fig. 1.1 shows the digital battlefield communication architecture elements [4]. Therefore, Defense Advanced Research Projects Agency (DARPA) launched a program called Global Mobile Information Systems (GloMo) to satisfy these requirements. Apparently, PCS plays a very important role.

PCS has become one of the hottest topics in the telecommunication industry since 1991. As stated above, PCS is not restricted to one technology, one system and/or one service; rather, it includes many different kinds of technologies, systems, and services. According to FCC's description, PCS encompasses a broad range of new radio communication service that will free individuals from the limitations of the wireline public switched telephone network and will enable individuals to communicate when they are from their home or office telephone lines [5]. Thus, FCC defined PCS as "radio communications that encompaes mobile and ancillary fixed communications that provide services to individuals and businesses and can be integrated with a variety of competing networks [6]. Table 2 lists the PCS air interface standards [7].

No matter for defense purposes or commercial applications, we are pushed to reach the new frontier of telecommunication: location-independent communication during the last decade of the twentieth century and the first decade of the twenty-first century. The wireless personal communicator is as common as the wireline telephone used to be. By using just one telephone number, a person can receive many formats of information: telephone, fax, and data. It provides reliable and affordable communication, anywhere and anytime; in the aircraft, carrier, submarine, office, car, or on the mountains. To bring this vision to fruition, two major improvements should be made: deploying enough satellites in the Earth orbits and improving the current state of wireless technology.

1.2 The CDMA Revolution

In PCS, two major techniques are applied: Code Division Multiple Access (CDMA) and Time Division Multiple Access (TDMA). In USA, digital cellular communication system IS-54 deployed in 1993 is based on TDMA; another system IS-95 deployed in 1995 is based on CDMA. In Europe, Global System for

Table 2 PCS air interface standard

Parameter	TAG-1	TAG-2	TAG-3	TAG-4	TAG-5	TAG-6	TAG-7
Heritage	New	IS-95 based	PACS	IS-136-based	DCS-based	DCT based	New
Access method	CDMA/ TDMA/ FDMA	DS-CDMA	TDM/ TDMA	TDM/ TDMA	TDMA	TDMA	D-CDMA
Duplex method	TDD	FDD	FDD	FDD	FDD	TDD	FDD
Bandwidth	5 MHz	1.25 MHz	300 KHz	30 KHz	200 KHz	1.25 MHz	5 MHz
Bit rate (kbit/s)	8 kbits/s	8 and 1.3	32	7.95	13	32	32
Voice channel per carrier	32	20	8	3	8	12	64
Modulation	QCPM	OQPSK/ QPSK	π/4 D-QPSK	π/4 D-QPSK	GMSK	π/4 D-QPSK	QPSK
Error control (voice)	None	FEC	None	FEC	FEC	None	FEC
Frequency reuse (N)	3	1	16 x 1	7 x 3	7 x 1 and 3 x 3	9	1
Max avg subscriber power	10 mW	200 mW	25 mW	200 mW	125 mW	20.8 mW	200 mW
Time frame length	20 ms	20 ms	2.5 ms	40 ms	4.615 ms	10 ms	-_
Time slot length	625 us	-	312.5 us	6.7 us	577 us	417 us	-
End-to-end speech delay	80 ms	50 ms	9 ms	110 ms	90 ms	10 ms	9 ms
Equalizer	No	No	No	Yes	Yes	No	No
Vocoder	CELP ADPCM	Variable-rate QCLP	ADPCM	VSELP	RPE-LTP	ADPCM	ADPCM

Mobile Telecommunications (GSM) is based on TDMA. GSM was first deployed in Germany in 1992. IS-54 and GSM can be catorized as the second generation of cellular communication systems. Though TDMA plays an important role in the second generation of digital cellular systems, CDMA is receiving more and more attention and the development commercial market is booming.

During the late 1980's and the early 1990's, rapid growth of the mobile subscribers made strong demand on system capacity and cost-effective systems for cellular and PCS. It encouraged engineers to consider the CDMA spread spectrum technology for commercial applications. CDMA changes the nature of the subscriber station from a predominately analog device to a predominately digital device. Old-fashioned radio receivers separate stations or channels by filtering in the frequency domain. CDMA receivers do not eliminate analog processing entirely, but they separate communication channels by means of a pseudo-random modulation that is applied and removed in the digital domain, not on the basis of frequency. Multiple users occupy the same frequency band. This universal frequency reuse is crucial to the very high spectral efficiency that is the hallmark of CDMA. CDMA is altering the face of cellular and PCS communication by:

- Dramatically improving the telephone traffic (Erlang) capacity due to universal frequency reuse;
- Dramatically improving the voice quality and eliminating the audible effects of multipath fading;
- Reducing the incidence of dropped calls due to handoff failures;
- Providing reliable transport mechanism for data communication, such as facsimile and internet traffic;

- Reducing the number of sites needed to support any given amount of traffic;
- Simplifying site selection;
- Reducing deployment and operating costs because fewer cell sites are needed;
- Reducing average transmitted power;
- Reducing interference to other electronic devices; and
- Reducing potential health risk.

Chief among these advantages stated above is universal frequency reuse. In TDMA or FDMA, frequency planning has become a key issue in the current scenario, with exceedingly high growth rates in many countries which compel operators to re-configure networks virtually on a monthly basis. Therefore, the search for smart techniques, which may considerably alleviate planning efforts becomes extremely important for operators in a competitive market [8]. In CDMA, all users occupy a common frequency sprctrum allocation. It not only increases the efficiency of spectrum usage, but also eliminate the complex work of planning for different frequencies. The key factor for such a novel performance is the use of noise-like carrier waves, as was first suggested decades ago by Claude Shannon [9].

The CDMA system has been adopted by the Telecommunication Industry Association TR-45 committee as TIA/EIA IS-95 standard for cellular and by the Alliance for Telecommunications Industry Solutions committee T1P1 and TIA-TR46 joint standard J-STD-008 for PCS. Today, several equipment manufacturs offer CDMA systems for PCS applications [10].

In USA, Pentagon's GPS, IRIDIUM from Motorola Inc., Odessey from TRW's inc., ELLIPSAT from Ellipsat Inc., GLOBAL STAR from Qualcomm Inc., etc, all make use of CDMA for efficient and high quality communication. Table 3 lists some low Earth orbit (LEO) mobile satellite communication systems proposed to the 1992 International Telecommunication Union World Administrative Radio Conference [11].

Table 3 Low Earth orbit (LEO) mobile satellite communication systems proposed to the Telecomm. Union World Administrative Radio Conference

Characteristics of proposed systems	IRIDIUM (Motorola)	ODYSSEY (TRW)	ELLIPSAT (Ellipsat)	GLOBAL (Loral and Qualcomm)	ARIES (CCI (Constellation Comm.)
Satellite no.	77	12	6	24	48
Class	LEO	MEO	LEO	LEO	LEO
Lifetime (yr.)	5	10	3	7.5	5
Orbit attitude (km)	755	10,6000	2903/426	1390	1000
Orientation	Circular	Circular	Elliptical	Elliptical	Circular
Initial geographical coverage	Global	CONUS, offshore United States, Europe, Asia-Pacific region	CONUS, offshore United States	CONUS	CONUS, offshore United States
CDMA	CDMA	CDMA	CDMA	CDMA	CDMA

Fig. 1.2 shows the global personal communication system (PCS) mobile communication vision by TRW's ODESSEY. Fig. 1.3 shows NASA's Advanced Communications Technology Satellite (ACTS) broadband aeronautical experiment setup [11]. The development of CDMA is booming.

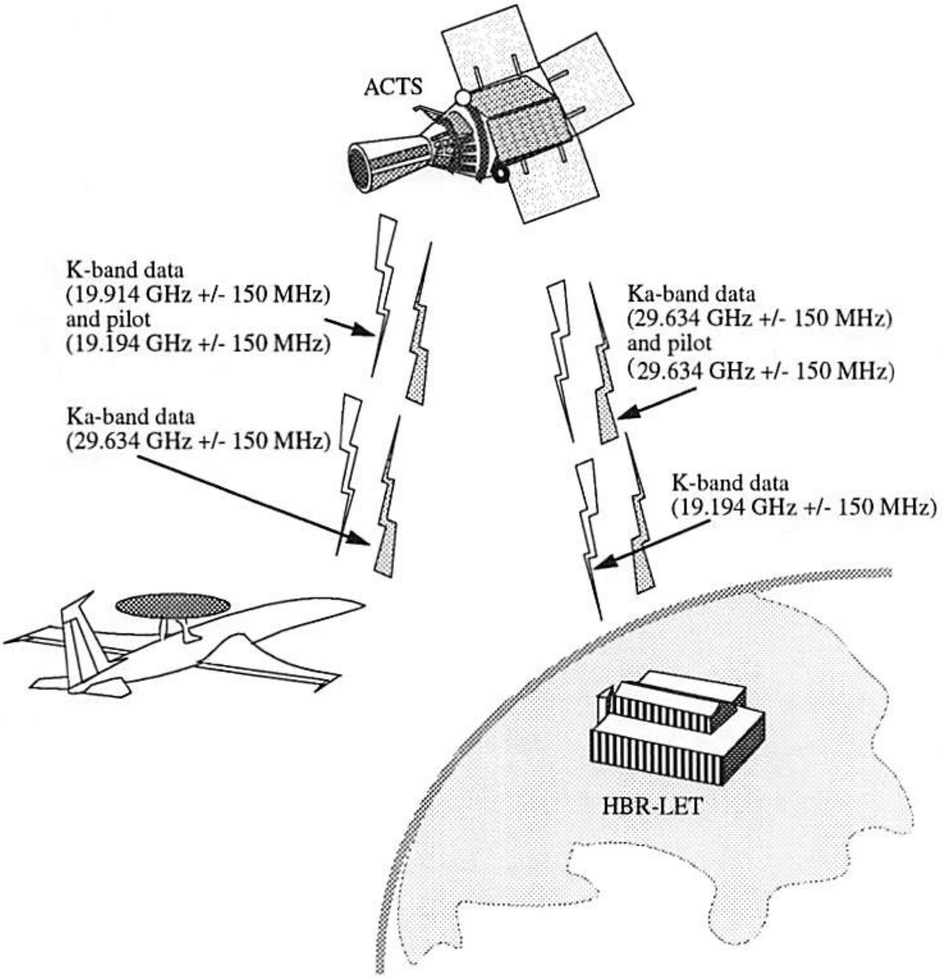


Figure 1.2: Global personal communications system-based mobile communication vision by a constellation of medium attitude orbit (MEO) satellites of TRW Inc. known as ODYSSEY [11].

1.3 Detection Problems in Communication Receivers

No matter in CDMA system or other systems, it is very important to have excellent detection techniques for high performance communication. In CDMA, one important detection issue is near-far problem. In other communication systems,

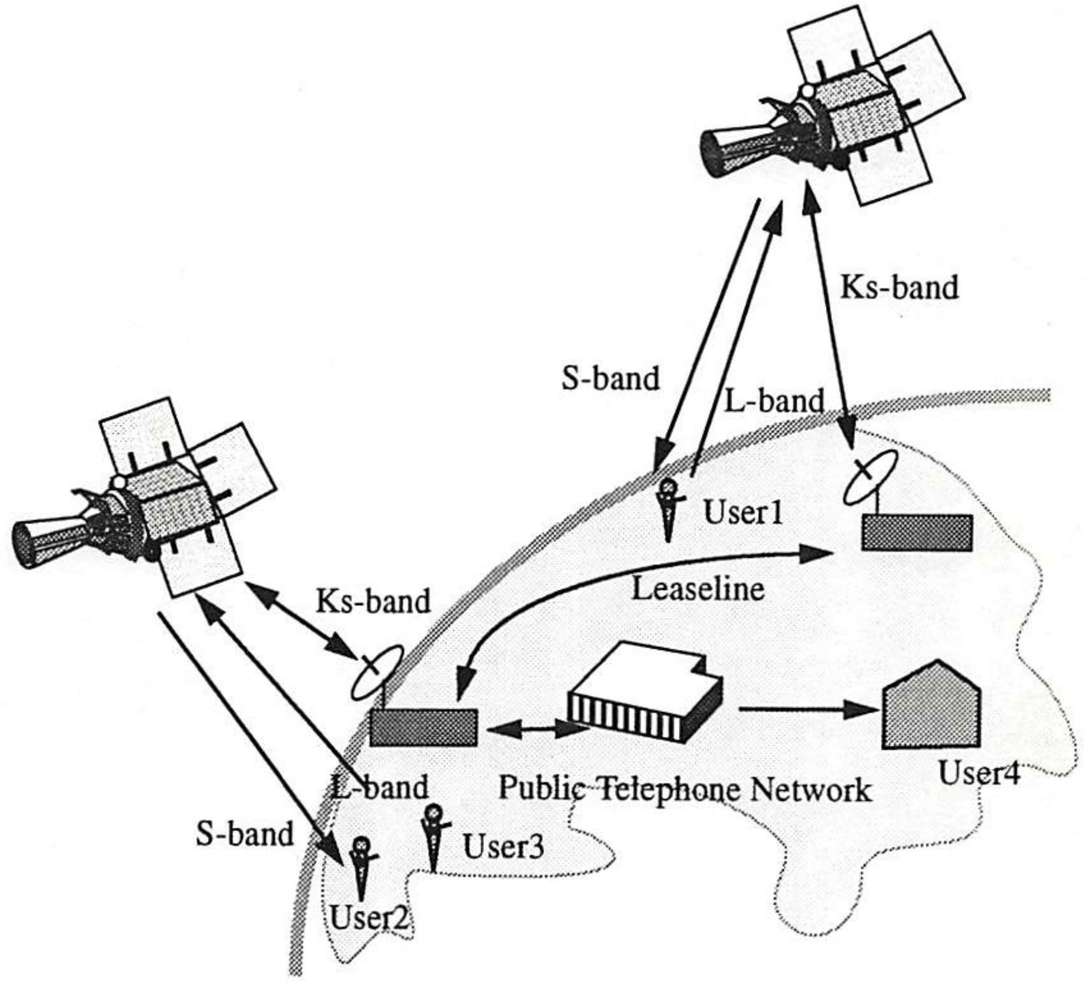


Figure 1.3: ACTS broadband aeronautical experiment setup. (From Abbe et al., 1993 [11].)

intersymbol interference (ISI) is also an important problem. These two problems will be briefly reviewed below.

1.3.1 Near-Far Problem

To maintain high quality and high spectral efficiency in CDMA systems, controlling signal power of users is very important. In an environment where propogation law for the intensity decay of signals is R^{-4}, the total dynamic range of path loss is on the order of 80 dB. Here, R is the distance. With a typical link budget for an IS-95A system, this means that the mobile transmitter must vary its power from about 2.5 nW to 0.25 W. In addition to the gross path loss dependence on distance, the loss may also vary rapidly due to multipath induced Rayleigh fading. In satellite communication, the high-power and low-power transmitters co-exist. In ground communication, some users may be near the base-station and some may be far away. A 60 dB or more signal power difference at the base-station for two mobiles is quite possible. When an unwanted user's received signal is much larger than the received signal power contributed by the desired user, the perfromance of CDMA is seriously impaired in the radio environment. This is called near-far problem and is a major technical problem in CDMA. In 1986, S. Verdu [12] showed that the optimal near-far resistant detector could be achieved by minimizing an integer quadratic object function. It means multiuser detection in CDMA can be converted into an optimization problem.

1.3.2 Intersymbol Interference (ISI)

Intersymbol interference is mainly caused by the suppression of interchannel interference and multipath propagation. The function to reduce or remove ISI is called equalization. For clear explanation, biphase shift keying (BPSK) with non-return-to-zero (NRZ) pulse is used. In frequency domain, BPSK with NRZ pulse has tails extending through out the frequency range, i.e., from $f = -\infty$ to $f = +\infty$. The tails will interfere with the neighboring channels. It is called interchannel interference. Since efficient spectrum utilization is extremely important, the Federal Communication Committee (FCC) and CCITT require that the side-lobes produced in BPSK be reduced below certain specific levels. To accomplish this requirement, a filter is employed to restrict the bandwidth allowed to the NRZ baseband signal. Thus the signal is distorted and there is a partial overlap of a bit (symbol) and its adjacent bits in a single channel. This overlap is called intersymbol interference (ISI). Any unfiltered baseband signal may cause interchannel interference. If it is band-limited by passing through a filter, the interchannel interference can be reduced at the expense of ISI. Fig. 1.4 shows the example of intersymbol intereference.

1.4 Biologically Inspired Compact Neural Networks

During the past several years, quite a few computing paradigms and architectures based on artificial neural networks were reported. Research results demonstrate that neural networks are very promising due to their capabilities in modeling and solving many complex scientific and engineering problems hardly approachable by traditional methods such as statistical pattern recognition, and conventional artificial intelligence. With the superior performance, neural networks are widely adopted for use in a variety of industrial, scientific, and commercial applications which range from signal processing, communication, economic tendency prediction, to resource scheduling [13].

One of most attractive advantages of neural networks is the efficient architectures for hardware implementation of these novel algorithms by microelectronic technology. When implemented in microelectronic hardware, neural networks exhibit high degree of fault tolerance to system damage, high data throughput rate due to their ability of parallel data processing, and even low power consumption. Therefore, development of advanced communication receivers by employing the neural network paradigm is one important research topic.

Generally speaking, the architectures of artificial neural networks can be classified into three categories:

- feedforward (multilayer) network,
- feedback & recurrent networks, and
- cellular & compact networks.

Fig. 1.5 shows the block diagrams of basic schemes of three kinds of neural networks.

In the feedforward neural network, each neuron may receive an input from the external environment and/or from neurons in the preceeding layer, but no feedback loop exists; while feeaback neural networks contain the feedback loop(s). The 2-dimensional cellular or the 1-dimensional compact neural networks [14] [15] are similar to the cellular automata and made of a massive aggregate of regularly spaced circuit cells, which interact with each other through the nearest neighbors. The cellular neural networks are suitable for the high-speed parallel image processing, especially for pattern recognition. Their compact characteristic is also very attractive for application-specific integrated circuit (ASIC) design. In communication systems, equalization or detection function in receivers can be viewed as operation of pattern recognition. Therefore, the 1-dimensional compact neural networks are well suited to implement high-speed communication receivers. The 1-dimensional compact neural network was biologically inspired. It was developed by Sheu, et al. [16] [17] and is a recurrent network. It is different from Hopfield neural network but very similar to the celular neural network.

In this dissertation, result on the biologically inspired compact neural network for implementation of the optimal multiuser detector [17] for CDMA system is described. A detailed analysis of employing this neural network to build the maximum likelihood sequence estimation (MLSE) detector for GSM system [16] is also addressed.

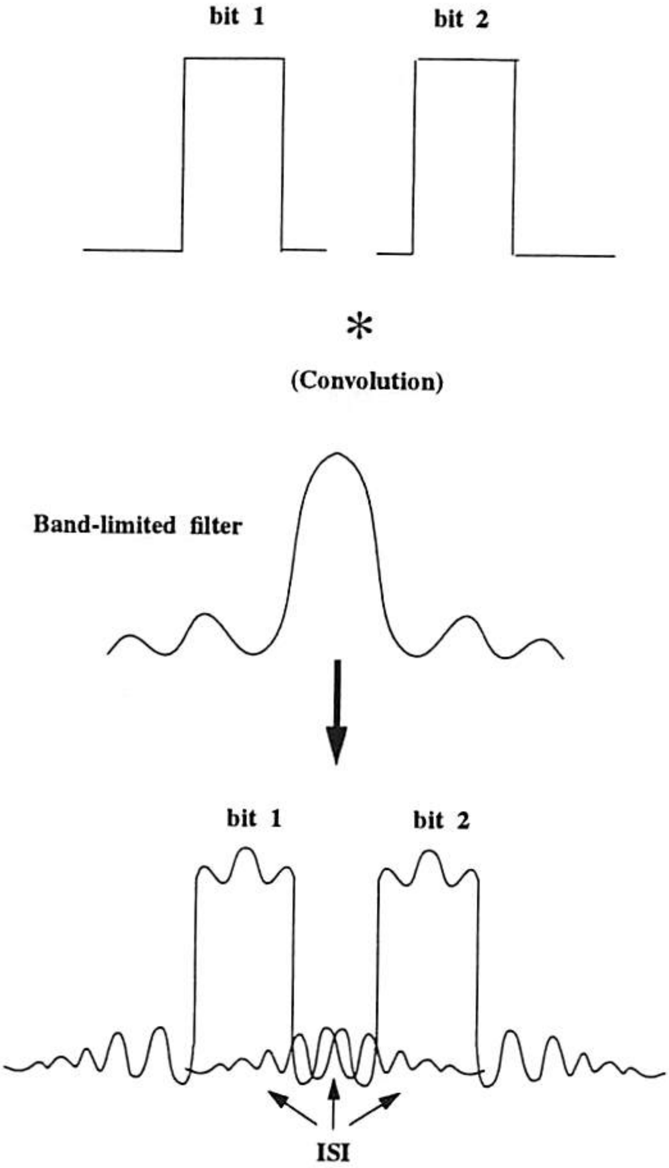
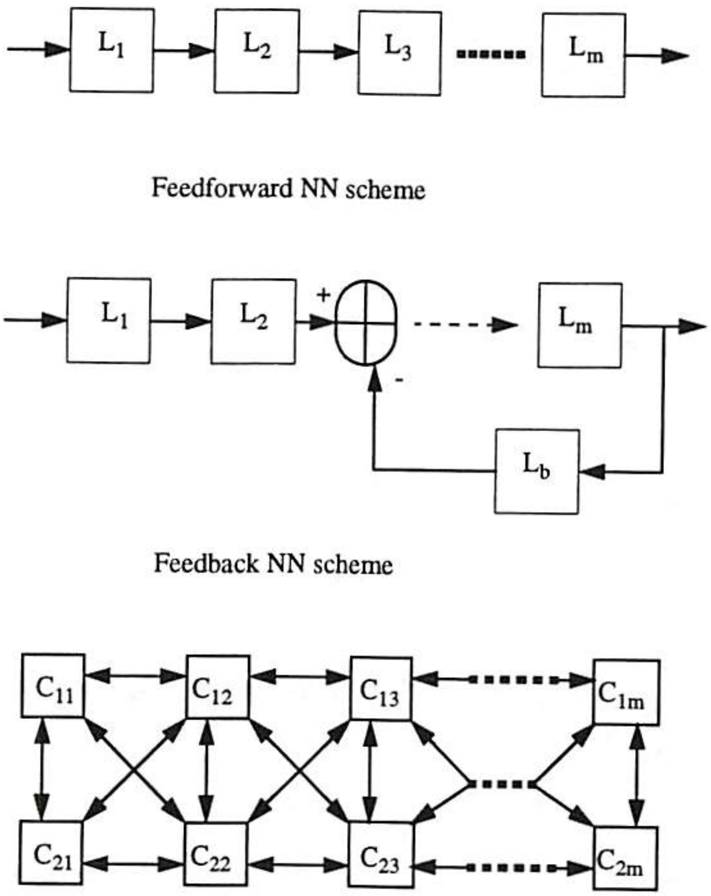


Figure 1.4: Example of intersymbol intereference.



Feedforward NN scheme

Feedback NN scheme

Cellular (Compact) NN scheme

Figure 1.5: Examples of three neural network architectures.

Reference List

[1] D. C. Cox, "Wireless personal communications: What is it?" *IEEE Personal Communications*, vol. 2, no. 2, pp. 20-35, Apr. 1995.

[2] RACE Annual Report, DGXIII-B, European Commission, Feb. 1994.

[3] J. S. Dasilva and B. E. Fernandes, "The European Research Program for Advanced Mobile Systems," *IEEE Personal Communications*, vol. 2, no. 1, pp. 14-19, Feb. 1995.

[4] B. M. Leiner, et al., "Goals and challenges of the DARPA GloMo program," *IEEE Personal Communications*, vol. 3, no. 6, pp. 34-43, Dec. 1996.

[5] Memorandum Opinion and Order, GEN Docket No. 90-314, June 9, 1994.

[6] Second Report and Order, ET Docket No. 92-9, 7 FCC Rcd 6886 at App. A, 1992.

[7] L. Goldberg, "PSC: Technology with fractured standards," *Electronic Design Report*, pp 65-78, Feb. 1995.

[8] M. Frullone, et al., "Advanced planning criteria for cellular systems,"*IEEE Personal Communications*, vol 3, no. 6, pp. 10-15, Dec. 1996.

[9] R. L. Peterson, R. E. Ziemer, D. E. Borth, *Introduction to Spread Spectrum Communications,* Prentice Hall PTR: Upper Saddle River, NJ, 1995.

[10] V. K. Garg, et al., *Applications of CDMA in Wireless/Personal Communications,* Prentice Hall PTR: Upper Saddle River, NJ, 1997.

[11] K. Feher, *Wireless Digital Communications,* Prentice Hall PTR: Upper Saddle River, NJ, 1997.

[12] S. Verdu, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. on Infonation Theory,* vol. IT-32, no. 1, pp. 85-96, Jan. 1986.

[13] B. J. Sheu and J. Choi, *Neural Information Processing and VLSI*, Kluwer Academic Publishers: Boston MA, 1995.

[14] L. O. Chua, L. Yang, "Cellular neural networks: Theory," *IEEE Trans. on Circuits and Systems,* vol. 35, pp. 1257-1272, Oct. 1988.

[15] L. O. Chua, L. Yang, "Cellular neural networks: Application," *IEEE Trans. on Circuits and Systems,* vol. 35, pp. 1273-1290, Oct. 1988.

[16] S. H. Bang, B. J. Sheu, "A neural network for detection of signals in communication," *IEEE Trans. on Circuit and Systems, Part I,* vol. 43, pp. 644-655, Aug. 1996.

[17] D. C. Chen, B. J. Sheu, W. Young, "A compact neural network based CDMA receiver," *IEEE Trans. on Circuit and Systems for Video Technology,* vol. 3, no. 4, Aug. 1997.

Chapter 2

Cellular and Compact Neural Networks

Many complex scientific problems can be formulated with a regular 1-D, 2-D and 3-D grid. Direct interaction between the signals on various grid points is allowed within a finite local neighborhood, which is sometimes called the receptive field. The original cellular neural network (CNN) paradigm was first proposed by Chua and Yang in 1988 [1] [2]. Later, Sheu et al. developed the 1-dimensional compact neural networks for communication receivers [7] [4]. The two most fundamental ingredients of the cellular/compact neural network paradigms are: the use of analog processing cells with continuous signal values, and local interaction within a finite radius. Many results on the algorithm development, VLSI implementation of cellular neural network systems were reported in the first four IEEE International Workshops on Cellular Neural Networks and Their Applications (Budapest, Hungary, 1990; Munich, Germany, 1992; Rome, Italy, 1994, Seville, Spain, 1996).

Due to its regular structure and parallelism, a 10×10 mm^2 cellular neural network microchip in a 0.5 μm CMOS technology can achieve the equivalence of about 1 tera operations per second. The cellular neural network architecture matches well with the paradigm that biologists have been seeking for many years

[5]. It provides a unifying model of many complex neural network architecture, especially for various forms of sensory modality.

The extented cellular neural networks can be viewed as cellular nonlinear networks. A cellular neural network has many important features [5]:

- 2-, 3-, or n-dimensional array of
- mainly indicated dynamic cells, which satisfies two properties:

  (1) interactions are local within a finite radius r, and

  (2) all state variables are continuous valued signals.

A weighting template specifies the interaction between each cell and its neighborhood cells in terms of their input, state, and output variables.

Each cell is identified by 2, 3, or n integers, $(i, j, \cdots, n)$. The time variable t may be continuous or discrete. The weighting template may be a linear or a nonlinear function of the state, input, and output variables of each cell. It could contain time-delay or time-varying coefficients. The dynamic systems may be perturbed by some noise sources of known statistics.

The heat equation, which is a typical partial differential equation, can be mapped onto a cellular neural network as reported in [1]. If a capacitor is added to the output node, wave-type equations can also be processed by a cellular neural network [6]. At equilibrium, the Laplace equation can be effectively handled [5]. Hence, the cellular neural network can be used to solve all three basic types of PDEs: the diffusion equation, the Laplace equation, and the wave equation.

2.1 Basic Theory and Computation Paradigm

2.1.1 Genaral Architecture

A cellular neural network is a continuous- or discrete-time artificial neural network that features a multi-dimensional array of neuron cells and local interconnections among the cells. The basic cellular neural network proposed by Chua and Yang [1] in 1988 is a continuous-time network in the form of an n-by-m rectangular-grid array where n and m are the numbers of rows and columns, respectively. Each cell in a cellular neural network corresponds to an element of the array. However, the geometry of the array needs not to be rectangular and can be such shapes as triangle or hexagon.

A multiple of arrays can be cascaded with an appropriate interconnect structure to construct a multi-layered cellular neural network. The r-th neighborghood cells consists of $C(k,l), 1 \leq k \leq n, 1 \leq l \leq m$, for which $|k-i| \leq r$ and $|l-j| \leq r$. The cell $C(i,j)$ has the direct interconnections with $N_r(i,j)$ through two kinds of weights, i.e., the feedback weights $A(k,l;i,j)$ and $A(i,j;k,l)$ and feedforward weights $B(k,l;i,j)$ and $B(i,j;k,l)$, where the index pair (k,l;i,j) represents the direction of signal from $C(i,j)$ to $C(k,l)$. The cell $C(i,j)$ communicates directly with its neighborhood cells $C(k,l) \in N_r(i,j)$. Since the cells $C(k,l)$ have their own neighborhood cells too, they also communicate with all other cells indirectly. Fig. 2.1(a) shows an n-by-m cellular neural network with $r = 1$. The cells filled with dashed lines represent the neighborhood cells $N_1(i,j)$ of $C(i,j)$, including $C(i,j)$ itself.

The block diagram of a cell $C(i,j)$ is shown in Fig. 2.1(b). The external input to the cell is denoted by $v_{uij}(t)$, and typically assumed to be constant $v_{uij}(t) = v_{uij}$

over an operation interval $0 \leq t < T$. The input is connected to $N_r(i,j)$ through the feedforward weights $B(i,j;k,l)$s. The output of the cell, denoted by v_{uij}, is coupled to the neighborhood cells $C(k,l) \in N_r(i,j)$ through the feedback weights $A(i,j;k,l)$s. Therefore, the input signals consist of the weighted sum of feedforward inputs and weighted sum of feedback inputs. In addition, a constant bias term is added to the cell. Fig. 2.2 shows the biologically inspired 1-dimensional compact neural network most suitable for the communication systems including the GSM and CDMA sysyems. If the weights represent the transconductance

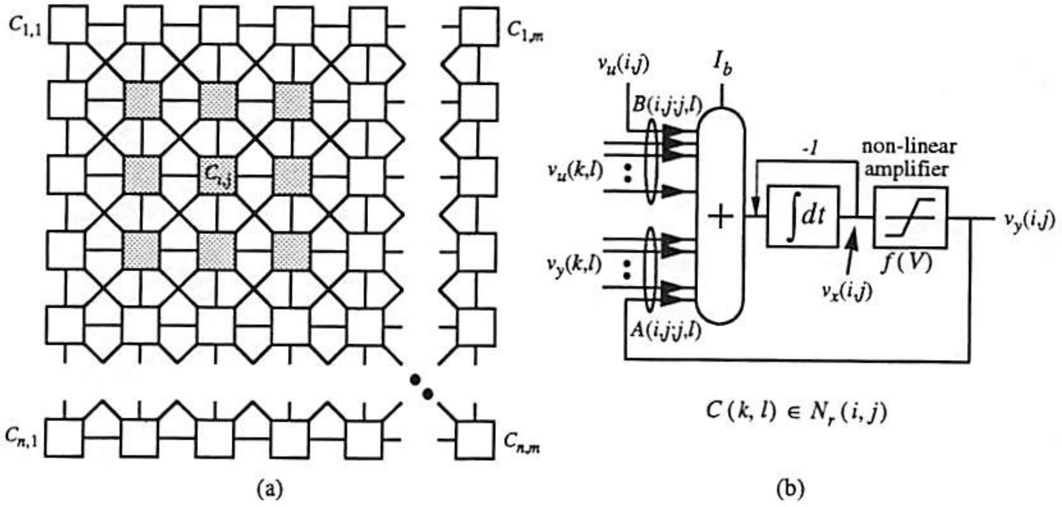


Figure 2.1: Cellular neural network. (a) An n-by-m cellular neural network on rectangular grid (shaded boxed are the neighborhood cells of C(i,j). (b) Functional block diagram of neuron cell.

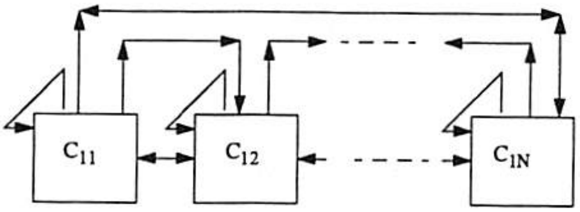


Figure 2.2: 1-dimensional compact neural most suitable for communication systems.

values among the cells, the total input current i_{xij} to the cell is given by

$$i_{xij}(t) = \sum_{C(k,l)\in N_r(i,j)} A(i,j;k,l)v_{ykl}(t) + \sum_{C(k,j)\in N_r(i,j)} B(i,j;k,l)v_{ukl}(t) + I_b, \quad (2.1)$$

where I_b is the bias current. R_x and C are the equivalent resistanace and capacitance of the cell, respectively. For simplicity of illustration purpose, I_b, R_x, and C_x are represented by dependent current sources and summed at the state node. Due to the capacitance C_x and resistance R_x, the state voltage v_{xij} is established at the summing node and satisfies a set of differential equations

$$\begin{aligned} C_x\frac{dv_{xij}(t)}{dt} &= -\frac{1}{R_x}v_{xij}(t) + i_{xij}(t) &(2.2)\\ &= -\frac{1}{R_x}v_{xij}(t) + \sum_{C(k,l)\in N_r(i,j} A(i,j;k,l)v_{ykl}(t) \\ &+ \sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l)v_{ukl}(t) + I_b; ..1 \le i \le n, 1 \le j \le m. \end{aligned}$$

The cell contains a nonlinearity between the state node and the output; and its input-output relationship is represented by $v_{yij}(t) = f(v_{xij}(t))$. The nonlinear function used in a cellular neural network can be any differentiable, non-decreasing, function $y = f(x)$, provided that $f(0) = 0, df(x)/dx \ge 0,\ f(+\infty) \to +1$ and $f(-\infty) \to -1$. Two widely used nonlinearities are the piecewise-linear and sigmoid functions as given by

$$\begin{aligned} y(x) &= f(x) \\ &= \begin{cases} \frac{1}{2}(|x+1| - |x-1|) & \text{piecewise-linear} \\ \frac{1-e^{-\lambda x}}{1+e^{-\lambda x}} & \text{sigmoid.} \end{cases} \qquad (2.3) \end{aligned}$$

Here, the parameter λ is proportional to the gain of the sigmoid function. For a unity neuron gain at $x = 0, \lambda = 2$ may be used for the sigmoid function. The gain of neurons in a Hopfield neural network is very large so that the steady-state outputs are all binary-valued. However, the positive feedback in the cellular neural network cell is so strong that the gain of the cell needs not to be large for guaranteed binary output in the steady state. Typically, a unity gain $df(x)/dx|_{x=0} = 1$ is used in cellular neural networks. The transfer characteristics of the piecewise-linear function is shown in Fig. 2.3.

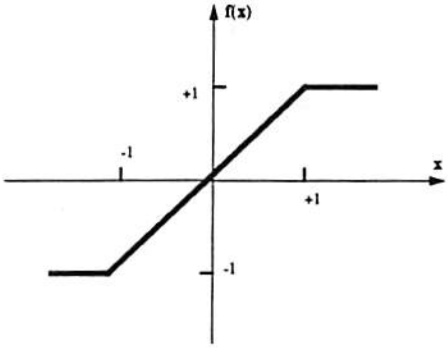


Figure 2.3: Piecewise-linear function.

The piecewise-linear function provides a mathematical tractability in the analysis, while the sigmoid-like nonlinearity can be easily obtained as a by-product of electronic circuits such as operational amplifier. The shift-invariant cellular neural networks have the interconnections that do not depend on the position of cells in the array except at the edges. The shift-invariant property of a cellular neural network is the most desirable feature when implementing a large-size electronic network such as a very large-scale integration (VLSI) chip. The weights of a shift-invariant cellular neural network can be represented by the $(2r+1) \times (2r+1)$ feedforward and feedback weighting templates

$$\begin{aligned} \mathbf{T_A} &= [a_{p,q}, -r \leq p, q \leq r], \\ \mathbf{T_B} &= [b_{p,g}, -r \leq p, q \leq r]. \end{aligned} \tag{2.4}$$

Let $N = n \times m$ be the number of cells in a compact neural network. By using the vector and matrix notations, (2.3) can be re-written as

$$C_x \frac{d\mathbf{x}}{dt} = -\frac{1}{R_x}\mathbf{x} + \mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{u} + I_b\mathbf{w}, \tag{2.5}$$

where

$$\begin{aligned}
\mathbf{x} &= [x_1 x_2 \cdots x_N]^T = [v_{x1}(t)|v_{x2}(t)|\cdots|v_{xn}(t)]^T, \\
\mathbf{y} &= [y_1 y_2 \cdots y_N]^T = [v_{y1}(t)|v_{y2}(t)|\cdots|v_{yn}(t)]^T, \\
\mathbf{u} &= [u_1 u_2 \cdots u_N]^T = [v_{u1}|v_{u2}|\cdots|v_{un}]^T, \\
\mathbf{A} &= \mathit{toeplitz}((\mathbf{A_0}|\mathbf{A_1}|\cdots|\mathbf{A_r}|0|\cdots), (\mathbf{A_0}|\mathbf{A_{-1}}|\cdots|\mathbf{A_{-r}}|0|\cdots) \\
\mathbf{B} &= \mathit{toeplitz}((\mathbf{B_0}|\mathbf{B_1}|\cdots|\mathbf{B_r}|0|\cdots), (\mathbf{B_0}|\mathbf{B_{-1}}|\cdots|\mathbf{B_{-r}}|0|\cdots)) \\
\mathbf{w} &= [1, 1, \cdots, 1]^T,
\end{aligned} \tag{2.6}$$

where,

$$\begin{aligned}
\mathbf{v_{xk}} &= [v_{xk1}(t) v_{xk2}(t) \cdots v_{xkm}(t)], \\
\mathbf{v_{yk}} &= [v_{yk1}(t) v_{yk2}(t) \cdots v_{ykm}(t)], \\
\mathbf{v_{uk}} &= [v_{uk1} v_{uk2} \cdots v_{ukm}], \\
\mathbf{A_k} &= \mathit{toeplitz}((a_{k,0} a_{k,1} \cdots a_{k,r} 0 \cdots), (a_{k,0} a_{k,-1} \cdots a_{k,-r} 0 \cdots)), \\
\mathbf{B_k} &= \mathit{toeplitz}((b_{k,0} b_{k,1} \cdots b_{k,r} 0 \cdots), (b_{k,0} b_{k,-1} \cdots b_{k,-r} 0 \cdots)),
\end{aligned} \tag{2.7}$$

and $\mathit{toeplitz}(\mathbf{a}, \mathbf{b})$ is defined as the Toeplitz matrix with $\mathbf{a}$ in the first row and $\mathbf{b}$ in the first column. Note that the submatrices $\mathbf{A_k}$ and $\mathbf{B_k}$ are Toeplitz, but

A and **B** are not. The elements of $\mathbf{T_A}$ and $\mathbf{T_B}$ are often normalized to the scale of $\mathbf{T_x}$, e.g., 10^{-3}. The notation of voltage $v_x(t)/v_y(t)$ and the state variables $\mathbf{x}/\mathbf{y}$ will be used interchangeably hereafter. Because $-1 \leq y_k \leq +1, \forall k$, the output variable $\mathbf{y}$ is confined within the N-dimensional hyeprcube so that $\mathbf{y} \in \mathbf{D^N} = \mathbf{y} \in \mathbf{R^N} : -1 \leq y_k \leq 1; k = 1, 2, \cdots, N$. The weighting templates are called symmetric if $A(i,j;k.l) = A(k,l;i,j)$ and $B(i,j;k,l) = B(k,l;,i,j)$. In this case, **A** and **B** are symmetric matrices and the stability of the network is guaranteed. In fact, the symmetry of **A** is a sufficient condition for stability. Under the constraint conditions $|v_{uij}(0)| \leq 1$ and $|v_{uij}| \leq 1, \forall i,j$, the shift-invariant cellular neural network always produces a stable output in the steady state. Moreover, if $A(i,j;k,l) > 1/R_x$, then the saturated binary outputs are guaranteed.

In any cellular neural networks, all states $v_{xij}(t), \forall \geq 0$, are bounded and the bound $v_{x,max}$ can be determined by [1]

$$\begin{aligned} v_{x,max} &= 1 + R_x|I_b| \\ &+ R_x \cdot max(\sum_{C(k,l)\in N_r(i,j)} (|A(i,j;k,l)| + |B(i,j;k,l)|)). \qquad (2.8) \end{aligned}$$

The terms in (2.8) account for the initial value, bias, feedback, and feedforward interactions, respectively. Therefore, the operating range of the circuits for summing and integration in Fig. 2.1(b) must be at least $-v_{x,max} \leq v_{xij}(t) \leq v_{x,max}$.

2.1.2 Stability

The stability of a nonlinear dynamic system including cellular neural networks is described by Lyapunov [7] or generalized energy function. For a cellular neural network with the piecewise-linear function, the energy function is given by [1]

$$\begin{aligned} E(t) &= -\frac{1}{2}\sum_{i,j}\sum_{C(k,l)\in N_r(i,j)} A(i,j;k,l)v_{yij}(t)v_{ykl}(t) + \frac{1}{2R_x}\sum_{i,j}(v_{yij}(t))^2 \\ &- \sum_{i,j}\sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l)v_{yij}(t)v_{ukl} - \sum_{i,j} I_b v_{yij}(t). \end{aligned} \tag{2.9}$$

For the sigmoid nonlinearity, the second term of (2.9) is replaced by

$$\frac{1}{R_x}\sum_{i,j}\int_0^{v_{yij}(t)} f_{ij}^{-1}(v)dv. \tag{2.10}$$

The expression (2.10) can be used for arbitrary nonlinearity $y = f(x)$ if its inverse function $x = f^{-1}(y)$ can be well-defined over the range of x. It can be interpreted as the area of the function $x = f^{-1}(y)$ when integrated from $y = 0$ to $y = v_{yij} \leq 1$. The piecewise-linear function used in (2.9) is a special case of this general expression (2.10). For the piecewise-linear function, $x = f^{-1}(y) = y, -1 \leq y \leq 1$, and

$$\int_0^{v_{yij}}(t)f^{-1}(y)dy = \int_0^{v_{yij}}(t)ydy = \frac{1}{2}(v_{yij}(t))^2, \tag{2.11}$$

which is consistent with the one in (2.9). In the vector and matrix forms, (2.9) is a scalar-calued quadratic function of output vector $\mathbf{y}$,

$$\begin{aligned} E &= -\frac{1}{2}\mathbf{y^T A Y} + \frac{1}{2R_x}\mathbf{y^T y} - \mathbf{y^T B u} - I\mathbf{y^T w} \\ &= -\frac{1}{2}\mathbf{y^T M y} - \mathbf{y^T b}, \end{aligned} \tag{2.12}$$

where $\mathbf{M} = \mathbf{A} - (1/R_x)\mathbf{I}$ and $\mathbf{b} = \mathbf{Bu} + I_b\mathbf{w}$. The stability of the network can be tested by checking the behavior of the energy function after the network is activated at time $t = t_0$. By using the chain rule, the time derivative of E can be given by a scalar product of the two vectors

$$\frac{dE}{dt} = \frac{\partial E}{\partial \mathbf{y}}\frac{d\mathbf{y}}{dt} = \sum_{k=1}^{N} \frac{\partial E}{\partial y_k}\frac{dy_k}{dt}, \tag{2.13}$$

where $\partial E/\partial \mathbf{y} = \nabla_{\mathbf{y}} E$ is the gradient of E with respect to $\mathbf{y}$. From (2.12), we have

$$\nabla_{\mathbf{y}} = \frac{\partial}{\partial \mathbf{y}}(-\frac{1}{2}\mathbf{y}^{\mathbf{T}}\mathbf{M}\mathbf{y} - \mathbf{y}^{\mathbf{T}}\mathbf{b}) = -\frac{1}{2}(\mathbf{M}\mathbf{y} + \mathbf{M}^{\mathbf{T}}\mathbf{y}) - \mathbf{b}. \tag{2.14}$$

If $\mathbf{A}$ is symmetric, so is $\mathbf{M}$ and $\mathbf{M} = \mathbf{M}^{\mathbf{T}}$. Therefore,

$$\nabla_{\mathbf{y}} E = -(\mathbf{M}\mathbf{y} + \mathbf{b}) = -(\mathbf{A}\mathbf{y} - \frac{1}{R_x}\mathbf{y} + \mathbf{b}). \tag{2.15}$$

Assumed that the network is activated at $t = t_0$ and the constraint condition $|v_{xij}(t_0)| \leq 1$ is satisfied. Then, it begins to operate in the linear region because $v_{yij}(t_0) = v_{xij}(t_0)$ and as time increases, some of the cells become saturated such that $|y_\rho| = 1$ for some ρ. Note that if $|x_\rho| > 1$, then $dy_\rho/dt = 0$ and the corresponding terms in (2.12) vanish. If we consider only nonzero terms, then for $k \neq \rho, y_k = x_k, dy_k/dt = dx_k/dt$, and $\partial E/\partial y_k = -C_x(dx_k/dt)$. Therefore,

$$\frac{dE}{dt} = -C_x \sum_{k \neq \rho} \frac{dx_k}{dt}\frac{dy_k}{dt} = -C_x \sum_{k \neq \rho} \left(\frac{dx_k}{dt}\right)^2. \tag{2.16}$$

Since $C_x > 0$, the energy E decreases as time elapses such that $dE/dt < 0, \forall t \geq t_0$. When all the cells become saturated, $dE/dt = 0$ and the network results in a stable binary output for which the energy function (2.12) is locally minimized. Note that the state $\mathbf{x}$ is stablized after the stable equilibrium is reached because $dE/dt = 0$.

If we use other neuron transfer characteristics $y = f(x)$, for which the inverse function $x = f^{-1}(y)$ is well defined in the range of x, (2.12) can be written as

$$E = -\frac{1}{2}\mathbf{y}^{\mathbf{T}}\mathbf{A}\mathbf{y} + \frac{1}{R_x}\int_0^y f^{-1}(\mathbf{v})d\mathbf{v} - \mathbf{y}^{\mathbf{T}}\mathbf{b}, \tag{2.17}$$

where the second term is simply an $N - by - 1$ vector with the integral expression in each element. In this case, $f^{-1}(\mathbf{y}) = \mathbf{x}$ and

$$\nabla_{\mathbf{y}} E = -\mathbf{A}\mathbf{y} + \frac{1}{R_x}f^{-1}(\mathbf{y}) - \mathbf{b} = -\mathbf{A}\mathbf{y} + \frac{1}{R_x}\mathbf{x} - \mathbf{b} = -C_x\frac{d\mathbf{x}}{dt}, \tag{2.18}$$

from which it follows that

$$\frac{dE}{dt} = \nabla_y E\frac{dy}{dt} = -C_x\frac{dx}{dt}\frac{dy}{dt} = -C_x\sum_{k=1}^{N}\frac{\partial f}{\partial x_k}\left(\frac{dx_k}{dt}\right)^2 \leq 0. \tag{2.19}$$

The information to be processed can be passed into the network in a form of the input $v_{uij}(0)$, or the initial values of the state variable. In any cases, the initialization of the state voltage $v_{xij}(t)$ is required at the beginning of each operation, such that $|v_{xij}(0)| \leq 1, \forall i, j$. Otherwise, the undesirable situation $E(t = 0) < E(t = +\infty)$ may occur. Local interconnection and simple synaptic weights are the most attrcative features of the cellular neural network for VLSI implementation in high-speed, real-time applications.

2.2 Discrete-Time Compact Neural Networks

Discrete-time cellular neural networks are a special type of feedback threshold network where the local interconnections and the shift-invariant weights are transferred from the continuous-time cellular neural networks. They are completely described by a recursive algorithm. The dynamic behavior is based on the feedback of clocked, binary outputs and a single cell is influenced by the inputs and outputs of neighboring cells. The architecture is closely related to the cellular automata, but different from them in having continuous-valued inputs and weights.

The discrete-time cellular neural network is the discrete-time version of (2.3) and defined by the state equation

$$x_{ij}(k) = \sum A(i,j;k,l)y_{k}l(k) + \sum B(i,j;k,l)u_{kl} + I_b, \tag{2.20}$$

and the output equation

$$\begin{aligned} y_{ij}(k) &= sgn(x_{ij}(k-1)) \\ &= \begin{cases} +1 & \text{if } x_{ij}(k-1) < 0, \\ -1 & \text{if } x_{ij}(k-1) \geq 0 \end{cases} \end{aligned} \tag{2.21}$$

for a cell $C(i,j)$, $i = 1,2,\cdots,n, j = 1,2,\cdots,m$, in an $n \times m$ rectangular-grid array. Here, $y_{ij}(k) \in +1,-1, u_{ij} \in \mathbf{D}, A(i,j;k,l) \in \mathbf{R}$, and $B(i,j;k,l) \in \mathbf{R}$, is the description of the next output through a set of linear inequalities in the discrete-time fashion. The fact that it does not include a sophisticated integration algorithm, allows simple implementation of the algorithm on a general-purpose digital computer. In a hardware-design viewpoint, the operation is quite insensitive to noise and parameter variations caused by fabrication tolerances and environmental effects. Thus, the interconnections among cells in a network or among several microchips in a large scale, multi-chip system, can be simplified.

The energy function of a discrete-time cellular neural network can be defined by the use of the Lyapunov theory for discrete-time systems [8] [9]

$$\begin{aligned} E(k) &= -\sum_{i,j}\sum_{k,l} A(i,j;k,l)y_{kl}(k-1)y_{ij}(k) \\ &- \sum_{i,j}\sum_{k,l} B(i,j;k,l)y_{kl}(k)(y_{ij}(k)+y_{ij}(k-1)) \\ &- \sum_{i,j} I_b(y_{ij}(k)+y_{ij}(k-1)). \end{aligned} \tag{2.22}$$

By assuming the symmetric feedback weights $A(i,j;k,l) = A(k,l;i,j)$, the differential energy $\triangle E = E(k+1) - E(k)$ is given by

$$\triangle E = -\sum_{ij} |x_{ij}(k)|(y_{ij}^2(k+1) - (y_{ij}(k-1)y_{ij}(k+1))). \tag{2.23}$$

Therefore, $\triangle E = 0$ if $y_{ij}(k-1) = y_{ij}(k+1), \forall i,j$, and $\triangle E \neq 0$ otherwise [10]. The energy function E decreases as k increases and the condition $\triangle E = 0$ for a stable state can be reached. However, for the condition $\triangle E = 0$ there exist two possible cases, i.e., $\forall i,j, y_{i,j}(k) = y_{i,j}(k-1)$ and $y_{ij}(k-1) = y_{ij}(k+1)$. The first case obviously corresponds to a stable state, while the second case represents a two-cycle oscillation between two different outputs. Thus, the stable operation is not guaranteed in a discrete-time cellular neural network with symmetric feedback templates. For some other classes of templates, the discrete-time cellular neural network is shown to be always stable [11] [12].

Reference List

[1] L. O. Chua, L. Yang, "Cellular neural networks: Theory," *IEEE Trans. on Circuits and Systems*, vol. 35, pp. 1257-1272, Oct. 1988.

[2] L. O. Chua, L. Yang, "Cellular neural networks: Applications," *IEEE Trans. on Circuits and Systems,* vol. 35, pp. 1273-1290, Oct. 1988.

[3] S. H. Bang, B. J. Sheu, "A neural network for detection of signals in communication," *IEEE Trans. on Circuit and Systems, Part I,* vol. 43, pp. 644-655, Aug. 1996.

[4] D. C. Chen, B. J. Sheu, "A compact neural network based CDMA receiver," *IEEE Trans. on Circuit and Systems, Part II,* accpted for publication in 1997.

[5] L. O. Chua, T. Roska, "The CNN paradigm," *IEEE Trans. on Circuits and Systems, Part I,* vol. 40, pp. 147-156, Mar. 1993.

[6] Proc. of Special Session on Cellular Neural Networks, European Conf. on Circuit Theory and Design (ECCTD), Sep. 1991.

[7] R. E. Kalman, J. E. Bertram, "Control system analysis and design via the 'second method' of Lyapunov Part I: Continuous-time systems," *Trans. on ASME,* pp. 371-393, June 1960.

[8] R. E. Kalman, J. E. Bertram, "Control system analysis and design via the 'second method' of Lyapunov Part I: Discrete-time systems," *Trans. on ASME,* pp. 394-400, June 1960.

[9] H. Harrer, J. A. Nossek, Discrete-Time Cellular Neural Networks, John Wiley & Sons: New York, NY, 1993.

[10] M. Minsky, S. Papert, *Percentrons - An Introduction to Computational Geometry,* MIT Press: Cambridge, MA, 1988.

[11] H. Magnussen, J. A. Nossek, L. O. Chua, "The learning problem for discrete-time cellular neural networks as a combinatorial optimization problem," Tech. Rep. UCB/ERL M93/88, Electronics Research Laboratory, College of Engineering, University of Berkeley, CA, 1993.

[12] H. Magnussen, J. A. Nossek, "A geometric approach to properties of the discrete-time cellular neural network," *IEEE Trans. on Circuits and Systems, Part I,* vol. 41, no. 10, pp. 625-634, Oct. 1994.

Chapter 3

MLSE and VLSI Architecture for Viterbi Algorithm

3.1 Receivers with Equalization Function

In high-speed digital transmission, unintentionally introduced intersymbol interference (ISI) and noise are the major impediments for the data detection. The performance of digital communication system can be severely degraded. When a digital pulse stream is transmitted, if one transmitted pulse is not allowed to decay away completely befor the transmission of the next one, ISI arises. The ISI is caused not only by channel distortion but also by multipath effects. Figure 3.1 shows one example of multipath effect on missile control. It is critical to design an optimum receiver which takes into account both the existence of ISI and additive noise. In 1972, G. D. Forney, Jr. proposed a recursive structure [1] to detect the received digital data symbols in the presence of ISI and additive white Gaussian noise (AWGN). This structure is a maximum-likelihood estimator of the entire transmitted sequence and known as Maximum-Likelihood Sequence Estimate (MLSE), which is an efficient decision rule on the received *sequence* rather than *symbol-by-symbol* detection [2]. In the presence of AWGN alone, the

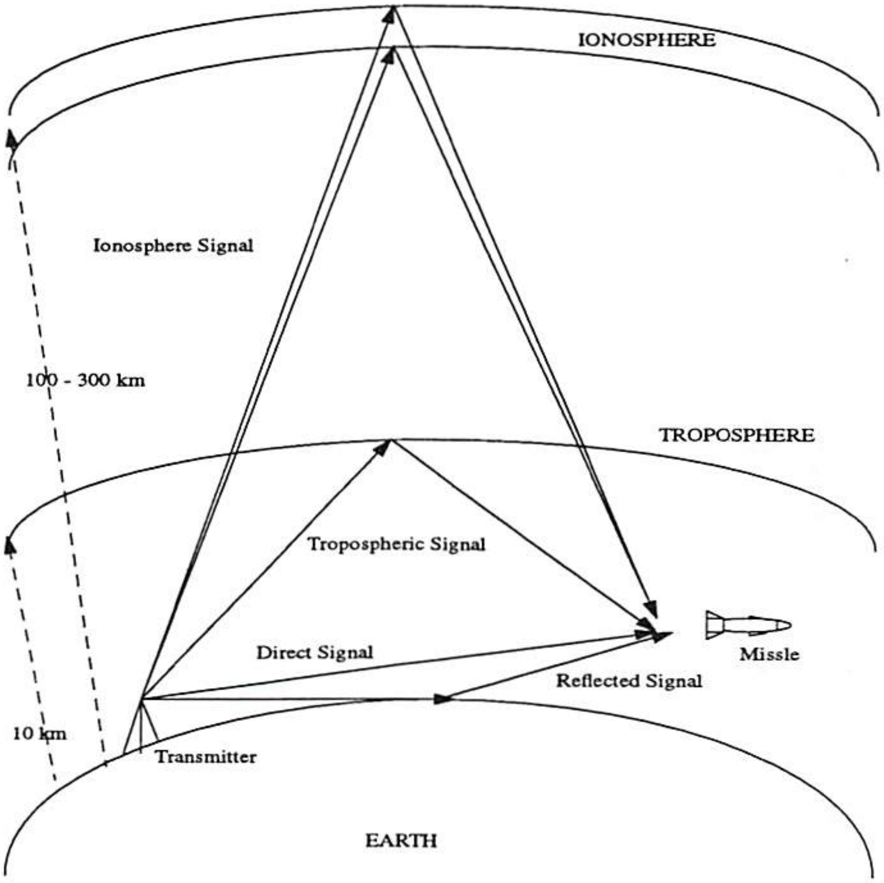


Figure 3.1: Example of multipath effect on missile control.

MLSE has the performance approaching that of an optimum symbol-by-symbol counterpart.

From the optimization point of view, the MLSE is a combinatorial maximization or minimization of the cost function over all possible sequences of a certain length. The signaling alphabet $\alpha = \alpha_k, k = 1, 2, \cdots, M$, and sequence $\mathbf{s_n} = \{s_i\}, i = 0, 1, \cdots, n-1$, corresponding to a finite set of numbers and the degree of freedom, respectively. There are M^n possible combinations over which MLSE computes the cost function.

3.2 Maximum-Likelihood Sequence Estimation

A complex data sequence $\{a_n\}$ is sent over a band-limited channel whch is corrupted by Gaussian noise. The MLSE determines the best estimate of a_n, i.e. $\alpha_n = \hat{a}_n$ that maximizes the likelihood function [3]. Let the channel be characterized by h(t). The received signal r(t) can be represented as:

$$r(t) = \sum_n \alpha_n h(t - nT) + w(t|\alpha_n), \tag{3.1}$$

where $w(t|\alpha_n)$ is the stationary Gaussian noise. Let the transmitted sequence have N data symbols and the channel memory be $L \cdot T$. The received signal will be observed during the time period $I = [0, F], F > (N + L)T$. Owing to the assumption of Gaussian noise, the likelihood function becomes

$$\begin{aligned} & p[r(t), t \in I|\alpha_n] = p[w(t|\alpha_n)] \\ \sim \quad & exp\{\frac{-1}{2N_0}\int_I\int_I w^*(t_1|\alpha_n)K^{-1}(t_1 - t_2) \times w(t_2|\alpha_n)dt_1dt_2\} \end{aligned} \tag{3.2}$$

where $K^{-1}(\tau)$ is the inverse of $K(\tau)$, and

$$K(\tau) * K^{-1}(\tau) = \delta(\tau). \tag{3.3}$$

Here $*$ represents convolution operation. By substituing (3.1) into (3.2) and discarding the terms independent of α_n, (3.2) becomes

$$p[r(t), t \in I|\alpha_n] \sim exp\{\frac{-1}{4N_0}[\sum_{nT\in I} 2Re(\alpha_n{}^* z_n) - \sum_{iT\in I}\sum_{kT\in I} \alpha_i{}^* s_{i-k}\alpha_k]\} \tag{3.4}$$

where

$$z_n = \int_I \int_I h^*(t_1 - nT)K^{-1}(t_1 - t_2)r(t_2)dt_1dt_2$$

$$s_l = \int_I \int_I h^*(t_1 - iT)K^{-1}(t_1 - t_2)h(t_2 - kT)dt_1dt_2 = s^*_{-l}. \tag{3.5}$$

Thus (3.4) can be further expressed as

$$p(w(t|\alpha_n)) \sim exp\{\sum_{i=1}^{M} 2\alpha_i z_i - \sum_{i=1}^{M}\sum_{i=1}^{M} \alpha_i s_{i-k} \alpha_k\}. \tag{3.6}$$

Under maximum-likelihood criteria, the estimated sequence is that for which (3.6) is maximized. Since (3.6) is a monotonically increasing function of the term in brace, given by

$$J_M(\{\alpha_M\}) = \sum_{i=1}^{M} 2\alpha_i z_i - \sum_{i=1}^{M}\sum_{i=1}^{M} \alpha_i s_{i-k} \alpha_k, \tag{3.7}$$

maximizing (3.6) is equivalent to maximizing (3.7). The notation $J_M(\alpha_M)$ indicates that the cost function for the sequence $\alpha_1, \alpha_2, \cdots, \alpha_M$. (3.7) will be referred to as the MLSE cost function.

The estimation procedure using direct evaluation of the MLSE cost function requires that (3.7) be evaluated for all possible sequences of length M that can be formed from data symbol $+1$ and -1. Thus, (3.7) must be evaluated 2^M times to obtain an estimate of the sequence $\{a_n\}$. To perform the estimate in real time, which is required by most communication links, the 2^M computations of (3.7) must be performed within MT time span. In most cases, direct evaluation of MLSE cost function is too computation intensive to be of practical use.

The number of computations required can be greatly reduced by the use of the Viterbi algorithm (VA), which requires on the order of 2^{L+1} comparison-and-add operations during each signaling interval T. Generally, M is about 40 to 60 and L is about 1 to 3.

3.3 Viterbi Algorithm and VLSI Implementation

The Viterbi algorithm (VA) was originally invented to decode convolution codes. G.D. Forney found it could be used to implement the MLSE efficiently. Almost at the same time, VA's applicability to partial-response systems was noticed by Omura and Kobayashi at UCLA independently [4] [5]. VA has the following properties [6] [7].

- Implementability: Like the best of the earlier "optimum" nonlinear processors, the VA is a recursive structure that does not grow with the length of the message sequence, and that has the complexity to be proportional to m^L, where m is the size of the input alphabet and L is the length of the impulse response $h(t)$. It is superior in not requiring any multiplications, but only m^L additions and m^{L-1} m-ary comparisons per received symbol, which greatly simplifies hardware implementation; It also requires only m^{L-1} words of memory (of the order of 10-30 bits per word).

- Analyzability: The ease with which performance can be analyzed is in significant contrast to all earlier work with nonlinear processors. At moderate-to-high signal noise ratios, the symbol-error probability is accurately upper-bounded and estimated by

$$P_r(e) \leq K_1 Q(d_{min}/2\sigma) \tag{3.8}$$

where d_{min}^2 is the minimum energy of any nonzero signal, σ is the spectral density of the noise, and Q is the probability of error function

$$Q(x) \equiv (2\pi)^{-1/2} \int_x^\infty e^{-y^2/2} dy. \tag{3.9}$$

- Optimality: This structure is optimum for maximum likelihood estimation of the entire transmitted sequence provided unbounded delay is allowed at the output, and effectively optimum in the same sense for reasonable finite delays.

In VA, every possible sequence is represented as a path in the treliss structure. VA is to find the shortest path, which is equivalent to find the maximum-likelihood sequence estimate. Figure 3.2 shows the diagram for the recursive evaluation of the shortest path in the treliss diagram. For more understanding about the VLSI implementation of the VA, Figure 3.3 shows the block diagram of a Viterbi decoder

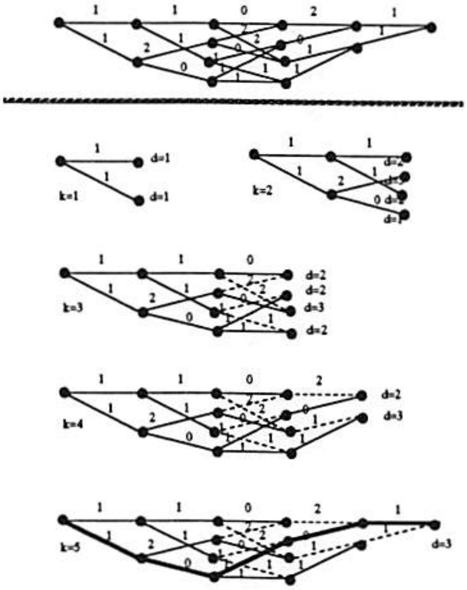


Figure 3.2: Recursive determination of shortest length path for four-state treliss diagram.

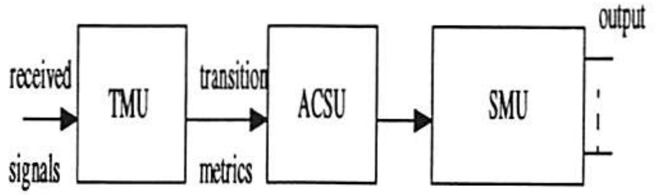


Figure 3.3: Chip architecture of Viterbi decoder [8].

chip [1]. As mentioned in the previous paragraph, the VA finds the most likely sequence of state transitions (a path through the treliss) through a finite state trellis by assigning in a first step a transition metric to all possible state transitions. These state transition metrics are computed from the received input samples in the so-called Transition Metric Unit (TMU) [9]. Subsequently, from two paths which end in the same state, the path with the smallest sum of the transition mertic is selected as the most likely. This decisions are required for each possible states. They are taken in the Add Compare Select Unit (ACSU), which accumulates the transition metrics recursively and outputs a decision bit accordingly for each state and each trellis cycle. These decisions are then processed in the Survivor Memory Unit (SMU) of the decoder, which keeps track of the history of decisions. Consequently, the content of the SMU allows the reconstruction of the paths that areassociated with the states. The problem of finding the most likely path through the trellis can then be solved by tracing all paths back in time unit they have all merged into one path. This path is called the final survivor path.

Reference List

[1] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. on Information Theory,* vol. IT-18, no. 3, pp. 363-378, May 1972.

[2] M. Simon, S. Hinedi, and W. Lindsey, *Digital Communication Techniques,* Prentice Hall PTR: Upper Saddel River, NJ, 1995.

[3] G. Ungerboeck, "Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems," *IEEE Trans. on Commun.,* vol. COM-22, no. 5, pp. 624-636, May 1974.

[4] H. Kobayashi, "A survey of coding schemes for transmission or recording of digital data," *IEEE Trans. Commun.,* vol. COM-19, no. 11, pp. 1087-1100, Dec. 1971.

[5] H. Kobayashi, "Correlation level coding and maximum-likelihood decoding," *IEEE Trans. Info. Theory,* vol. IT-17, no. 9, pp. 586-594, Sep. 1971.

[6] A. J. Viterbi, "Error bounds for conventional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Info. Theory,* vol. IT-13, no. 5, pp. 260-69, Apr. 1967.

[7] A. J. Viterbi, J. K. Omura, *Principles of Digital Communication and Coding,* McGraw-Hill: New York, NY, 1982.

[8] O. J. Joeressen, H. Meyr, "A 40 Mb/s soft-output Viterbi decoder," *IEEE. J. of Solid-State Circuits,* vol. 30, no. 7, pp. 812-818, July, 1995.

[9] P. G. Gulak, E. Shewdyk, "VLSI structures for Viterbi receivers: Part I - general theory and applications," *IEEE J. Selected Areas in Commun.,* vol. 4, pp. 142-154, Jan. 1986.

Chapter 4

1-D Compact Neural Network Based Deector

In the previous chapter, we review the most widely used maximum likelihood sequence estimation (MLSE) and its implementation, Viterbi algorithm (VA), for communication receivers. In a particular digital chip example reported in the literature [1], metric calculation for every signal path, metric comparison for every possible state and tracking of the decisions from the output are needed. These units will 1) comsume a lot of power, 2) need significant amount of memory for storage and, 3) take considerable CPU time for the calculation and comparison. With the increase of channel memory, i.e. the increase of the ISI length, the power comsumption, calculation time and storage memory will incresae. To develop high speed and low power communication systems as stated in Chapter 1, we need to explore new algorithms or techniques to provide better or alternative solutions to these problems.

Artificial neural networks provide us the opportunity to develop new design techniques. Artificial neural networks have shown great promise in solving many complex signal processing and optimization problems that can not be satisfactorily addressed with conventional approaches. The surpervised or unsupervised learning methods [2] [3] and massively parallel architectures of artificial neural networks provide attractive properties of optimization and fast problem solving. The neural networks in communication have been motivated by the adaptive learning capability and the collective computational properties to process real world signals. Well designed neural networks have the ability to perform the error correction of error-control codes, equallization of transmission channels, crossbar switch control, and wireless/networking control.

Therefore, some other architectures of receivers with blind equalization by applying feedforward neural networks, radial basis function (RBF) networks mapping of the optimal Bayesian equalizer solution, or multilayer perceptron neural networks with high-order cumulants were reported in the literature [4] [5] [6]. However, these architectures require time-comsuming training algorithm and corresponding VLSI implementations are complicated. Besides, the stencouragingructures of the entire receiver system are very complex. Hence, they are not attractive for the high speed and lower power communication systems.

Due to the inconvenience stated above, we are motivated to find new scheme for high speed and lower power equalizers. There are two encouraging facts that make us explore the possibility of implementation of MLSE by compact neural networks as described below.

- In estimation theory, MLSE is the estimate method with best performance. The only issue of MLSE is the deferring decision-making, i.e., the output is with latency delay. With the advances of VLSI technology, it is possible to implement the MLSE algorithm by very high speed circuits and the delay is greatly shortened.

- Compact neural networks use analog processing cells and local interaction within a finite radius. Performimg maximum likelihood decoding of linear block error-correcting codes is shown to be equivalent to finding a global minimum of the energy function associated with a compact neural network. Given a code, a neural network can be constructed in such a way that there exists one-to-one correspondence between every codeword and every local minimum of the energy function.

In this chapter, parallel architecture of the compact neural network for the MLSE implementation is described. This algorithm was first developed by Bang and Sheu [7]. The compact neural network has collective computational properties and can be used to solve difficult optimization problem with signals represented in 1-dimensional array format. The cost function to be minimized in the MLSE has the same quadrratic form as the Lyapunov function associated with the compact neural network. If the cost function is properly mapped onto the network, then the desired estimate is obtained at the output. Optimal or optimized solutions can be obtained by applying the paralleled hardware annealing method which is a deterministic process for searching a globally minimum energy state in a short period of time.

For convenient discussion and study, algorithm described in [7] is briefly reviewed in this chapter. Further analysis about the constraint functions and error analysis were shown.

4.1 Digital Communication and Compact Neural Networks

The actual ISI channel together with baseband Nyquist filters in the transmitter and receiver can be modeled as a finite impulse response (FIR) filter of length $L+1$

whose impulse response is given by $h(k) = h_k$ with the corresponding z-transform $H(z)$. Here, L is the number of symbol intervals over which the ISI spans and hence $h(k) = 0$ for $k < 0$ and $k > L$. The received signal $r(t)$ is produced by the convolution of $u(k) = \sum_i u_i \delta(k-i)$ with $h(k)$ where $\delta(k)$ is the Kronecker delta function, plus white Gaussian noise $n(k)$ of zero-mean and finite variance σ^2,

$$r_k \equiv r(k) = \sum_{i=0}^{L} u_i h(k-i) + n(k). \tag{4.1}$$

Here $\mathbf{r_n} = \{r_0, r_1, \cdots, r_{n-1}\}$ and $\mathbf{u_n} = \{u_0, u_1, \cdots, u_{n-1}\}$ are the received and transmitted sequences of length n, respectively. For a sufficiently large n, the MLSE algorithm is to choose a sequence that maximizes a scalar cost function

$$J = -\sum_{k=0}^{n-1} \left| r_k - \sum_{i=0}^{L} h_i u_{k-i} \right|^2 = -\sum_{k=0}^{n-1} \left| r_k - \sum_{i=0}^{n-1} h_{k-i} u_i \right|^2 \tag{4.2}$$

for all possible combinations of sequences of length n. We define some variables for easy use:

$$\begin{aligned}
x_l &\equiv \sum_{k=0}^{n-1} h_k^* h_{k+l} = \sum_{k=0}^{L} h_k^* h_{k+l}, && (4.3)\\
y_i &\equiv \sum_{k=0}^{n-1} r_k h_{k-i}^*, && (4.4)\\
\mathbf{u} &= [u_0 \; u_1 \cdots u_{n-1}]^T = \mathbf{u_I} + j\mathbf{u_Q}, && (4.5)\\
\mathbf{y} &= [y_0 \; y_1 \cdots y_{n-1}]^T = \mathbf{y_I} + j\mathbf{y_Q} && (4.6)\\
\mathbf{X} &= \begin{bmatrix} x_0 & x_{-1} & \cdots & x_{-n+2} & x_{-n+1} \\ x_1 & x_0 & \cdots & x_{-n+3} & x_{-n+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n-2} & x_{n-3} & \cdots & x_0 & x_{-1} \\ x_{n-1} & x_{n-2} & \cdots & x_1 & x_0 \end{bmatrix} = \mathbf{X_I} + j\mathbf{X_Q}, \qquad \mathbf{X} \in C^{n \times n}.
\end{aligned}$$

By keeping the items related to the original signals, we can obtain

$$\begin{aligned}
\hat{J}_n &= \frac{1}{2}(\mathbf{u}_\mathbf{I}^\mathbf{T}\mathbf{X}_\mathbf{I}\mathbf{u}_\mathbf{I} + 2\mathbf{u}_\mathbf{Q}^\mathbf{T}\mathbf{X}_\mathbf{Q}\mathbf{u}_\mathbf{I} + \mathbf{u}_\mathbf{Q}^\mathbf{T}\mathbf{X}_\mathbf{I}\mathbf{u}_\mathbf{Q}) - (\mathbf{u}_\mathbf{I}^\mathbf{T}\mathbf{y}_\mathbf{I} + \mathbf{u}_\mathbf{Q}^\mathbf{T}\mathbf{y}_\mathbf{Q}) \\
&= \frac{1}{2}[\mathbf{u}_\mathbf{I}^\mathbf{T}|\mathbf{u}_\mathbf{Q}^\mathbf{T}]\begin{bmatrix} \mathbf{X}_\mathbf{I} \,|\mathbf{X}_\mathbf{Q}^\mathbf{T} \\ ---- \\ \mathbf{X}_\mathbf{Q}|\,\mathbf{X}_\mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{u}_\mathbf{I} \\ -- \\ \mathbf{u}_\mathbf{Q} \end{bmatrix} - [\mathbf{u}_\mathbf{I}^\mathbf{T}|\mathbf{u}_\mathbf{Q}^\mathbf{T}]\begin{bmatrix} \mathbf{y}_\mathbf{I} \\ -- \\ \mathbf{y}_\mathbf{Q} \end{bmatrix} \\
&\equiv \frac{1}{2}\bar{\mathbf{u}}^\mathbf{T}\bar{\mathbf{X}}\bar{\mathbf{u}} - \bar{\mathbf{u}}^\mathbf{T}\bar{\mathbf{y}}. \qquad (4.7)
\end{aligned}$$

4.2 The Compact Neural Network

A compact neural network is a continuous-time artificial neural network that features a one-dimensional array of neuron cells and local interconnections among the cells. The cellular neural network proposed by Chua and Yang [8, 9] in 1988 is a 2- and 3-dimensional network in the form of an N-by-M rectangular-grid array where N and M are the numbers of rows and columns, respectively. The compact neural network study is focused on the one-dimensional structures. Each cell in a nonlinear network corresponds to an element of the array. Fig. 4.1 shows an example of the block diagram of a typical nonlinear network and the circuit diagram of a neuron cell. With $N_r(i,j)$ being the r-th neighborhood cells of cell $C(i,j)$, the dynamics of a nonlinear network can be described by a set of nonlinear differential equations

$$\begin{aligned}
C\frac{dv_{xij}(t)}{dt} &= -\frac{1}{R_x}v_{xij}(t) + \sum_{C(k,1)\in N_r(i,j)} A(i,j;k,l)v_{ykl}(t) \\
&+ \sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l)v_{ukl}(t) + I_b;\; 1 \le i \le n, 1 \le j \le m, \qquad (4.8)
\end{aligned}$$

where $v_{xij}(t), v_{yij}(t)$, and $v_{uij}(t)$ are the state, output, and input voltage of the cell $C(i,j), A(i,j;k,l)$ and $B(i,j;k,l)$ are the feedback and feedforward synaptic weights between cells $C(i,j)$ and $C(k,l) \in N_r(i,j)$. Here C and R_x are the equivalent capacitance and resistanceat the state node, and I_b is the bias current to the cell. The magnitude of neuron output voltage is often normalized to the unity so that $-1 \leq v_{yk} \leq +1, \forall k$. The cell includes a nonlinearity between the

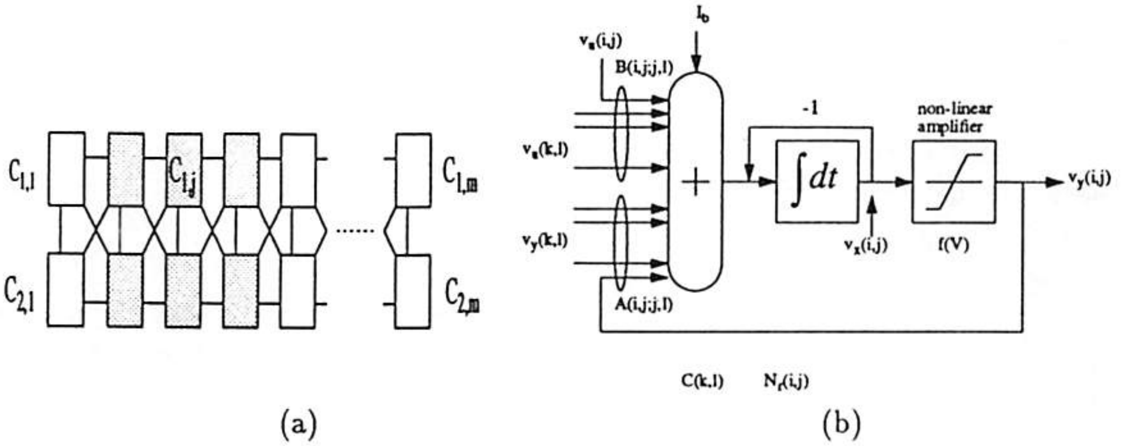


Figure 4.1: Compact neural network. (a) An $2-by-m$ compact neural network on rectangular grid (shaded boxes are the neighborhood cells of $C(i,j)$). (b) Functional block diagram of neuron cell.

state variable and the output result and its transfer function can be represented by $v_{yij}(t) = f(v_{xij}(t))$. The transfer function used in a nonlinear network [8] is the piecewise-linear function and can be described by

$$y = f(x) = \frac{1}{2}(|x+1| - |x-1|). \tag{4.9}$$

The shift-invariant nonlinear networks have the interconnections that do not depend on the position of cells in the array except at the edges. The shift-invariant property of a nonlinear network is a very attractive feature when implementing a large-size electronic network on a VLSI chip.

For the nonlinear network with the piecewise-linear function, the Lyapunov or generalized energy function is given by [8]

$$\begin{aligned} E(t) &= -\frac{1}{2}\sum_{i,j}\sum_{C(k,l)\in N_r(i,j)} A(i,j;k,l)v_{yij}(t)v_{ykl}(t) + \frac{1}{2R_x}\sum_{i,j}(v_{yij}(t))^2 \\ &- \sum_{i,j}\sum_{C(k,l)\in N_r(i,j)} B(i,j;k,l)v_{yij}(t)v_{ukl} - \sum_{i,j} I_b v_{yij}(t). \end{aligned} \tag{4.10}$$

Let n be the number of cells in a nonlinear network. In vector and matrix forms, (4.10) is a scalar-valued quadratic function of the output vector $\mathbf{y}$,

$$\begin{aligned} E &= -\frac{1}{2}\mathbf{v_y^T A v_y} + \frac{1}{2R_x}\mathbf{v_y^T v_y} - \mathbf{v_y^T B v_u} - I_b\mathbf{v_y^T w} \\ &= -\frac{1}{2}\mathbf{v_y^T M v_y} - \mathbf{v_y^T b}, \end{aligned} \tag{4.11}$$

where $\mathbf{M} = \mathbf{A} - (1/R_x)\mathbf{I}$ and $\mathbf{b} = \mathbf{Bv_u} + I_b\mathbf{w}$ for an n-by-1 unity vector $\mathbf{w}$.

4.3 System Mapping and Optimization

Fig. 4.2 shows the block diagram of the neural network MLSE receiver. The received signal $r(t)$ is first separated into two baseband signals, i.e., in-phase signal $r_I(t)$ and quadra-phase signal $r_Q(t)$. The signals are then sampled at $t = kT$ where T is the duration of a symbol, and the resulting discrete-time signal $r_I(k)$ and $r_Q(t)$ are correlated with the channel impulse response $h(k)$. The correlation filter matched to channel impulse response $h(k)$ is approximated by an FIR filter, whose tab coefficients are updated sequence by sequence.

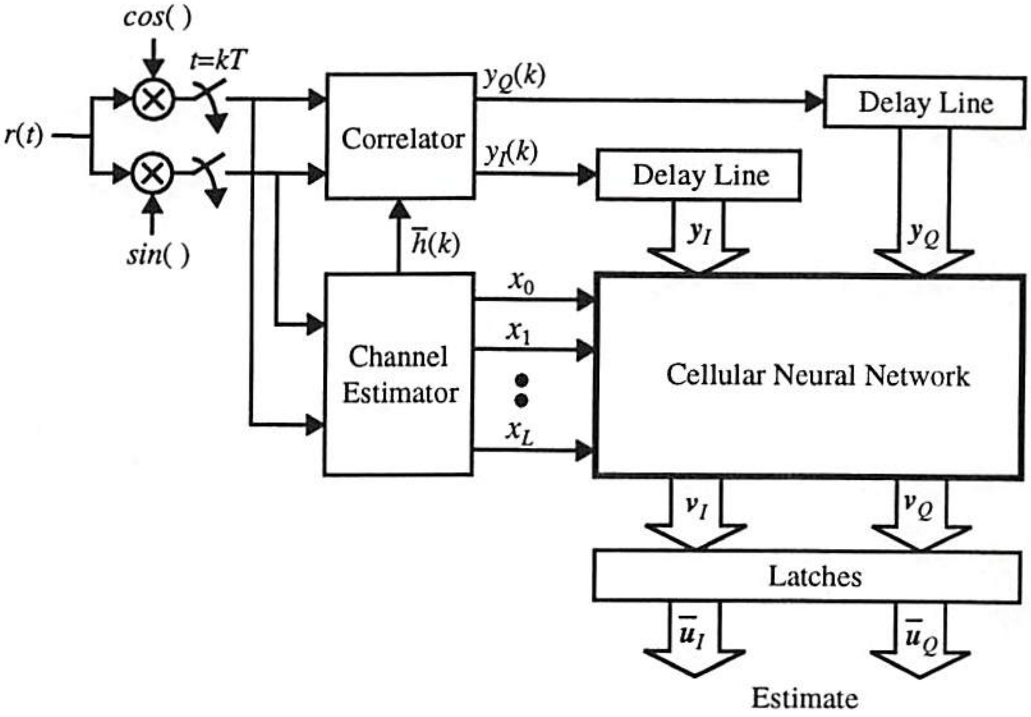


Figure 4.2: Block diagram of neural network MLSE receiver.

A compact neural network can be used as the core of nonlinear signal processing for the MLSE as shown in the figure. The desired estimate $\hat{\mathbf{u}}_{\mathbf{n}}$ can be obtained at the output of a nonlinear network if

$$\mathbf{M} = -\bar{\mathbf{X}} = -\begin{bmatrix} \mathbf{X_I} \ |\mathbf{X_Q^T} \\ ----- \\ \mathbf{X_Q}|\,\mathbf{X_I} \end{bmatrix} \text{ and } \mathbf{b} = \bar{\mathbf{y}} = \begin{bmatrix} \mathbf{y_I} \\ -- \\ \mathbf{y_Q} \end{bmatrix}. \qquad (4.12)$$

In other words, the cost function $\hat{J}_n$ is mapped onto a neural network constructed by the transconductance matrix $\bar{\mathbf{A}} = -\bar{\mathbf{X}} + T_x\bar{\mathbf{I}}$ and input vector $\mathbf{b} = \bar{\mathbf{y}}$. Here, the constant term $T_x\mathbf{I}$ represents a positive unity feedback in each cell. If the neural network produces saturated binary or multi-level values in the steady state, the output represents the MLSE of received sequence, i.e., $\hat{\mathbf{u}}_{\mathbf{n}} = \{\hat{u}_0 \ \hat{u}_1 \ \cdots \hat{u}_{n-1}\}$. After n symbols are shifted into the delay lines, the network performs the MLSE of an n-symbol sequence through an autonomous evolution of its internal state for $0 \leq t < T_C$ where T_C is the convergence speed of the network.

4.3.1 Hardware Annealing

Even with a correct mapping of the MLSE function onto a neural network, the desired optimal or optimized solutions are not guaranteed because a combinatorial optimization problem always involves a large number of local minima [12]-[13]. Therefore, in addition to the basic structure of the network, the annealing capability is provided to obtain the global minimum of the cost function over all possible combinations of sequence. The hardware annealing [10] [11] is a dynamic process for finding the optimum solutions in the recurrent associative neural networks such as Hopfield networks and nonlinear networks. The optimized solutions can be obtained by applying the hardware annealing technique to avoid local minima problems which are inherent in combinational optimizations.

The hardware annealing is performed by controlling the gain of the neuron, which is assumed to be the same for all neurons throughout the network. After the state is initialized to $v_x = v_x(0)$, the initial gain at time $t = 0$ can be set to an arbitrarily small, positive value such that $0 < g(0) \ll 1$.

It then increases continuously for $0 < t \leq T_A$ to the nominal gain of 1. The maximum gain $g_{max} = 1$ is maintained $T_A < t \leq T_C$, during which the network is stabilized. When the hardware annealing is applied to a nonlinear network by increasing the neuron gain $g(t)$, the transfer function can be described by

$$v_{yij}(t) = f(g(t)v_{xij}(t)) = \begin{cases} +1 & ; g(t)v_{xij}(t) \geq 1 \\ g(t)v_{xij}(t) & ; -1 < g(t)v_{xij}(t) < 1 \\ -1 & ; g(t)v_{xij}(t) \leq -1 \end{cases} \qquad (4.13)$$

or simply $y = f(gx)$. Note that the saturation level is still $y = +1$ or -1 and only the slope of $f(x)$ around $x = 0$ varies.

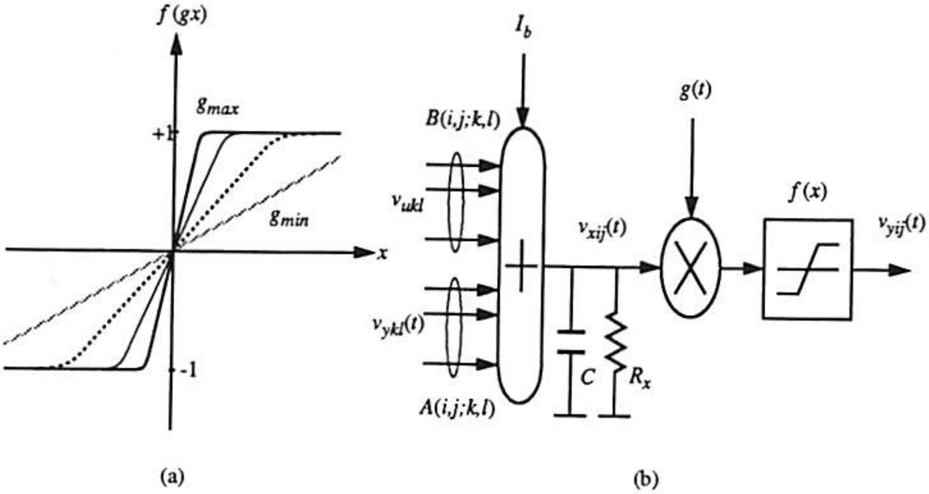


Figure 4.3: Variable-gain piecewise-linear neuron cell. (a) Transfer curves for several gain values. (b) Block diagram of variable-gain cell with two-quadrant analog multiplier.

4.4 Simulation Results

The simulation of a simple binary communication system with several ISI channels is performed by solving the differential equations. Random data sequence $\mathbf{u_n} = \{u_k\}, k = 0, 1, \cdots, n-1$, is generated and convolved with a channel response $h(k)$ which is assumed to be known exactly. The simulation result were conducted on a binary communication system with the ISI channel given by

$$H_m(z) = \frac{1}{\sqrt{1.25}}(1.0 + 0.5z^{-1}). \tag{4.14}$$

In this case, $x_0 = 1.0, x_1 = x_{-1} = 0.4, x_k = 0$ for $|k| \geq 2$. Here, 100 simulation runs were performed independently on the sequences of length 1000 ($n = 100$) for each signal-to-noise (SNR) value.

Next, the error rates of MLSE by unannealed and annealed networks are shown in Fig. 4.4(a) for a two-ray minimum-phase channel (4.14) and in Fig. 4.4(b) for a two-ray nonminimum-phase channel $H_n(z) = (0.5 + z^{-1})/\sqrt{1.25}$, respectively. Errors were accumulated and then divided by 10,000. Thus, probability of error is obtained. For comparison, the results of Viterbi algorithm (VA) are also shown in the figures. In the simulations of the channel $H_n(z)$, it is assumed that the decisions are made in reference to direct received samples which are half of the magnitude of delayed versions. It might be worthwhile mentioning that the NN-MLSE for a minimum-phase channel $H_m(z)$ is less efficient than the VA at the moderate values of SNR, but does not suffer much from the nonminimum-phase characteristics of the channel. Fig. 4.5 show the performance of three different annealed gains: $g(t) = (t/2), (t/2)^{1/3}, (t/2)^2$ applied to the neural network equalizer. It means the annealed gain changed linearly with time has the best chance to find the global minimum in digital computation. Fig. 4.6 and Fig. 4.7 show that the error bit positions in VA and neural network receiver with piecewise linear function are almost the same at low signal to noise ratio. In Fig. 4.6, the logic 1 index of the original signal is 3, 4, 7, 12, 13, 14, 15, 16, 17, 20, 22, 23, 24, 25, 29, 30, 32, 35, 41, 43, 44, 46, 50, 51, 53, 57, 63, 64, 65, 66, 67, 68, 69, 72, 73, 74, 77, 83, 85, 86, 88, 89, 93. The error index for compact neural network is 76, 88 and that for VA is 71, 76, 88. In Fig. 4.7, the logic 1 index of the original signal is 7, 8, 10, 12, 13, 17, 18, 23, 24, 30, 31, 34, 35, 39, 40, 41, 42, 43, 52, 56, 58, 59, 62, 65, 68, 69, 70, 71, 74, 75, 76, 77, 80, 81, 82, 85, 86, 89, 91, 93, 94, 95, 97. The error index for compact neural is 11, 20, 74, 97 and that for VA is 20, 74, 81. The errors are obtained by taking the difference of original signals and the estimates. Therefore the error amount is either +2 or −2.

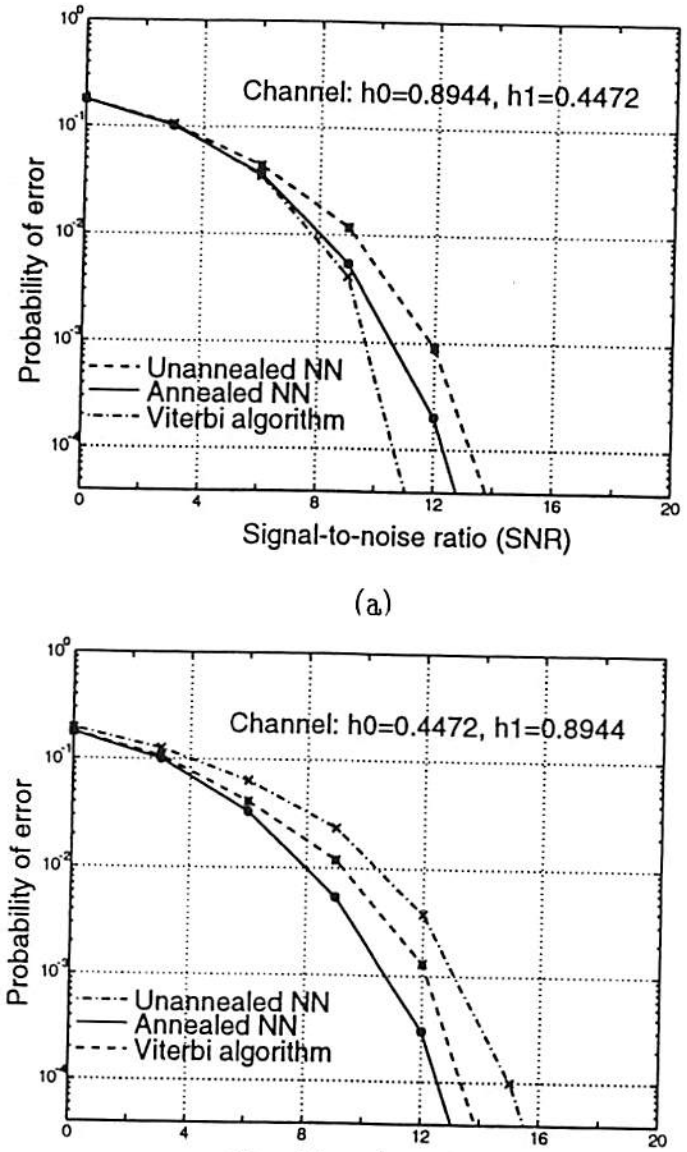
Fig. 4.8 shows statistical comparison results of the error bit positions between the outputs of four neural networks and the VA. Suppose there are $E1$ errors in one NN receiver and $E2$ errors in the VA. And there are $S3$ errors have the same index in the NN receiver and VA. Then percentage of same error bits P is

$$P = \frac{S3}{E2}. \tag{4.15}$$

The numbers of same error bit position for annealed neural networks and VA are larger than that for unannealed neural networks and VA. In addition, the numbers of same error bit positions for annealed neural networks and VA are larger than that for the unannealed neural networks and VA. From the above simulation data, we find a compact neural network with piece-wise linear function and annealing ability has best performance and most similar error behavior of the VA over the others.

4.5 Discussion

The collective computational behavior of a compact neural network is used to solve the maximum-likelihood estimation of signals in the presence of inter-symbol interference and white Gaussian noise. It is demonstrated that compact neural network is an efficient way of realizing the MLSE receiver. Therefore, the compact neural network MLSE (NN-MLSE) presented here can be thought of as an alternative to the VA. Unlike VA implementation, the NN-MLSE does not reqiure a vast amount of memory for storage and the computation time does not incresae with the increasing channel memory. In summary, it complies with the requirements of high speed and low power.



(a)

(b)

Figure 4.4: Error performance of different methods. (a) Two-ray minimum-phase channel. (b) Two-ray nominimum-phase channel.

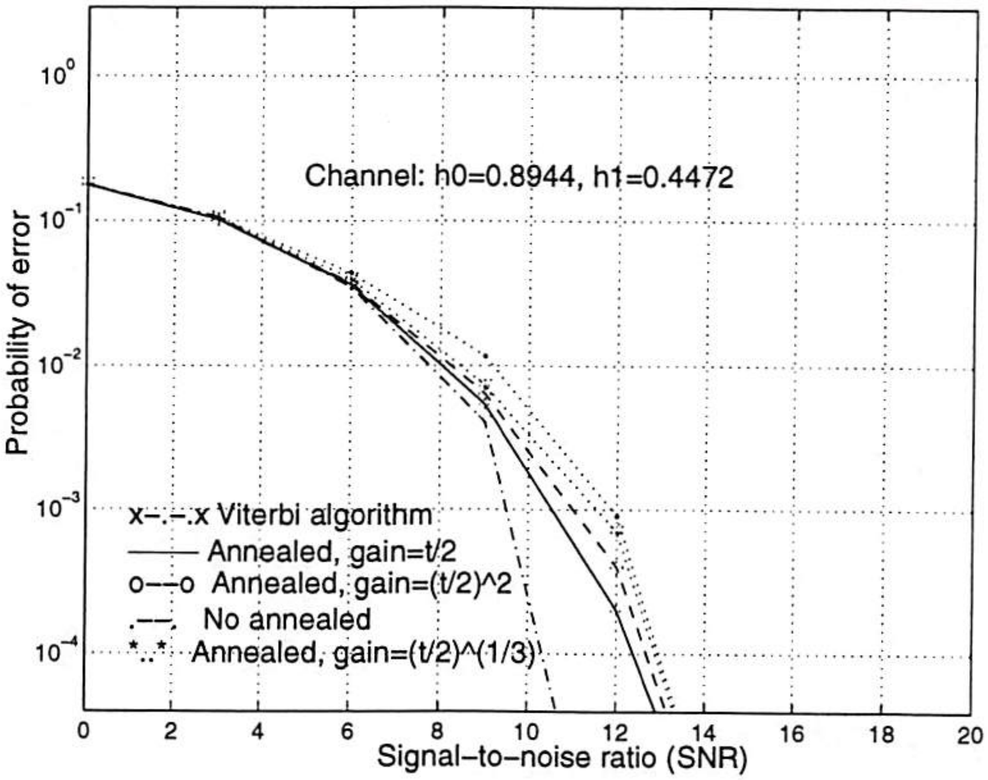


Figure 4.5: Performance of different annealing gains.

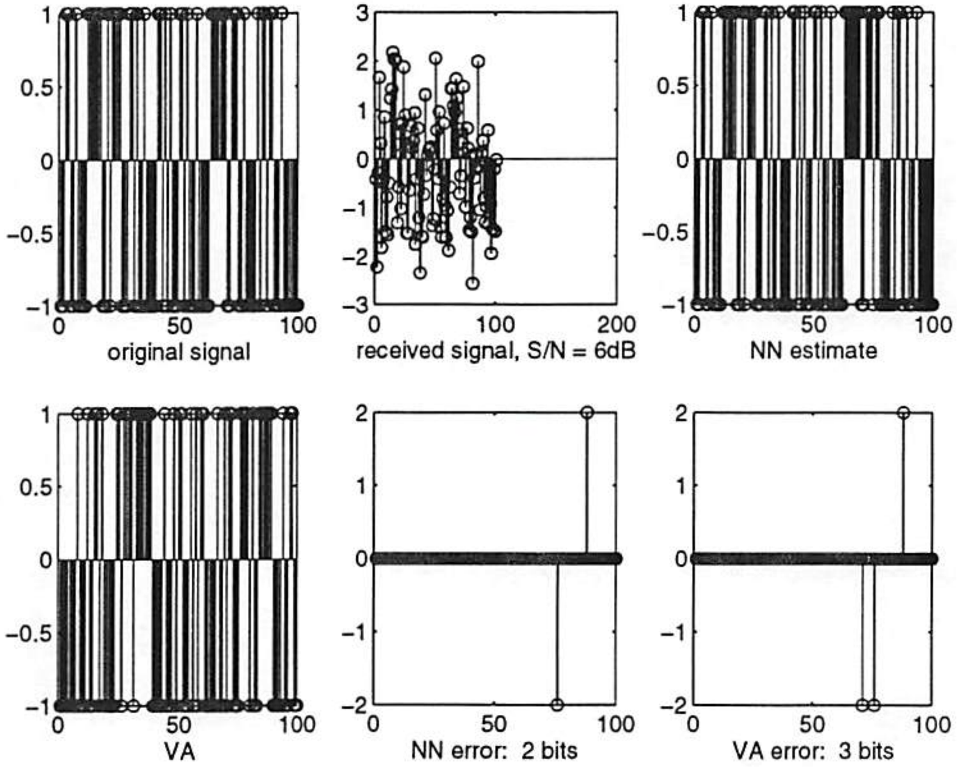


Figure 4.6: Error bits comparison between VA and 1-D receiver: case 1.

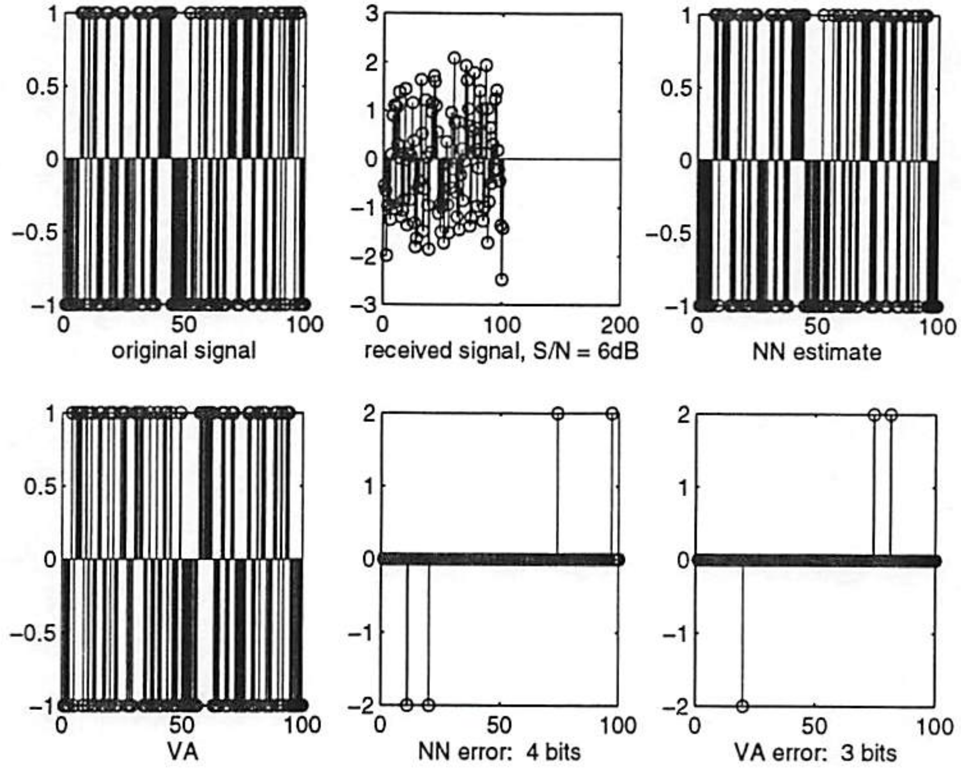


Figure 4.7: Error bits comparison between VA and 1-D receiver: case 2.

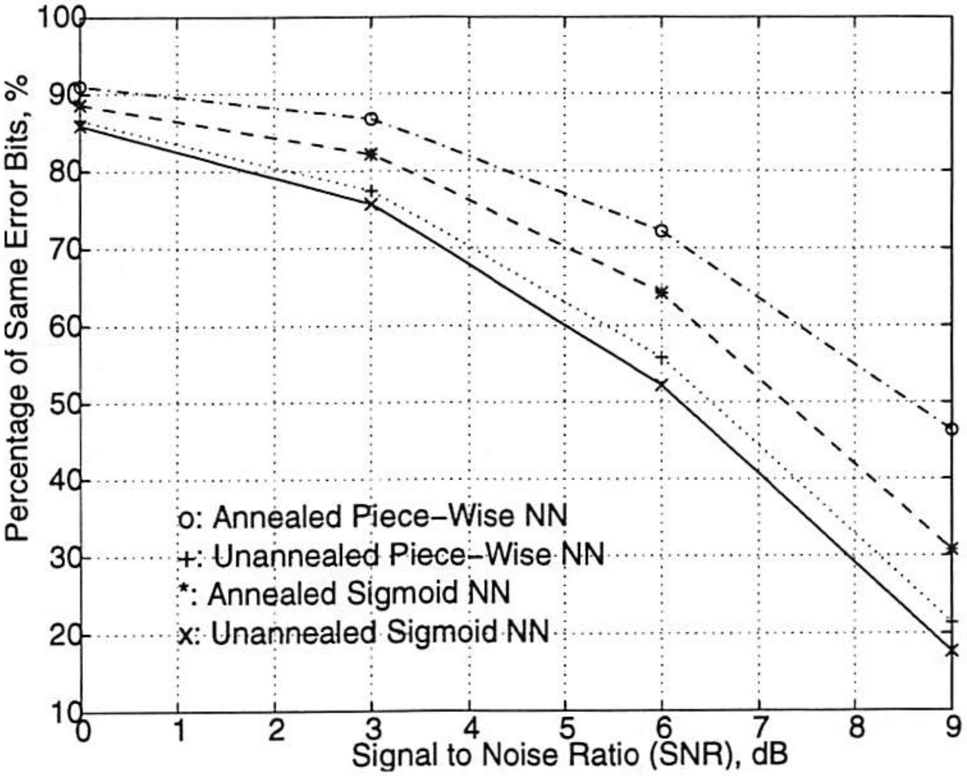


Figure 4.8: Plot of percentage of same position error bits vs. signal-to-noise ratio.

Reference List

[1] O. J. Joeressen, H. Meyr, "A 40 Mb/s soft-output Viterbi decoder," *IEEE. J. of Solid-State Circuits,* vol. 30, no. 7, pp. 812-818, July, 1995.

[2] Bing J. Sheu, Joongho Choi, *Neural Information Processing and VLSI,* Kluwer Academic Publishers: Boston, MA, 1995.

[3] A. C. Cichocki, R. Unbehauen, *Neural Networks for Optimization and Signal Processing,* John Wiley & Sons: New York, NY, 1993.

[4] S. Mo, B. Shafai, "Blind equalization using higher order cumulants and neural networks," *IEEE Trans. on Signal Processing,* vol. 42, no. 11, pp. 3209-3217, Nov. 1994.

[5] S. Chen, C. F. N. Cowan, P. M. Grant, "Orthogonal least square learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks,* vol. 2, no. 2, pp. 302-309, Mar. 1991.

[6] S. Chen, B. Mulgrew, S. McLaughlin, "Adaptive Baysian equalizer with decision feedback," *IEEE Trans. on Signal Processing,* vol. 41, no. 9, pp. 2918-2927, Sep. 1993.

[7] S. H. Bang, B. J. Sheu, "A neural network for detection of signals in Communication," *IEEE Trans. on Circuits and Systems, Part I,* vol. 43, pp. 644-655, Aug. 1996.

[8] L. O. Chua, L. Yang, "Cellular neural networks: Theory," *IEEE Trans. on Circuits and Systems,* vol. 35, pp. 1257-1272, Oct. 1988.

[9] L. O. Chua, L. Yang, "Cellular neural networks: Application," *IEEE Trans. on Circuits and Systems,* vol. 35, pp. 1273-1290, Oct. 1988.

[10] B.W. Lee, B.J. Sheu, "Modified Hopfield neural networks for retrieving the optimal solution," *IEEE Trans. Neural Networks*, vol. 2, pp. 137-142, Jan. 1991.

[11] S. Bang, B.J. Sheu, Eric Y. Chou, "A hardware annealing method for optimal solutions on cellular neural networks," *IEEE Trans. on Circuits and Systems, Part II,* vol. 43, no. 6, pp. 409-421, June 1996.

[12] Y. Takefuji, K.C. Lee, "Artificial neural networks for four-coloring map problems and K-colorability problems," *IEEE Trans. Circuits Systs.*, vol. 38, pp. 326-333, Mar. 1991.

[13] P.M. Pardalos, J.B. Rosen, *Constrained Global Optimization: Algorithms and Applications*, Springer-Verlag: Berlin, Germany, 1987.

Chapter 5

Another Application: 1-D Compact Neural Network Detector for Hard Disk Drive

5.1 Application and Advantages in Hard Disk Drive

In this chapter we will apply our detector built by compact neural network to the hard disk driver for personal computers. Continuous advances of personal computers, including the desktop, laptop, or notebook computers, have created the need to reduce the size and weight of the products. According to statisics in hard disk drivers, technology advances have made the annual increase of data storage capacity and data rate up to 60% and 40% respectively during 1990-1996, while the annual decrease of cost per mega-byte storage is 40%. To maintain such an improvement, it is desirable to explore new design techniques to improve hard disk driver electronics with high performance, low cost and low power dissipation. Presently, most systems use run-length limited (RLL) coding and peak detection (PD) [3] to achieve high reliability and high storage densities on the order of 100,000 $bits/mm^2$ in rigid-disk drivers. The application of the more advanced technique consisting of partial-response (PR) signaling in combination with maximum-likelihood sequence detection (MLSD) makes much more progress. This technique is called PRML and was first introduced by H. Kobayashi at IBM Corp. in 1971 [4]. PR signaling allows the data rate to be increased for a given bandwidth channel, but the signal-to-noise ratio (SNR) is degraded. The MLSE allows part of the lost SNR to be regained. PRML was utilized in an experimental system and achieved a density of 1.8 million $bits/mm^2$ [1]. In 1994, IBM Corp. employed PRML detection in 5.25-inch disk drive in the new RISC 6000 workstations to become the first company to use the technique for commercially available disc drive [1]. High cost of digital implementations was the primary reason that this technique was delayed for the commercial use for such a long time. This provides good motivation for the analog implementation of PRML. Analog implemenations may prove to be cheaper, lower and faster.

5.2 Magnetic Recording Channel

Fig. 5.1 shows the read channel block diagram of magnetic recording. Saturation recording is used to record the binary data. At the receiver, the signal is first passed through an automatic gain control (AGC) amplifier, an analog-to-digital converter (A/D), a finite-impluse-response (FIR) filter to shape the spectrum, then a detector [5, 6]. To understand the analog detector for the read channel

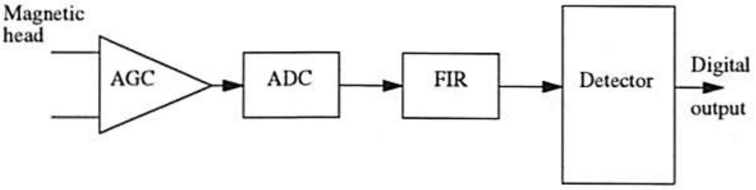


Figure 5.1: Read channel block diagram.

block diagram shown in Fig. 5.1, we have to study the behavior of the magnetic recording system first.

Fig. 5.2 shows the read/write waveforms for saturation magnetic recording. In the magnetic storage, the binary data are recorded by changing the direction of magnetization of the media, which is fully saturated for all disc drivers [7]. As the read head passes over a transition in the direction of magnetization, it responds to the rate-of-change of the magnectic flux. According to the Lentz's law, the flux change will produce a voltage. This means the data information is contained in the orientation of transitions in the magnetization [2]. The readback voltage can be viewed as the convolution operation between the channel step response and the derivative of magnetization pattern. In Fig. 5.2, the width of the channel step response is narrow compared to the transition spacing so that it does not cause significant ISI. Because the output of the channel at time k is the weighted sum of the present channel input and the previous channel input, i.e.,

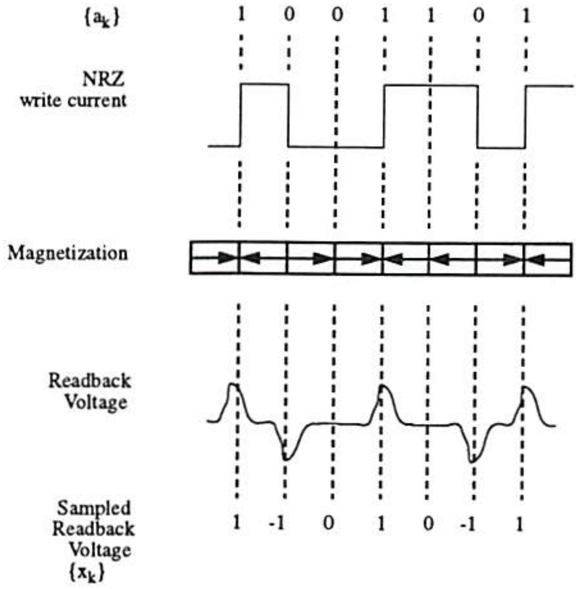


Figure 5.2: Read/Write waveforms for digital magnetic recording [7].

$x_k = b_k - b_{k-1} = (1 - D)b_k$, where D is a unit delay operator, a digital magnetic recording system is inherently partial response. The channel $1 - D$ is called dicode response.

The transition spacing determines the storage density. Generally, the transition spacing is about two to three times of the width of the channel step response in a practical system, and ISI appears. Therefore, the partial response is not so simple as $1 - D$. When the rest part of frequncy response of a magnectic recording system is considered, a good choice for the PR polynomial is either PR IV, represented by $1 - D^2$, or one of the class of extended PR IV (EPR IV), represented by $1 + D - D^2 - D^3$. Both polynomials exhibit spectral nulls at dc and at the Niquist frequency. These characteristics are appropriate for the type of band-pass channels encountered in the magnetic recoding system. PR IV is more attractive than EPR IV due to its simplicity to implement and good representation to the actual system. From the circuit perspective, the output of a PR IV channel included two independent interleaved sequences, each with a dicode response having a two-baud delay. The detector can be implemented as two parallel dicode detectors operating at half the symbol rate. From the system perspective, PR IV has two major advantages: reasonably matching to the unequalized channel so the required equalizer has acceptable SNR degradation and shown to be much more robust than some other polynomials in the presence of off-track interference, gain errors, and misequalization.

5.3 PRML System for Digital Magnetic Recording

Fig. 5.3 shows the PRML systems for digital magnetic recording. The combined channel includes the magnetic recording channel Q(t); the analog filter F(t) and the finite impulse response filter C(t) [8]. For detailed understaning, two different representations with the same function are considered. Fig. 5.4 shows the discrete-time system without error in gain and sampling time. The spectrum of the combined channel $\bar{H}$ is also shown. This spectrum can be viewed as the production of $1+D$ and $1-D$. Fig. 5.5 represents the PR IV sequence which includes two independent sub-sequences which can be represented as the dicode sequences.

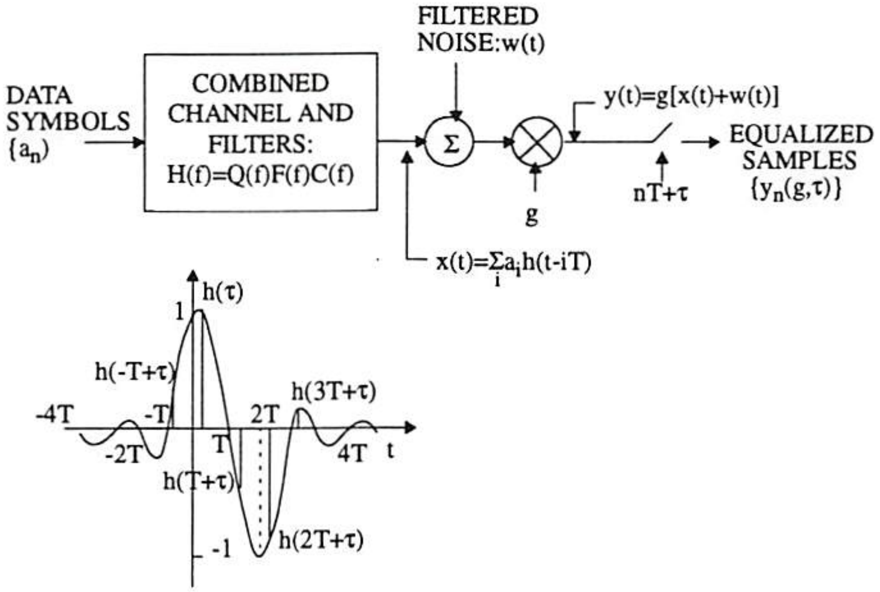


Figure 5.3: Equivalent model of PR signal system [8].

In Fig. 5.3, $h(t)$ can be written as

$$h(t) = d(t) - d(t - 2T) \tag{5.1}$$

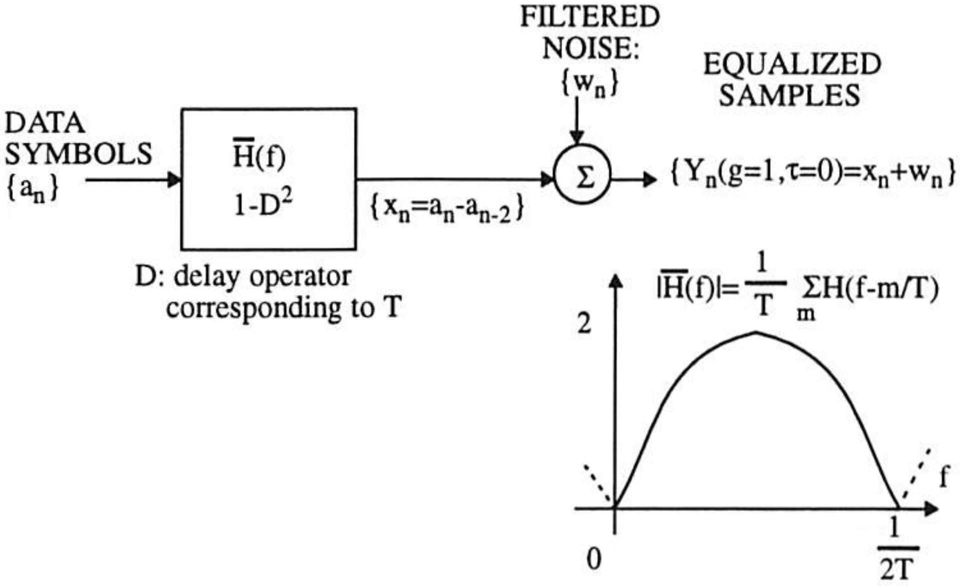


Figure 5.4: Discrete-time system and channel spectrum of PR IV recording scheme [8].

where $d(t)$ is any Nyquist signal element. At the sampling time nT,

$$p(nT) = \begin{cases} 1 & for \quad n = 0 \\ 0 & for \quad n < 0. \end{cases} \tag{5.2}$$

Therefore, $\{h(nT)\} = \{\cdots, 0, h(0) = +1, h(2T) = -1, 0, \cdots\}$. x(t) is the output of the combined channel and y(t) is the signal corrupted by Gaussian noise w(t),

$$y(t) = g(x(t) + w(t)) = g(\sum_k a_k h(t - kT) + w(t)). \tag{5.3}$$

At time $nT + \tau$, (5.3) becomes

$$\begin{aligned} y_n(g, \tau) &\equiv y(nT + \tau) \\ &= g(\sum_k (a_k - a_{k-2}) \cdot d((n-k)T + \tau) + w(nT)). \end{aligned} \tag{5.4}$$

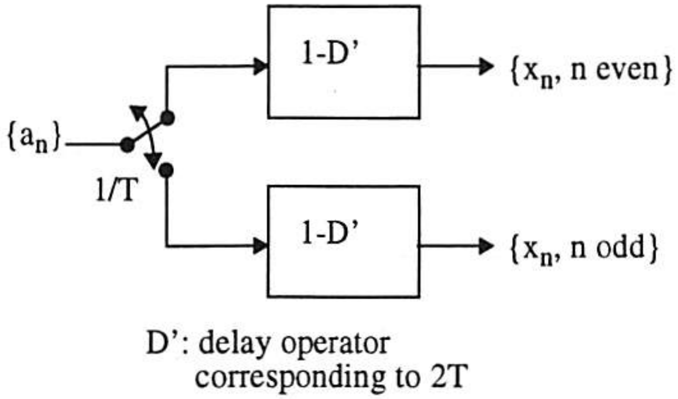


Figure 5.5: PR IV sequence includes two independent dicode sequences [8].

For the correct overall gain and at the ideal sampling time nT, $g = 1$ and $\tau = 0$, the equalized sample can be represented as

$$y_n(g = 1, \tau = 0) = x_n + w_n. \tag{5.5}$$

From (5.2) and (5.4), the sampled output of the combined channel and filters, x_n, is equal to

$$x_n = a_n - a_{n-2}, \qquad x_n \in \{-2, 0, +2\}. \tag{5.6}$$

The block diagrams of (5.5) and (5.6) are shown in Fig. 5.4. The samples x_n are the outputs of a discrete channel characterized by the PR IV polynomial $1 - D^2$, where D is the unit delay T operator. The frequency response of this discrete channel is given by

$$\bar{H}(f) = \frac{1}{T}\sum_m H(f - m/T) = 1 - e^{\frac{-j4\pi}{T}}. \tag{5.7}$$

The spectrum of $\bar{H}$ as shown in Fig. 5.4 has nulls at the dc and Nyquist frequency $\frac{1}{2T}$. This characteristic meets the requirement described in the previous section and therefore is well-suited for the type of band-pass channel encountered in magnetic recording system.

From (5.6), it is clear that see the data sequence x_n is only dependent on the input symbol $\{a_n$, n odd(even)$\}$ for n odd(even). The PR IV sequence can be viewed as two independent and interleaved dicode PR sequences with polynomial 1-D$'$. Here, D$'$ represents the operator with 2T delay. Fig. 5.5 shows the block diagram of this characteristic and the output sequence $\{x_n\}$ will be fed to the Viterbi detector which implements the MLSE decision rule.

5.4 Circuit Block Diagram for PRML System

The circuit block diagram for PRML Hard-Disk Drive (HDD) is shown in Fig. 5.6 [9]. There are three main parts: analog front-end signal processor, companion digital ASIC and Viterbi detector. The basic operation principle is described here. The analog read signal first goes through an automatic gain control (AGC)

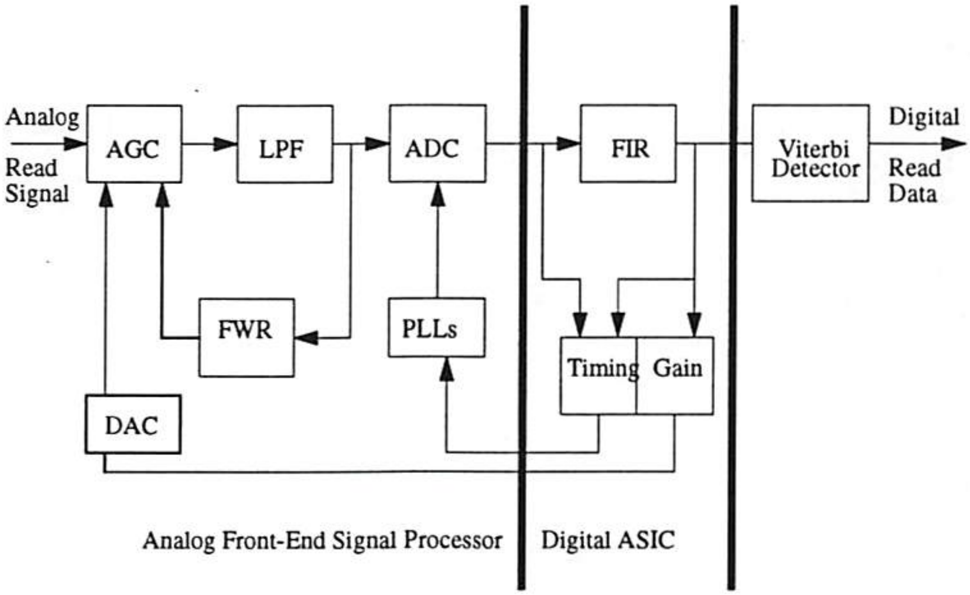


Figure 5.6: Circuit block diagram of hard disk drive.

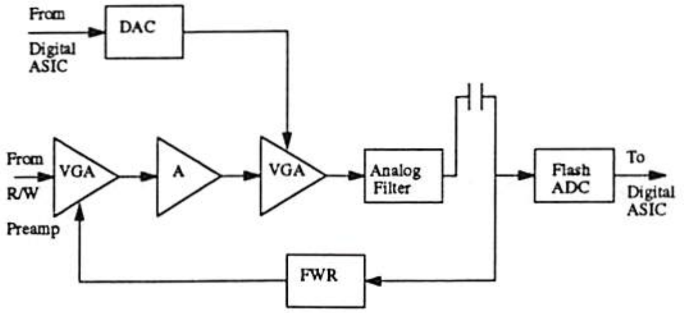


Figure 5.7: Block diagram of AGC.

circuit and a low-pass filter which generally is of 5th or 7th order. Then the signal is synchronously sampled by an A/D converter (ADC) whose clock is provided by the synchronizer (SNC) phased-looked loop (PLL). The output samples of ADC is fed to a finite impulse response (FIR) filter to equalize these samples to a PR IV spectrum shape. Finally, the signal is passed through a Viterbi detector which implements the MLSE decision rule. The output of the Viterbi detector is the final data.

The AGC can be implemented as a variable gain amplifier (VGA) whose gain is controlled by the feedback loop from the output of the FIR filter to the VGA through a digital-to-analog converter (DAC) block. More delicate design approach using three separate Gilbert multipliers was also used [10]. Fig. 5.7 shows the block diagram of AGC, which consists of 3 fully-differential gain stages that amplify the low-amplitude signals from the read head preamplifier to a signal level suitable for the rest of the read channel circuitry. The gain of AGC can be independently set by two separate loops: analog loop and digital loop. The analog loop controls the gain through the full-wave rectifier. The digital loop closes the AGC via a digital-to-analog converter (DAC) from the digital block.

The synchronizer phase-locked loops (SNC PLL) are used to generate the READ clock for hard disk drive. The SNC PLL block will generate a low-gitter clock to sample the signal retrieved from the hard disk. In the READ mode the SNC-CLK is used to synchronize the ADC which samples and converts the amplified and filtered signal. The READ clock phase is established by phase-locking the SNC-CLK to a synchronization preamble field written at the start of each data sector.

The ADC is a key block for the magnetic recording process. To reach the high speed requirement, flash ADC is employed. In addition to the speed concern, a flash ADC can help to avoid the stability problem caused by the excess delay from the time recovery loop. Fig. 5.8 shows an example of flash ADC [10]. A fully differential 6 bit 72 MHz flash converter is appropriate. It consists of a differential sample and hold, a difference reference, a comparator array, a CMOS ROM encoder and output buffers. An ADC without sample-and-hold (S/H) circuitry

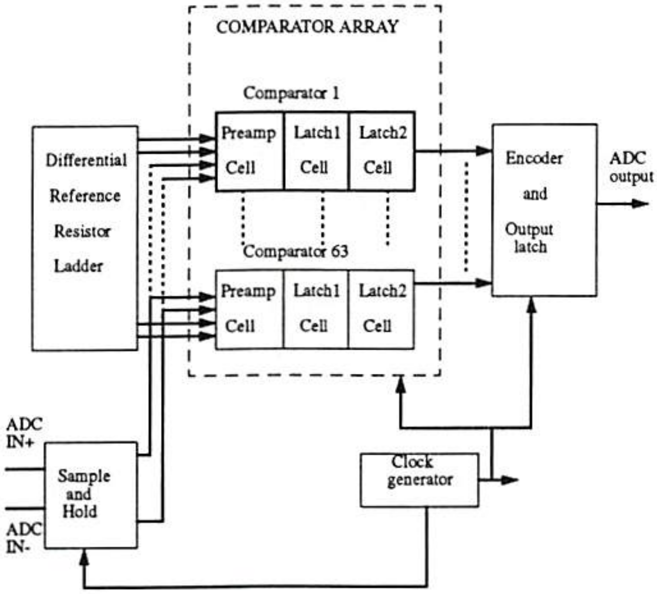


Figure 5.8: An example of flash ADC [10].

at their input node has strict requirements placed on their comparators and clock generation circuitry in order to insure that each comparator senses the same input voltage at the same time. Placing the S/H circuitry at the input node helps to relax the requirements on the comparators and clock generation. Because of the severe requirements imposed on the input slew rate and sampling jitter of ADC, the S/H circuitry is placed at the input node to the ADC.

5.5 Discussion

From the previous subsections, the magnetic recoding channel $1 - D^2$ is usually implemented by two circuit modules: analog front-end processor and digital ASIC. The Viterbi detector is included in the digital ASIC module. To achieve low cost and low power operation, research interest on using analog circuits to replace the digital ASIC including the use of analog Viterbi detector, has been very high. Our 1-D compact neural network design is a very promising solution.

Reference List

[1] R. W. Wood, "Magnetic megabits," *IEEE Spectrum*, pp. 32-38, May 1990.

[2] T. W. Matthews, R. R. Spencer, "An integrated analog CMOS Viterbi detector for digital magnetic recording," *IEEE J. of Solid-State Circuits,* vol. 28, no. 12, pp. 1294-1302, Dec. 1993.

[3] P. Pai, A. A. Abidi, "A 40-mW 55 Mb/s CMOS equalizer for use in magnetic storage read channels," *IEEE J. of Solid-State Circuits,* vol. 29, no. 4, pp. 489-499, Apr. 1994.

[4] H. Kobayashi, "Application of probablistic decoding to digital magnetic recording systems," *IBM J. Res. Develop.,* pp. 64, Jan. 1971.

[5] C. A. Laber, P. R. Gray, "A 20-MHz sixth-order BiCMOS parasitic- intensive continuous-time filter and second- order equalizer optimized for disk-drive read channels," *IEEE J. of Solid-State Circuits,* vol. 28, no. 4, pp. 462-470, Apr. 1993.

[6] S. Mita, et al., "A 150Mb/s PRML chip for magnetic disk drive," *Proc. of IEEE Int. Solid-State Circuits Conference,* pp. 62-63, San Francisco, CA, Feb. 1996.

[7] R. R. Spencer, "Simulated performance of analog Viterbi detectors," *IEEE J. on Selected Areas in Commun.,* vol. 10, no. 1, pp. 277-288, Jan. 1992.

[8] R. D. Cideciyan, et al., "A PRML system for digital magnetic recording," *IEEE J. on Selected Areas in Commun.,* vol. 10, no. 1, pp. 201-216, Jan. 1992.

[9] G. Tyson, et al., "A 130Mb/s PRML read/write channel with digital-servo detection," *Proc. of IEEE Int. Solid-State Circuits Conferenc,* pp. 64-65, San Francisco, CA, Feb. 1996.

[10] C. S. Wong, et al., "A 50 MHz eight-tap adaptive equalizer for partial-response channels," *IEEE J. of Solid-State Circuits,* vol. 30, no. 3, pp. 228-234, Mar. 1995.

Chapter 6

Compact Neural Network Based CDMA Detector with Robust Near-Far Resistance

The use of the spread spectrum communication technology originated in the unique needs of military communication. This technology grew out of research efforts during World War II for the purpose of providing secure means of communication in hostile environments, i.e., hidding the transmitted signal from eavesdropper and overcoming the intentionally strong interference [1]. By spreading the spectrum of the transmitted signal, this goal can be achieved. With the fast development of electronic technology and highly worldwide commercial demand in the quality and quantity of mobil cellular communication systems and personal communications service (PCS), spread spectrum digital technology becomes more popular and important. There are two most important schemes of spread spectrum technology: direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (HPSS) [2] [3]. Most spread spectrum communication systems have been developed based upon the two schemes. Among the various systems for the spread spectrum technology, Code Division Multiple Access (CDMA), which is a DSSS method, has received significant attention and become the most booming approach. Accroding to Management Consultants International, in Washington, D.C., CDMA carriers have the potential to generate annual revenues of \$10 billion by the year 2000 [4].

In CDMA communication, each user is given a unique and distinctive code. These codes are almost uncorrelated with one another and used to spread the transmitted signals to the full availble bandwidth. It means signal collisions are not destructive and each of the signals involved in a collision only results in a slight increase in error rate. Many more subscribers are allowed to share the same frequency band and the efficiency is increased. Besides, the issues of allocating different frequencies to different users or cells are eliminated.

To ensure high transmission quality, it is very important to control the signal power rapidly and accurately. However, it is not an easy task. In satellite communication, there may be high-power and low-power transmitters. In ground communication, one user may be closer to the receiver while the other user may be far from the receiver. When an unwanted user's received signal power is much larger than the received signal power presented by the desired user, the performance of CDMA system is largely degraded. This problem is referred to as the near-far problem and is one major technical issue in CDMA systems. To upgrade the system quality, it is necessary to develop a technology to reduce or overcome the effects of the near-far problem.

Recently, B. Aazhang et al. [5] and U. Mitra [6] proposed feedforward neural networks based detectors for multiuser accessing. They achieved good performance when the number of users were small. However, the hardware complexity appeared to increase exponentially as the number of users incresed. In 1986, Sergio Verdu [7] showed that an optimized detector for near-far resistant multiuser demodulation was possible by minimizing a quadratic objective function, assuming that users' signals were uncorrelated and their spreading codes (i.e. pseudorandom codes) were known. This was a very encouraging result because our proposed compact neural network with the hardware annealing function will be a powerful tool to find the optimized solution by minimizing a quadratic objective function. In this chapter, a compact neural network for CDMA detector with robut resistance to the near-far effect and the optimized solution is described.

6.1 Traditional Multiple Access Communication

Traditionally radio communication systems have separated users by either frequency channels, time slots, or both. These concepts date from the earliest days of radio technology. Even spark transmitters used resonant circuits to narrow the spectrum of their radiation. Scheduled net operation was probably the first manifestation of time slotting. Modern cellular systems began with the use of channelized analog FM. More recently several hybrid FDM-TDM digital systems have been developed for service quality and capacity. In all these systems, each user is assigned a particular time-frequency slot.

In large systems the assignments to the time-frequency slots cannot be unique. Slots must be reused in multiple cells in order to cover large service areas. Satisfactory performance in these systems depends critically on control of the mutual interference arising from the reuse. The reuse concept is familiar even in television broadcasting, where channels are not reused in adjacent cities.

The cellular telephone system used in North America allocates approximately AMPS 416 channels to each operator (30kHz spacing, with a total allocation of 12.5 MHz in each direction). The same frequency obviously cannot be reused in any adjacent pair of cells because a user on the boundary between those cells would receive both signals with equal amplitude, leading to an unacceptably high interference level. A plane can be tiled with hexagonal cells, labeled in accordance with the seven-way pattern shown in Fig. 6.1. Therefore, if a unique set of channels is assigned to each of the seven cells, then the pattern can be repeated without violating the adjacent requirement. Although this idealized pattern is not strictly applicable in all real systems, the seven-way reuse pattern is very desirable. The capacity of systems built in this way is determined by the bandwidth per channel and the seven-way reuse pattern. In an AMPS, therefore, the maximum capacity per cell is approximately 416/7 = 59. For three-way sectored cells, the same $K = 7$ reuse applies over all three sectors, that is, only about 59/3 = 19 channels are available in each sector. In an ideal geometry the reuse pattern looks like Fig. 6.1, representing channel sets by distinct numbers.

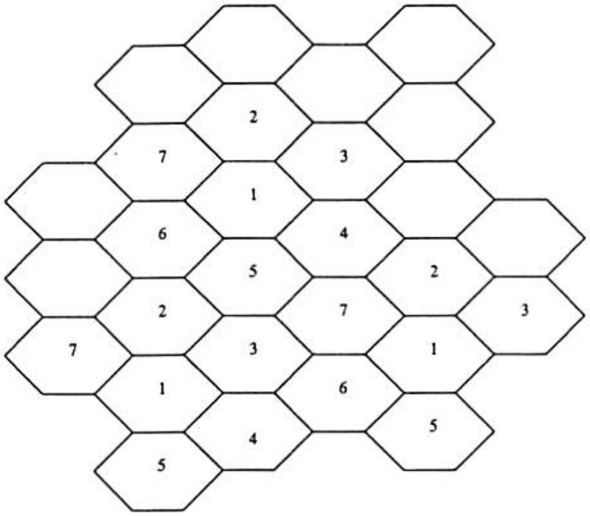


Figure 6.1: Frequency reuse map.

It should be noted that achievement of the $K = 7$ reuse, rather than an even larger number, depends on the fact that the effective propagation decay law is faster than free space. In a vacuum space, electromagnetic radiation decays in intensity like R^{-2}. However measurements have consistently shown that the effective propagation law exponent is typically between -3.5 and -5 in the ground mobile environment.

CDMA offers an answer to the capacity problem. The key to its high capacity is the use of noise-like sequences, i.e., pseudorandom sequences, as first suggested decades ago by Claude Shanon [1]. Instead of partitioning either the frequency spectrum or time into disjoint "slots", each user is assigned a different pseudorandom sequence,

6.2 CDMA Communication

6.2.1 Pseudorandom Sequence

Pseudorandom sequence (or pseudonoise code, PN code) is one of the "standard components" in the CDMA system [8]. Let's review the generation of pseudorandom sequences and some interesting and useful properties of these sequences.

A pseudo-random sequence generator is shown in Fig. 6.2. $S_1, S_2, \cdots, S_n$ are the symbols for the shift registers. Not all of the shift registers should be connected to the parity generator as indicated by the dashed lines in the figure. The

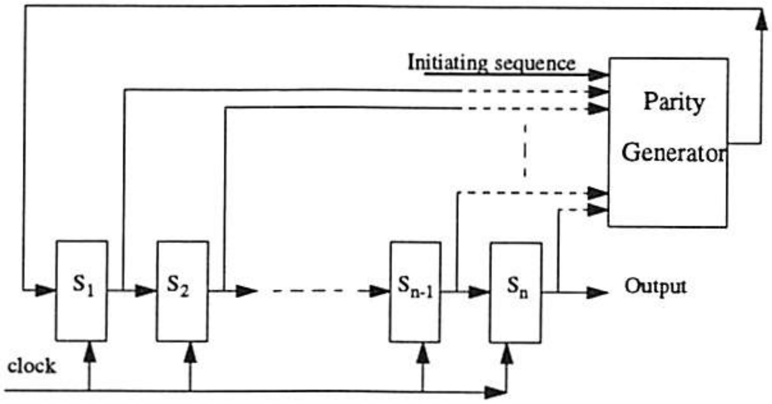


Figure 6.2: A example of pseudo-noise generator.

initialization sequence and the outputs of the shift registers are sent to the parity generator. If the inputs of the parity generator are even number of logic 1, the output of the parity generator is logic 0. Otherwise, it is logic 1.

The output is also called linear feedback shift register (LFSR) sequence. The period of the output is dependent on the initialization sequence. Every LFSR sequence is periodic with period $P \leq 2^n - 1$. There exist initialization sequences which result in an LFSR sequence with period $P = N = 2^n - 1$. LFSR sequence whose period being equal to N is called maximum length (linear) shift register (MLSR) sequence. MLSR sequences have some interesting and useful properties for the spread spectrum communication applications:

- The number of one's in an MLSR sequence is $\frac{1}{2}(N+1)$, which is one more than the number of zero's.

- The binary sum of an MLSR sequence and its phase-shift version is another phase-shift version of the original MLSR sequence.

- The autocorrelation function $R_{PN}(\tau)$ has only two values: 1 or $\frac{-1}{N}$.

Fig. 6.3 shows the plot of autocorrelation of pseudorandom sequence. When N becomes very large, the autocorrelation function of a pseudorandom sequence is very much similar to that of white noise. That's the reason why pseudorandom noise is also called PN codes.

Suppose $A_1 = [0, 1, 1, 1, 0, 0, 1]$ is an MLSR sequence and A_2 is the circularly phase-shift version of A_1, i.e., $A_2 = [1, 0, 1, 1, 1, 0, 0]$. Therefore, A_1 and A_2 are all MLSR sequences. Now let's check the properties of MLSR sequences described above.

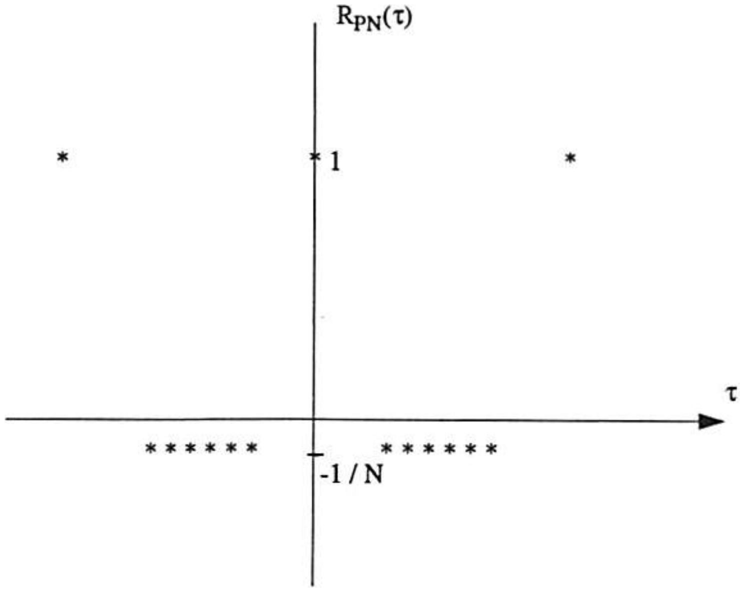


Figure 6.3: Autocorrelation of pseudorandom sequence.

- The number of one's in each sequence is 4, which is one more than the number of zero's.

- The binary sum of A_1 and A_2 is $A_3 = [1, 1, 0, 0, 1, 0, 1]$, which is another phase-shift version of A_1.

- Here, conversion from unipolar representation to bipolar representation is made first. In the new representation, let $A_1 = [-1, 1, 1, 1, -1, -1, 1]$ and $A_2 = [1, -1, 1, 1, 1, -1, -1]$ first. The correlation of A_1 and A_1 is $((-1) \cdot (-1)+1 \cdot 1+1 \cdot 1+1 \cdot 1+(-1) \cdot (-1)+(-1) \cdot (-1)+1 \cdot 1)/7 = 1$. The correlation of A_1 and A_2 is $((-1) \cdot 1+1 \cdot (-1)+1 \cdot 1+1 \cdot 1+(-1) \cdot 1+(-1) \cdot (-1)+1 \cdot (-1))/7 = -1/7$.

6.2.2 Short Code

The short code is a pair of period 2^{15} sequences that are used for spreading the CDMA Forward Channel. They are also used in conjuction with the Long Code for spreading the CDMA Reverse Channel. They are both derived from period $2^{15}-1$ Linear Feedback Shift Register (LFSR) shown in the subsection 6.2.1. The short code LFSR tap polynomials are, for the I-sequence

$$P_I(x) = x^{15} + x^{13} + x^9 + x^8 + x^7 + x^5 + 1, \tag{6.1}$$

and for the Q-sequence

$$P_Q(x) = x^{15} + x^{12} + x^{11} + x^{10} + x^6 + x^5 + x^4 + x^3 + 1. \tag{6.2}$$

The extra zero bit is inserted in each sequence immediately after the occurence of 14 consecutive zeros from the generator register. This occurs once per period.

6.2.3 Long Code

A long code is a period $2^{42}-1$ LFSR sequence that is used for spreading the reverse link. There is only one long code sequence. Diffreent stations are distinguished not by the sequence itself but by its relative phase. The long code is added to each of the two (I and Q) short code sequences to ensure that cross correlations between the signals from distinct stations are always small.

The long code LFSR tap polynomial is

$$\begin{aligned} G(x) &= x^{42} + x^{35} + x^{33} + x^{31} + x^{27} + x^{26} + x^{25} + x^{22} + x^{21} + x^{19} \\ &+ x^{18} + x^{17} + x^{16} + x^{10} + x^7 + x^6 + x^5 + x^3 + x^2 + x^1 + 1. \end{aligned} \tag{6.3}$$

The different phases of the long code are generated by use of one of the well-known properties of LFSR sequences. Any modulo-2 sum of two different phases of LFSR sequences results in a new phase-shift version of the two sequences. A corollary of this property is the fact that all internal nodes of any LFSR generator also run through the same sequences at the generator output, but with different phases.

The additional property of LFSR sequences is exploited in the long code generation process for the reverse link spreading. A 42-bit number, the Long Code Mask, is used to select particular bits of the 42-bit long code generator register. The selected nodes are processed by the summation and modulo 2 operation. The resultant of the sum, that is, the modulo-2 inner product of the generator state with the mask, is the generator output corresponding to that mask.

6.2.4 Basic Principles of CDMA Communication

Fig. 6.4 shows the diagram of a modem of the CDMA communication system for the kth user. CDMA system is of the DSSS shceme. At the transmitted end, the input data are first multiplied by the PN code. Each user is assigned a different code. The bit rate of the PN code should be much higher than the information bit rate. After being multiplied by the PN code, the input data are again modulated by the sinusoidal carrier with frequency f_o and then sent out by the tramsmitter. Thus, the signals are modulated twice. At the receiving end, the received data is first demodulated by the sinusoidal carrier coherently and then multiplied by the same PN code synchronously. Fig. 6.5 shows the diagram of modulation of the input signal waveform by the PN code. The information bit rate is smaller than the bit rate of PN code. Therefore, the spectrum of the modulated signal $b(t)s(t)$ is much wider than that of the data signal.

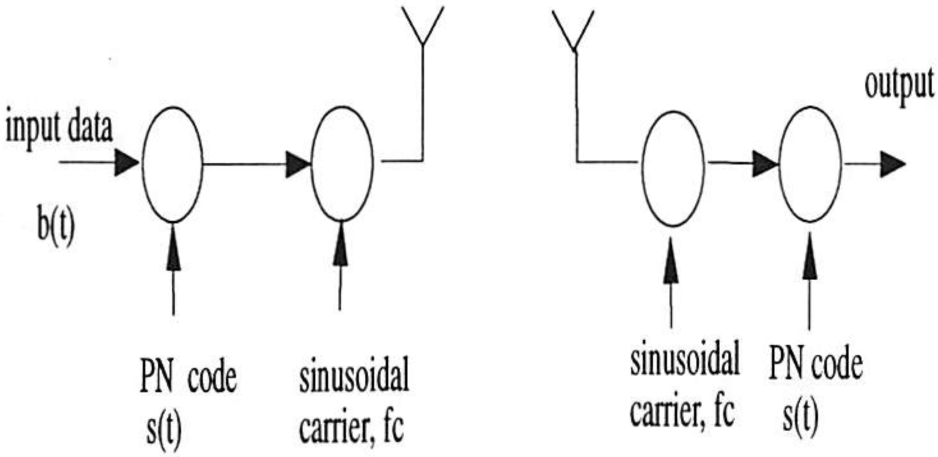


Figure 6.4: Modem of CDMA communication system.

6.2.5 Reverse CDMA Channel

The reverse CDMA channel handles the mobile-to-base direction of communication. The mobile device communicates with the base station over access channel or the reverse traffic channel. The access channel is for origination, process orders, and responding to paging. After voice or data communication is established, the traffic channel is used. Any particular reverse channel is active only for calls to the associated mobile station, or when access channel signaling is taking place to the associated base station.

In the IS-95A cellular service, the transmit frequency of mobile station is 45 MHz below that of base station. In the ANSI J-STD-008 PCS, the transmit frequency of mobile station is 80 MHz below that of base station. Permissible frequency assignments are in 30 KHz increments in cellular and 50 kHz in PCS.

There are $2^{42}-1$ reverse CDMA channels [3]. Every mobile station is assigned uniquely and permanently to one of these logical channels. When transmitting traffic, every mobile station uses one logical channel. That channel is used by

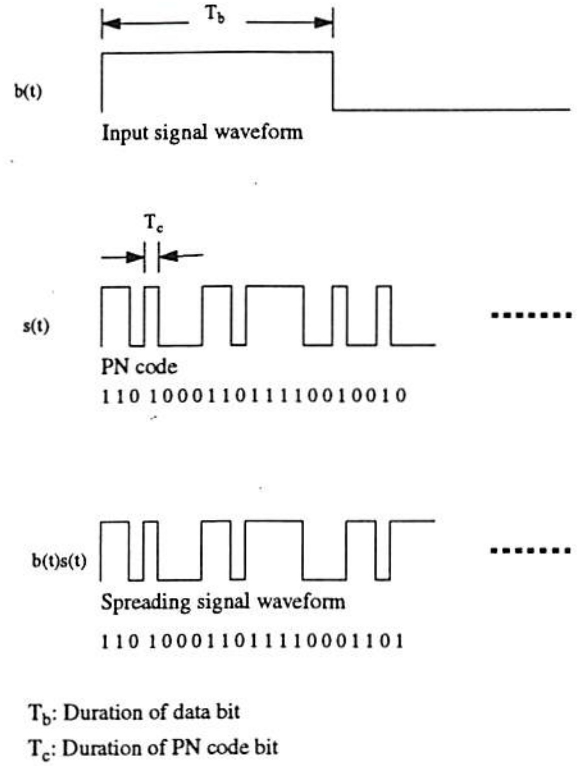


Figure 6.5: Modulation waveforms.

the mobile device whenever it transmitts traffic. The channel does not change upon handoff. Other logical channels are associated with base stations for system access. This reverse link addressing is accomplished through manipulation of period $2^{42} - 1$ Long Code, which is part of the spreading process. Fig. 6.6 shows the block diagram of the core processing that generates one Reverse CDMA traffic channel.

6.2.6 Forward CDMA Channel

The CDMA forward channel is for base-to-mobile communication. It includes a pilot channel, an operational sync channel, optional paging channels, and several forward traffic channels. The pilot is a spread, but otherwise unmodulated direct

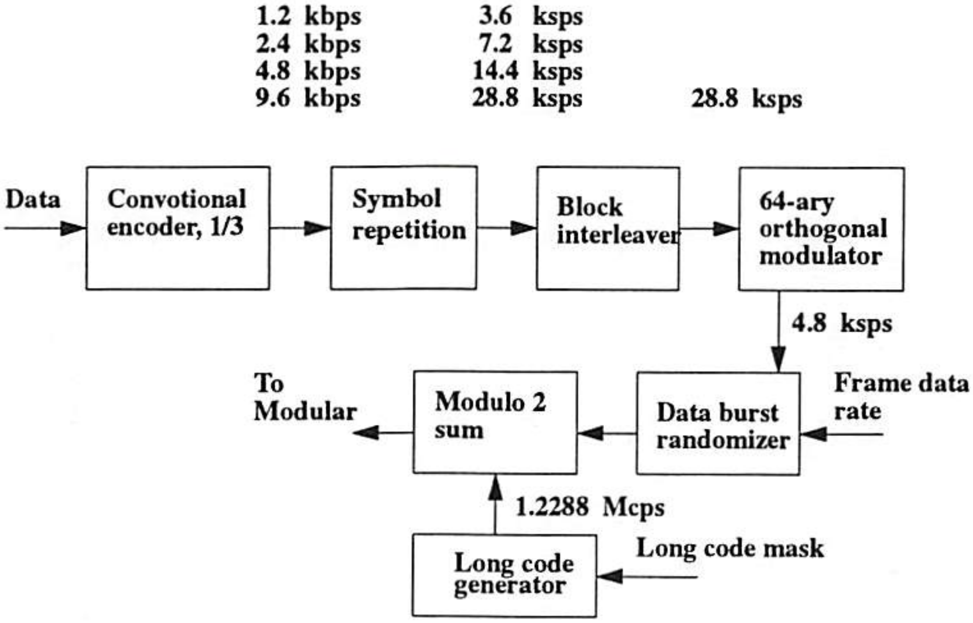


Figure 6.6: Reverse CDMA traffic channel [2].

sequence spread spectrum (DSSS) signal. These channels are orthogonalized, spread, added together and then sent to the modulator. The pilot and overhead channels establish the system timing and station identity. The pilot channel is also used in the mobile-assisted handoff (MAHO) process. Processes of constructing the forward channels are very similar to those of the reverse channels. Fig. 6.7 shows the forward CDMA traffic channel.

6.2.7 The Importance of Power Control

The key to the high capacity of commercial CDMA is extremely simple. If, rather than using constant power, the transmitters can be controlled in such a way that the received powers from all users are roughly equal, then the benefits of spreading are realized. For controlled received power, then the subscrbers can occupy the same spectrum, and the benefits of interference averaging accrue.

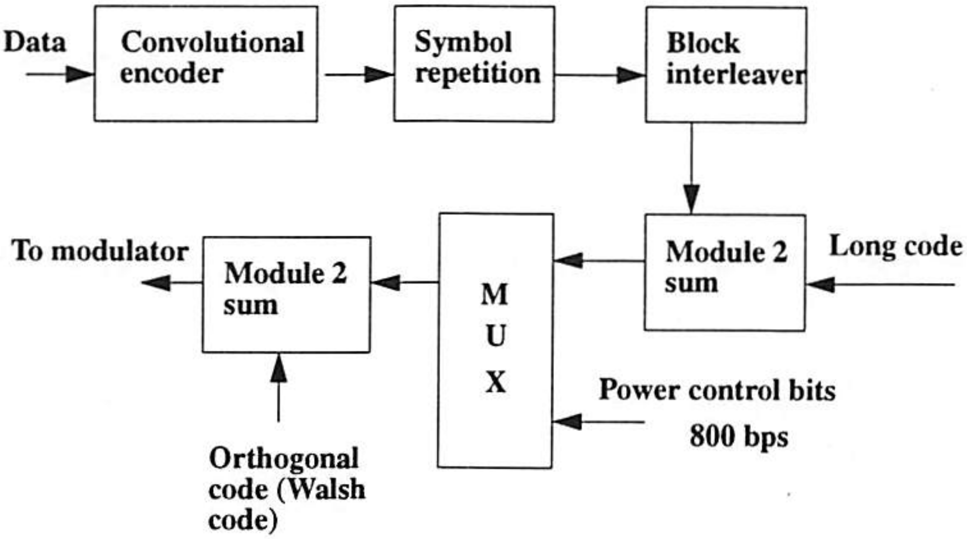


Figure 6.7: Forward CDMA traffic channel [2].

By assuming perfect power control, the noise plus interference can be expressed as

$$N_0 + I_0 = n_0 + (N-1)P_s, \tag{6.4}$$

where N is the total number of users and N_0 is the thermal noise. The Signal-to-noise ratio (SNR) is

$$\frac{E_b}{N_0 + I_0} = \frac{P_s/R}{N_0 + (N-1)P_s/W} \,. \tag{6.5}$$

Maximum capacity is achieved if the power control is adjusted so that the SNR is exactly what it needs to be for an acceptable error rate. If we set the left handside of (6.5) to the targeted SNR value and solve for N, then basic capacity equation for CDMA becomes

$$N - 1 = \frac{W/R}{(E_b/(N_0+I_0))_{target}} - \frac{N_0}{R} \,. \tag{6.6}$$

Using the numbers for IS-95A CDMA with the 9.6 kbps rate operation, the N value approximates 15.1 dB or $10^{1.51} = 32$ users. The targeted 6 dB SNR value is a nominal estimate. Once power control is available, the system designer and operator have the freedom to trade quality of service for capacity by adjusting the SNR value. Note that capacity and SNR are reciprocal. A three dB improvement in SNR incurs a factor of two loss in capacity, and vice-versa.

According to (6.6), a capacity in the neighborhood of 16-64 users corresponds to the E_b/N_0 being in the 3-9 dB range. In the same bandwidth and the targeted SNR value, a single sector of a signle AMPS cell has only 2 channels available.

6.2.8 Near-Far Problem

CDMA was always dismissed as unworkable in the mobile radio environment becasue of what was called the "near-far problem." Suppose at time t, k users transmit data at the same carrier frequency f_0. Then the received signal is

$$r(t) = \sum_{i=1}^{k} \sqrt{2P_s} s_i(t) b_i(t) cos(w_0 t + \theta_i). \tag{6.7}$$

Each user's signal power P_s is the same. The bit rate of the PN code $s_i(t)$ is f_c and the data rate of $b_i(t)$ is f_b. θ_i is an independent random phase. For detector one, the received signal $r(t)$ is multiplied by $s_1(t)$ and by $cos(w_0 t + \theta_1)$ to generate the signal r_{01}. If the items which will not pass through the decision device are dropped, then (6.7) becomes

$$\begin{aligned} r'_{01} &= \sum_{i=1}^{k} \sqrt{P_s} s_1(t) s_i(t) b_i(t) cos(\theta_i - \theta_1) \\ &= \sqrt{P_s} b_1(t) + \sum_{i=2}^{k} \sqrt{P_s} s_1(t) s_i(t) b_i(t) cos(\theta_i - \theta_1) \end{aligned}$$

$$= \quad \sqrt{P_s} b_1(t) + \sum_{i=2}^{k} \sqrt{P_s} s_{1i}(t) cos\theta_{1i} \tag{6.8}$$

where $\cos(\theta_{1i}) \equiv \cos(\theta_1 - \theta_i)$ and $s_{1i}(t) \equiv s_1(t)s_i(t)$. In (6.8), the first item is the desired signal and the second item is the k-1 independent interfering signals. The error probability at the output is

$$P_e = \frac{1}{2} erfc\sqrt{2\frac{1}{k-1}\frac{f_c}{f_b}}. \tag{6.9}$$

To achieve low probability of error, the following condition is important

$$\frac{f_c}{f_b} \gg \frac{k-1}{2}. \tag{6.10}$$

In (6.9), every user is assumed to have has the same signal power. If the unwanted user's signal power is much larger than the desired user's signal power, the probability of error will increase. This issue is referred to as the near-far problem [9]. Our proposed compact neural network based detector has robust resistance to this problem.

6.3 Decision Rules of CDMA Detectors

Suppose there are K active users sharing the same Gaussian channel at a given time t. The kth user is assigned a signature waveform $s_k(t)$ (PN code), $t \in [0, T]$, and a string of bits $\{b_k(i) \in \{-1, +1\}\}$ is transmitted. In a CDMA system, the signal at the detector is the superposition of K transmitted signals and the noise $n(t)$ citemp,

$$r(t) = \sum_{k=1}^{K} b_k(i) s_k(t - iT) + n(t), \qquad t \in [iT, (i+1)T]. \tag{6.11}$$

If we focus on one symbol interval in (6.11), the function of a receiver is to recognize every active user's symbol at the specified interval. There are three kinds of detectors used in a receiver. One is called the conventional detector which is widely used, another is called suboptimal multiuser detectors and the other is the optimal multisuer detector (OMD) with optimized solution in detection.

6.3.1 Conventional Detector

A conventional detector consists of a bank of filters matched to the signature waveforms of K users. Simple decision devices following the matched filters provide every user's symbol estimates based upon the signs of the output of the matched filters at the specific time interval,

$$\begin{aligned} y_k^{(i)} &= \int_{iT}^{(i+1)T} r(t) s_k(t - iT) dt \\ \mathbf{b}_{\mathbf{CD}}^{(\mathbf{i})} &= sign(\mathbf{y}^{(i)}) \end{aligned} \qquad (6.12)$$

where $\mathbf{b}_{CD}^{(i)} = [b_0^{(i)} b_1^{(i)} \cdots b_{K-1}^{(i)}]^T$ and $\mathbf{y}^{(i)} = [y_0^{(i)} y_1^{(i)} \cdots y_{K-1}^{(i)}]^T$. The block diagram of a conventional detector is shown in Fig. 6.8.

The error probability of the kth user for the conventional detector is

$$\begin{aligned} P_k^c &= P[y_k > 0 | b_k = -1] \\ &= \sum_{\mathbf{b} \in \{-1,+1\}^K} P[y_k > 0 | \mathbf{b}] P[\mathbf{b} | b_k = -1] \\ &= 2^{1-K} \sum_{\mathbf{b} \in \{-1,+1\}}^{K} \sum_{\mathbf{b} \in \{-1,+1\}^K} Q(\frac{w_k - \sum_{i \leq k} b_i H_{ik}}{\delta \sqrt{w_k}}). \end{aligned} \qquad (6.13)$$

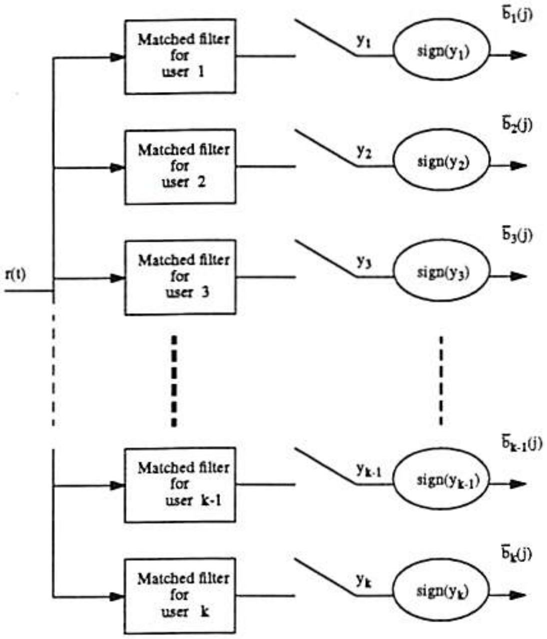


Figure 6.8: Block diagram of a conventional detector.

The asymptotic efficiency of the conventional detector in the low noise region becomes

$$\begin{aligned}
\eta_k^c &= sup\{0 \leq r \leq 1; \lim_{\sigma \to 0} P_k^c / Q(\frac{\sqrt{r w_k}}{\sigma}) < +\infty\} \\
&= max\{0, 1 - \sum_{i \neq k} \frac{|H_{ik}|}{w_k}\} \\
&= max\{0, 1 - \sum_{i \neq k} |R_{ik}| \frac{\sqrt{w_i}}{\sqrt{w_k}}\} \qquad (6.14)
\end{aligned}$$

where $\mathbf{R}$ is the matrix of normalized cross correlation, i.e.,

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2} \qquad (6.15)$$

where $\mathbf{W} = diag\{w_1, \cdots, w_K\}$. It follows from (6.14) that the conventional kth user detector is near-far resistant (i.e., its asymptotic efficiency is bounded away from zero as a function of the interfering users' energies) only if $R_{ik} = 0$ for all

$i \neq k$, i.e., only if the kth user's signal is orthogonal to the subsapce spanned by the other signals [11].

6.3.2 Optimal Multimuser Detector (OMD)

An OMD produces an estimate based on the maximization of the logarithm of the likelihood function. In the synchronous case, the matrix representation of $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{H}\mathbf{b} + \mathbf{n} \tag{6.16}$$

where $\mathbf{n} = [n_0^{(i)} n_1^{(i)} \cdots n_{K-1}^{(i)}]^T$ and $\mathbf{H} \in R^{K \times K}$ is a cross correlation matrix of the signature waveforms,

$$h_{ij} = \int_0^T s_i(t) \cdot s_j(t) dt. \tag{6.17}$$

$\mathbf{H}$ is nonnegative definite. Given the observation $r(t)$, the OMD is to generate an estimate $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_{K-1})$ to minimize the cost function [7],

$$\hat{\mathbf{b}}_{OMD}^{(i)} = arg \min_{b \in \{-1,+1\}^K} \int_0^T \left[r(t) - \sum_{k=1}^{K} b_k s_k(t) \right]^2 dt. \tag{6.18}$$

After manipulation, (6.18) can be written in a matrix form,

$$\hat{\mathbf{b}}_{OMD}^{(i)} = arg \min_{b \in \{-1,+1\}^K} \frac{1}{2} \mathbf{b}^T \mathbf{H} \mathbf{b} - \mathbf{b}^T \mathbf{y}. \tag{6.19}$$

The square of the difference between the received signal and the original signal is [11]

$$\min_{\mathbf{b} \in \{-1,1\}^K} \min_{\mathbf{d} \in \{-1,1\}^K} |\sum_{i=1}^{K} b_i s_i(t) - \sum_{i=1}^{K} d_i s_i(t)|^2$$

$$= 2 \min_{\epsilon \in \{-1,0,1\}^K|_{\epsilon_k=1}} \epsilon^T \mathbf{H} \epsilon. \tag{6.20}$$

Therefore the asymptotic efficiency of the optimum multiuser detector is equal to

$$\eta_k = \frac{1}{w_k} \min_{\epsilon \in \{-1,0,1\}^K|_{\epsilon_k=1}} \epsilon^T \mathbf{H} \epsilon. \tag{6.21}$$

When $\sigma \to 0$, the optimum multiuser detector achieve the minimum probability of error for each user. The highest asymptotic efficiency is

$$\eta_k = \frac{1}{w_k} \min_{\epsilon \in \{-1,0,1\}^K} \epsilon^T \mathbf{H} \epsilon. \tag{6.22}$$

In the two-user case, (6.22) becomes

$$\eta_1 = min\{1, 1 + \frac{w_2}{w_1} - 2|\rho|\frac{\sqrt{w_2}}{\sqrt{w_1}}\}, \tag{6.23}$$

where $\rho = R_{12}$ is the cross-correlation of $s_1(t)$ and $s_2(t)$. There isn't any explicit expression for (6.22). This combinational optimization problem is an NP-complete prblem [11] [16].

6.3.3 Suboptimal Detectors for Multiuser Detection

Many multiuser detectors with suboptimal solutions were proposed [6] [10] [11]. Suboptimal detectors are not as complex as the optimal detector but its performance is inferior. It is still very challenging to implement the suboptimal detectors by compact microelectronic circuits. Here, some suboptimal detectors and their characteristics are briefly reviewed. In general, these detectors can be classified into two categories: subtractive interference cancellation detectors and linear multiuser detectors [10].

(A) Linear Detectors

In linear multiuser detectors, linear mapping technique to the soft output of the conventional detectors to reduce the Multiple Access Interference (MAI) is employed.

(A-1) Decorrelating Detector

Suppose **H** is the K x K correlation matrix. The decorrelating detector is the inversion of the correlation matrix,

$$\begin{aligned}\hat{\mathbf{x}} &= sgn(\mathbf{H}^{-1}\mathbf{y}) \\ &= sgn(\mathbf{W}^{-1/2}\mathbf{R}^{-1}\mathbf{W}^{-1/2}\mathbf{y}) \\ &= sgn(\mathbf{W}^{-1/2}\mathbf{R}^{-1}\mathbf{y}) \\ &= sgn(\mathbf{R}^{-1}\mathbf{y}). \qquad (6.24)\end{aligned}$$

Numerical example is taken in section 6.5. This detector does not require knowledge of the energies of any of the active users. In the absence of noise, the output vector **y** of the matched filters is equal to **Hx** and the solution is optimum. In the noisy environment, the noise component in the $\mathbf{H}^{-1}$ are correlated and $\hat{\mathbf{x}}$ is not the optimium decision. But this detector completely eliminates the multiple access interference (MAI). It is very similar to the zero-forcing equalization which is used to completely eliminate Intersymbol Interference (ISI). Though it has better performance than a conventional detector, some significant disadvantages exist.

One disadvantange of this detector is that it causes noise enhancement (similar to the zero-forcing equalizer) [11]. The power associated with the noise term at the output of the decorrelating detector is always greater than or equal to the power associated with the noise term at the output of the conventional detector for every bit. The second disadvantage is the complexity of the inversion of matrix **R**. It is very difficult to perform the inversion by compact electronic circuits.

(A-2) Minimum Mean-Squared Error (MMSE) Detector

Based upon the knowledge of the received signal powers and background noise, mean-squared error $E[|\mathbf{d} - \mathbf{L}\mathbf{y}|^2]$ between the real data and the soft output of the conventional is minimized. An MMSE detector is to implement this minimization by linear mapping. As reported in [12] [13], it can be expressed as

$$\mathbf{L}_{MMSE} = [\mathbf{R} + (N_0/2)\mathbf{A}^{-2}]^{-1}. \tag{6.25}$$

Therefore, the estimate of this detector is

$$\hat{\mathbf{x}}_{MMSE} = \mathbf{L}_{MMSE} \cdot \mathbf{y}. \tag{6.26}$$

Because this detector takes into account the background noise, its performance is better than that of the decorrelating detector. When the backgorund noise becomes zero, decorrelating detector and MMSE detector have the same performance.

The MMSE detector needs the estimation of the amplitudes of the received signals. Its performance also depends upon the powers of the interference users [12]. Its near-far resistance is worse than that of the decorrelating detector. Like the decorrelating detector, it also needs the operation of matrix inversion.

(B) Substractive Interference Cancellation

In substractive interference cancellation detectors, interference was estimated and then substracted out. Here, the Substractive Interference Cancellation (SIC) operation is briefly described. SIC cancels the interference step by step. At each stage of this detector, one additional user's direct-sequence is decided, regenerated and then canceled out from the received signal. Therefore, the remaining users can have less MAI for their transmission in the next stage [14] [15]. Fig. 6.9 shows the block diagram of the first stage. At the first stage, the received signals are ranked in descending order of received powers. Although the SIC detector has the potential to provide significant improvement over the conventional detector, it is not straight-forward to implement. First, the signals need to be reordered whenever the power profiles changes. Second, if the initial data estimate is not reliable enough, the intereference effect will become quite serious. Implementation by compact microelectronic circuits is also very challenging.

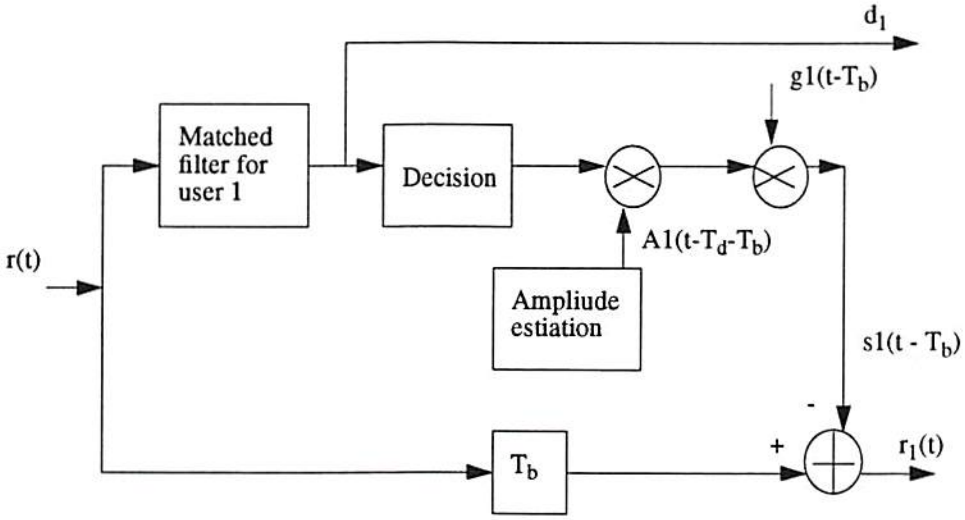


Figure 6.9: Block diagram of the first stage of subtractive interference cancellation detector.

6.4 Implementation of Optimal Multiple-Access Detector by Biologically Inspired Compact Neural Network

In (6.19), the multiuser detection problem becomes a quadratic optimization problem. Let's recall the expression of the energy function of a compact neural network,

$$E = -\frac{1}{2}\mathbf{v}_{\mathbf{y}}^{\mathbf{T}}\mathbf{M}\mathbf{v}_{\mathbf{y}} - \mathbf{v}_{\mathbf{y}}^{\mathbf{T}}\mathbf{b}. \tag{6.27}$$

Hence, the output of a compact neural network will be the desired estimate $\hat{\mathbf{b}}$ if

$$\mathbf{M} = -\mathbf{H} \quad \text{and} \quad \mathbf{b} = \mathbf{y}. \tag{6.28}$$

It means if the synapse weight matrix $\mathbf{M}$ is equal to $-\mathbf{H}$ and the output of the matched filters feeds into the input of a compact neural network, the desired estimate $\mathbf{b}$ will be obtained at the output of the compact neural network. According to this mapping, every neuron has one self-feedback synapse and $K-1$ synapses connecting to the other $K-1$ neurons. It is a combined version from the cellular neural network and the Hopfield neural network. This structure is very similar to the structure of the real neuron systems. Therefore, it is a biologicaly inspired neural network. Matrix $\mathbf{H}$ is symmetric and positive semidefinite. The symmetirc property of $\mathbf{H}$ is a sufficient condition for guaranteed stable operation of this network.

Fig. 6.10 shows the functional block diagram of a compact neural network based CDMA detector. Fig. 6.11 shows the structure of the compact neural network core used in Fig. 6.10.

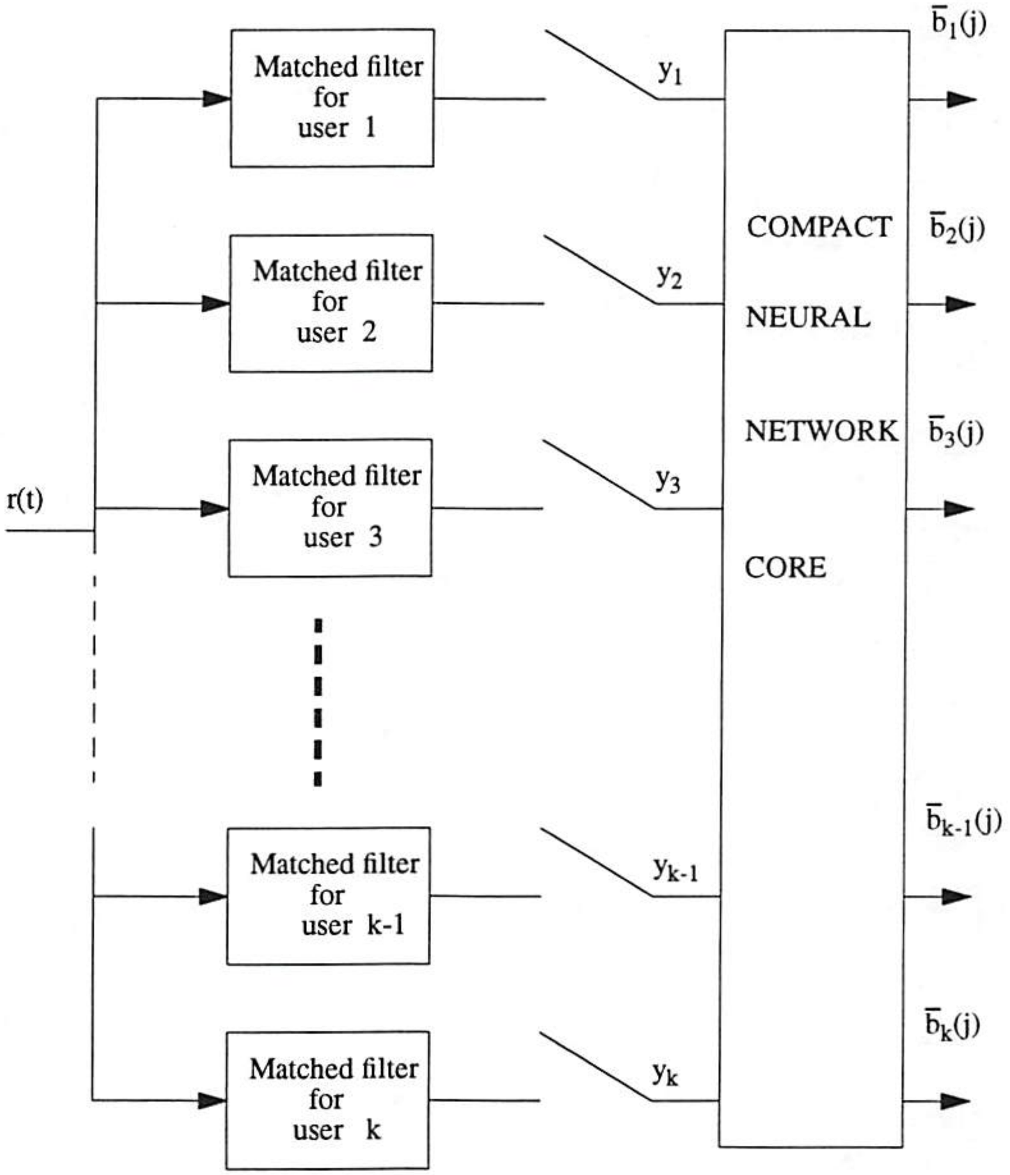


Figure 6.10: Function block diagram of compact neural network based CDMA receiver.

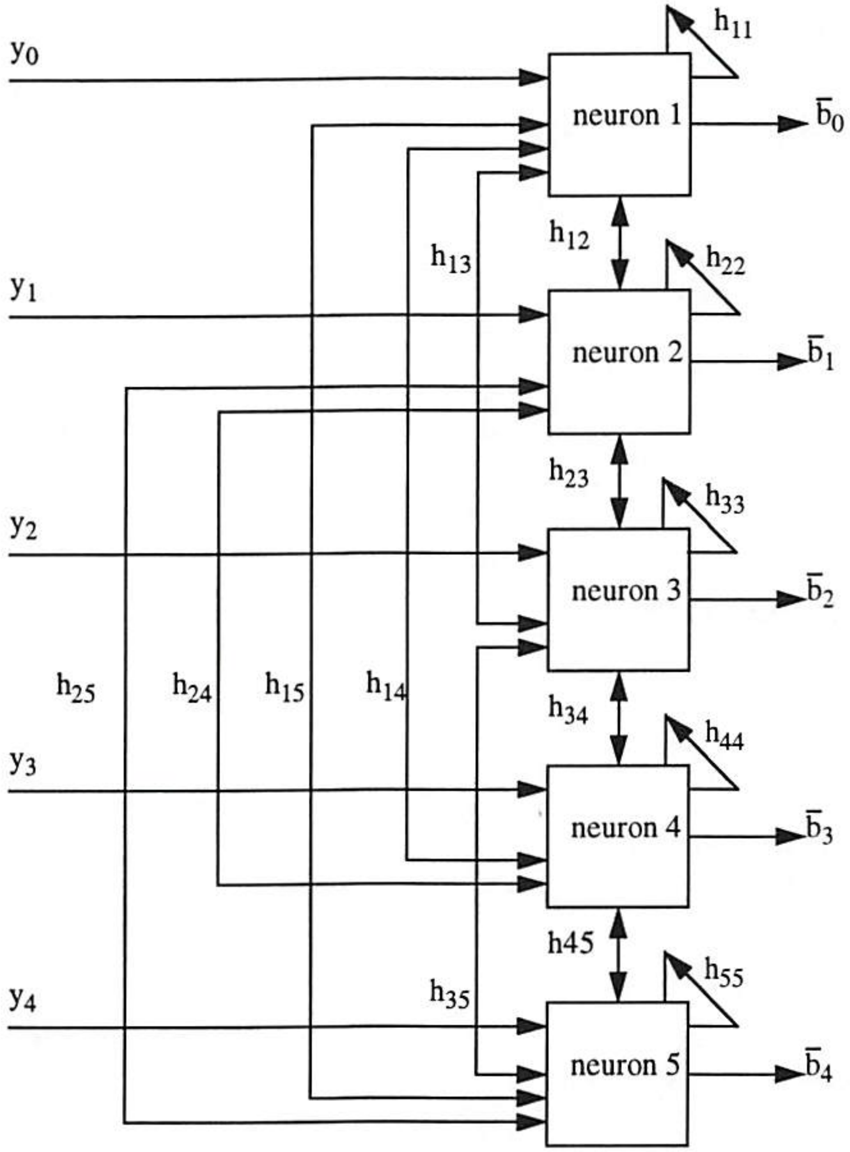


Figure 6.11: Structure of the neural network core in CDMA detector for K=5 case.

6.5 Simulation Results

In this section, comparison of the performance of the conventional detector and that of the biologically inspired compact neural network based CDMA detector is presented. First, Consider a $K = 2$ synchronous and noiseless case. The original transmitted signal is $b_{sent} = [-1, -1]^T$ at a given time. The near-far ratio r_{nf} is defined as

$$r_{nf} = \frac{\int s_i^2 dt}{\int s_j^2 dt} \tag{6.29}$$

and the normalized cross-correlation of signature waveforms h is defined as

$$h = \frac{\int s_i s_j dt}{\sqrt{\int s_i^2 dt \int s_j^2 dt}}. \tag{6.30}$$

Therefore, the synapse template **H** can be written as

$$\mathbf{H} = \begin{bmatrix} r_{nf}^{1/2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & h \\ h & 1 \end{bmatrix} \begin{bmatrix} r_{nf}^{1/2} & 0 \\ 0 & 1 \end{bmatrix}. \tag{6.31}$$

The outputs of the matched filters **y** were sent to the neuron core for detection. The failure points (FPs) were recorded. Fig. 6.12 and Fig. 6.13 show the distribution of failure points. Each failure point is determined if $b_{detected} \neq b_{sent}$. The result of OMD is obtained by full search as the reference method. The simulation range for the near-far ratio r_{nf} is $[1, 10^{0.01}, 10^{0.02}, \cdots, 10]$ and the range for the correlation function h is $[-0.9, -0.89, -0.88, \cdots, 0]$.

The constraint energy function was not used in the simulation results as shown in Fig. 6.12. Lots of failure points are produced by the conventional detector. But only some failure points are generated by the simple compact neural network based CDMA detector with or without hardware annealing. Simulation results which include the constraint energy function are shown in Fig. 6.13. There was no failure point from the compact neural network based CDMA receiver with piecewise linear function and the constraint energy function. Fig. 6.14 shows the

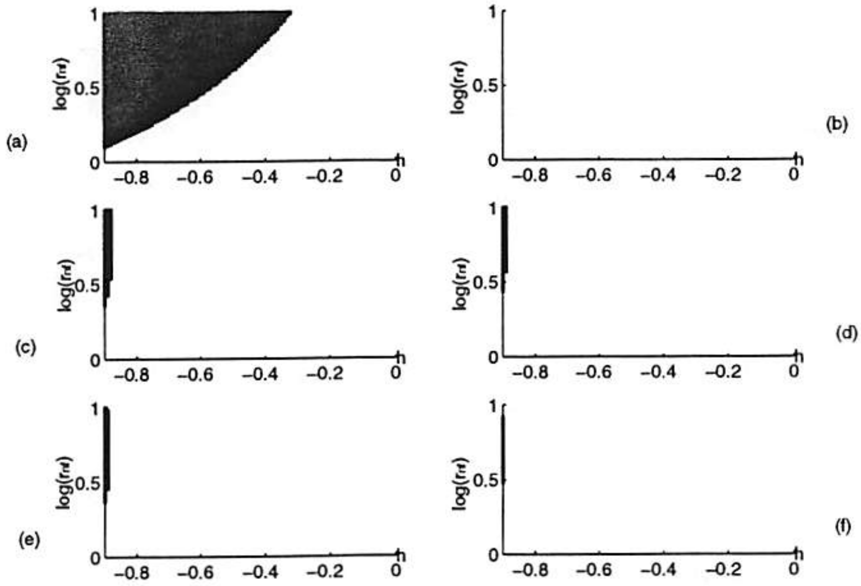


Figure 6.12: Distribution of failure points. Constraint energy function is not used. The logarithmic function is with base 10. (a) Conventional detector. (b) Optimal multiuser detector. (c) Compact NN based CDMA receiver with sigmoid function. (d) Compact NN based CDMA receiver with sigmoid function and hardware annealing. (e) Compact NN based CDMA receiver with piecewise linear function. (f) Compact NN based CDMA receiver with piecewise linear function and hardware annealing.

results of signals corrupted by neighboring user's interefrence and Gaussian noise. The signal-to-noise ratio for user 1 is fixed at 10 dB. The compact neural network based CDMA receiver with piecewise linear function was employed. Notice that a compact neural network based CDMA receiver could have better performance than optimal multiuser detector in the noise-corrupted cases. Data were obtained from 10,000 cases. Fig. 6.15 shows the results of three synchronous users transmitting their signals spreading by the Gold codes [2]. The Gold codes for user 1, user 2

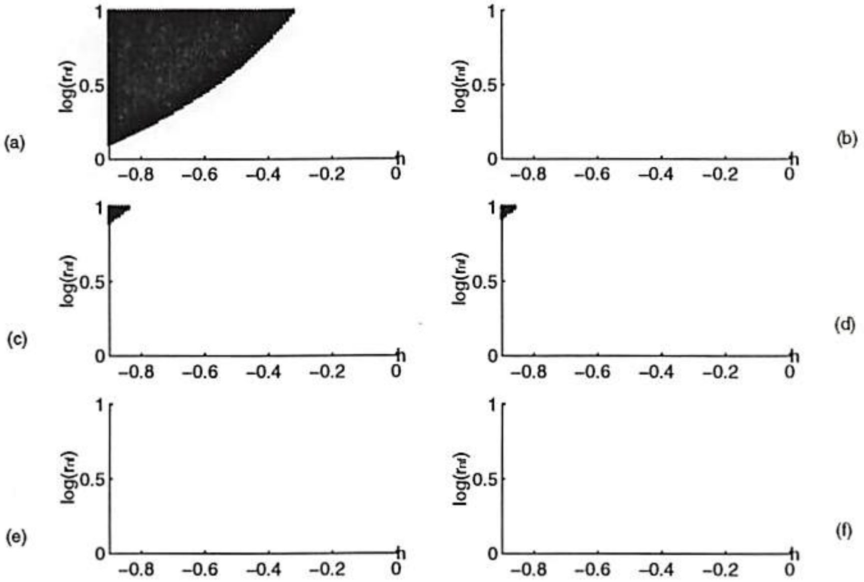


Figure 6.13: Distribution of failure points. Constraint energy function is added. Conditions (a) to (f) are the same as in Fig. 6.12.

and user 3 are 000110, 0011011 and 1010101, respectively. The power of the first user is 2 dB stronger than that of the other two users. Phase difference is also considered. Suppose the signal amplitudes of user 2 and user 3 are all equal to 1. Then the signal amplitude of user 1 is $10^{(2/20)} = 1.259$. With signal-to-noise ratio of 6 dB, the output $\mathbf{y}$ of the matched filter bank can be [-0.1098, 1.0536, 0.0359], [0.929, 0.5273, -0.4376], [0.1621, -0.0202, -0.9309],etc. Errors were cumulated and divided by 10,000 to determine the probability of error. The biologically inspired compact neural network detector performs almost as excellent as the optimal multiuser detector.

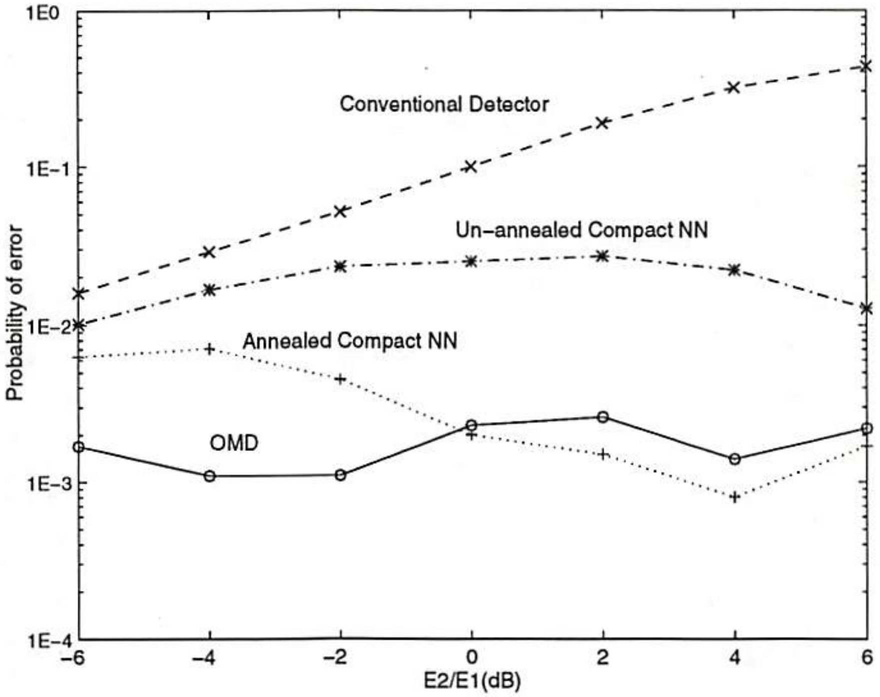


Figure 6.14: Error probability of conventional detector, compact NN based CDMA detector with or without hardware annealing function, and optimal multiuser detector. Signal-to-noise ratio for user 1 is fixed at 10 dB.

6.5.1 Summary

The collective computational behavior of a compact neural network is used to implement the optimal multiuser detection (OMD) for CDMA system in the presence of multiple access interference (MAI) and Gaussian noise. It is demonstrated that compact neural network is an efficient architecture for realizing the OMD. In addition, the performance of the neural network based OMD can be enhanced by paralleled hardware annealing technique which is suitable for high-speed operation of neural networks. The important properties are emphasized as follows:

- No electronic implementation has been developed yet.

- Achieving optimized solution of near-far resistance performance and thus providing better performance/capacity gains over the conventional detector.
- Having good potential to recover signals corrupted by Gaussian noises.
- Not requiring matrix inversion.
- Parallel processing.

The converged solution is obtained in one single cycle of the neural network operation, which can be realized around a microsecond or less. The throughput rate of NN-OMD supports the required speed of the peripheral circuitry such as matched filters which have to run at a symbol rate.

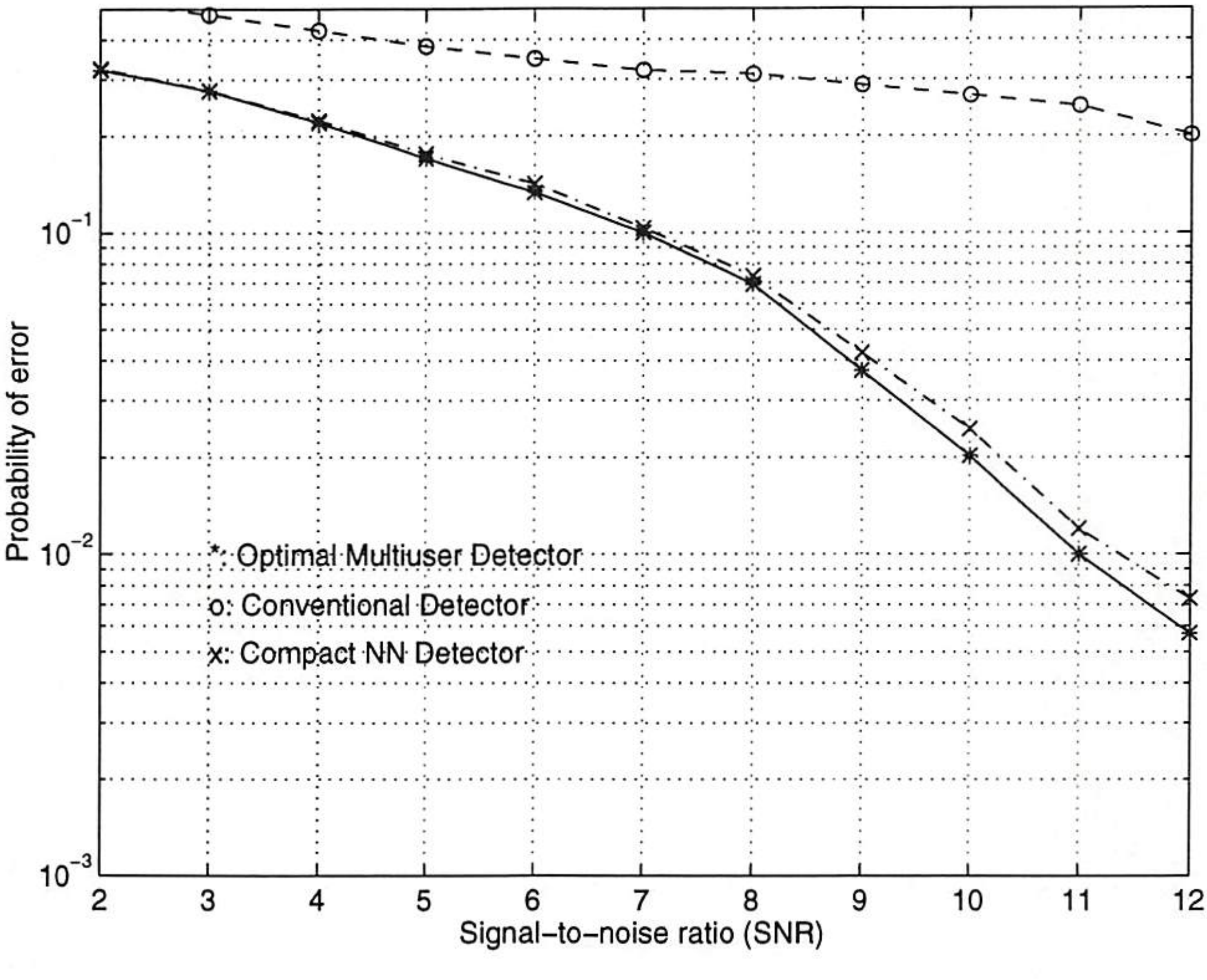


Figure 6.15: Error probability of compact NN based CDMA detector with hardware annealing function, optimal multiuser detector, and conventional detector. Maximum near-far ratio is 2 dB. A 3-user case is considered.

Reference List

[1] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication,* Addision Wesley: Reading, MA, 1995.

[2] R. L. Peterson, R. E. Ziemer, D. E. Borth, *Introduction to Spread Spectrum Communications,* Prentice Hall PTR: Upper Saddle River, NJ, 1995.

[3] V. K. Garg, K. Smolik, J. E. Wilkes, *Applications of CDMA in Wireless/Personal Communications,* Prentice Hall PTR: Upper Saddle River, NJ, 1997.

[4] T. E. Bell, et al., "Communications," *IEEE Spectrum,* pp. 30-41, Jan. 1996.

[5] B. P. Paris, B. Aazhang, G. Orsak, "Neural networks for multiuser detection in CDMA communication," *IEEE Trans. on Commun.,* vol. 40. pp. 1212-1222, July 1992.

[6] U. Mitra, H. V. Poor, "Neural network techniques for adaptive multiuser demodulation," *IEEE J. Selec. Areas in Commun.*, vol. 12, pp. 1460-1470, Dec. 1994.

[7] S. Verdu, "Computational complexity of optimum multiuser detection," *Algorithmica*, vol. 4, pp. 303-312, Springer Verlag: New York, NY, 1989.

[8] J. Proakis, *Digital Communications,* Prentice-Hall: Englewood Cliffs, NJ, 1988.

[9] T. Schilling, *Principles of Communication Systems*, McGraw-Hill, New York, NY, 1986.

[10] S. Moshavi, "Multi-user detection for DS-CDMA communications," *IEEE Communications Magazine,* pp. 124-136, Oct. 1996.

[11] R. Lupas, S. Verdu, "Linear multiuser detectors for synchronous code-division multiple access channels," *IEEE Trans. on Information Theory,* vol. 35, no. 1, pp. 123-136, Jan. 1989.

[12] Z. Xie, R.T. Short, C. K. Rushforth, "A family of suboptimum detector for coherent multi-user communications," *IEEE J. Selec. Areas in Commun.*, vol. 8, no. 4, pp. 683-90, May 1990.

[13] M. Hong, U. Madhow, S. Verdu, "Blind adaptive multiuser detection," *IEEE Trans. on Commun.,* vol. 42, no. 12, pp. 3178-88, Dec. 1994.

[14] A. J. Viterbi,"Very low rate conventional codes for maximum theoritical performance of spread-spectrum multiple-access channels," *IEEE J. Selec. Areas in Commun.*, vol. 8, no. 4, pp. 641-49, May 1990.

[15] R. Kohno, et al., "Combination of an adaptive array antenna and a canceller of interference for direct-sequence spread sprectrum multiple access system," *IEEE J. Selec. Areas in Commun.*, vol. 8, no. 4, pp. 675-82, Apr. 1990.

[16] M. R. Garey, D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* Freeman: San Francisco, CA, 1979.