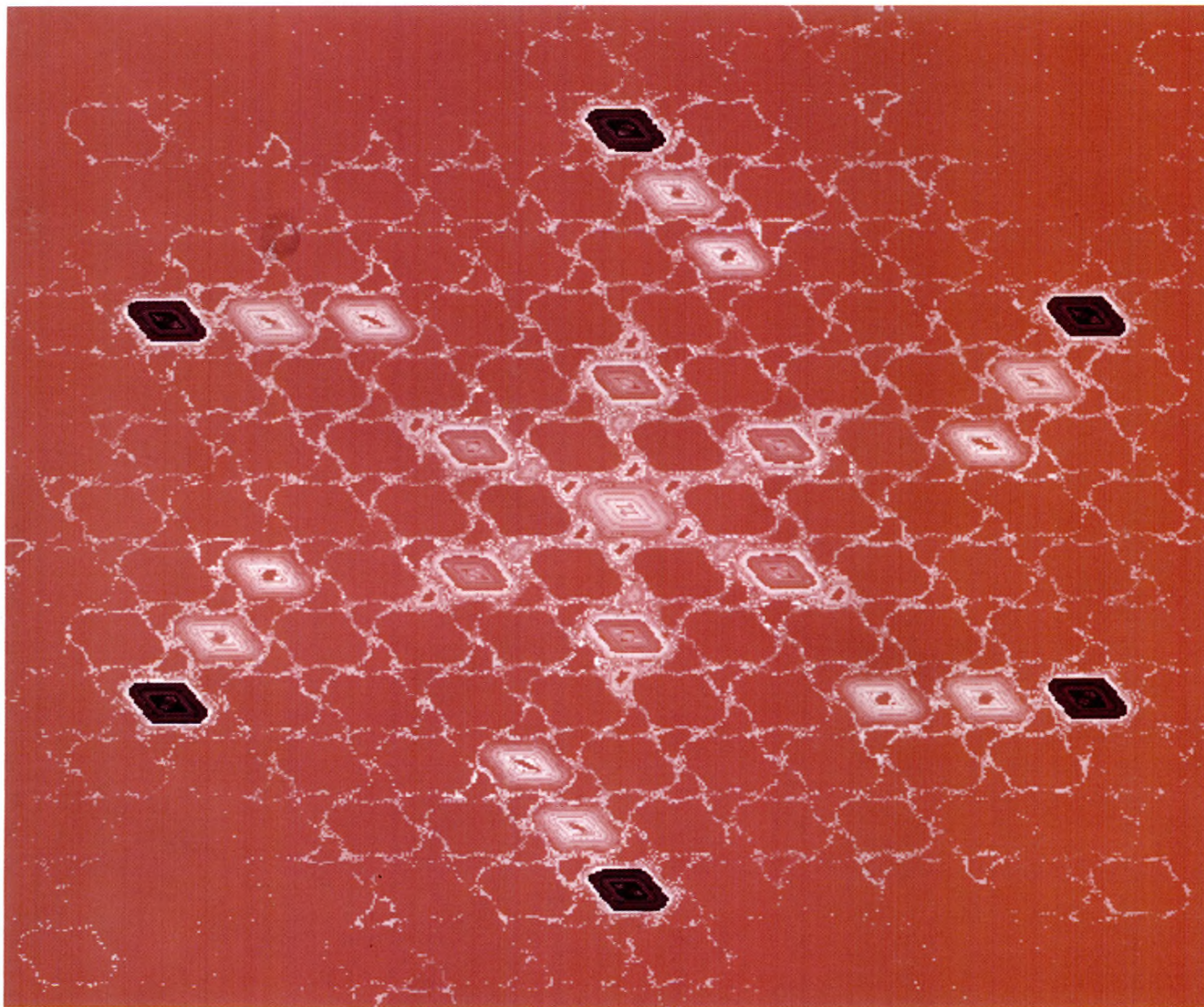


Naval Research Reviews

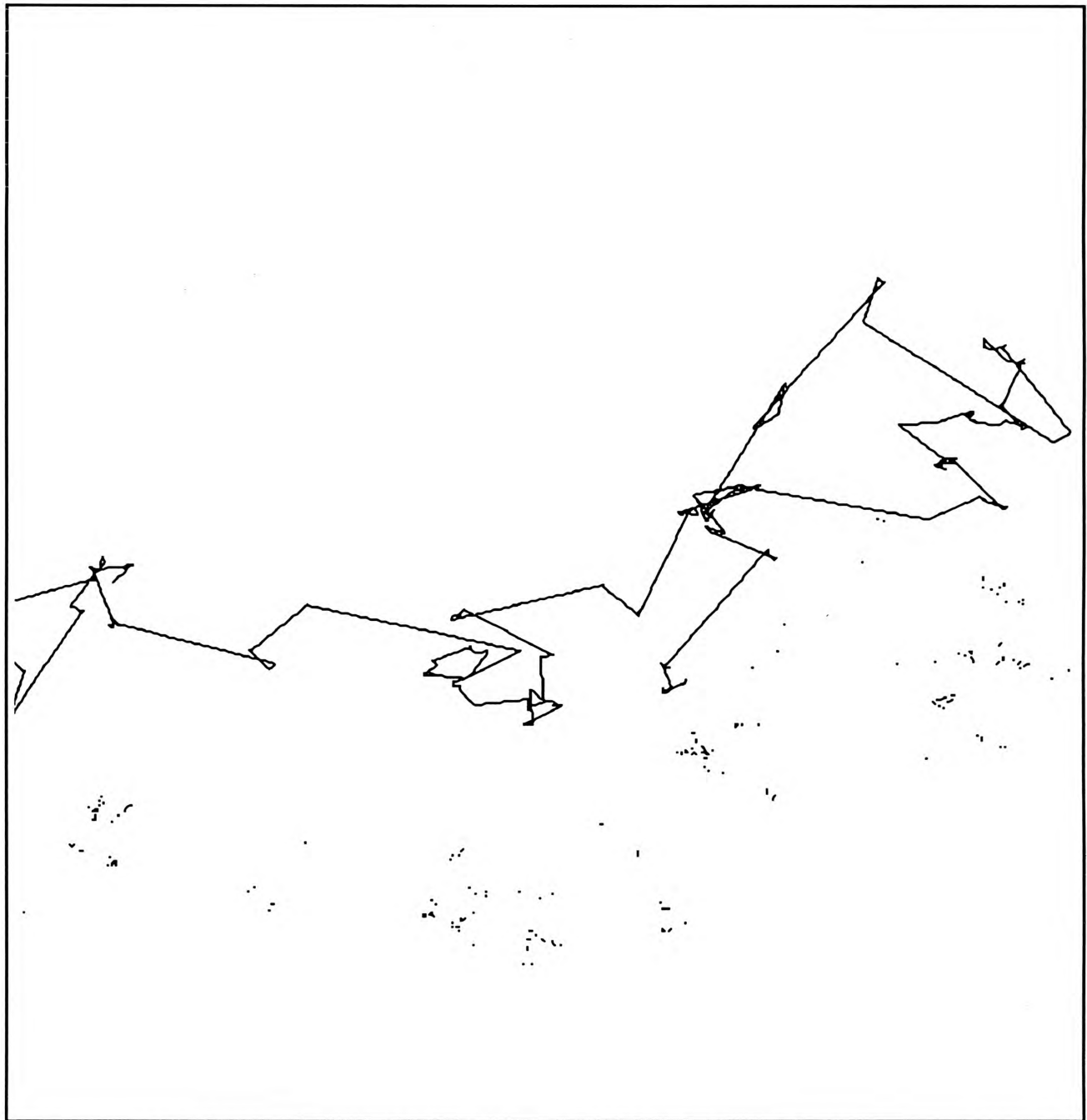
Office of Naval Research
Three / 1993
Vol XLV



NOV 01 1993
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Digitized by Google

Original from
UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN



A random walk is a basic concept in probability theory. Brownian motion can be viewed as a random walk where the average jump size is well defined. Fractal random walks do not have an average jump size, but jumps of all length scales enter, and none of them dominates. In this figure the connected line segments show the trajectory of a fractal random walk. The dots below the trajectory show the fractal set of turning points of the random walk. When a velocity field is associated with the path lengths this type of space-time random walk has been used to model turbulence. According to Kolmogorov scaling, a trajectory segment should have a velocity proportional to its length raised to the $1/3$ power. This theory was developed by M. Shlesinger, B. West, and J. Klafter in *Physical Review Letters*, Vol. 58, 1100-1103 (1987). The theory was successfully applied to a numerical simulation of turbulent pipe flow with a Reynolds number of 100,000 by F. Hayot in *Physical Review A* 43, 806-810 (1991). The fractal random walk approach showed the flattened (relative to laminar flow) velocity profile across the pipe which reflects the enhanced mixing due to turbulence. With ONR support Dr. Hayot is exploring numerically the simulation of turbulent flow past a cylinder using the fractal space-time random walk approach. The above figure was calculated by M. Shlesinger using True Basic for the Mac.

Naval Research Reviews

Office of Naval Research
Three / 1993
Vol XLV

Articles

2

Applications of Chaos and Fractals

Michael F. Shlesinger

12

Controlling Chaos

Mark L. Spano

William L. Ditto

4

Communicating with Chaos

Thomas L. Carroll

Louis M. Pecora

23

Nonlinear Resonance: Noise-Assisted Information Processing in Physical and Neurophysiological Systems

Adi Bulsara

John K. Douglass

Frank Moss

38

Dialogue on Noisy Chaos

Robert Cawley

Guan-Hsong Hsu

49

Attractors of Iterated Systems and Their Application To Image Compression

E. W. Jacobs

R. D. Boss



CHIEF OF NAVAL RESEARCH
RADM Marc Pelaez, USN

DEPUTY CHIEF OF NAVAL RESEARCH
TECHNICAL DIRECTOR
Dr. Fred Saalfeld

CHIEF WRITER/EDITOR
William J. Lescure

SCIENTIFIC EDITORS
Dr. Robert J. Nowak
Dr. J. Dale Bulman

MANAGING EDITOR
Norma Gerbozy

ART DIRECTION
Typography and Design
Desktop Printer, Arnold, MD

Departments

22

Profiles in Science

About the Cover

The trajectories (coded by color) of several hundred initial conditions are followed for a two-dimensional Hamiltonian system. The underlying equations are called a Zaslavsky map and were chosen here to exhibit 6-fold symmetry. Some trajectories stay localized with periodic orbits while others exhibit a complex global dynamics traveling through the 6-fold symmetric stochastic web. The dynamics are related to the behavior of charged particles in electromagnetic fields. Computer calculations by M. Shlesinger, using True Basic for the Mac.

Naval Research Reviews publishes articles about research conducted by the laboratories and contractors of the Office of Naval Research and describes important naval experimental activities. Manuscripts submitted for publication, correspondence concerning prospective articles, and changes of address, should be directed to Code OPARI, Office of Naval Research, Arlington, VA 22217-5000. Requests for subscriptions should be directed to the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20403. *Naval Research Reviews* is published from appropriated funds by authority of the Office of Naval Research in accordance with Navy Publications and Printing Regulations. NAVSO P-35.

Original from
UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN
Three/1993

Applications of Chaos and Fractals

*Dr. Michael F. Shlesinger, Director
ONR Physics Division
Guest Editor
e-mail: mike@onr-hq.navy.mil*

Introduction

There have been two great trends in 20th century physics; investigating the very small (atomic, nuclear, particle physics) and investigating the very large (astronomy and cosmology). It is only in the last decade that a third trend has been recognized; investigating the very complex. Complexity is all around us, but it is only recently that discoveries and methods have arisen which gives us hope to treat it as a science. The two words which are the most recognizable from this new discipline are chaos and fractals.

Chaos arises as a common behavior in nonlinear dynamical systems. Nonlinear dynamics arose with Newton's equations of motion and the scale-invariant nonlinear $1/r$ gravitational potential. While the earth-moon system was solvable (integrable) the 3-body problem was not. We now know that chaos can appear in the 3-body problem. Newton worried about the 3-body problem because he wondered how orbital earth-moon stability could be maintained in the face of additional bodies, such as comets, since there is no friction to damp out perturbations. This question, in fact, was not answered until the 1950's (both stability and instability can occur) with the work of Kolmogorov. He showed that in phase space that both islands of stability and a chaotic sea coexist. Which of these regions you explore depends on the energy of the system and your initial conditions. A strong Russian school, steeped in tradition, advanced our knowledge of such Hamiltonian systems. The Zaslavsky map on the cover is one such example.

Fractals are beautiful hierarchical geometric figures which are by now familiar to many around the world. The field was mainly developed by Mandelbrot over the past 40 years. He was able to use the scaling properties of fractals to define a continuous dimension, called the fractal dimension. Fractals can be self-similar with a single dimension; multifractal with a distribution of dimensions; or self-affine with independent scaling in different directions. Perhaps, the most famous fractal is the Mandelbrot Set which now even appears as a popular screen saver on computers. We honor Professor Mandelbrot with a brief biographical note and a sketch of his likeness, in this issue.

The set of pre-Mandelbrot fractals is nonzero, but while the early examples of fractals were interesting, the results were isolated, and held up as pathologies and not as the seeds of a new field of scientific inquiry. Examples included Weierstrass'

continuous, but non-differentiable function; Levy's theory of the algebra of random variables with infinite moments; and Richardson's discovery that the length of a country's border depended, in a scaling manner, with the resolution of measurement. It is only in recent times that the deep power of fractal geometry is appreciated and its intricacies can be revealed through the use of computers. Fractals have led to a paradigm shift which has changed our view of nature. It is now realized that many shapes in nature from turbulent dissipation fields to mountain profiles are better served by fractal geometry than by standard Euclidean geometry. Fractal shapes also abound in the phase space of nonlinear dynamical trajectories. The historical use of the term strange attractor obscures the dominant feature that the trajectories are fractal, and highlights the typical reaction of a field to encountering its first fractal.

Before the chaos revolution the main conclusion one drew about nonlinear systems is that they were difficult to analyze. The computer revolution has eased this perception. Usually, one just separated stable from unstable regions and chose to stay well within the stable bounds. Another early lesson was that the study of one nonlinear equation would not shed light on the properties of another. A major development in nonlinear dynamics is discovering how wrong the above statement is.

In the early 1970's nonlinear dynamics turned to seeking stable particle-like solutions (called solitons) to integrable nonlinear equations with an infinite number of conservation laws. Solitons are the antithesis of chaos. The mathematics behind solitons was quite complex, and kept the field only open to experts. In 1963, Lorenz published an example of a nonlinear system of three ordinary differential equations whose solutions was extremely sensitive to initial conditions. His equations were related to a meteorological problem of the convection of heated air. The trajectory of his solutions traced out a highly intricate figure with what would now be called a fractal structure. Such structures were named strange attractors by Ruelle and Takens in 1971. They were studying the mathematics of strange attractors which arise after a two frequency (quasi periodic) response becomes unstable with a change in a parameter which adds a third frequency to the mix. All of these advances still left complexity as a small technical area of science.

The field of nonlinear dynamics exploded in the late 1970's with the experimental verification of Feigenbaum's period doubling universality in Libchaber and Maurer's liquid helium convection experiments. Not only did arcane theoretical notions about mappings of the unit interval into itself reveal themselves as fundamental in a real 3D fluid with untold degrees of freedom, but also the renormalization group found an extension into the realm of dynamical systems. The trickle of research in dynamical systems soon became a flood covering diverse fields with a common language and phenomena

Original from

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

based on the twin notions of instability and fractal dimension. The field of nonlinear dynamics is now represented by over half a dozen specialized journals, a large number of texts, a best selling historical account by James Gleick called "*Chaos*", and it is even well covered in high school science fairs. The basis of the success of nonlinear dynamics and fractals lies in being able to understand very complex behaviors from very simple laws. This fortunate turn of events allows a student to learn a fair amount rapidly without great effort. This simplicity can be misleading with the field being too easy for beginners and too hard for experts. Research is now proceeding from temporal chaos to spatio-temporal patterns and chaos, self-organization, and adaptive systems.

The focus on applications is the theme of this issue of the Naval Research Reviews. Each article is based on research at a Navy laboratory. This is not by accident, but reflects the lead the Navy has taken in turning the state-of-the-art in chaos and fractals into applications. This is not just the judgment of ONR, but was recognized, as well, by Science in its May 10, 1991 issue where they stated that for applications of nonlinear dynamics that

"its no coincidence that most of the pioneering work is being done in Navy laboratories. Since 1983 the Office of Naval Research has been the only government agency with a funding program specifically for chaos studies." Our selection looks at synchronization of chaotic circuits with applications from communication to robotics; controlling chaos with applications from pacemakers to actuators; exploring the interaction of nonlinear dynamics with noise for novel detectors based on noise-induced resonance processes; employing nonlinear dynamics for signal processing in phase space; and novel techniques for employing fractals for image compression. This is just a sampling of the research underway in Navy laboratories. I leave it now to the authors of the articles to give you a glimpse into the frontiers they are inventing and exploring in the exciting world of chaos and fractals.

Communicating with Chaos

Thomas L. Carroll and Louis M. Pecora
Naval Research Laboratory

The study of chaos, a complex form of motion, is relatively new. It is only now that possible uses for chaos are beginning to emerge. We believe that many new technologies useful to the Navy will eventually come from this field. One possible application that we have been studying is the use of chaos as a broad band signal for communications.

Until recently, the study of classical mechanics proceeded in two different directions. On the one hand were deterministic mechanical systems, where one could write down all the laws of motion. Deterministic systems were believed to proceed in an orderly fashion, like clockwork; it was from these studies that the nineteenth century concept of God as a clock maker arose. Small perturbations in the motion were possible, but it was believed that they would be damped out over time and have little effect.

On the other hand, statistical mechanics developed as the study of stochastic, or random, systems. No laws of motion existed, so these systems were not predictable. If one studied a large number of identical stochastic systems, one could find that their motion did fit certain statistical distributions. The study of statistical mechanics therefore involved the study of long time averages of the motion of a system, of the average motion of many identical systems.

As early as the time of Newton, it was suspected that not all motion fit into these two neat classes. Newton had speculated that the deterministic motion of a planet in orbit might become highly irregular if it was perturbed by a third body, such as a comet. A foundation for the study of complex nonlinear dynamical systems was not laid until the late nineteenth century by the mathematician Poincare. This work was not built on until the 1960's, when a more serious study of nonlinear dynamics and chaos began.

Communications depend on both deterministic and stochastic mechanics. Deterministic mechanical laws govern the behavior of such common devices as oscillators or filters. Stochastic mechanics is used in the study of noise and in random number generators for encryption or spread spectrum communications. The type of motion known as chaos has some of the properties of both of these types of motion. Chaos is a highly irregular, nonperiodic form of deterministic motion which is extremely sensitive to initial conditions. In principle, one could calculate the future motion of a chaotic system, but a small error in specifying the initial conditions will grow exponentially with time. Since all measurement involves some error, it is practically impossible to predict the future motion of a chaotic system. The unique properties of chaos have not yet been exploited for any useful systems; most applied research has concentrated on suppressing chaos. We have been studying how to apply chaos to problems of interest to the Navy, such as communications. After a brief introduction to chaos, we will describe one possible application of chaos.

Introduction to Chaos

The description of chaos came about as a result of advances in the field of nonlinear dynamics¹. A dynamical system is a system that changes over time in a way that may in principle be described by some set of rules, such as differential equations or recursion relations. A pendulum is an example of a dynamical system, as is an electrical oscillator. If one knows the rules that govern a dynamical system, one may predict the future behavior of the system knowing only its starting point. For this reason, the type of dynamical systems that we describe here are known as deterministic.

Original from
UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

The behavior of a dynamical system is typically described by plotting its trajectory in a phase space, where each direction corresponds to one dynamical variable. The systems we are concerned with are dissipative, so their motion in phase space will eventually settle down to a finite region known as an attractor. A system with periodic motion will follow a closed path in phase space. A system whose motion possesses incommensurate periods will move on the surface of a torus.

Complex motion is possible if one or more points in the phase space are unstable, so that motion in regions of phase space near these points diverges. Since all real systems are finite, this motion can not diverge forever. It may be that the system variables reach some maximum possible value, such as a power supply voltage, and stay there. In other systems, some mechanism for folding the motion back in towards the unstable region may exist. In this type of attractor, trajectories repeatedly are stretched apart near the instability and folded back in. This repeated stretching and folding in phase space produces the complex motion known as chaos. The stretching and folding also cause the motion to depend sensitively on the initial conditions, making prediction a practical impossibility.

The rate of this stretching is described by the Lyapunov exponents. One may measure what happens to small perturbations to the trajectory in phase space. The average change in a perturbation is measured over the entire attractor. The natural logs of the components of the average change vector are the Lyapunov exponents. A positive exponent means that perturbations are increasing along some direction, a negative exponent means that perturbations are decreasing, while a zero exponent means that perturbations are staying the same size. A chaotic attractor has at least one Lyapunov exponent greater than zero, while the largest exponent for a periodic attractor is zero.

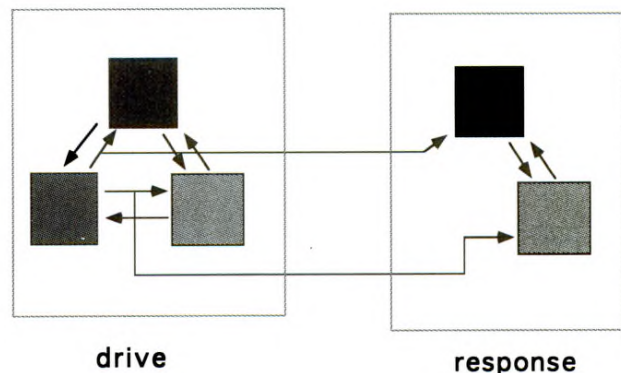
Chaotic Synchronization

Since chaos looks like noise, most applied work on chaos has been directed toward eliminating it. We saw these noise-like properties as useful. One might think of a chaotic system as a noise source that, because of its deterministic nature, may be easily characterized. It occurred to us that if we could produce a chaotic signal at one location and reproduce an identical chaotic signal at a remote location, then we could encode information onto the original chaotic signal and decode it using the reproduced signal. Because of their great sensitivity to initial conditions, isolated chaotic systems not only will not synchronize with each other, their outputs will not even be correlated with each other. It is necessary to send information from one chaotic system to the other in order to synchronize them.

Our approach was to view a chaotic system as a group of interconnected subsystems. We could then reproduce one of the subsystems and drive it with whatever chaotic signals it

Figure 1.

Block diagram of synchronizing chaotic systems. The response system is a duplicate of part of the drive system. The response system is driven by the signal or signals that come from the missing part of the system.



originally saw from the full chaotic system². Figure 1 is a schematic of this arrangement. We call the full chaotic system the drive system, and we call the driven subsystem the response system. The output of the response system depended on its Lyapunov exponents. If all the Lyapunov exponents of the driven response system were less than zero, then the response system will follow the drive signal, so that a particular signal in the response system is synchronized with the corresponding chaotic signal in the drive system. Because the response system is stable, this method is not sensitive to small amounts of noise or mismatch in the drive and response systems. The range of initial conditions over which this synchronization will occur varies from system to system, and depends mainly on the presence of other basins of attraction in the system.

We may demonstrate this synchronization with a numerical experiment. We use the Lorenz equations, which are well known in the field of nonlinear dynamics:

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y - x) \\ \frac{dy}{dt} &= -xz + rx - y \\ \frac{dz}{dt} &= xy - bz\end{aligned}$$

When $\sigma = 10.0$, $r = 60.0$ and $b = 8/3$, these equations produce chaos. We used a numerical integration routine on a workstation to integrate these equations. We chose a subsystem consisting of the x and z equations and drove them with the y variable from the full chaotic system. Figure 2 shows how the z variable in the response system converged to the z variable in the drive system within a few cycles of the drive.

Cascading Synchronized Chaotic Systems

We may actually choose other response systems for the Lorenz equations above. The y and z equations together also form a stable subsystem that may be used as a response system when driven by the x variable. Since there are two possible response systems, we may also cascade these response systems as shown in Fig. 3. One may think of the cascaded response systems as a black box where a chaotic signal from the drive signal goes in and a chaotic signal comes out. If the response systems are synchronized to the drive system, the output chaotic signal matches the input signal. If we change a parameter in the drive system, then the output chaotic signal does not match the input chaotic signal. In this way we may send

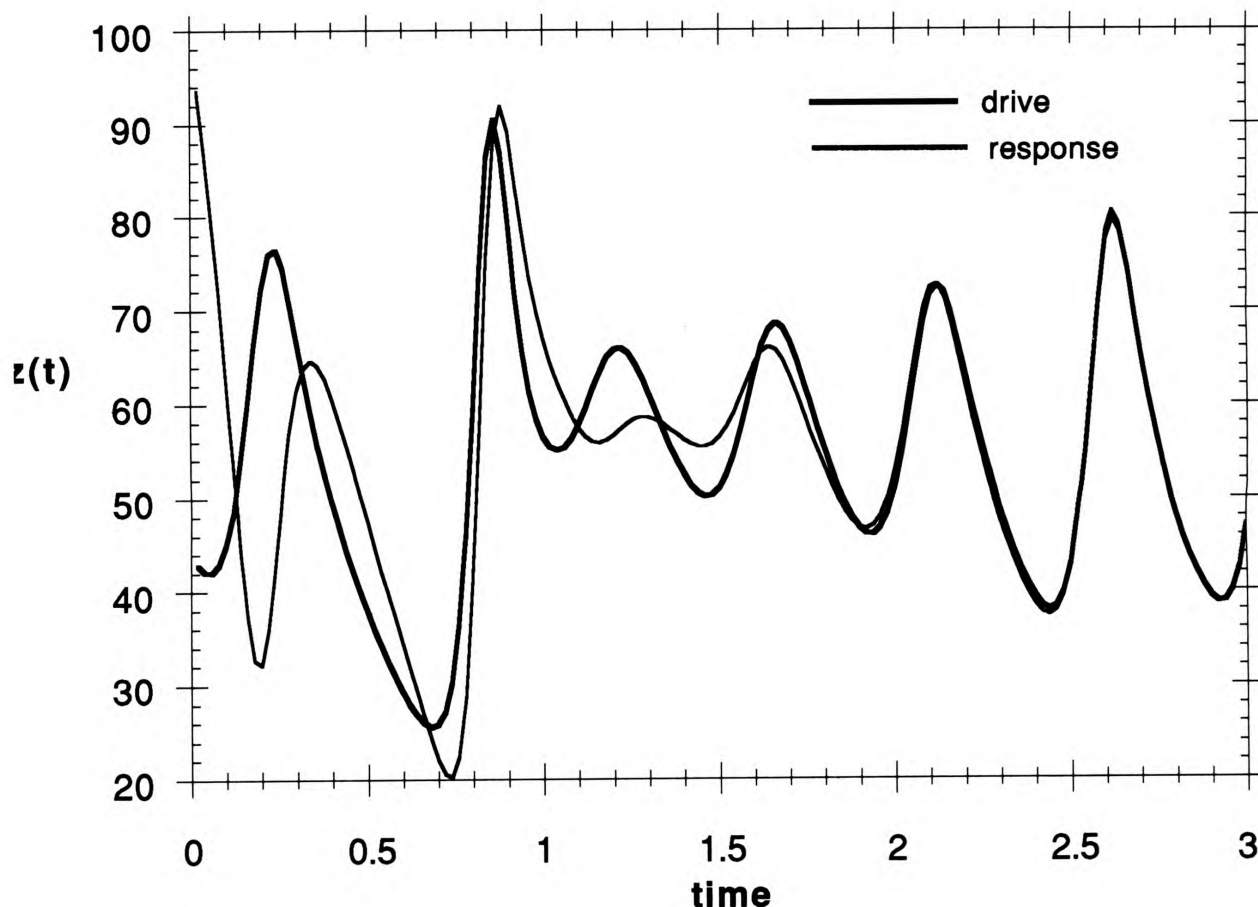
information using the chaotic signal as a carrier.

We wanted to show that this chaotic synchronization worked in real systems, so we built the simple chaotic circuit shown in fig. 4³. This circuit becomes chaotic through a period doubling cascade as shown in fig 5 (a)-(d), which shows a series of attractors seen as resistor R_{12} is decreased. Figure 6 shows the power spectrum of the voltage measured at the x_1 point in the circuit. This power spectrum is broad band with a few prominent frequency peaks. More complex chaotic systems can have power spectra that resemble colored noise.

Designing even a simple chaotic circuit such as this was a challenge. There are no general laws that determine when a given system will be chaotic. We could merely use our intuition and experiment to find chaotic circuits. We also had to find a chaotic circuit that could be divided into at least two response subsystems. There is no general method for

Figure 2.

Time series from the computer simulation of the Lorenz equations showing how the response system converges to the drive system.



designing a circuit so that it has stable subsystems. We were able to come up with a few guidelines that work in simple cases, but again much intuition was necessary.

The hardest problem to solve was matching the response subsystems to the full circuit. Matching the circuits required finding nonlinear elements whose behavior matched to within 1 or 2 % over a wide range. Typical semiconductor devices by themselves never come close to these figures. Some manufacturers produce chips that perform analog multiplication, but these still were not reproducible enough. We finally solved the problem of producing reproducible nonlinearities by using an arrangement from analog computer technology known as a diode function generator. In this type of circuit, diodes are used to switch different linear characteristics on or off to produce a piecewise linear approximation to any function. The boxes in fig. 4 labeled g1 and g2 contain diode function generators.

This circuit could be divided into two subsystems, one containing the x_4 variable and the other containing the x_1, x_2, x_3 variables. Figure 7 shows schematically how the circuits were cascaded, with the full system driving the x_4 subsystem, which then drove the x_1, x_2, x_3 subsystem. With this arrangement, the x_1 signal used as an input and the x_1 output match to within about 2%.

Figure 4.

Schematic of a circuit used to study the cascading of synchronized chaotic systems. Voltages were sampled at the points labeled x_1, x_2, x_3, x_4 .

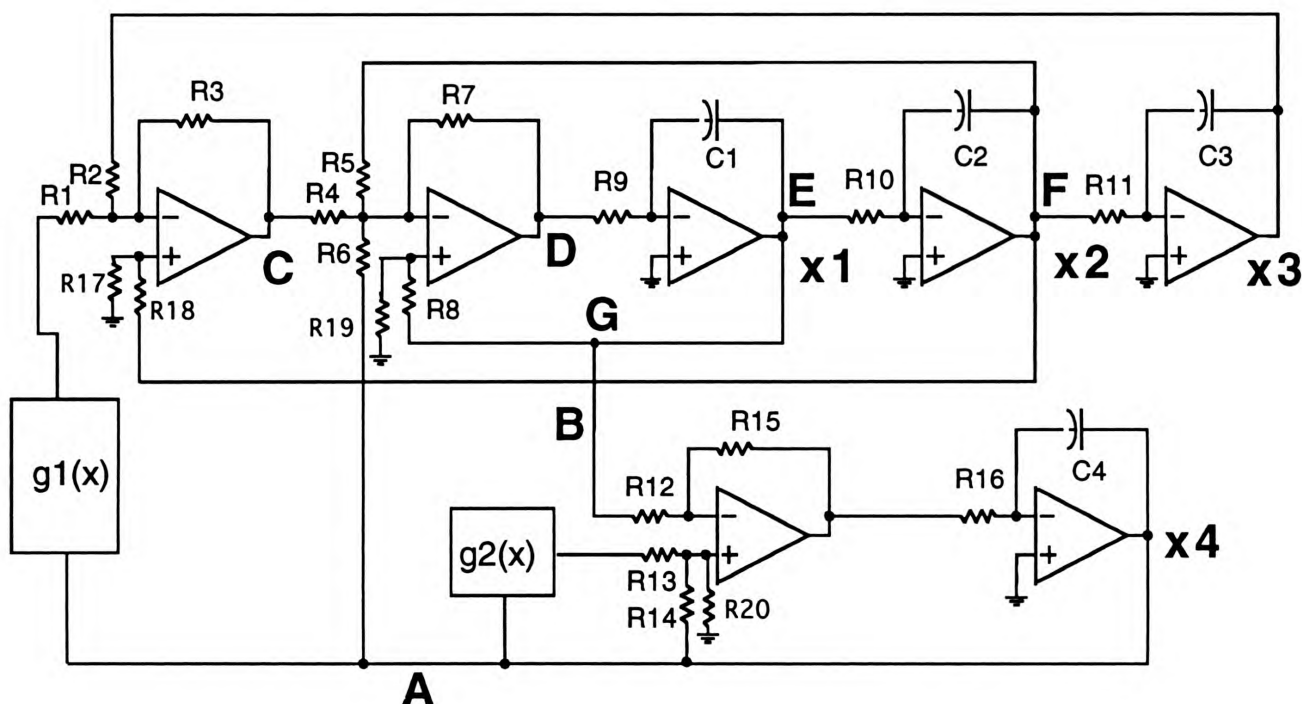
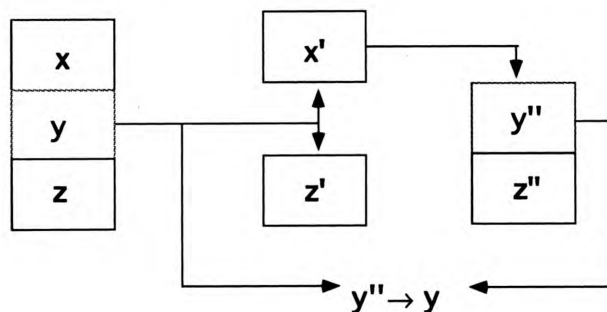


Figure 3.

Block diagram of cascading the synchronized chaotic Lorenz equations.

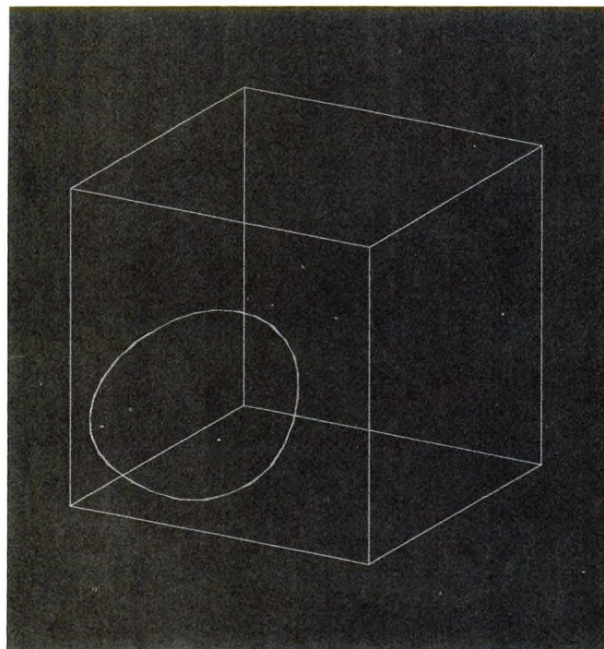


Applying Chaotic Synchronization

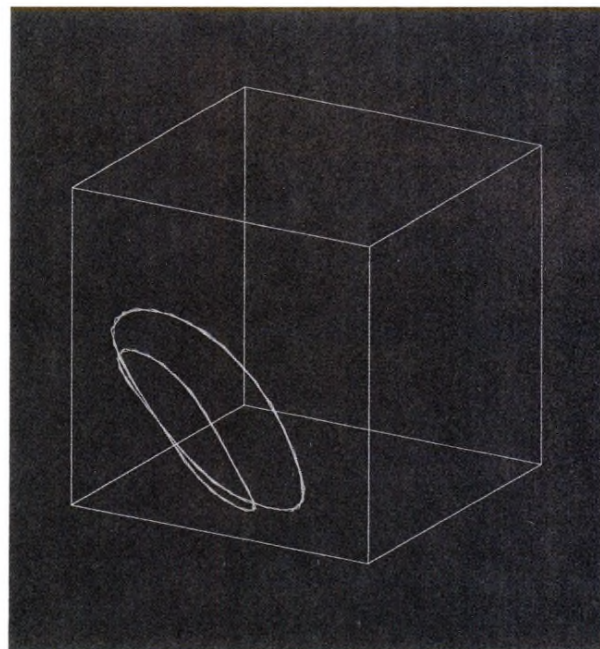
We have tried several methods for sending information with our chaotic carrier signal. The simplest method involves using a single non-cascaded response system. In this arrange-

Figure 5.

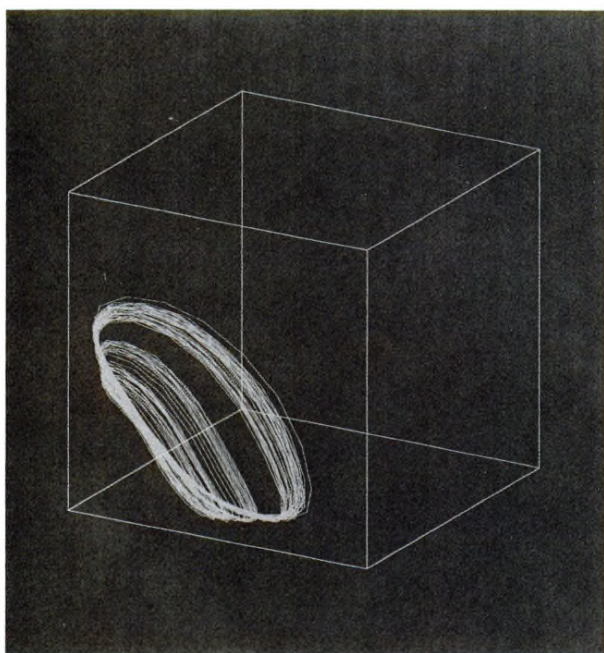
(a)-(d) show a series of attractors seen in the circuit as resistor R_{12} is decreased.



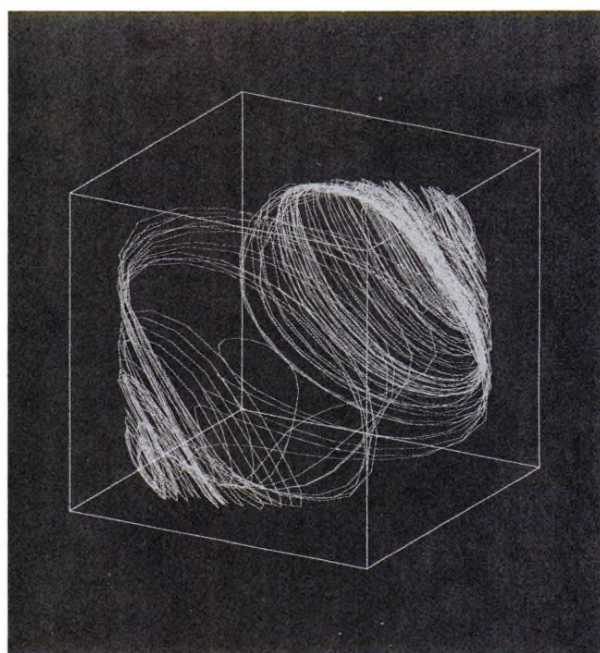
(a)



(b)



(c)



(d)

ment, the drive signal is used to synchronize the response system to the drive system. In the Lorenz system, for example, the x-z subsystem would be driven by the y signal. Information would then be added to another of the chaotic signals, such as the z signal. This signal would also be transmitted to the response system. The synchronized z signal produced by the response system could then be subtracted from the transmitted z+information signal to yield the information. We have successfully demonstrated this method in a circuit.

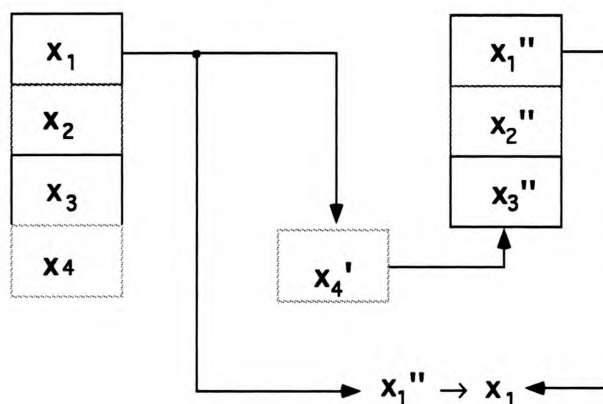
This approach is simple, but it does require that the user send two signals. A more elegant approach is to use our cascaded synchronized chaotic systems to send information with only one signal. We have tried two simple methods: one that is analogous to the amplitude modulation that is already in use, and another that resembles phase modulation. We applied these ideas to both numerical experiments and the circuit described above. In this space we will describe only the circuit results.

In our amplitude modulation experiment, we used an analog multiplier chip to multiply the output of the drive system by $1 + 0.1 \sin(\omega t)$. We used a modulation frequency ω of 6 Hz, which was about 1% of the most prominent frequency in the power spectrum of the chaos, which was about 700 Hz. The effect of the modulation was easily detected by taking the difference between the modulated driving signal and the output of the response circuit. There was a limit to the fastest modulation signal that could be used. The rate at which the response system converged to a drive signal was determined by the least negative of the Lyapunov exponents, which was about -600 s^{-1} for this system. If a modulation signal of 70 Hz was used, the response circuit could not follow the drive signal as well.

An alternate method for sending a signal on a chaotic carrier with the cascaded synchronized chaotic circuits was to modulate a parameter in the drive circuit. We put an analog multiplier chip in front of one of the operational amplifiers in

Figure 7.

Block diagram of cascading the synchronized chaotic circuit.

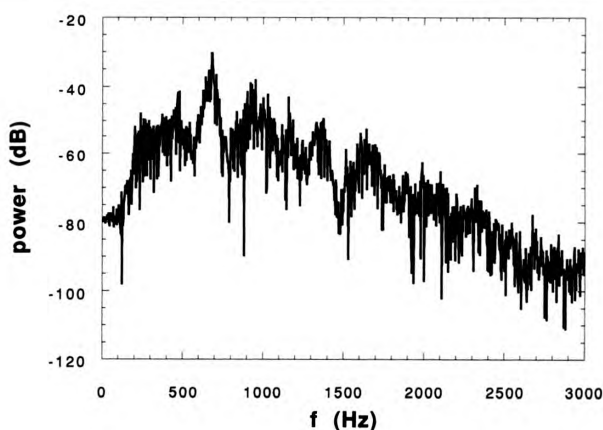


our chaotic circuit and used a potentiometer to set the multiplication factor. Changing this multiplication factor would change the output of the drive circuit, so that the input to and output from the response circuit no longer matched. Using this difference, one could calculate the difference between the multiplication factor in the drive circuit and the multiplication factor applied to a multiplier in the corresponding part of the response circuit. The multiplication factor in the response circuit could then be corrected to track the multiplication factor in the drive circuit. Used in this fashion, the cascaded circuits resembled the chaotic version of a phase locked loop.

Extracting the parameter difference from the difference between two chaotic signals was not as straightforward as the same procedure for two periodic signals, where one could compare frequency or phase. A chaotic signal does not have a

Figure 6.

Power spectrum of the x_1 signal from the circuit when the circuit is in the chaotic attractor seen in Fig. 5(d).



single frequency or a phase. In this case, there is a quantity resembling an average phase for the chaotic signal that can be used. When all circuits are synchronized, if we strobe the output of the response system when the drive signal crosses zero, the result will be zero. If the drive and response are not synchronized, then the output of the response will not be zero when the drive crosses zero. The average of the output signal at this zero crossing varies approximately linearly for small parameter differences. An analog integrator circuit is used to create this average from the output of a sample and hold circuit. This average is then used with an analog control circuit to correct the value of the multiplication factor in the response circuit.

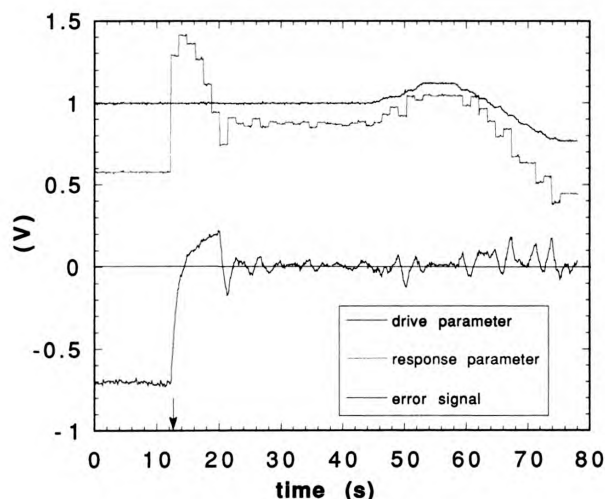
Figure 8 shows that this technique works in a real circuit. This figure shows the value of the multiplying factor in the drive circuit and in the response circuit. At $t=12$ s, the control is turned on. The multiplication factor in the response circuit is immediately affected, and soon settles to a value near the multiplication factor for the response circuit. The response does not match the drive more closely because these parameters are set with analog multiplier chips, which do not match very well, but were convenient to use for this experiment. The response multiplication factor is able to track changes in the drive multiplication factor, as is shown in the figure.

Conclusions

These simple experiments merely serve as proof that these concepts work. These examples may not correspond to the way

Figure 8.

Results of using a chaotic carrier from the circuit to send the value of a parameter to a cascaded response circuit.



that these principles would actually be used, but that is not important. The important part of this work is that it starts engineers, physicists and mathematicians thinking differently about using chaotic systems. The idea that chaotic systems may be taken apart and recombined to form new, useful systems is completely new. So is the idea of driving a system with a chaotic signal. There are many variations on these ideas that we have not discussed.

In our experiments, we also have the luxury of no noise. Chaotic signals are transmitted by wires. In a real communications system, the chaotic signals would be broadcast though a noisy background. Separating a nonperiodic signal from noise is not an easy task. There are several new software algorithms being developed that might work adequately with our system to allow us to detect chaotic signals in a noisy background.

Summary

We have shown how one may take apart and reconnect chaotic systems so that one may drive one system with another, which then synchronizes to the drive. We have used these techniques to create chaotic versions of am or fm radios, where the periodic carrier is replaced with a chaotic carrier. These ideas may have applications in encryption or in spread spectrum communications, technologies which are important both to the Navy and to civilian industries.

Biographies

Thomas L. Carroll received a B. S. degree in physics from Carleton College in 1981 and M. S. and Ph D. degrees in physics from the University of Illinois at Urbana-Champaign in 1983 and 1987, where he did spectroscopic studies of molecules under high pressure. He came to NRL in 1987 as an Office of Naval Technology postdoctoral fellow and became a permanent lab employee in 1991. His initial research at NRL concerned nonlinear dynamics and chaos in ferromagnetic resonance experiments in yttrium iron garnet. Since that time his research has involved coming up with new applications for chaos and building analog electronic circuits or computer models to test these circuits. He has published approximately 30 papers and has applied for four patents.

Louis M. Pecora is a research physicist at the Naval Research Laboratory in the Materials Physics Branch. In 1977, he received a Ph.D. from Syracuse University and was awarded an NRC postdoctoral fellowship at NRL where he worked on applications of positron annihilation techniques in determining electronic states in copper alloys. This work continued into the areas of reconstruction of momentum densities using novel tomographic techniques. Dr. Pecora moved into the field of nonlinear dynamics in solid state systems; his work led to the discovery and characterization of chaotic transients

in spin-wave behavior in yttrium iron garnets. More recently, he has focused on the applications of chaotic behavior, especially the effects of driving systems with chaotic signals. This has resulted in the discovery of synchronization of chaotic systems and pseudoperiodic driving of nonlinear systems. Dr. Pecora has published over 50 papers and applied for four patents pertaining to the applications of chaos. Address: Naval Research Laboratory, Code 6341, Washington, D.C., 20375-5000, e-mail: pecoara or carroll@zoltar.nrl.navy.mil.

References

1. P. Berge, Y. Pomeau and C. Vidal, *Order within Chaos* (Wiley, New York, 1984).
2. L. M. Pecora and T. L. Carroll, "Driving Systems with Chaotic Signals", *Phys. Rev. A* **44** 23742383 (1991).
3. T. L. Carroll and L. M. Pecora, "A Circuit for Studying the Synchronization of Chaotic Systems", *Int. J. Bifurcations and Chaos* September 1992.

Controlling Chaos

Mark L. Spano
Naval Surface Warfare Center
and
William L. Ditto
Georgia Institute of Technology

Abstract

The concepts of chaos and its control are reviewed. Both are discussed from an experimental as well as a theoretical viewpoint. Examples are then given of the control of chaos in a diverse set of experimental systems. Current and future applications are discussed.

Introduction

Chaos – A rough, unordered mass of things.

Ovid, "Metamorphoses"

Shakespeare once asked, "What's in a name?" The answer is nothing and everything. Nothing because "A rose by any other name would smell as sweet." And yet, without a name Shakespeare would not have been able to write about that rose or to distinguish it from other flowers that smell less pleasant. So also with chaos. The dynamical phenomenon we call chaos has always existed, but until its naming we had no way of distinguishing it from other aspects of nature such as randomness or order. And lacking any way to treat it as a separate entity, we would not even be able to conceive of manipulating it or controlling it. Thus the naming of chaos by James Yorke¹ in 1975 was an essential first step towards identifying it and learning to control it.

From this identification came the recognition that chaos is pervasive in our world. Chaos was first suspected in the weather patterns² that govern our lives. It is easy to make simple electronic circuits that are chaotic³. Mechanical systems on such wildly different scales as laboratory pendula⁴ and orbiting planets⁵ have been shown to exhibit chaos. Laser emissions⁶ can fluctuate chaotically. The human heart shows

evidence⁷ of beating to chaotic rhythms. Chemical reactions^{8,9} oscillate in chaotic fashion. Of these diverse systems, we have learned to control all of those that are on the smaller scale. Systems on a more universal scale such as the weather and the planets remain beyond our control.

In what follows we will discuss the concept of chaos, from both a theoretical and a laboratory viewpoint, always with an eye towards methods of controlling and exploiting it. Afterwards we will look at the application of these methods to the systems mentioned above.

The Concept of Chaos

*A violent order is disorder; and
A great disorder is an order. These
Two things are one.*

Wallace Stevens, "Connoisseur of Chaos"

Before its discovery and naming, chaos was inevitably confused with randomness and indeterminacy. Because many systems *appeared* random, they were actually thought to *be* random. This was true despite the fact that many of these systems seemed to fall into spans of almost periodic behavior before returning to more "random" motion. Indeed this observation leads to one of the definitions of chaos: the superposition of a very large number of periodic motions. A chaotic system may dwell for a brief time on a motion that is very nearly periodic and then may change to another motion that is periodic with a period that is perhaps five times that of the original motion and so on. This constant evolution from one (unstable) periodic motion to another produces a long-term impression of randomness, while showing short-term

Original from

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

Figure 1.

Phase space trajectories (a, c, e) and their corresponding Poincaré sections (b, d, f) for period 1, period 2 and chaotic orbits.

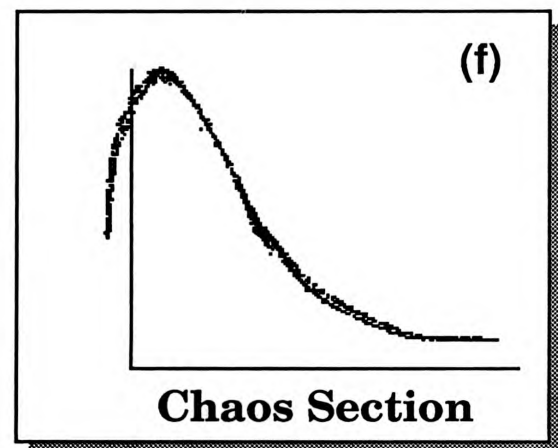
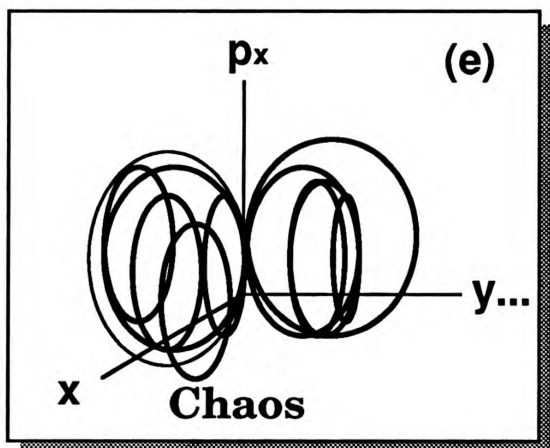
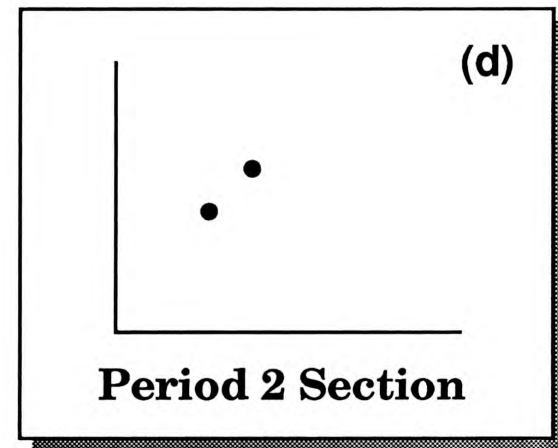
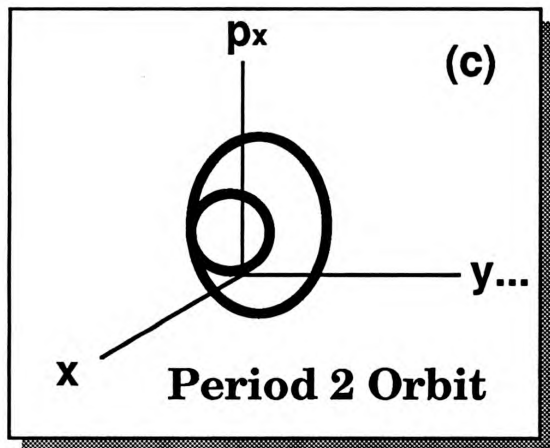
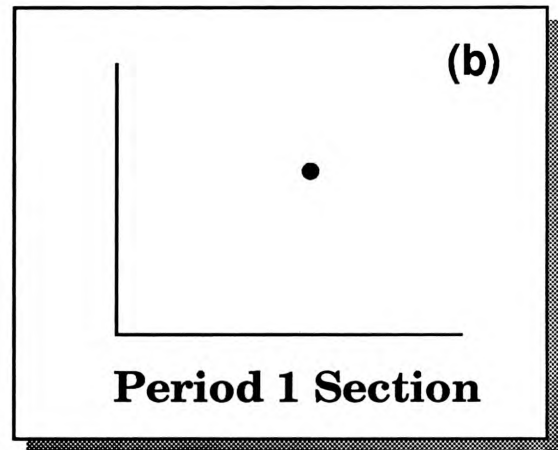
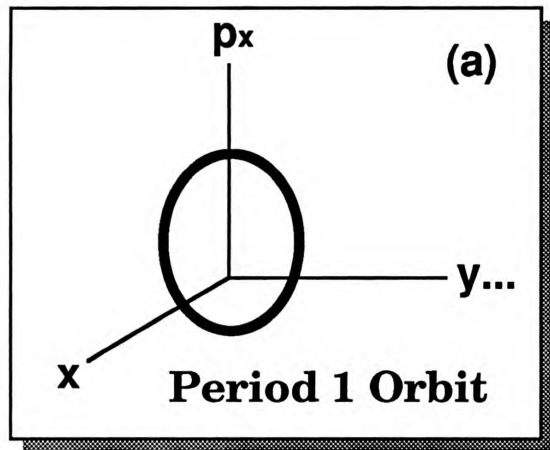
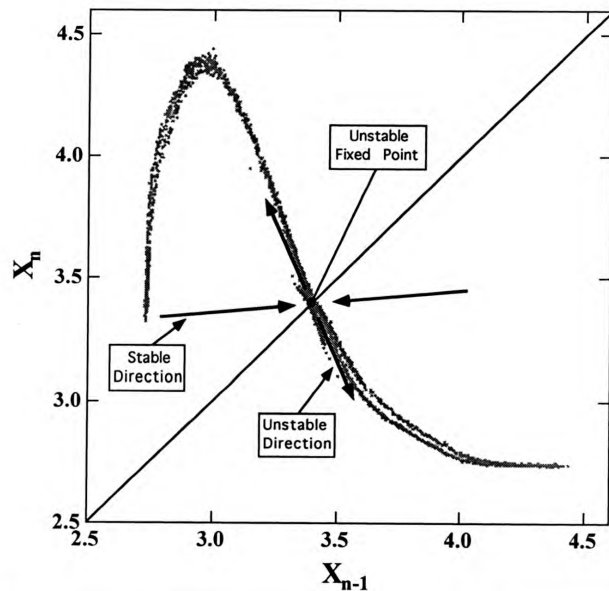


Figure 2.

The geometry around an unstable fixed point on a chaotic attractor. The shape is that of a saddle.



glimpses of order. These glimpses are not misleading, since chaos is deterministic and not random in nature.

A second definition of chaos is this: the sensitivity of a system to small changes in its initial conditions. This is the “butterfly effect” of legend. The story goes as follows: the small breeze created by a butterfly’s wings in China may eventually mean the difference between a sunny day and a torrential rainstorm in California. If you visualize the motion of a chaotic system in phase space (the space composed of the usual spatial dimensions plus their accompanying momentum dimensions), you soon see that all these unstable periodic motions have corresponding trajectories in phase space. And since there are a very large number of these unstable motions in a bounded space, the trajectories corresponding to each unstable periodic motion must be packed quite closely together. Thus only a very small perturbation may push the system from one unstable periodic motion to any of a large number of others.

This last is both the hope and the despair of those who have to deal with chaotic systems. It is the despair because it effectively renders long term prediction of these systems impossible. For instance, in a computer simulation a small perturbation may be introduced by numerical round-off, while, in the real world, systems are invariably perturbed by noise. Thus, if a system is chaotic, these small perturbations quickly (indeed, exponentially) grow until they completely change the behavior of the system.

Paradoxically, the cause of the despair is also the reason to hope. Because if a system is so sensitive to small changes, could not small changes be used to control it? This realization led Ott, Grebogi and Yorke¹⁰ (OGY) to propose an ingenious and versatile method for the control of chaos, the details of which are given in the following section.

The Idea of Control

*At which the universal host up sent
A shout that tore hell's concave, and beyond
Frighted the reign of Chaos and old Night.*

John Milton, “Paradise Lost”

The starting point for the control of chaos is the phase space of the system. Dynamical systems trace smooth trajectories in phase space as they proceed forward in time. For periodic systems these trajectories are curves that are traced repeatedly every period of the system. For example, the phase space of a driven simple pendulum consists of one dimension for its position and one dimension for its momentum. The trajectory of the pendulum in this space is simply a circle which is traced over and over, as in Fig. 1a. This trajectory contains all the information that is needed to predict the future

Figure 3.

Determining the response of a chaotic attractor to a change in a system parameter, H_{dc} . The change in the unstable fixed point position with a change in the parameter yields the vector quantity $\mathbf{g}$.

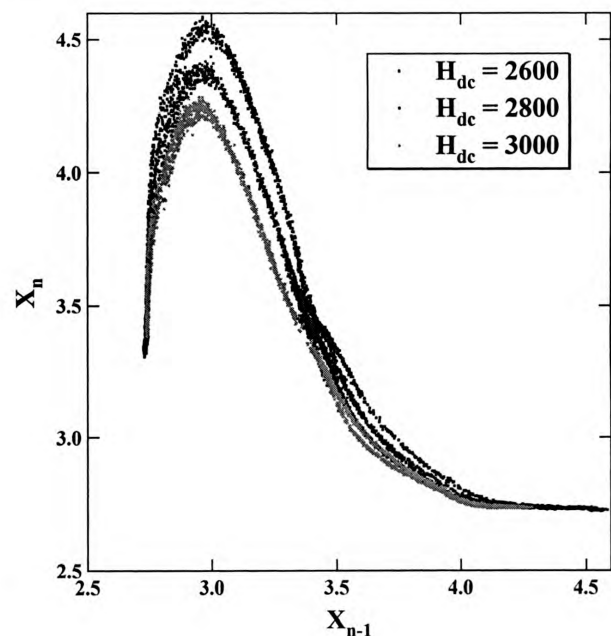
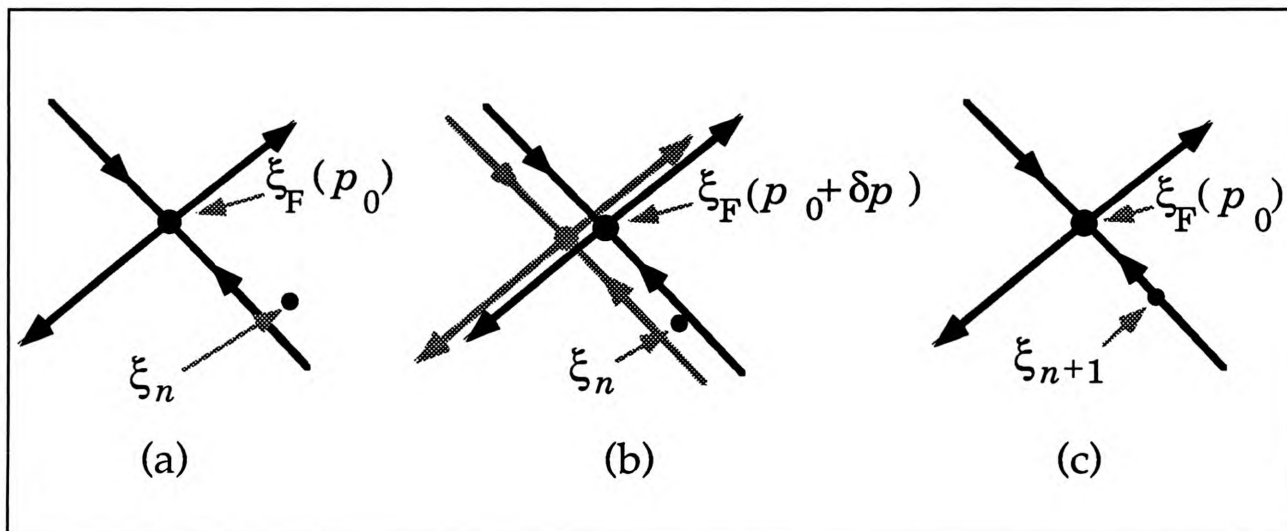


Figure 4.

An outline of the OGY control method: (a) the system state point approaches the unstable fixed point. It will naturally tend to move in along the stable direction and out along the unstable direction; (b) the saddle point (unstable fixed point) is positioned so that the system state point tends to move back toward the original unstable fixed point; (c) the system state point reaches the original unstable fixed point and the perturbation is turned off. In real-world systems with noise, the control would have to be applied continuously to balance the state point on the saddle.



dynamics of the system. In fact it contains too much information. A more useful representation can be obtained by cutting through this phase space with a plane which intersects the circle in two places. The infinite number of points on the circle trajectory has been reduced to merely two. If we further confine ourselves to *directed* piercings of the plane, we are left with only a single point (Fig. 1b). Such a *Poincaré section* reduces our information to a manageable level.

If our experiment were more complex, perhaps a driven pendulum with two masses placed at different points on a flexible string or the simple pendulum driven well beyond its linear regime, more complex motions might occur. One such might be a motion in which the trajectory traced a double loop in its phase space (Fig. 1c). A Poincaré section through this trajectory would show two points (Fig. 1d). This pendulum could, of course, still execute a simple, single loop trajectory given the right initial conditions. This single loop defines the *period one trajectory* of the system. The double loop trajectory takes twice as long and is accordingly dubbed *period two*. The usefulness of the Poincaré section is now apparent: the number of points in the section immediately reveals the underlying periodicity of the system. Generally, any periodic system of period n will have a section which consists of a finite set of n distinct points which reflect the fundamental periodicity of the system.

By way of contrast, a truly random system will behave in such a way that the trajectories in phase space wander over the entire volume of space available to the system. Its corresponding section will be filled uniformly and densely.

Chaos falls between these two extremes. Since chaos is a superposition of a number of periodic motions, one might expect to see a finite number of points indicating several periodic motions characteristic of the chaotic section. This is true, as far as it goes. However, since chaos is the superposition of a large (read *infinite*) number of periodic motions, the number of points in the section is also infinite. In general, these points form an extended geometric structure, called the system's *chaotic attractor*, which is not a finite set of points (*i.e.*, does not represent periodic motion) and which also does not fill space (*i.e.*, is not random), as shown in Fig. 1f. (Generally a chaotic attractor is a fractal object.) Knowledge of this attractor and of its response to small perturbations of the system are the only ingredients that are necessary for the control of chaos.

In order to control chaos according to the OGY scheme, it is only necessary to identify an unstable periodic point in the attractor, to characterize the shape of the attractor locally around that point and to determine the response of the attractor at that point to an external stimulus. Let's look at each of these three steps in detail.

Original from

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

Three/1993

Figure 5.

Schematic of the magnetoelastic ribbon experiment.

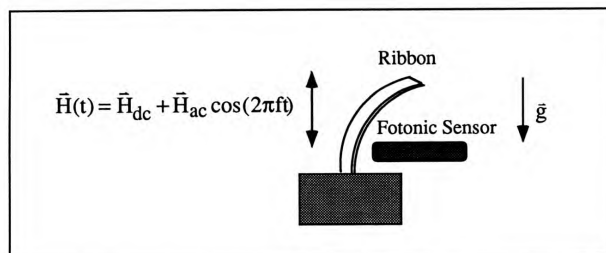
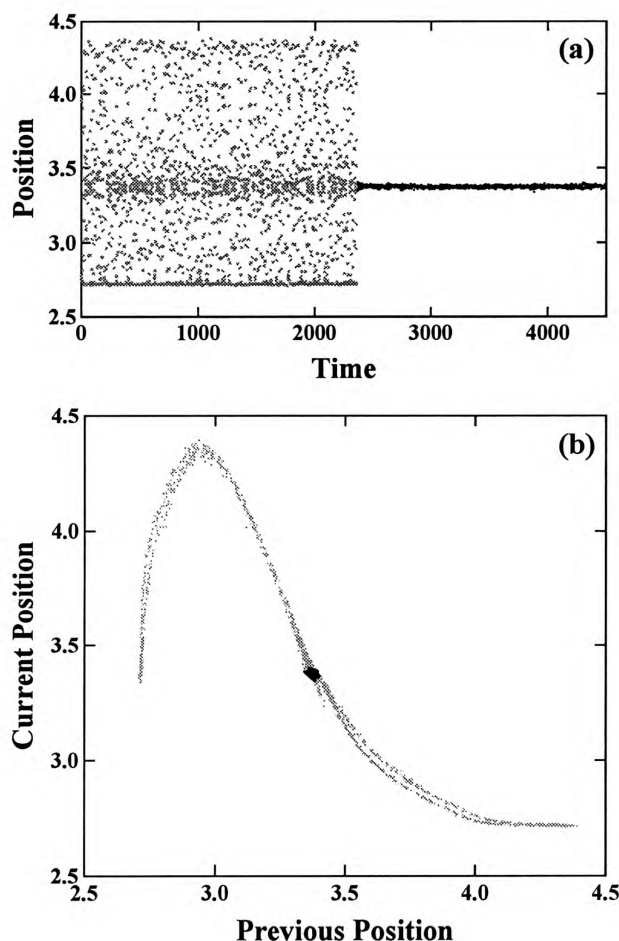


Figure 6.

(a) Ribbon position as a function of time for chaotic motion (gray) followed by controlled motion (black); (b) The same data presented as a delay coordinate embedding.



The identification of an unstable periodic point is easy since every point in the attractor represents an unstable periodic motion with some period n . (Of course it is easier to find points with low order periodicity (small n), but chaos control has been successfully implemented to select periodic motions of order up to about 30!) This is one of the strengths of the OGY method of control: the ability to control on any periodic motion from the infinity of motions present in the system. This capability would allow an engineer to design highly flexible systems employing chaos control.

Characterizing the shape of the attractor is also easy. Once we have determined which unstable fixed point we wish to control about, we observe the motion of the point representing the current state of the system (*system state point*) on the attractor. In low dimensional chaotic systems, this point will frequently approach the vicinity of our chosen unstable fixed point and then move away again. It turns out that, in the neighborhood of the unstable fixed point, the approach is consistently along the same direction (called the *stable direction*) and the departure along another direction (called the *unstable direction*), as indicated in Fig. 2. These two directions, one of which is stable (incoming) and the other of which is unstable (departing), form a saddle around the unstable fixed point. These *eigenvectors*, along with the speed with which the system state point approaches or departs the vicinity of the unstable fixed point (*stable or unstable eigenvalues* respectively), are all that is necessary to characterize the shape of the attractor locally around our chosen point.

The final requirement, determining the response of the attractor to an external stimulus, is perhaps the most difficult of the three to learn. Here we must introduce a change into the system by varying one of the system parameters. The difficulty lies in first identifying the system parameters and then locating one that may be quickly changed. (More on this will be discussed in the experimental section below.) Once such an identification has been made, the position of the unstable fixed point is measured for several values of the parameter just slightly different from the nominal value (Fig. 3). These measurements tell us the change of the attractor position with respect to a change in the system parameter and complete the knowledge of the system needed for control.

All of this information is fed into the OGY control equations:

$$\delta p = C (\xi_n - \xi_f) \cdot f_u \quad (1)$$

where

$$C = \frac{\lambda_u}{\lambda_u - 1} \frac{1}{g \cdot f_u} \quad (2)$$

δp is the amount one needs to change the system parameter from its nominal value in order to control the system. The perturbation δp is updated once every period of the system. It

is proportional to the distance of the system state point, ξ_n , from the fixed point, ξ_f , projected onto the unstable direction, f_u . The constant C is itself calculable from the change of the attractor position with respect to a change in the system parameter, g , projected onto the unstable direction and from the value of the unstable eigenvalue, λ_u . All of these are known from our earlier measurements.

An example may serve to clarify the matter. Fig. 4a shows an unstable fixed point with its stable and unstable directions. The system state point, ξ_n , is shown approaching the fixed point near the stable manifold. If left to itself, the natural motion of the system would be to approach the unstable fixed point along the stable direction and then move away from the fixed point along the unstable manifold (towards the upper right). However, if we introduce the OGY perturbation into the system, we move the old unstable fixed point and its accompanying manifolds to the other side of the system state point (Fig. 4b). The new motion of the system state point will still be along the stable manifold with motion away from the *new* fixed point along the unstable manifold (towards the lower left). However this is motion towards the old unstable fixed point. Thus the OGY scheme has controlled the system exactly onto the old unstable fixed point (Fig. 4c). In a perfect world, it would remain there forever. However if there is even the slightest error in the calculation of C , it will only move close to the fixed point. Or in the presence of noise it will be perturbed from the fixed point after it has reached it. Both of

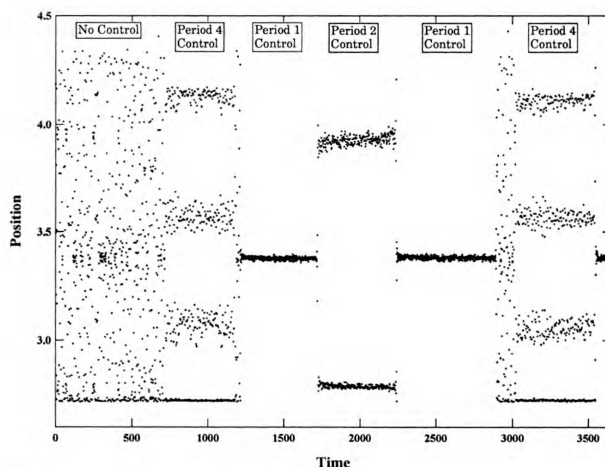
these problems are easily handled by repeating the OGY control once each period of the system.

This act of moving the saddle point to control the motion of the system state point is similar to balancing a baseball on a saddle or a baseball bat on the palm of your hand. The former case is directly analogous, with one stable and one unstable direction. Small movements of the saddle keep the ball at the point of unstable equilibrium; it is not necessary to change the shape or nature of the saddle itself.

Many methods of "control" have been proposed. Those related to the OGY method^{9,12} use perturbations that are typically small (a few percent of the magnitude of the accessible system parameter). Thus, it is easy to design systems that implement this control. This is in sharp contrast to the model-based methods¹³ which seek to fundamentally change the nature of the chaotic system under study by introducing "perturbations" to the adjustable parameter that are comparable to its size. Such methods are more properly named the "removal of chaos" rather than "control" because they fundamentally change the nature of the system under study. The gentle nudges prescribed by OGY control leave the fundamental dynamics of the system essentially unchanged. The fixed point and its manifolds are only shifted slightly by the perturbation and, between perturbations, the motion of the system remains chaotic, moving on an attractor that differs little from the original. It is only in the long term, when *averaged* over many periods and over motion on several slightly different attractors, that the motion is non-chaotic. In this sense the OGY method is truly a *control* of chaos and hence it is able to select any of the periodic orbits that are embedded in the chaotic motion.

Figure 7.

The ability of the OGY method to select out various periodic motions from those comprising the chaotic motion. (Note that the fact that the motions presented here are all a power of two is purely coincidental; the control could have as easily been performed about the period 3 or period 5 motion instead.)



Experimental Realities

In all chaos there is a cosmos, in all disorder a secret order.

Carl Jung

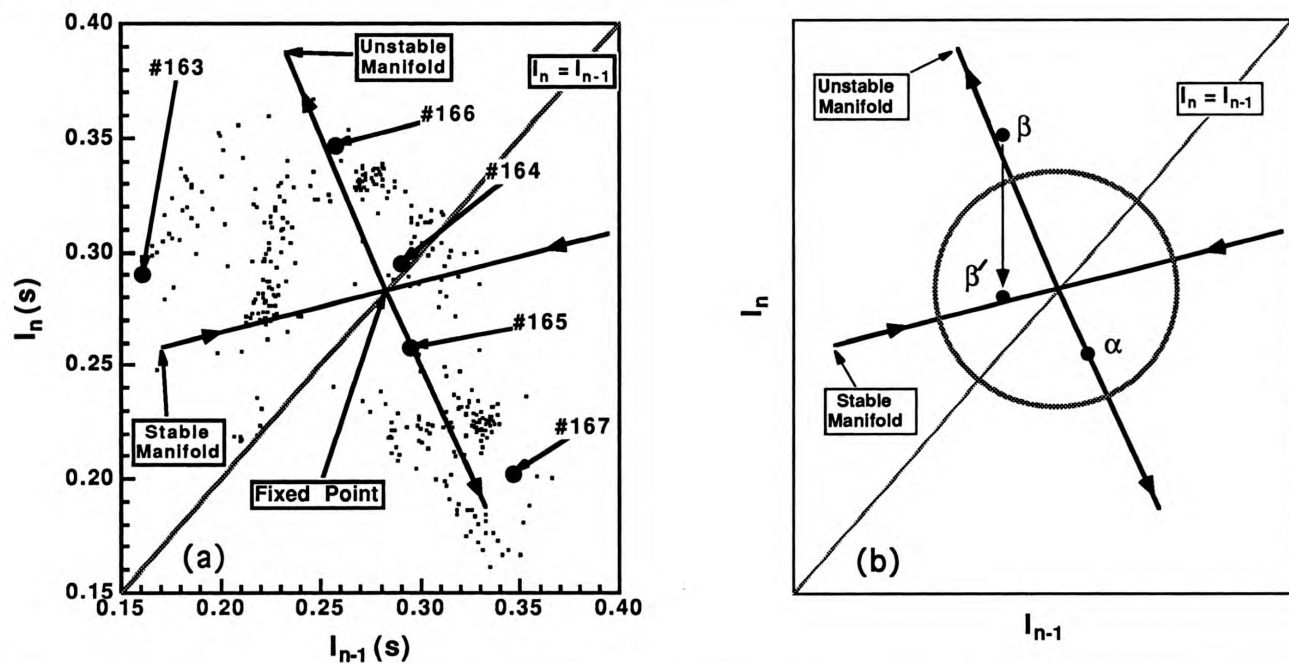
An experimental physicist's first reaction to the OGY scheme might well be one of horror; in any real-world system the number of degrees of freedom (positions plus momenta) may be quite large, possibly even infinite and the OGY scheme requires the measurement of all of them! Two facts conspire to make the OGY method not only possible but also simple to implement.

The first is that many physical systems, while technically infinite dimensional, actually only express a small number of these dimensions. Thus even though the magnetomechanical ribbon (discussed below) is a distributed mass system and therefore technically infinite dimensional, its motion is quite low dimensional and its attractor is well represented in a three dimensional phase space and hence a two dimensional Poincaré section. Thus only two experimental variables need to be measured to correctly characterize its chaotic attractor.

The second fact of importance to the experimentalist is that the Poincaré section may be replaced in the OGY scheme

Figure 8.

Current interbeat interval, I_n , versus previous interbeat interval, I_{n-1} , for the rabbit heart experiment. (a) Starting with point 163, the system approaches the unstable fixed point along the stable direction and then leaves it along the unstable directions. (The unstable eigenvalue is negative, thereby causing subsequent points to alternate above and below the unstable fixed point. However the distance from the fixed point increases in an exponential fashion, as is typical of chaotic motion.) (b) An idealized drawing of the control. The point a should be identified with point 165 in (a). The next point would naturally tend to be b (point 166). However a stimulus is applied to the heart causing it to beat early and thereby shortening the next interbeat interval to b' on the stable manifold. Thus the next interval will tend to move along the stable manifold toward the fixed point.



by a *delay coordinate embedding*. This embedding relies on the intuition that the information obtained by measuring all the system variables (positions and momenta) at one given time may also be obtained by measuring only one system variable at several subsequent times. The system state vector (the set of all the positions and momenta, $\xi_n = (x^1(t_n), x^2(t_n), \dots, p^1(t_n), p^2(t_n), \dots)$), is replaced by a delay coordinate vector, $(x^i(t_n), x^i(t_n - \Delta t), x^i(t_n - 2\Delta t), x^i(t_n - 3\Delta t), \dots)$, where the superscript i indicates one particular experimental measurable and for some appropriately chosen delay, Δt . (If the system is periodically driven, Δt is usually chosen to be the period of the driving force. In non-driven systems the appropriate delay is usually related to some fundamental time scale inherent in the system.) In fact it may be shown that the attractor as represented in the phase space and the attractor reconstructed from the delay coordinate embedding have the same topological properties, thereby allowing one to be substituted for the other with (relative) impunity.¹⁴

With these two simplifications in hand, control of low dimensional chaotic systems becomes simple. In fact it turns

out that the OGY method is also extremely robust. By this we mean that, as long as the perturbation δp is sufficient to move the saddle point to the opposite side of the system state point as shown in Fig. 4b, it does not matter if the saddle is moved well past the state point or just barely past it. The corollary to this is that the control scheme is quite resistant to external noise and to errors arising from imprecision in the measurement of the position and shape of the attractor.

We note in passing that eqs. 1 and 2 give rise to two variant methods of control. In the first method, one completely measures all the items mentioned above, calculates the constant C , and then proceeds to control the system. The alternative arises from the fact that, although many values (λ_n , f_n , and g) are needed to characterize the attractor, they are all folded into the single constant C . Thus, if a period of adjustment can be tolerated, one might simply guess a value of C and then adjust it empirically until control is achieved. This latter method, requiring knowledge only of the location of the desired fixed point and an educated guess at C makes this quite suitable for incorporation into automatic control systems.

Implementations of Control

*Not chaos-like together crushed and bruised,
But, as the world, harmoniously confused*
Alexander Pope, "Windsor Forest"

Mechanics

The first experimental control of chaos was achieved in a magnetoelastic ribbon experiment⁴. This system is sketched in Fig. 5. The ribbon, an amorphous magnetic material about 100 mm long by 3 mm wide by 0.025 mm thick, is extremely nonlinear in its response to a magnetic field. In zero field it is stiff enough to support its own weight in an upright position. When a small magnetic field is applied, its stiffness decreases by an order of magnitude and the ribbon buckles under its own weight. At still larger fields, the material again becomes stiff and resumes an upright position. If a magnetic field, H_{dc} , is applied to start the ribbon in the soft regime and to that is added an ac magnetic field, H_{ac} , that alternately takes the ribbon from soft to stiff and back again at some frequency, f , the ribbon will oscillate about its equilibrium point. At high enough values of H_{ac} and f , the ribbon's motion will become chaotic.

The position of the ribbon is measured once each period, $T = 1/f \approx 1\text{Hz}$, of the driving field by means of a beam of light directed at a small spot on the ribbon. The intensity of the reflected light varies as $1/r^2$, yielding highly accurate and noise-free measurements of the ribbon position. These are sent to the computer controlling the experiment, which also control the strengths of the magnetic fields. In theory, either of the three system param-

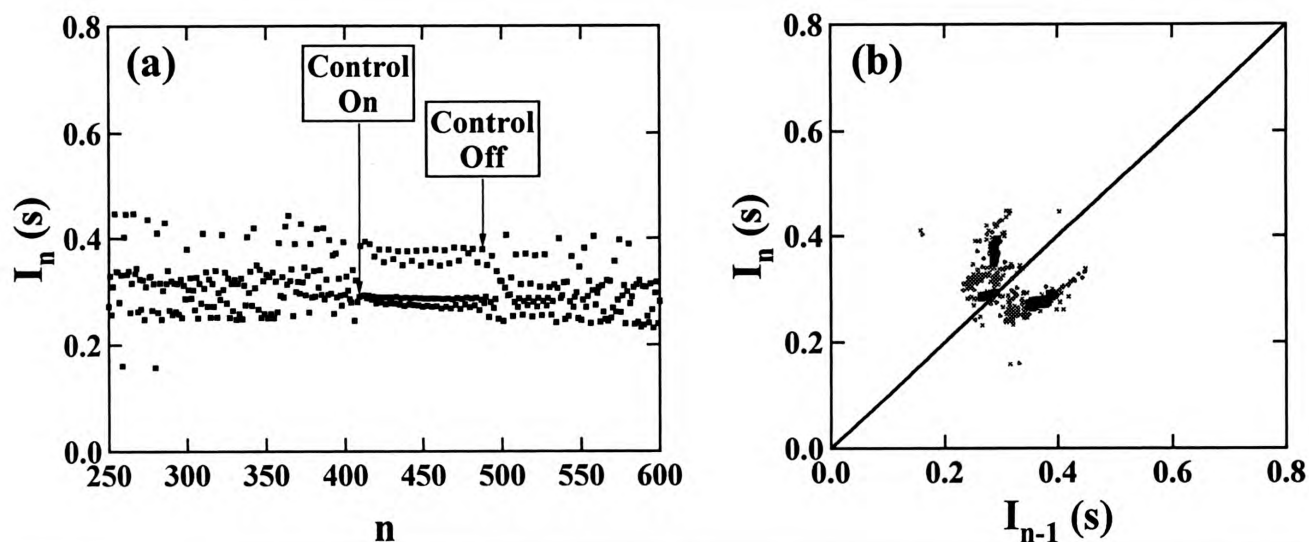
eters, H_{dc} , H_{ac} or f , could be used to control the chaos. We chose to use H_{dc} . The gray data in Fig. 6a show the position of the ribbon as a function of time when the ribbon is moving chaotically. When the computer initiates the period 1 control, the position of the ribbon (measured once each period T ; i.e., stroboscopically) is shown in black in Fig. 6a. The corresponding attractor (plotted as a delay coordinate embedding and with the same color coding) is shown in Fig. 6b.

As mentioned above the OGY method may select any of a number of periods from the chaotic signal. Fig. 7 shows the same system switching from chaos to period 4 to period 1 to period 2 to period 1 and finally back to period 4. There are interludes of chaos between each section of control because the computer must wait for the ribbon's position to approach the vicinity of the new unstable fixed point. This time can actually be decreased by a new technique called *targeting*^{15,16}, in which the sensitivity of the chaotic system to small perturbations plus a *global* (but still empirical) knowledge of the chaotic attractor are employed to direct the system rapidly to any desired point on the attractor.

The question arises as to whether one might control about a point that does not lie on the chaotic attractor. It turns out that the difficulty in such a case lies not in the control itself, but in getting the system to the vicinity of the desired point. For if the point in question is not on the attractor, it will never be visited. However, once control is established at a "normal" unstable fixed point, it is often possible to adjust some system parameter (in this case the temperature) so that the attractor smoothly changes to one that does not include the point in question. Yet it has been shown experimentally that the control

Figure 9.

(a) Interbeat intervals as a function of time for the rabbit heart experiment. (b) Delay coordinate embedding for the same data. The pre- and post-control data is gray, while the controlled data is shown in black.



will be maintained about this point. The reason is that, although the point is no longer part of the attractor, the local geometry about the point has not changed appreciably and thus the algorithm still calculates the proper control perturbations. This is an important result for understanding the robustness of the method even under large changes to the system itself.

Electronics

In an experiment³ of great potential importance for our electronics-based world, Earle Hunt of Ohio University used the empirical version of the OGY scheme to control the chaotic oscillations of a diode resonator. The significance of his experiment lies in the extension of the applicability of the method to high frequencies (53 kHz) and in his ability to select from the chaotic attractor periods up to period 23.

Lasers

In conjunction with Hunt, Rajarshi Roy et al.⁶ of Georgia Tech controlled the chaotic oscillations of a chaotic multimode laser. This laser would autonomously become chaotic at pumping levels exceeding a certain threshold. The system is somewhat higher dimensional than the previous systems, making the control more difficult to obtain. Additionally the system is not periodically driven, making it necessary to choose a "natural periodicity characteristic of the system" as the time scale for analyzing their data and applying the control signals. A third intricacy is that the natural frequency mentioned above is on the order of hundreds of kilohertz. They developed a method related to the OGY method to control of the system up to period 9. Thus the laser could be operated stably at power levels exceeding by as much as a factor of 15 those which might be attained without chaos control.

Biology

The authors have recently extended the techniques of chaos control to the field of biology.⁷ In this experiment a section of a rabbit's heart was induced to beat chaotically by the injection of the drug ouabain. The measurable quantity here is the interval between heartbeats, which is about 0.8 sec in the healthy tissue. The effect of the drug is to accelerate the heartbeat and to cause the interbeat intervals to vary chaotically. The chaotic attractor for this situation is shown in Fig. 8a.

There are several factors that distinguish this experiment from the earlier successes at chaos control. The first is that the data for the delay coordinate embedding are no longer equally spaced in time, but are rather intervals between events. The second factor is that there are no adjustable system parameters that may be modified quickly enough to implement the unmodified OGY control scheme. The authors accordingly developed a method whereby the perturbation is applied directly to the system variables, rather than to some system parameter.

This method employs the same local linearization around the unstable fixed point as used by OGY, but instead of moving the saddle point to the system state point, this method moves the system state point back to the saddle by giving the state point small nudges equal in magnitude but opposite in sign to those yielded by eqs. 1 and 2. To calculate these nudges it was necessary to use the observed chaotic attractor to predict the time of the next heartbeat. Then an electrical stimulus was used to induce a heartbeat before the predicted natural beat could occur, thereby shortening the next interbeat interval. The amount of time to advance the next beat was calculated so as to place the system state point (Fig. 8b) directly onto the stable direction of the attractor. Thus the succeeding beat would tend to naturally move toward the desired unstable fixed point rather than away from it.

The results are shown in Fig. 9. It was not possible in these early experiments to achieve a good period 1 (i.e., "normal") heartbeat. But it was possible to control the chaos consistently into a period 3 beat, which, while not optimal, is better for pumping blood than chaotic beating. As for the relevance to human heart arrhythmias, there is evidence that atrial and ventricular fibrillation may be examples of chaos. Thus future work may be able to present strategies for dealing with these serious and widespread arrhythmias.

Chemistry

The best known oscillatory chemical reaction, the Belousov-Zhabotinsky reaction, may exhibit either periodic or chaotic signatures, depending on the flow rates of the reactants into the tank. These rates comprise the system parameters to be used for control. The variable that is measured is the voltage of a bromide electrode inserted into the tank. Using the OGY method, Petrov et al.⁸ at West Virginia University were able to stabilize both period 1 and period 2 oscillations in their reactor.

Rollins et al.⁹ at Ohio University have also applied chaos control to a chemical system. They developed a recursive method (based on the Dressler and Nitsche modification of the OGY method) that makes available a wider choice of accessible system parameter for use in controlling the chaos.

Future of Control

Chaos often breeds life, when order breeds habit.

Henry B. Adams,

"The Education of Henry Adams"

Theory

Two fundamental questions govern future chaos control theories. The first is the problem of controlling higher dimensional chaos in physical systems. A start has been made in this direction in a recent paper by Auerbach, Grebogi, Ott and Yorke¹⁷ on controlling chaos in systems of arbitrarily high dimensions. The method can be "implemented directly from

time series data, irrespective of the overall dimension of the phase space." Although applied in numerical studies, this method has yet to be tested experimentally.

The second question has yet to be addressed: the problem of control in a spatio-temporal system. Many fluid dynamical and biological systems may be modeled as a multi-dimensional array of oscillators connected to each other in some complex fashion. Such systems exhibit both spatial as well as temporal chaos. The study of such systems, while of obvious importance for real-world applications, is as yet in its infancy.

Experiment

The importance of chaos control in electronic circuits was mentioned above and should be stressed once again. In a world dominated by electronics, the ability not only to remove chaos where it is not wanted, but also to make more flexible circuitry by exploiting chaos and its control, could have a tremendous impact on our lives. The control of chemical reactions is no less important, since an increase of only a few percent in the efficiency of a commercial chemical reaction could conceivably save millions of dollars. The 15-fold increase in the power at which a laser can be operated stably promises new applications for these modern tools. And the ability to control one of the great diseases of modern times could be a boon beyond reckoning. The challenge for experimentalists now lies not only in achieving control of higher dimensional and spatio-temporal chaos, but also in guiding the achievements of the past three years out of the laboratory and into our lives.

Biographies

Dr. Mark L. Spano graduated *Summa cum Laude* from Saint Joseph's College in Philadelphia with a Bachelor of Science degree in Physics. He proceeded to graduate school at the University of Maryland, where he received his Master of Science degree and subsequently a Doctorate of Philosophy in experimental Physics in 1980. At present he is a research physicist at the White Oak Laboratory of the Naval Surface Warfare Center. Address: Naval Surface Warfare Center, White Oak, Silver Spring, MD 20903, e-mail: mark@chaosnswc.navy.mil.

Dr. William L. Ditto graduated from the University of California at Los Angeles with a Bachelor of Science degree in Physics. In 1988 he received his Ph.D in theoretical Physics (Field Theory) from Clemson University. Upon graduation Dr. Ditto accepted an Office of Naval Technology Postdoctoral Fellowship with the White Oak Laboratory of the Naval Surface Warfare Center. At present he is a professor of Physics at the Georgia Institute of Technology. Address: Georgia Institute of Technology, School of Physics, Atlanta, GA 30332.

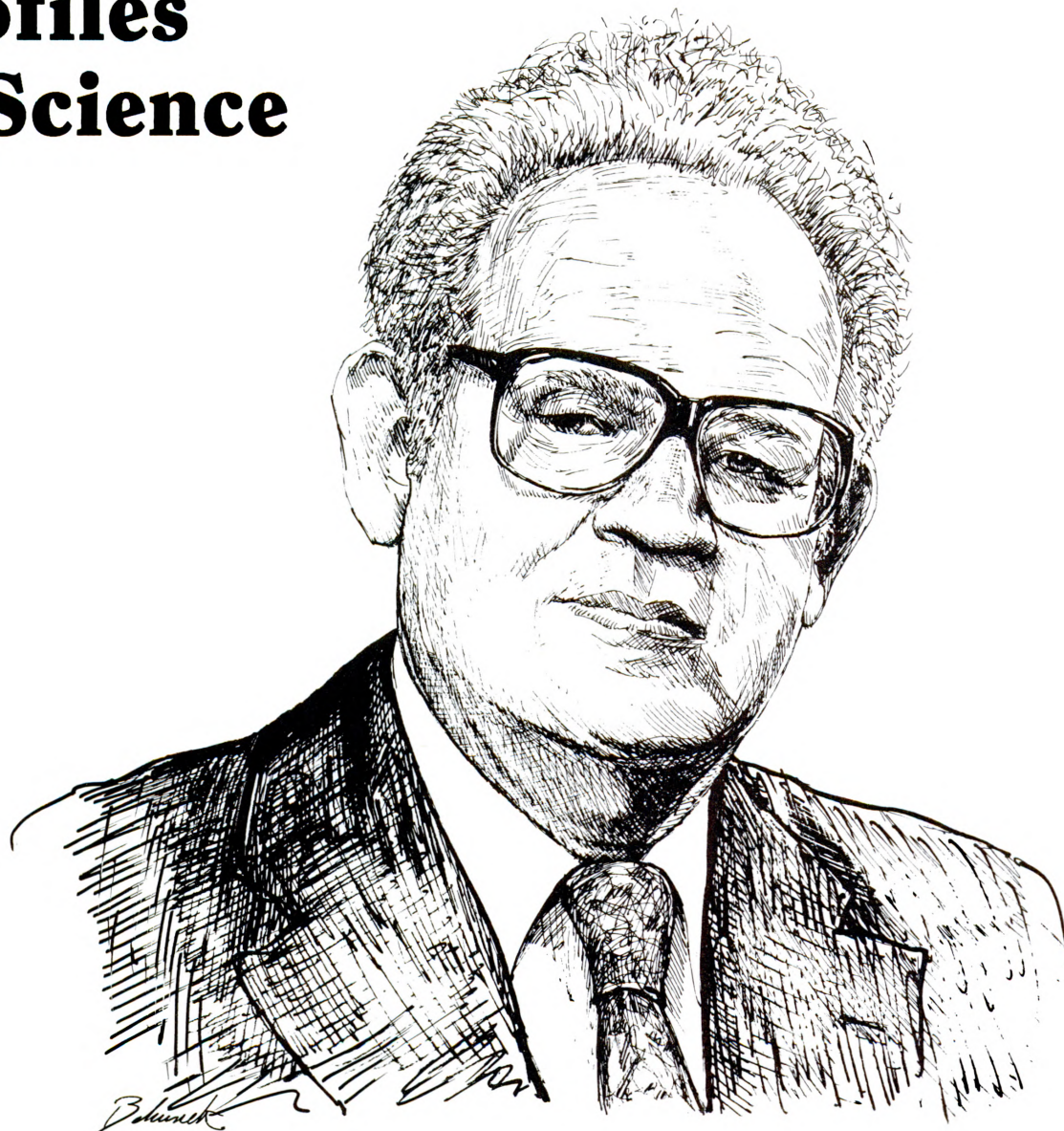
Their work has twice appeared on the cover of *Science News*. Both have been honored with the NSWC Research Department's Outstanding Young Professional of the Year

award, Dr. Spano in 1985 and Dr. Ditto in 1991. Their work on the control of chaos in a magnetoelastic ribbon received the Navy's Independent Research Excellence Award for his FY 91 and they at present have a patent application pending on the Control of Cardiac Chaos.

References

1. James Gleick, *Chaos* (Viking, New York, 1987), p. 65; see also T-Y. Li and J. A. Yorke, *Amer. Math. Monthly* 82, 985 (1975).
2. E. N. Lorenz, *J. Atmos. Sci.* 20, 130 (1963).
3. E. R. Hunt, *Phys. Rev. Lett.* 67, 53 (1991).
4. W. L. Ditto, S. N. Rauseo and M. L. Spano, *Phys. Rev. Lett.* 65, 3211 (1990).
5. Gerald Jay Sussman and Jack Wisdom, *Science* 241, 433 (1988).
6. Rajarshi Roy, T. W. Murphy, Jr., T. D. Maier and Z. Gills, *Phys. Rev. Lett.* 68, 1259 (1992).
7. Alan Garfinkel, Mark L. Spano, William L. Ditto and James N. Weiss, *Science* 257, 1230 (1992).
8. Valery Petrov, Vilmos Gaspar, Jonathan Masere and Kenneth Showalter, *Nature* 361, 240 (1993).
9. R. W. Rollins, P. Parmananda and P. Sherard, *Phys. Rev. E* 47, R780 (1993).
10. E. Ott, C. Grebogi and J. A. Yorke, *Phys. Rev. Lett.* 64, 1196 (1990).
11. Actually these are the covariant eigenvectors, e_s and e_u , from which must be derived the contravariant eigenvectors, f_s and f_u . This is because the two directions are, in general, not orthogonal. See Ref. 10 for more detail.
12. B. Peng, V. Petrov and K. Showalter, *J. Chem. Phys.* 95, 4957 (1991).
13. A. Hubler, *Helv. Phys. Acta* 62, 291 (1989).
14. However, see Ute Dressler and Gregor Nitsche, *Phys. Rev. Lett.* 68, 1 (1992).
15. T. Shinbrot, E. Ott, C. Grebogi and J. A. Yorke, *Phys. Rev. Lett.* 65, 3215 (1990); T. Shinbrot, E. Ott, C. Grebogi and J. A. Yorke, *Phys. Rev. A* 45, 4165 (1992).
16. Troy Shinbrot, William Ditto, Celso Grebogi, Edward Ott, Mark Spano and James A. Yorke, *Phys. Rev. Lett.* 68, 2863 (1992).
17. Ditzia Auerbach, Celso Grebogi, Edward Ott and James A. Yorke, *Phys. Rev. Lett.* 69, 3479 (1992).

Profiles in Science



B. B. Mandelbrot

Benoit B. Mandelbrot was born in Poland in 1924 and later moved to France, where his life and education were torn by World War II. These conditions contributed to his being a self-taught and independent thinker. Over the last 40 years he has developed the field of Fractals, a geometric-based concept, where dimension is a continuous variable. His ideas have become so compelling, that the citation for Dr. Mandelbrot's 1993

Wolf Prize for Physics credits him for changing our view of Nature. Dr. Mandelbrot has spent most of his professional life at IBM. For the past several years he has split his time between IBM and Yale. At Yale, with Office of Naval Research sponsorship, he is delving deeper into research and training a new generation of students in the world of fractals. Dr. Mandelbrot is the recipient of numerous awards and honors and is a member of the National Academy of Sciences.

Original from
UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

Nonlinear Resonance: Noise-Assisted Information Processing in Physical and Neurophysiological Systems

Adi Bulsara
Naval Command, Control & Ocean Surveillance Center

John K. Douglass and Frank Moss
University of Missouri at St. Louis

Introduction

The distinctive hiss of radio “noise” is familiar to all of us old enough to have listened to vintage a.m. radio shows and certainly to all those personnel who ever tried to “pull in” a weak signal during a tour in military communications. Indeed, virtually since the birth of the telephone and radio and extending to today’s sophisticated telecommunications designs, engineers have devoted prodigious efforts to eliminating or minimizing the effects of noise. In consequence, an entire discipline, known as *linear* filter theory, has evolved and is familiar to every electrical engineering and/or communications student, usually as one or a set of rather rigorous courses. By contrast, this paper is concerned with a *nonlinear* filtering process known as *Stochastic Resonance* (SR), research on which has been supported by the Physics Division of the Office of Naval Research (ONR) for several years and more recently by the Cognitive and Neural Sciences Division of the ONR.

To those schooled in linear doctrine, filtering with SR begins with a most radical premise: that the noise, which may either be inherent or generated externally, can be used to

enhance the flow of information through certain, carefully designed, nonlinear systems. SR has now been demonstrated in a variety of physical experiments ranging from ring lasers through various solid state devices including SQUIDs (superconducting quantum interference devices) to noise-driven chaotic attractors. Moreover, there is accumulating, though quite preliminary, evidence that nervous systems may have evolved in such a way as to make use of SR in the transmission and possibly also the processing of sensory information. Should this notion withstand the rigors of detailed scientific scrutiny, it would constitute the *discovery* of SR as a natural process as opposed to the demonstration of the phenomenon in experiments designed by man. This possibility is quite exciting, since it promises to teach us something fundamentally new. Moreover, it is not difficult to imagine what such a development would imply for the many signal processing applications of importance to the Navy. Readers who wish to learn more about this new field of study may consult the proceedings of a recent international conference devoted to SR and jointly sponsored by the ONR, NCCOSC-RDT&E Division and by NATO¹.

The idea of SR is not difficult to understand. The basic requirement is for a bistable system (though other configurations are also admissible) which is most easily imagined by picturing an energy potential with two wells separated by a barrier such as the one shown in Fig. 1. Suppose that the ball shown near the bottom of the right hand well represents the state point of the system. Information is transmitted through the system in the form of switching events between the wells. The output is the coordinate $x(t)$, and our interest is only in whether $x(t) > 0$ or $x(t) < 0$ (that is, whether the ball is in the right or left well). Imagine that a weak, time-periodic input signal, for example a sinusoidal wave, is applied in such a way as to rock the potential as shown by the three images in Fig. 1, alternately raising and lowering each well in relation to the barrier. The signal is assumed to be "weak" in the sense that its amplitude is never sufficient by itself to cause the state point to switch wells. Thus, the signal by itself cannot cause information to flow through the system. Now consider that noise, in the form of a random, time dependent, horizontal force, is applied to the ball. If the noise is distributed like, for example, a Gaussian, there will always be a non zero probability for a switching event to take place. Of course, the probability of a switching event becomes greatest when the state point is in the elevated well, so that it "sees" a barrier reduced in height separating it from the lower, or more stable, well. This condition occurs at the time when the signal is at a maximum, so that the probability of a transition is coherent with the input signal. In consequence, the sequence of switching events, while largely random, will also be coherent to some degree with the input signal. Information about the input signal is, therefore, manifest in the switching events which themselves occur only by virtue of the presence of the noise.

The statistical physics governing such processes (however, without external signals) was put forth over 50 years ago by H.A. Kramers². The rate at which switching events take place, we call it the "Kramers rate", is given by

$$\omega_K = C \exp \left\{ -\frac{\Delta U_0}{D} \right\}, \quad (1)$$

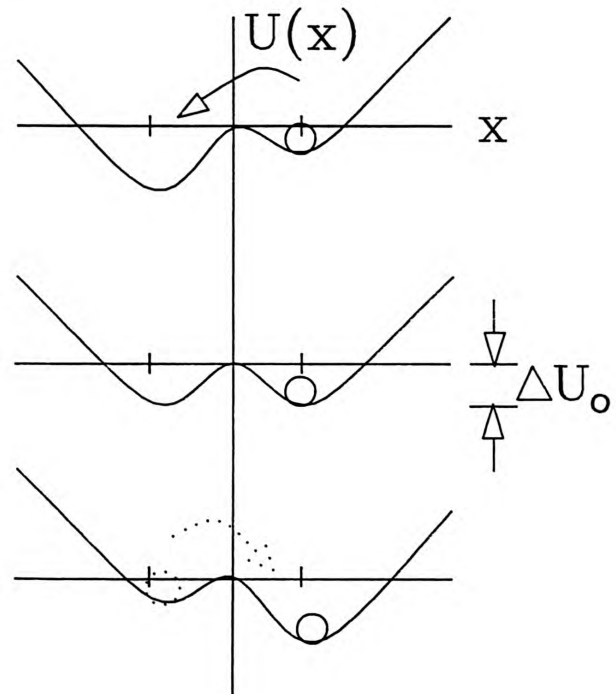
where ΔU_0 is the barrier height, D is the noise intensity (the variance) and C is a constant which depends on the detailed shape of the potential and is of no interest to us here. The key point, and the single most important property of this formulation from a signal processing point of view, is that the Kramers rate is an *exponential* function of the instantaneous barrier height. This causes the rate to be extremely sensitive to small changes in the barrier height, and to transmit those changes to the output in the sequence of switching events. In the presence of the weak sinusoidal signal,

$$\Delta U_0(t) \rightarrow \Delta U_0 + f(\delta \sin \omega t), \quad (2)$$

and the Kramers rate becomes periodic, $\omega_K \rightarrow \omega_K(t)$. Thus, we can easily understand the exponential sensitivity of $\omega_K(t)$

Figure 1.

A representation of a bistable potential modulated by a cosine function of phase 0 radians, top; $\pi/2$, middle and π , bottom. The unmodulated barrier height is ΔU_0 as shown by the middle view.



on the signal strength δ . The switching events can be thought of as sampling the input waveform at more-or-less random times. Obviously, in order for the output to convey a maximum amount of information about the input signal, the "sampling rate" must be sufficiently large. This means that the signal frequency should be small compared to the Kramers rate, $\omega \ll \omega_K$, a condition called the "adiabatic approximation" which is characteristic of most current statistical theories of SR. A modern theory of SR was first put forth by McNamara and Wiesenfeld³. Indeed, recent research in this laboratory (J.D. & F.M.) has shown that, just as is the case with digital signal processing, a Nyquist frequency, $\omega_N \equiv (1/2) \omega_K$, can be identified, which plays the role of a cutoff frequency in SR, just as it does when the waveform is periodically sampled in the more familiar way.

The output is thus a sequence of switching events between, for example, +1 and -1, as shown in Fig. 2(a). [We completely neglect the intra well motion and replace the raw output by a sequence of ± 1 's denoting the well in which the ball is instantaneously confined.] The sequence of time intervals, $T_0, T_1 \dots T_i$, identify the time intervals between like

switching events, say from the -1 to the +1 well. Two objects can be assembled and averaged from such a time series, and from each a signal-to-noise (SNR) can be extracted. The first is the residence time probability density, $P(T)$, which is a histogram of the residence times, T_i . An example histogram is shown in Fig. 2(b). Note the exponentially decaying sequence of peaks at all integer multiples of the signal period⁴⁻⁶. An SNR can be defined by counting the number of events in some narrow range centered on each peak, summing these for all peaks, and counting and summing all counts in the same range centered on each minimum between peaks. The SNR is formed from the ratio of the peak count to the valley count. The second object is more familiar. It is the power spectrum, an example of which, obtained from analog simulation, is shown in Fig. 2(c). This power spectrum shows a sharp signal feature at the signal frequency (and other much weaker ones at the odd harmonics) riding on a Lorentzian background due to the noise. The strength, S ($\equiv$ the area under the peak) is obtained by locally integrating the neighborhood including the peak. The strength represents that fraction of the total number of switching events (or neural spikes) which are coherent in frequency with the signal (or stimulus). The incoherent fraction shows up in this spectrum as the fluctuating, wide band, Lorentzian background curve. The noise intensity, N , is given by the amplitude of this background at the base of the fundamental signal feature, i.e. the value of the noise background at the signal frequency. The SNR is then defined in the standard way familiar from communications engineering: $SNR = 10 \log_{10} \left\{ \frac{S}{N} \right\}$ in decibels (dB).

These two SNR's, of course, depend upon, and change with, the external noise intensity D . The essence of SR is the remarkable property that both SNR's attain maxima at specific values of D , that is, *there exists an optimal noise intensity which maximizes the transmission of information through the system*. An example of this behavior, which happens to have been obtained from an analog simulation of a simple single neuron model, is shown in Fig. 3.

Though the current applications of SR are very exciting, and consequently have stimulated a vigorous research enterprise currently in progress at various institutes scattered around the globe, the concept of SR is not new. We therefore begin this paper with a brief historical background. Though SR has been demonstrated in a variety of experimental settings, including the sensory, or peripheral, nervous system as indicated below, it has not been experimentally observed in collections of coupled neurons as are found in the central nervous system. In order to explore the feasibility of this possibility, we have included a section devoted to a unique theoretical approach to modeling the dendritic system with a set of coupled bistable systems. In this model, the important and interesting case of nonlinear coupling is treated for the first time. This is followed by a section which describes an experimental search for SR in crayfish mechanosensory receptors. In addition, we discuss the results of an experiment

designed to demonstrate SR in a bistable SQUID. We conclude with a summary of our research and discuss a number of speculative questions.

Historical Background

The mechanism of SR was first propounded by Vulpiani and his co workers⁷ as an interesting stochastic effect in nonlinear dynamics which might have useful applications in a variety of fields. C. Nicolis⁸, independently, and the aforementioned authors, together with Parisi⁹, proposed SR, in 1982, as a possible explanation of the observed periodicities in the recurrences of the earth's ice ages. In this view, the earth's climate is represented by a one dimensional bistable potential, one (meta) stable state of which represents a largely ice covered earth¹⁰. The external noise is assumed to come from short term fluctuations in the balance between radiative and convective transport processes, and the periodic modulation is most often supposed to originate from variations in the insolation resulting from a small observed oscillation in the eccentricity of the earth's orbit having a period of 100,000 years¹¹.

In 1983, Fauve and Heslot made detailed measurements on a noise driven, periodically modulated, bistable electronic system (a Schmidt trigger)¹². They measured the power spectrum of the output from which they extracted the SNR, and observed that this quantity passed through a maximum with increasing noise intensity, thus demonstrating SR for the first time in a laboratory experiment. The location of the maximum in the SNR was identified (roughly) with the specific value of the noise intensity for which the Kramers rate in the unperturbed potential becomes comparable to the modulation frequency. No theory was put forth by these authors. Instead, their experiment served to clearly demonstrate the observable, physical aspects of SR.

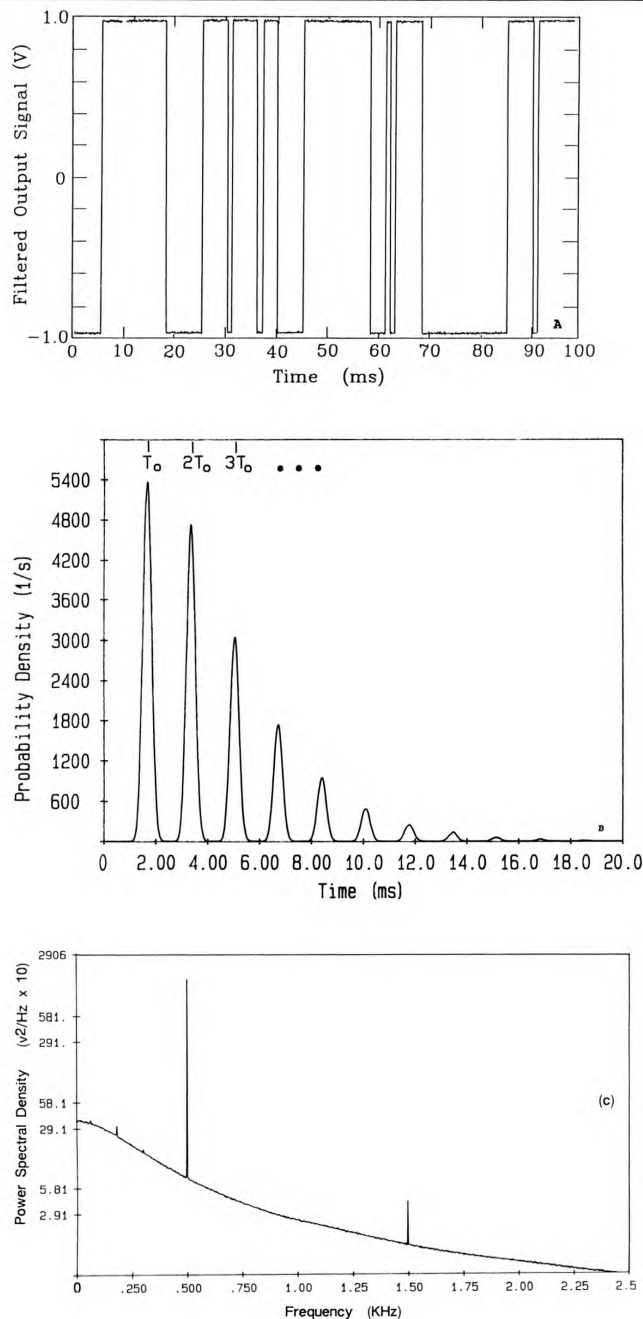
Interest in SR seems then to have waned until 1988 when McNamara, Wiesenfeld and Roy demonstrated it in an ingenious experiment with a ring laser¹³. This experiment immediately stimulated a rash of theoretical activity^{3,14-19} as well as a number of digital and analog simulations¹⁸⁻²¹.

Bifurcation and Stochastic Resonance in Coupled Neuro-Dendritic Processes (A. Bulsara)

Cooperative effects arising from the interaction of large numbers of nonlinear dynamical systems with *nonlinear* coupling, are of considerable importance in neural information processing systems. We can create a richer and more neurophysiologically realistic model of neural activity in the brain by developing a model of neural-dendritic coupling which expressly accounts for the way in which the many afferent

Figure 2.

Results obtained from analog simulations of typical bistable systems driven by weak sinusoidal signals buried in a Gaussian, quasi white noise, here displayed for illustrative purposes. (a) A time series showing noise driven switching between the two states represented by ± 1 . (b) A residence time probability density showing successive peaks at integer multiples of the period of the modulation, in this example $T_0 = 1.66$ ms. (c) The power spectrum obtained from a time series similar to the one shown in (a) for a modulation frequency $f_0 = 500$ Hz. Note the sharp signal features at f_0 and at $3f_0$. (The delta function-like features and the amplification of only the odd harmonics are signatures of SR.) The noise background curve is accurately a Lorentzian, $\frac{Q}{1 + \omega^2 \tau^2}$, where Q is a constant, and $\tau = \frac{1}{\omega_K}$.



connections into the cell body influence the somatic membrane potential. Such a model would begin to fill the need within the Artificial Neural Network Community for neural models which go beyond current “weighted sum” paradigm for artificial neuron connectivity. Such models also have use in engineering applications. Moving from simple “axonal” connection weight neural models to neural-dendritic models with a richer structure will allow investigation both of events at the neural level (e.g. interspike interval histograms and stochastic resonance) and also potentially at the neural systems level. This will also introduce the possibility of introducing cross-scale interactions into artificial neural systems.

In dealing with neurophysiologically realistic networks, one of the most interesting problems is the modeling of the relationship between the input to a neuron via its many afferent connections, and the corresponding output via action potentials generated at the axon hillock. Adequate models must take into account both the nature and the temporal properties of the input and output phenomena. Most models of interconnected neurons treat the input to the neuron as being related, in a functionally simple manner, to the output of the connecting neuron. Such “artificial” neural network models²²⁻²⁵ assume that the “connectivity” of neuron i to neuron j may be modeled by multiplying the axonal output activity of neuron i by a “connection weight”. Such simple models do not, however, accurately depict neural connectivity, which is mediated by a “feltwork” of dendro-dendritic synapses. Recent simplified models have made it possible to explore some interesting physics of coupled elements and to develop artificial neural networks with useful properties, however, more neurophysiologically realistic models require consideration of the time-dependent neural spiking activity as well as the influence of the surrounding dendro-dendritic “bath”.

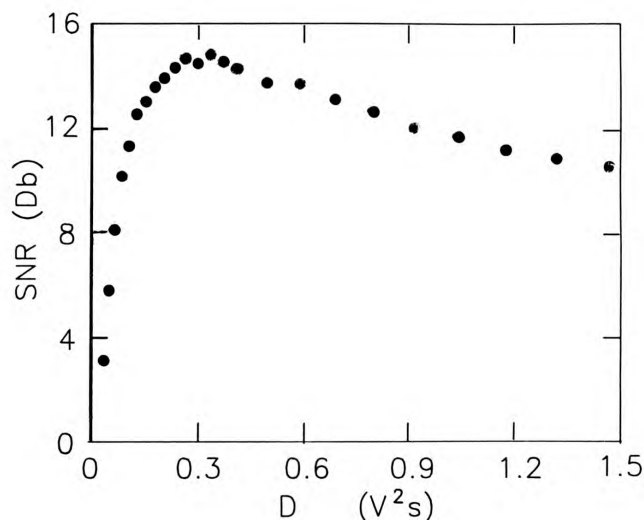
Recently, we have considered²⁵ the influence of a large number ($O(N^3)$) of dendritic connections on the dynamics of a single neuron. We include *a priori* self-coupling terms as well as a weak periodic external modulation. In addition, we assume the presence of background noise in each element i of the N -body system under consideration. We model the dynamics of both the neuron and a large number of small non-overlapping volume elements of “dendro-dendritic space” near the reference neuron as

$$C_i \dot{u}_i = \sum_{j=1}^N J_{ij} \tanh(u_j) - \frac{u_i}{R_i} + F_i(t) + q \sin(\omega t). \quad (3)$$

An equation of this form describes a set of N nonlinear coupled bistable oscillators (the coupling is also nonlinear). The $i=1$ index is taken to denote the cell body and the indices $i=2 \dots N$ (where N is large) denote the dendritic spaces. u_i represents the activation (i.e. the somatic membrane potential) at the trigger zone of the reference neuron. $R_1 C_1$ is the activation decay time due to repeated firings of action potentials

Figure 3.

A typical SR curve showing the maximum in SNR at an optimal noise intensity, in this case $\approx 0.3 \text{ V}^2 \cdot \text{s}$. The SNR values were obtained from the power spectra as discussed in the text. This example was obtained from a bistable single neuron model of the type treated here as a network. See Ref. 27.



down the axon. Alternately, we may think of C_1 and R_1 as the input capacitance and trans-membrane resistance, respectively. For both the neuron and dendritic spaces, $F_i(t)$ is taken to be Gaussian, delta-correlated (that is “white”) noise having zero mean and variance σ_i^2 (the noise sources for different indices i are uncorrelated). J_{1j} represents the coupling between the cell body and the dendritic spaces, and $J_{ij}(i,j > 1)$ represents the (local) coupling between the dendritic spaces. We allow the matrix $\mathbf{J}$ to be asymmetric with the excitatory or inhibitory nature of the couplings determined by the sign of the elements J_{ij} . The synaptic activity in the i^{th} dendritic volume is represented by u_i with ($i > 1$). We assume that the input into the cell as well as each dendritic volume is bounded and can be represented by a weighted sigmoid function. By modeling the interaction of the neuron with the tessellation of dendritic volume spaces in the manner described above, we create a more neurophysiologically realistic model.

In order to extract the dynamics of the neuron (described by the activation u_i) from the coupled system (3), we invoke the *ansatz* that the time scale for neural activation at the soma, is much longer than that for dendritic activity.

$$C_1 R_1 < C_i R_i, \quad (i > 1) \quad (4)$$

This assumption, justified in references 25, constitutes the cornerstone of our theory. However, while it is quite simple to

electronically implement a network subject to the constraint (4), one cannot apply this assumption to all classes of neurons. Networks of neurons and dendrites having time constraints that violate (4) must be treated via more general quasi-linear mean field theories. For the case when (4) is satisfied, however, we can²⁵ adiabatically eliminate the bath variables from (3) and write down an equation for the variable u_1 in terms of the bath variables. Specifically, an N-body Fokker Planck Equation (FPE) is constructed from (3). Haken's slaving principle²⁶ then permits us to factorize the N-body probability density function describing the system (3) into a product of a single-body density function for the slow variable u_1 and a conditional density function for the remainder of the system. This, in turn, leads to a separation of the N-body FPE into a FPE for the probability density function of u_1 (which contains the bath variables u_{i1}) and a FPE for the bath variables. The latter is solved in the long-time limit, after invoking a local equilibrium assumption for the bath variables (this is tantamount to a quasi-linearization of the bath dynamics). We are ultimately lead to a closed FPE for the slow variable u_1 , whence a stochastic differential equation may be readily written down by inspection:

$$\dot{u}_1 = -\alpha u_1 + \beta \tanh(u_1) + \delta \sin(\omega t) + \sqrt{\sigma_e^2} F(t), \quad (5)$$

where the coefficients α , β , δ and σ_e^2 are now functions of the dendritic bath parameters and the coupling matrix J with $F(t)$ being the noise. In carrying out the procedure leading to (5) we have assumed further that the modulation frequency is smaller than the Kramers frequency of the unmodulated system. This assumption is a cornerstone of the adiabatic theory of stochastic resonance³ on which our subsequent results are based. Further, we assume that

$$\sigma_i^2 R_{i-2} C_{i-1} (i > 1) \quad (6)$$

This assumption guarantees the convergence of the steepest descent techniques used to obtain the reduced dynamics (5) and places an upper limit on the noise strengths (with very large amounts of noise, the interesting cooperative behavior is lost). Note also the absence of terms involving coupling between pairs of bath oscillators; these terms are $O(R_i R_j)$ ($i, j > 1$) or higher and are assumed to be negligible.

The bifurcation properties of the reduced system (5) may be studied (in the absence of the noise and modulation terms) via the potential function,

$$U(u_1) = \frac{\alpha}{2} u_1^2 - \beta \ln \cosh(u_1). \quad (7)$$

For positive α and β , the potential is bimodal (for $\beta/\alpha > 1$) with minima located at $c \approx (\beta/\alpha) \tanh(\beta/\alpha)$. The transition to mimodality is accompanied by a pitchfork bifurcation in the most probable value of the activation u_1 with the two states

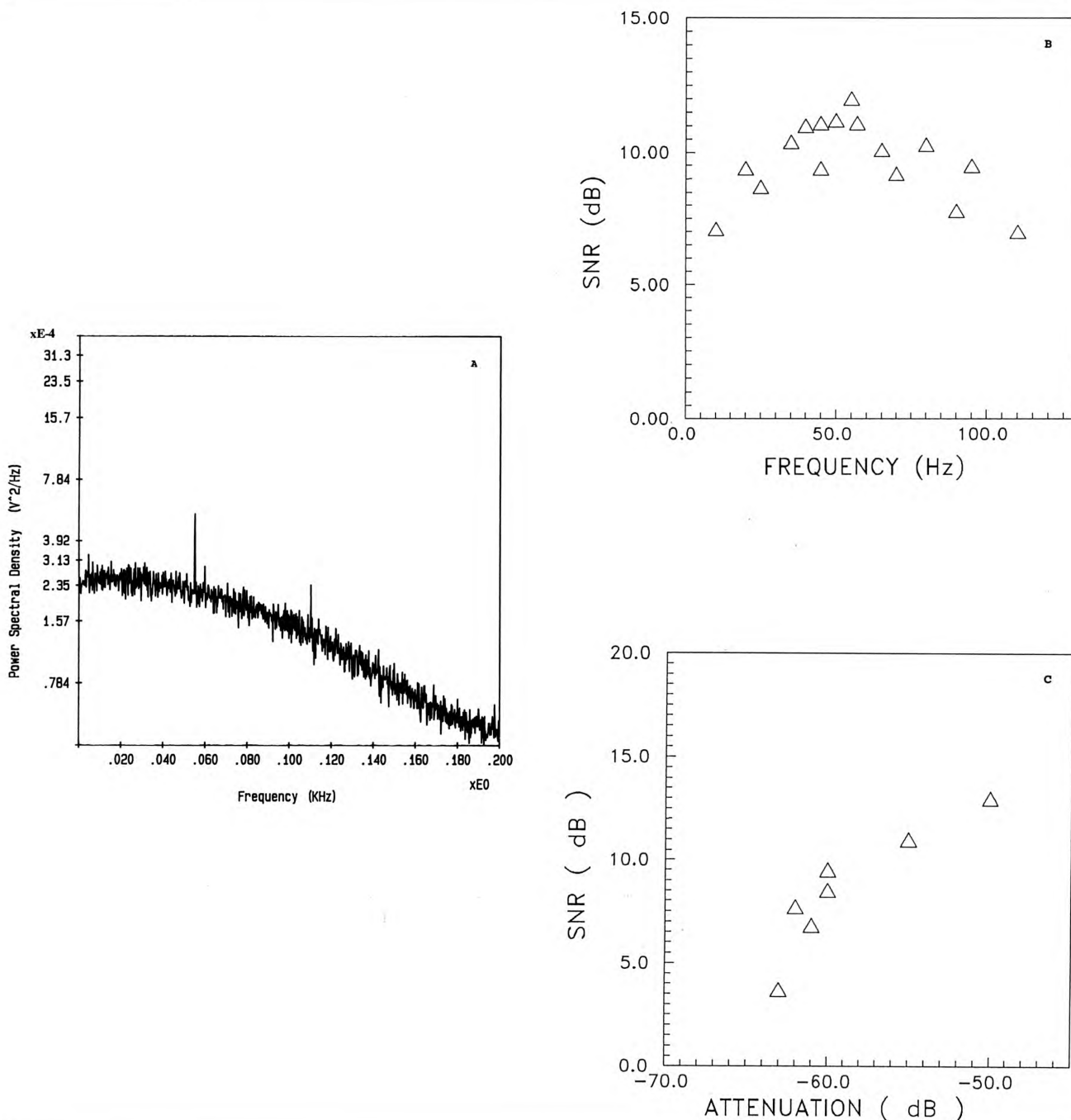
(attractors) corresponding roughly to the quiescent and firing states of the neuron. The flow (given by the derivative of the first term on the rhs of (7)) exhibits the characteristic N-shaped relationship known to exist in excitable cells over certain parameter ranges. The *magnitudes* as well as the *signs* of the elements of J govern (via the coefficients α and β) the transition to bimodality as well as the occurrence and magnitude of other cooperative effects (e.g. stochastic resonance) in the reduced system. Further, it has been found that for the case in which the isolated cell body admits of a monostable potential, the coupling to the dendritic bath can actually induce bimodality. The opposite effect can also occur: a potential that is bistable in the absence of the dendritic bath can be rendered monostable by the coupling to the bath.

We now consider the effects of the deterministic modulation, which leads to stochastic resonance wherein a small amount of noise can introduce correlated switching events in the dynamics (5), corresponding to the system described by the effective potential (7). We define the deterministic switching threshold as the critical value δ_c of the scaled modulation amplitude δ , above which one would obtain deterministic switching in the $\sigma_1^2 = 0 = \sigma_2^2$ case. Then, in order to satisfactorily explain stochastic resonance using the adiabatic theory of McNamara-Wiesenfeld³, we must ensure that $\delta < \delta_c$ and $\omega < \omega_K$, the Kramers rate for the unmodulated system. The adiabatic conditions can be satisfied in the reduced dynamics (5), if we ensure that there is no deterministic switching in the isolated ($J_{ij}=0$) case, and we operate within the realm of validity (defined by (6)) of the theory.

The signal-to-noise ratio (SNR) corresponding to the membrane potential variable u_1 has been calculated^{25,27} using the adiabatic theory and plotted as a function of the noise strength α^2 . The curve is the bell-shaped curve that characterizes stochastic resonance, with the maximum occurring at a critical noise value equal to the height of the barrier of the effective potential (5) (this maximum corresponds to a matching of the modulation frequency ω to twice the Kramers rate ω_K , hence the description of the effect as a "resonance"). For a given noise strength, the SNR is found to be enhanced over the value it would take in the absence of any coupling ($J_{ij}=0$) to the dendritic bath, with the maximum enhancement occurring for the case of a non-noisy bath. The important result is that the coupling to the bath enhances the SNR *even in the presence of noise* (recall that the inequality (6) imposes an upper limit on the noise). Similar effects have been observed recently in a mean field model of linearly coupled bistable oscillators²⁸. The enhancement of the SNR through the dendritic coupling is readily explained²⁵ on the basis of the renormalization of the potential barrier height and the modulation amplitude in the effective equation (5), by the coupling to the dendritic bath. Plotting the peak SNR vs. the bath resistance R_2 (we assume that the different dendritic volume elements have the same resistances $R_{i>1} \equiv R_2$ and the same

Figure 4.

(a) A power spectrum obtained from a crayfish mechanoreceptor cell with the associated hair stimulated at 55Hz. Compare with Fig. 2 (c), and note the similarities: delta function-like features and Lorentzian-like noise background; and the difference: a strong 2nd harmonic. More averaging, i.e. assembling the power spectrum from a longer time series, would result in reduced fluctuations on the noise background. (b) An example tuning curve used to obtain the best frequency of the cell, in this example at ≈ 55 Hz. (c) Threshold data. The horizontal axis is the relative attenuation off the stimulus signal. The threshold attenuation is where the SNR would cross zero. In practice we could measure down to the point where the signal feature, as shown in (a), became indistinguishable from the fluctuations on the noise background.



noise strengths $\sigma_{i>1}^2 \equiv \sigma_2^2$) yields a curve peaked at $R_2 \approx \sigma_2^2$. This provides a tool for achieving optimum SNRs in electronic analogs of the neuro-dendritic interaction.

Experiments on Single Sensory Neurons. (John K. Douglass and Frank Moss)

Introduction

Though noise assisted information transmission and stochastic resonance have been demonstrated in a variety of physical experiments^{1,29}, it would be of wide interest to discover a natural system which exhibits or exploits these processes. Certainly, the phenomenon of noise induced switching among two or several stable or metastable states, in the presence of a weak, information carrying function, is simple enough that one might expect to find it in a number of diverse natural dynamical situations, the onset of ice age climates coherent with a weak periodic orbital feature^{8,9} being only one possibility. A candidate system must have an inherent or external (natural) source of noise, and it could either be characterized by a fixed, multistable potential, as has been the case for all of the demonstrated experiments to date, or it could be a limit cycle or relaxation oscillator, where the noise induced switching is between a fixed (time-independent) point and the limit cycle or relaxation oscillation. In order for a system to exploit SR, for example, in order to enhance some sort of information transfer process an additional, and perhaps much rarer, ingredient is necessary: some type of self optimization process which would adjust either the noise intensity or the parameters of the potential in such a way that the coherence between the information carrying function and the response of the system is maximized. In other words, the system should be capable of adjusting itself so that it operates at the maximum signal-to-noise (SNR) as shown, for example, in Fig. 3.

These requirements point to biological systems, especially systems of neurons, as possible examples of naturally-occurring SR for the following reasons. First, noise is ubiquitous in nervous systems, and a paramount function of nervous systems is to detect very weak signals from the external world. Second, living creatures excel at self optimization, in the long term, via Darwinian evolution, and on much shorter time scales via physiological mechanisms. Just how much noise is present in biological sensory systems? A recent and growing literature suggests that the noise intensity observed within some sensory systems is larger (often several orders of magnitude larger!) than can be accounted for on the basis of equilibrium statistical mechanics, for example, the Brownian motion of the sensory element and the measured transduction of this motion into electrical impulses in hair cells of the vertebrate inner ear³⁰. Why is this noise so large? Could it be that mechanisms have evolved which exploit SR in order

to optimize the sensory perception of information? This is a heady and tantalizing question indeed. In order to further explore this possibility, we decided to perform experiments on one of the simplest of all sensory systems: single hair mechanoreceptor cells in the tailfan³¹⁻³³ of the crayfish, *Procambarus clarkii*.

In this interesting system, small motions (as small as 30 nanometers) of hairs located on the surface of the crayfish tailfan are transduced by their associated sensory neurons into trains of action potentials (spikes). These spikes then travel along sensory nerves to the caudal ganglion. Moreover, many such neurons show high spontaneous firing rates, indicating substantial sources of internal noise. In our experiments, a piece of the tailfan together with its nerve bundle and ganglions was excised. The nerve bundle and ganglion were exposed surgically, and the appendage was mounted on an electro-mechanical vibrator in a saline solution. The preparation, which typically persisted in good physiological condition for 8 to 12 hours, was moved by the vibrator thus stimulating the hair with a sinusoidal motion relative to the solution. An extracellular, suction pipette electrode was attached to the nerve bundle and moved with micromanipulators until an appropriate neuron was identified through its spontaneous spiking activity. The electrode was attached to a battery operated preamplifier. This apparatus was mounted on a commercial vibration isolation table and operated inside a Faraday cage. The preamplifier was connected to standard extracellular electronics located off the table and outside the Faraday cage. The outputs of the electronics (spike times and signal traces) were digitized and analyzed by computer. The following quantities were measured, either on line or from later analysis of the time series: power spectrum (PS), interspike interval histogram (ISIH), cycle histogram (CH), spike interval return map (SIM), and the spike phase return map (SPM). These quantities fall into two groups: those which are coherent with the stimulus frequency (PS, ISIH and SIM), and those which are coherent with the stimulus phase (CH and SPM). From these two groups, two SNR's were defined: one based on frequency coherence using the PS and ISIH and the other based on phase coherence using the vector strength (VS) obtained from the CH. Stimulus frequency and intensity could be adjusted as well as the temperature of the saline solution.

Experiments were performed in one of two general modes depending on the character of the stimulus and the source of the noise. In the first mode, the noise source was the neuron's internal (spontaneous) noise, and the stimulus was a pure sinusoidal motion. The noise was controlled by varying the temperature of the preparation, a technique already developed for similar experiments on the frog auditory system by P. Narins and his colleagues³⁴. For these experiments, we chose a neuron with a relatively high spontaneous spike rate, and measured SNR's as a function of the temperature, keeping the stimulus intensity and frequency constant. In the second type

Figure 5.

A stochastic resonance experiment on a crayfish mechanoreceptor cell for an individual acclimated at high temperatures (circles) and another individual acclimated at low temperature (triangles). (a) The SNR versus the temperature, which was used to change the internal noise of the receptor cell. (b) the SNR versus the noise background measured from the power spectrum at the stimulus frequency for the same data shown in (a). Note the absence of a maximum in the SNR. (c) The noise versus temperature calibration. Note the sharp maximum at $\approx 30^\circ\text{C}$ and the subsequent rapid decline of the noise for the high temperature individual.

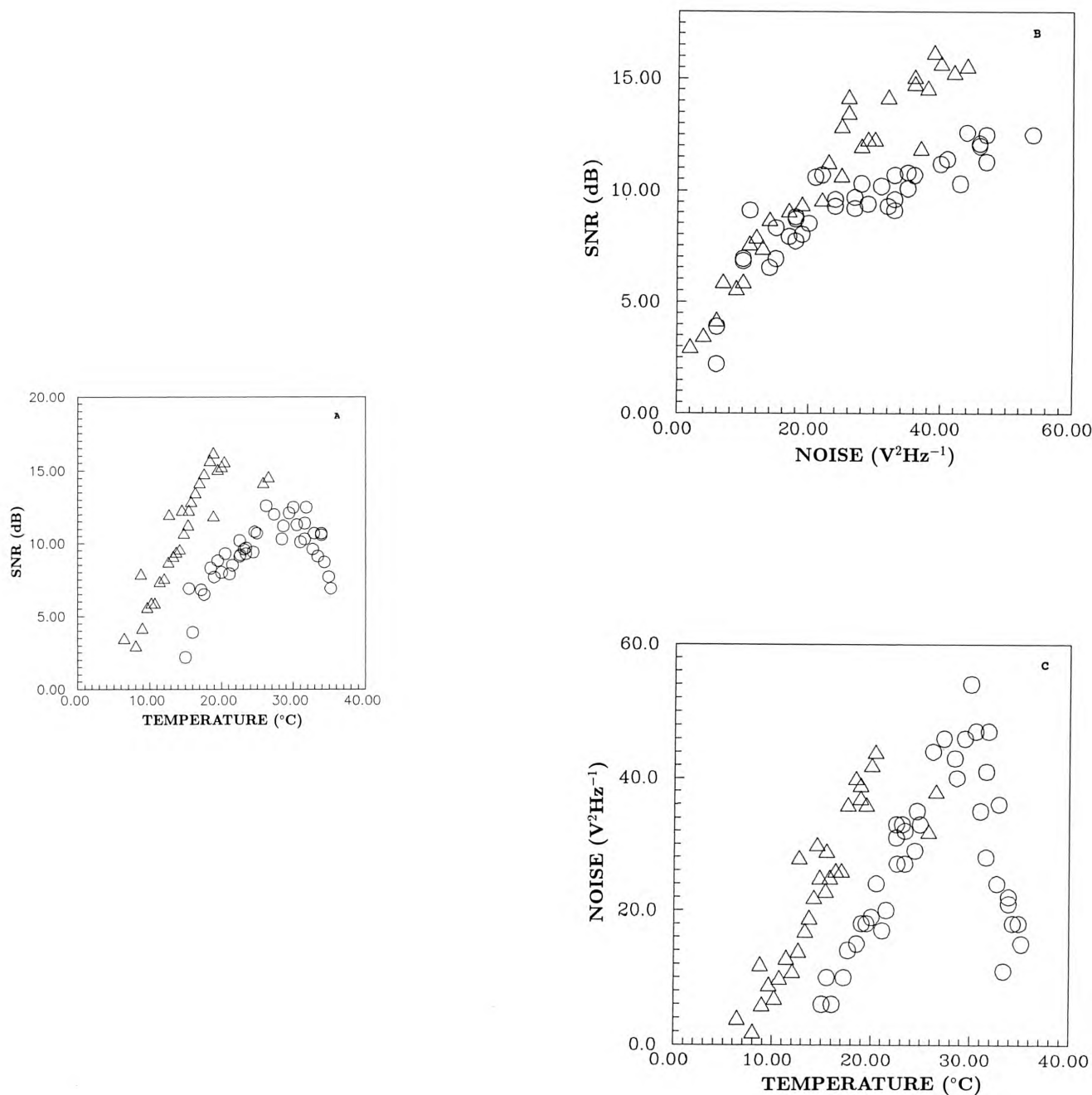
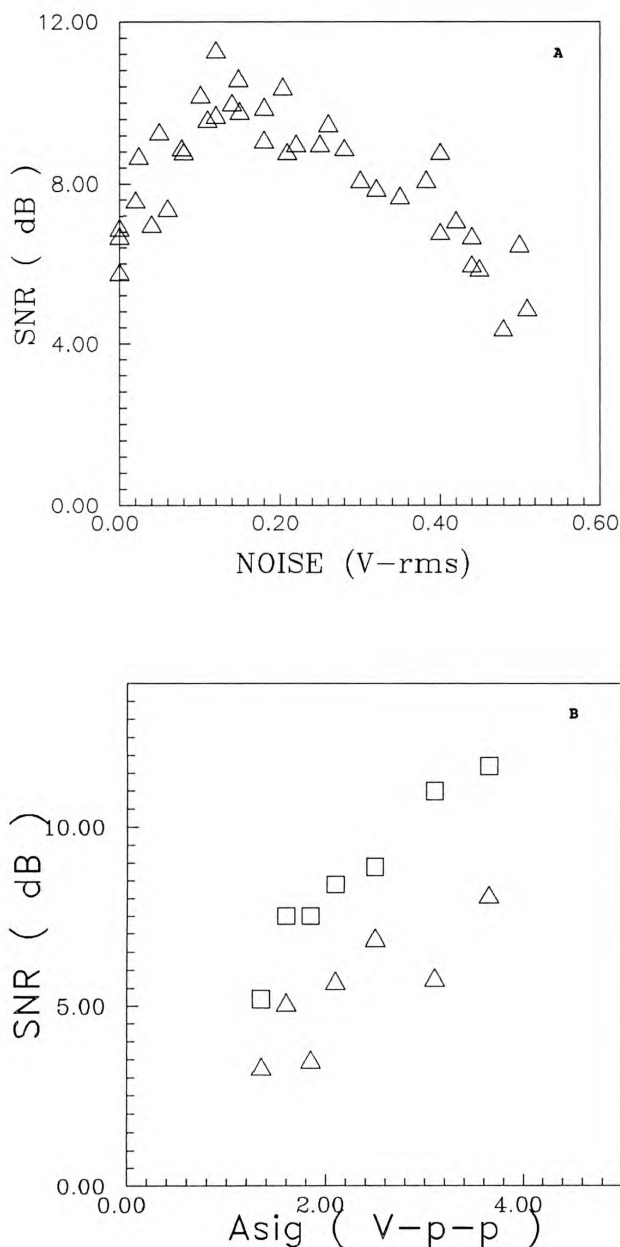


Figure 6.

An SR experiment on a crayfish mechanoreceptor with external noise added to the periodic signal. (a) The SNR versus the external noise voltage. (b) The SNR versus the periodic signal amplitude: with the optimal noise (0.14 V-rms)(squares), and without added noise (triangles).



of experiment, the noise source was external. A neuron with a relatively low spontaneous spike rate was chosen, and wide-bandwidth, Gaussian noise was added to the periodic signal to form the stimulus. The tailfan appendage was therefore moved through the solution more-or-less randomly, but with a weak periodic coherence. The SNR's were measured as a function of external noise intensity, with a constant periodic signal and a constant (room) temperature.

All experiments were begun by first locating the individual hair to whose cell the electrode was attached and then measuring its directional sensitivity³⁵. All subsequent measurements were performed in the stimulus direction corresponding to the maximum response. A tuning curve was then obtained by measuring the previously described VS's and the SNR's as a function of stimulus frequency (keeping the hair-fluid relative *velocity* constant). All subsequent measurements were obtained with the stimulus tuned to the characteristic or "best" frequency (the frequency of maximum response). The last set of measurements in the experimental preface provided the threshold curve, that is, the SNR and VS as a function of stimulus intensity. Subsequent measurements were then usually performed with the stimulus just above threshold (usually about 10 to 20 percent above the minimum detectable signal in the power spectrum) though in some of the experiments with external noise, the SNR's and VS's were measured with and without the (external) noise applied while varying the periodic stimulus intensity.

An example measured power spectrum is shown in Fig. 4(a). Note that the fundamental signal feature is a sharp delta-like peak at 55 Hz which was the stimulus frequency for this example. The SNR is obtained from the spectrum as described above. This particular spectrum was obtained from a spike train measured over an approximately 5 min. period and was representative of those used to measure all SNR's reported below. The tuning curve for this neuron is shown in Fig. 4 (b). The threshold was defined as the value of stimulus intensity for which the fundamental signal feature became indistinguishable from the fluctuations in the background noise curve. Example threshold data are shown in Fig. 4 (c).

Experiments using the internal neuron noise

In these experiments, the (internal) noise was controlled with the external temperature of the saline solution, which was typically varied between about 10°C and 35°C. Experiments have been performed with 8 separate preparations. Though the frequency and threshold characteristics of the individual neurons varied among the preparations, the temperature data from all these preparations were self consistent. The general features are represented by the data shown in Fig. 5 in which a crayfish mechanoreceptor received a constant intensity periodic stimulus while its temperature was varied. Figure 5 (a) shows the measured SNR, based on the power spectrum,

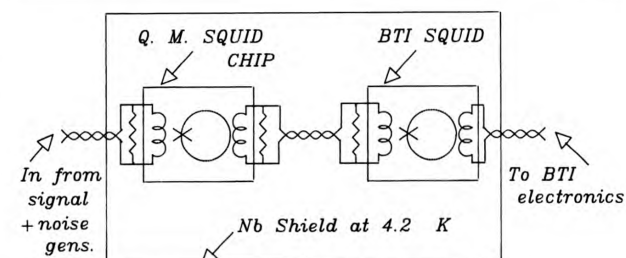
plotted versus the temperature. Consider first the open circles. These were obtained from a crayfish which had been acclimated for many weeks in a rather high temperature environment: ca. 28 to 35°C. The data show a range of SNR values with a maximum of about 12 dB near 28°C. While this result does show a maximum SNR at an optimum temperature, it cannot be interpreted as stochastic resonance. Fig. 5 (b) shows why. Here we have plotted the same data versus the noise background N. The open circles show a monotonic increase in SNR with spike (internal) noise intensity and specifically, no maximum at an optimum noise intensity (which we here assume as the defining characteristic of systems showing SR). Nevertheless, this result is significant. No analysis carried out within the framework of linear transform theory can account for an SNR which *increases* with internal noise intensity. Consequently we must seek an explanation of this result using nonlinear dynamics, and within this framework, the dynamics which accounts for SR seems the most likely to succeed.

Figure 5 (c) shows why there is no maximum with noise intensity. This is the "calibration" of the noise intensity versus the temperature. It is clear that, commencing with the coincident maxima shown by the open circles in (a) and (c), the noise intensity itself drops precipitously with further increases in the temperature. This behavior is unlike that of physical systems showing SR for which the noise intensity continues to increase beyond the maximum in the SNR. We do not have a physiological explanation of this behavior considering that our crayfish can live (were living) at temperatures up to 36°C. We can put forth two speculative explanations, one exotic and the other pedestrian. The exotic speculation is that there exists a self regulatory mechanism which attempts to maximize the SNR for every set of external conditions. In this interpretation, from the low temperatures, the internal noise intensity is continuously increased in order to increase the SNR in accord with the basic SR dynamics. However, when some limited dynamic range for this mechanism is approached, the strategy changes, and beyond the maximum, the noise drops rapidly in a (failing) efforts to maintain the SNR. This strategy works over a very limited range as shown by Fig. 5 (a), where the SNR on the high temperature side of the maximum does decline, but not nearly as precipitously as the noise. The pedestrian explanation is that there is only one signal transmission-noise generating mechanism which operates over the entire temperature range, albeit an essentially nonlinear one, but the nerve simply begins to die at about 28°C. We have noticed that if we exceed about 33°C, the nerve is damaged as evidenced by the fact that the low temperature data are no longer reproducible afterwards.

In an effort to increase the temperature range over which we could observe these effects (and hence possibly observe a maximum in the SNR versus N data), we acclimated a number of crayfish in a constant temperature environment at 9°C for at least six weeks³⁶. We have now performed experiments on

Figure 7.

Schematic of the bistable SQUID experiment showing the experimental SQUID chip supplied by Quantum Magnetics and the BTI measuring SQUID. Noise and signal voltages supplied by the external electronics were transformed into external magnetic flux in the coil C1.



four of the low temperature acclimated fish. Only one of these experiments was successful, in the sense that the data were easily reproducible and repeatable for substantial changes in temperature. In that experiment, the appendage was dissected in cold saline, and we began taking data at the low temperature end of the scale (at $\approx 10^\circ\text{C}$). The data from this experiment are shown by the triangles in Fig. 5. We can see that the qualitative behavior is the same as for the high temperature acclimated crayfish except shifted down in temperature by about 6°C. [Unfortunately, we cannot say whether the maximum in SNR versus temperature was also shifted down because of an apparatus failure during this experiment.] Our research with temperature-acclimated crayfish is still in progress, consequently we can draw no further conclusions at this time.

Experiments using external noise.

For these experiments, neurons with low spontaneous rates (low internal noise) were chosen. Wide bandwidth³⁷, Gaussian noise from a noise generator was then added to the periodic signal to make up the stimulus. In the course of an experiment, the external noise intensity was increased from zero to values considerably larger than the signal intensity (as measured by the rms amplitudes). The motivation for these experiments was that since sensory neurons in general behave like bistable systems operating between fixed point and firing states, and since SR is easily observable in physical bistable systems (and also in fixed point-limit cycle systems such as the FitzHugh-Nagumo model neuron³⁸) when exposed to external noise, then SR might be observable using a sensory neuron as the bistable element. We have performed only four such experiments, hence our results must be still considered as preliminary. Data from one experiment are shown in Fig. 6 (a) where we have plotted the SNR versus the external noise voltage. It is clear that there is an SR peak at a noise voltage of ≈ 0.14 V-rms.

We also measured threshold curves with and without noise for this neuron. The results are shown in Fig. 6 (b), where we have plotted the SNR versus the intensity of the periodic component of the stimulus (the signal). The squares represent measurements in the presence of external noise equal to 0.14 V-rms, which is near the optimal noise intensity as determined from the data in Fig. 6 (a). The triangles are for zero external noise. The squares fall consistently above the triangles, indicating that *we can detect the periodic signal more easily in the presence of external noise*. Of course, this does not mean that a living crayfish in a natural environment can improve the sensitivity of its perceptions by making use of external noise, though these results certainly do suggest that possibility. Careful behavioral studies would be necessary to confirm or refute this hypothesis. A step in this direction will be to make simultaneous recordings from sensory and motor neurons³⁹, experiments which we have planned for the near future.

We were able to find a peak in the SNR in only two experiments. In the other two experiments, for reasons that are unclear, the SNR was maximal at zero external noise and monotonically decreased for all larger values. The most obvious explanation for these results is that if the internal noise is itself large enough that the maximum is already exceeded, the external noise can only decrease the SNR. All of this underscores the mysteries associated with the internal noise but serves to emphasize the importance of its influence on the dynamics of stimulus detection.

Experiments with a Bistable SQUID Loop. (E.W. Jacobs, A. Hibbs, J. Bekkedahl, A. Bulsara and F. Moss.)

We have also demonstrated SR in a bistable SQUID loop, as a first step in stimulating interest in possible applications using superconducting devices. We begin with an equation governing the magnetic flux trapped within an rf-SQUID loop⁴⁰.

$$LC\ddot{\phi} + \tau_L\dot{\phi} + \phi + \frac{1}{2\pi}\beta\sin(2\pi\phi) = \phi_e, \quad (8)$$

where $\phi = \Phi(t)/\Phi_0$ is the normalized magnetic flux trapped within the loop, $\phi_e = \Phi_e(t)/\Phi_0$ is the normalized flux externally imposed on the loop, $\Phi_0 = h/2e$ is the flux quantum, L and C are the loop inductance and junction capacitance respectively, and $\tau_L = L/R_J$ is the junction resistance. The parameter which determines the shape of the potential governing the dynamics of (8) is $\beta = 2\pi L i_c / \Phi_0$, where i_c is the junction critical current. In our experiment, the external flux Φ_e was composed of DC, periodic and stochastic components:

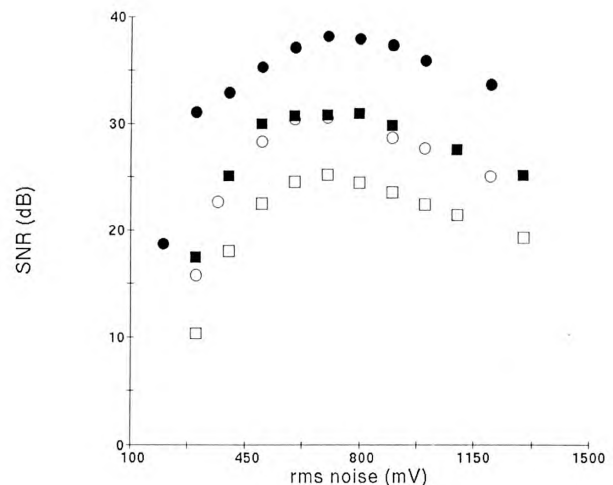
$$\Phi_e(t) = \Phi_{DC} + \Phi_M \sin(\omega_s t) + \Phi_N(t), \quad (9)$$

where the periodic component represents an audio frequency signal, and the stochastic component was a Gaussian noise whose bandwidth was in the audio range⁴¹. Equation (8) is bistable for certain values of β and Φ_{DC} . However, the quantity which shows the bistable dynamics is the flux trapped within the loop, $\phi(t)$. In order to experimentally observe the bistable dynamics, one must measure the trapped flux. This requires a second SQUID, either mounted coaxially with the loop of the first SQUID⁴², or coupled to it with a superconducting transformer. We chose the latter configuration. The primary SQUID was a thin film device mounted on a single chip with integrally mounted, superconducting transformer primaries supplied by Quantum Magnetics⁴³. The secondary, or measuring, SQUID was a standard BTI model⁴⁴. A schematic diagram of the experimental setup is shown in Fig. 7.

In our experiment, $\beta = 2.0$ and $\Phi_{DC} = 0.5\Phi_0$, values which guaranteed that the potential was bistable. Experiments were performed at two signal frequencies, 17.6 Hz and 100 Hz with signal peak voltages of 650 mV-pk and 475 mV-pk respectively. The noise, or stochastic, component was supplied by a standard noise generator and the noise voltage varied over the range from 100 to 1500 mV-rms. (1.0 V was equivalent to $0.1\Phi_0$ of applied external flux.). The power spectra of $\phi(t)$ were measured and averaged in the usual way at the output of the

Figure 8.

The SNR versus rms noise voltage for the bistable SQUID experiment, with $f_s = 17.6$ Hz, $v_s = 650$ mV-rms (filled circles) and $v_s = 237$ mV (open circles); and $f_s = 100$ Hz, $v_s = 475$ mV (filled squares) and $v_s = 237$ mV (open squares). For these data $1.0 V = 0.1\Phi_0$ at the coil C1



BTI SQUID electronics, and the SNR's were determined as described above. The results of this experiment are shown in Fig. 8 where the circles represent the results for the low signal frequency and the squares for the high frequency. At each frequency, data were collected for two different signal strengths. The maxima occur at a noise voltage of ≈ 700 mV which is equivalent to an rms fluctuation of $0.07\Phi_0$ within which a coherent signal equivalent to $0.0237\Phi_0$ peak at 17.6 Hz was easily detectable. This clearly demonstrates that bistable SQUIDs used in combination with SR, can be useful in detecting weak, coherent magnetic signals buried in external noise, an application of considerable importance to the Navy.

Discussion and Summary

In this paper we have presented some results accumulated during the last few years which indicate that noise can be used effectively in nonlinear systems to improve the information content of received signals. The prime example of this process is SR, which has now been demonstrated in a variety of physical situations. As a prelude to further research on noise processes in the central nervous system and within the brain, we have shown that collective effects in networks of coupled bistable elements using more realistic (nonlinear) coupling schemes and introducing model synaptic noise, can lead to enhancements of SR. However, convincing experimental research on noise-driven nonlinear processes in neurons probably must begin with the simplest possible systems. We have thus introduced our program on hair mechanoreceptor cells in the crayfish tailfan. We have shown that the internal neuron noise can be controlled with the temperature of the preparation, and we have demonstrated an exponentially increasing SNR with internal noise intensity. Though an optimum temperature was observed, no optimum internal noise intensity was discovered. However, using external noise combined with the signal, we were able to observe SR in some neurons. Finally, we demonstrated SR at the fractional flux quantum level in a bistable SQUID arrangement designed to detect weak, noisy magnetic signals.

A number of questions and speculations are prompted by this work. First, *are there self regulatory mechanisms in neurophysiological systems which are capable of maximizing the SNR in response to changing environmental conditions?* Such mechanisms are certainly conceivable. In a neuronal network such as we have described, one can imagine an internal regulating mechanism which continuously adjusts the coupling coefficients J_{ij} as well as other dendritic and soma parameters that influence the effective potential function. Similarly, at the peripheral or receptor level of nervous systems, membrane noise levels can be adjusted via efferent connections⁴⁵. Even in simple receptor systems lacking known efferents (e.g. crayfish hair mechanoreceptors), we can speculate that biochemical feedback mechanisms could adjust membrane noise or

other parameters in order to maximize the SNR. Second, *in view of the success of the SR experiment using external noise on the crayfish, is it possible that living animals can exploit external background noise in order to detect signals with greater sensitivity?* This will be a difficult question to answer satisfactorily, but a step forward is to look simultaneously at stimulated sensory and motor neurons connected to the caudal ganglion. Finally, *what is the future of the bistable SQUID as a practical implementation of SR?* The next steps in this direction might include the redesign of the experiment briefly described above, to make use of integrated thin film chip SQUID technology and beyond that perhaps to explore the possibilities for SR in bistable high temperature SQUIDs.

Acknowledgements

We are grateful to Lon A. Wilkens for many crucial discussions and for providing the laboratory space and apparatus for the crayfish experiments. Eleni Pantazelou, Dave Pierson and Chris Dames offered invaluable assistance with the experiments, simulations and data analysis. The SQUID experiment would not have been possible without the expertise and dedicated work of Bill Jacobs and Jill Bekkedahl of NCCOSC and Andy Hibbs and Lowell Burnett of Quantum Magnetics who generously supplied the thin film SQUID chip as well as much of their own time. The authors are especially grateful to Frank Gordon of NCCOSC for his continuing and enthusiastic support of the bistable SQUID project and of the international workshop cited in Reference 1. All research herein reported was supported by the Physics Division and the Cognitive and Neural Sciences Division of the ONR.

Biographies

Adi Bulsara received his Ph.D. in 1978 from the University of Texas at Austin in Physics. This was followed by postdoctoral study at the University of California at San Diego. He is currently a scientist in the RDT&E Division of the Navy Command, Control and Ocean Surveillance Center (NCCOSC) in San Diego. His current research interests include the theoretical study of noise and chaos in nonlinear dynamical systems with applications to problems in physics and neuroscience.

John Douglass received a B.A. in Biology in 1980 from Oberlin College and a Ph.D. in Zoology in 1986 from Duke University. He was an NIH postdoctoral fellow at Yale University until 1990, when he accepted a research and teaching position at the University of Missouri at St. Louis. He is currently a Research Associate in Physics and Biology at UM-St. Louis. His research interests are in mechanoreception, photoreception, and information processing in sensory systems.

Address: Naval Command, Control & Ocean Surveillance Center, RDT&E Division, San Diego, CA 92152-5000, e-mail: bulsara@peanuts.nosc.mil.

Frank Moss obtained his Ph.D. in Physics from the University of Virginia in 1964. This was followed by postdoctoral study on quantum fluid turbulence at the University of Rome, Italy. He was twice awarded senior visiting fellowships by the Science and Engineering Research Council of Great Britain for studies in low temperature physics and nonlinear dynamics at the Universities of Exeter and Lancaster. He joined the Physics Department at the University of Missouri at St. Louis in 1971 where he is currently Professor of Physics and Biology. His current research interests center on experimental studies of stochastic phenomena in nonlinear dynamical systems in physics and biology.

Address: Department of Physics, University of Missouri at St. Louis, St. Louis, Mo 63121, e-mail: c4646@slvaxa.umsi.ed.

References

1. *Proceedings of the N.A.T.O. Advanced Research Workshop on Stochastic Resonance in Physics and Biology*, edited by F. Moss, A. Bulsara and M.F. Shlesinger, special issue, *J.Stat. Phys.* **70** (January 1993).
2. (a) H. A. Kramers, *Physica* **7**, 284 (1940) (b) P. Hänggi, P. Talkner and M. Borokovec, *Rev. Mod. Phys.* **62**, 251 (1990).
3. (a) B. McNamara and K. Wiesenfeld, *Phys. Rev. A* **39**, 4148 (1989) (b) P. Jung and P. Hänggi, *Europhys. Lett.* **8**, 505 (1989).
4. This peak sequence is dependent on the choice of time intervals used to assemble $P(T)$. Instead of this example, if we had chosen the residence time in the well +1, that is the time between a $-1 \rightarrow +1$ event and the subsequent $+1 \rightarrow -1$ reset event, the resulting $P(T)$ would have shown peaks at the odd integer multiples of the half period. We call this the ABBA symmetry. The only other choice available in a bistable system is the time interval between a $-1 \rightarrow +1$ event and the next $-1 \rightarrow +1$ event, which we call the ABAB symmetry. The latter choice results in a similar $P(T)$, except that the peak sequence is located at all integer multiples of the full period of the signal, as shown by the example in Fig. 2 (b). It turns out that the ABAB symmetry is important in neurobiology. See Ref. 6.
5. T. Zhou, F. Moss and P. Jung, *Phys. Rev. A* **42**, 3161 (1990)
6. A. Longtin, A. Bulsara and F. Moss, *Phys. Rev. Lett.* **67**, 656 (1991)
7. R. Benzi, S. Sutera and A. Vulpiani, *J. Phys. A* **14**, L453 (1981)
8. C. Nicolis, *Tellus* **34**, 1 (1982)
9. (a) R. Benzi, G. Parisi, A. Sutera, and A. Vulpiani, *Tellus* **34**, 10 (1982) (b) R. Benzi, G. Parisi, A. Sutera, and A. Vulpiani, *SIAM, J. Appl. Math.* **43**, 565 (1983) (c) A. Sutera, *Quart. J. Roy. Meteorol. Soc.* **107**, 137 (1981)
10. see for example, M. Chil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics* (Springer-Verlag, Berlin 1987) Part IV.
11. The dynamical cause of the ice age recurrences, and the question of whether they are indeed periodic, continue as current and controversial topics for research. See, I. Winograd, T. Coplen, J. Landwehr, A. Riggs, K. Ludwig, B. Szabo, P. Kolesar and K. Revesz, *Science* **258**, 255 (1992); and C. Nicolis in Ref. 1.
12. S. Fauve and F. Heslot, *Phys. Lett.* **97A**, 5 (1983)
13. (a) B. McNamara, K. Wiesenfeld and R. Roy, *Phys. Rev. Lett.* **60**, 2626 (1988) (b) G. Vermuri and R. Roy, *Phys. Rev. A* **39**, 4668 (1989)
14. R.F. Fox, *Phys. Rev. A* **39**, 4148 (1989)
15. P. Jung, *Z. Phys. B* **16**, 521 (1989)
16. C. Presilla, F. Marchesoni and L. Gammaitoni, *Phys. Rev. A* **40**, 2105 (1989).
17. P. Jung and P. Hanggi, *Phys. Rev. A* **41**, 2977 (1990)
18. J.E. Fletcher, S. Havlin and G.W. Weiss, *J. Stat. Phys.* **51**, 215 (1988)
19. (a) L. Gammaitoni, F. Marchesoni, E. Menichella-Saetta, and S. Santucci, *Phys. Rev. Lett.* **62**, 349 (1989) (b) L. Gammaitoni, E. Menichella-Saetta, S. Santucci, F. Marchesoni, and C. Presilla, *Phys. Rev. A* **40**, 2114 (1989)
20. (a) G. Debnath, T. Zhou and F. Moss, *Phys. Rev. A* **39**, 4323 (1989) (b) T. Zhou and F. Moss, *Phys. Rev. A* **41**, 4255 (1990)
21. For reviews of the use of analog simulators in stochastic, nonlinear dynamics, see *Noise in Nonlinear Dynamical Systems*, edited by F. Moss and P.E.V. McClintock (Cambridge University Press, Cambridge, 1989) Vol. 3. pp. 159, 222 and 243.
22. J. Hopfield, *Proc. Nat. Acad. Sci.*, **79**, 2554 (1982); **81**, 3088 (1984).
23. S. Shamma; in *Methods in Neuronal Modelling*, eds. C. Koch and I. Segev (MIT Press, Cambridge, MA, 1989).
24. A. Maren, C. Harston and R. Papp; *Handbook of Neural Computing Applications* (Academic Press, N.Y., 1991)
25. (a) A. Bulsara, A. Maren and G. Schmera, *Biol. Cyb.* preprint; A. Bulsara and A. Maren; "Coupled Neural-Dendritic Processes, Cooperative Stochastic Effects and the Analysis of Spike Trains" in *Proc. First Appalachian Conference on Neurodynamics*, Radford, VA, 1992. (b) A. Bulsara and G. Schmera, *Phys. Rev. E* **47**, 3734 (1993).
26. H. Haken, "Synergetics" (Springer Verlag, Berlin, 1977).

27. A. Bulsara, E. Jacobs, T. Zhou, F. Moss and L. Kiss, J. Theor. Biol. **152**, 531 (1991)
28. P. Jung, U. Behn, E. Pantazelou and F. Moss, Phys. Rev. A **46**, R1709 (1992).
29. F. Moss, *Stochastic Resonance: From the Ice Ages to the Monkey's Ear* in: **Some Problems in Statistical Physics**, edited by G.W. Weiss, **Frontiers in Applied Mathematics** (SIAM, Philadelphia, in press)
30. See for example, W. Denk and W.W. Webb, Hear Res. **60**, 89 (1992) for elegant measurements of the Brownian motions of the sensory hair bundle simultaneously with the spontaneous fluctuations of the intracellular voltage in frog saccular hair cells.
31. B.M.H. Bush and M.S. Laverack in: *The biology of crustacea*, Vol. 3 edited by D. Bliss (Academic Press, N.Y., 1982) pp. 399-468
32. K. Wiese, J. Neurophysiol. **39**, 816 (1976)
33. D. Mellin, J. Exp. Biol. **40**, 137 (1963)
34. I. Stiebler and P. Narins, Hear. Res. **46**, 63 (1990), and P. Narins, private communications.
35. J. Douglass and L. Wilkens, in preparation.
36. Crayfish in general, and *Procambarus clarkii* in particular, are very robust and successful animals, with various species inhabiting the globe from the arctics to the tropics. Our crayfish, which were removed from the high temperature environment, were only mildly sluggish during the first few days in the 9°C room, after which time they appeared to fully resume their previous behavior patterns.
37. This signal plus the noise was then passed through an amplifier and routed to the electromechanical vibrator. This latter device certainly has a limited frequency response of its own (a characteristic which was taken into account when measuring the tuning curves in order to keep the hair-fluid relative velocity constant at different frequencies). No effort was made to compensate the power spectrum of the noise in order to account for the response of the vibrator.
38. A. Longtin in Ref. 1.
39. Motor neurons exiting the ganglion to which our sensory neuron provides input are connected to abdominal muscles which operate the familiar tail-flip escape reflex.
40. See for example, A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect*, (John Wiley & Sons, Inc., New York, N.Y., 1982).
41. rf-SQUIDS can respond in the frequency range from DC to gigahertz, and the external electronics in our experiment had a bandwidth to 30 kHz, consequently the SQUID and its external electronics can respond essentially instantaneously to both the signal and the noise.
42. A. Silver and J. Zimmerman, Phys. Rev. **157**, 317 (1967)
43. Quantum Magnetics, 11578 Sorrento Valley Road, Suite 30; San Diego, CA 92121.
44. BTI Model 420.
45. E. Kaplan and R.B. Barlow, Jr., Nature **286**, 393 (1980)

Dialogue on Noisy Chaos

Robert Cawley and Guan-Hsong Hsu
Naval Surface Warfare Center

I. Introduction

Genesis

What is chaos? Where does it come from? And so what? Where's the physics? Not only that, what use is it?

With the invention of calculus three centuries ago¹, came the sometimes hard problem of integration, of finding a function specified by giving only its derivative, or only slightly more generally – but usually harder – a relation among two or more of its derivatives. With the calculus came also Newton's laws of motion, and the usually very hard problem of integration to find the motion specified by just the expressions of those laws as relations among the fundamental observables of the motion, the positions and their derivatives, the velocities and accelerations. Each unique physical system had its own version of those laws, and with each came another sometimes easy, but more often hard, mathematics problem of integration. The motions of the planets and the moon, of massive objects connected by springs, of vibrating beams, of the complex swirling of fluids in a laboratory experiment, of the ocean and atmosphere, all present their own version of the problem of integration. As ubiquitous as the kinds of physical systems governed by Newton's laws, have been the kinds of systems of differential equations specifying each its own version of the problem of integration. Today we can add many more such systems, involving reaction rates among constituents in chem-

ical mixtures, nerve impulse propagation in biological systems, light propagation in optical laser cavities, and just about any kind of real world classical system properly modeled by relations among physical observables and their derivatives, that is to say, by systems of differential equations. In the terminology of mathematicians a system of differential equations is called a "dynamical system."

A dynamical system governing a real world system is normally derived by application of fundamental laws, such as those of mechanics, local thermodynamic equilibrium, electromagnetic theory, and the like, to the actual system at hand, or rather more often to a simplified model of the actual system. When a physical system is complex, the importance of modeling may become paramount, simply to be able to have a picture we can comprehend and hope to keep track of. But some physical, chemical, biological and engineering systems may resist modeling efforts, making the governing, or underlying dynamical systems that describe them, defy the imaginative efforts of theorists to find a tractable picture. An apocryphally familiar crucible of examples is the weather, where the laws and governing equations are pretty well-known, but the solution is not.

Is complicated behavior chaos then? What is Chaos? And what uses does it have?

For two centuries our scientific and engineering ancestors approached the clearly fundamental problem of integration, of solving differential equations, of quite literally getting the

answers, by hit and miss trickery, or by brute force. Out of this approach have come the classical elementary functions, such as elliptic integrals, Bessel functions and hypergeometric functions of all sorts, nonlinear equations and systems having proper names attached, like the Riccati equation and Kepler system, and even the whole framework of classical perturbation theory. No evident chaos here, however.

But about one century ago a central shift in approach was introduced by the French mathematician, Henri Poincaré.² Instead of attempting to solve given systems of differential equations, he in effect re-framed the problem as one of "what are some of the properties of the solutions that can occur?" Like Newton before him, whose ruminations about Nature had led him to invent the calculus, Poincaré was led to topology, and to the formulation of the first foundations of modern dynamical systems theory. Poincaré's immediate intellectual successor, the American mathematician, George David Birkhoff, expanded fundamentally on the approach of Poincaré and carried the program further.³ But the consequences of the early work of Poincaré and Birkhoff entailed such horrendous complexities and were so discouraging that research into dynamical systems came to a virtual halt for nearly half a century.

Is this where Chaos came from?! Poincaré and Birkhoff changed the subject?

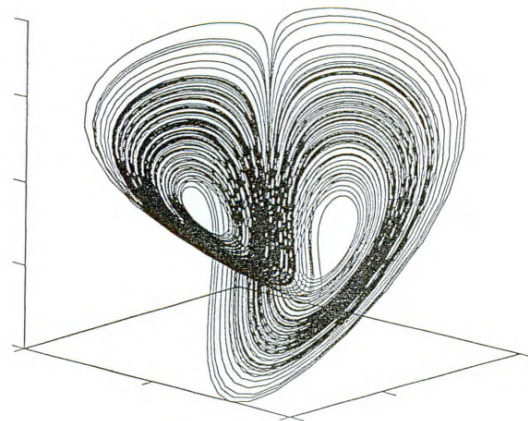
The shift in approach to the old calculus problem of integration, of *getting* the answers, engineered by Poincaré and Birkhoff, the shift from that of a classical algebra-like analysis to a qualitative geometrical, or topological, approach to characterizing logically possible solutions, deserves to be called a paradigm shift. This near still-birth was revived by mathematicians in the 1960s,^{4,5} much closer to a time when the computer would become available as a catalyst. The new paradigm was on its way to becoming a literally new calculus.

Revelations I-from the back of the Book

Alright, I assume you will tell us what chaos is pretty soon, if you can, and where it comes from, and all that. What you have said so far is something like this: Calculus is everywhere in the scientific and engineering world, and calculus doesn't give the answers even when I know the governing laws. It gives problems, systems of differential equations typically, that cry out for integration. Then you said that Poincaré and Birkhoff cheated by going to the back of the Book to see what kinds of answers there are, only they couldn't read some of them very well somehow because the answers are sometimes very complicated. Finally, you mentioned the computer, which I assume is going to have little trouble displaying the complicated answers; and you said the new paradigm, the view from the back of the Book, has by now become a veritable new calculus. But you still have not told us what chaos is. And, although I can imagine the technology importance a new calculus could have, I'd still like to hear about a specific or two.

Figure 1.

Fractal issuing from the Lorenz system. The trajectory $(x(t), y(t), z(t))$ quickly settles down to motion along the complicated curve shown.



In the back of the Book, instead of differential equations, and actually also their close cousins the difference equations (or, equivalently, maps,) one finds dynamical systems as flows and maps that act on manifolds. The manifold for the dynamics is called the phase space. A system of ordinary differential equations will typically possess trivial constant solutions,⁶ and periodic (limit cycle) solutions. The constant solution corresponds to a fixed point in the manifold for the corresponding flow, while the periodic solution corresponds to a simple closed curve. Owing to the geometric character of the phase space picture, periodic solutions are also called periodic orbits. But another kind of solution which now appears to occur typically in systems of differential equations, and which often used to be discarded when it was turned out by a computer,⁷ instead is simply aperiodic, that is, a sustained irregular oscillation. An aperiodic solution rendered in the phase space by a computer gives what looks like a pretty good fractal⁸ (see Figure 1). This kind of solution behavior is called chaos, and the phase space orbit a chaotic orbit. An oft cited example is the Lorenz system specified by the following equations in three dimensional space, R³

$$\frac{dx}{dt} = \sigma (y - x)$$

$$\frac{dy}{dt} = \rho x - y - xz, (x, y, z) \in \mathbb{R}^3 \quad (1)$$

$$\frac{dz}{dt} = -\beta z + xy$$

where σ , β and ρ are constants. For the parameter choice, $\sigma = 10$, $\beta = 10/3$ and $\rho = 28$ the fractal of Figure 1 results.

So, Chaos is a sustained irregular oscillation solution? With stress on the word solution, I presume.

With a few buts. There are issues we can't go into here due to lack of space. For instance, in a mathematical setting, we should insist upon a precise definition of chaos, so we can really say something about it, thus we might require there be a "Smale horseshoe" in the dynamics. At another level of refinement, we may want to require the existence of at least one positive Lyapounov exponent in order to distinguish chaotic from strange nonchaotic behavior. In the latter the phase portrait may be a fractal, but a time series issuing from the equations of motion, although aperiodic, doesn't possess extreme sensitivity to initial conditions, a feature usually required of chaos.

Gospel

Chaos isn't physics, then, it's a kind of kinematics. Suppose I observe chaotic behavior in an experiment, what then? Where's the science?

That's right, it isn't physics, or chemistry, or biology, or meteorology, or any of the scientific or engineering areas in which chaotic behavior may be found. Like periodic oscillation, it is a universal category of behavior, and at one level of description the "science of Chaos" is basically a kinematic science. The term "science of chaos" is nearly oxymoronic, however; the science in question is really that of dynamical systems theory, together with its applications and manifestations in the world. If I observe chaotic behavior in an experiment, it is a very strong indicator of the presence of an underlying dynamical system governing the behavior of the system under observation. This presence is arguably the most interesting thing at least initially, not the romance of finding chaos.

Well, suppose I observe periodic behavior. That should be an equally strong indicator of the presence of an underlying dynamical system, shouldn't it?

The two phenomena do have an equal status, in that respect. But there are a couple of important practical differences nonetheless. The first is that an underlying simplicity of the sort implicit in a real world system's being well modeled by just a few coupled ordinary differential equations, let us say, would not have been expected in general from aperiodic behavior until now. The evidence of the simplicity in the case of periodic behavior is immediate in the regular behavior of the observed data; in the case of aperiodic behavior one doesn't see that simplicity until one looks at a phase space representation, a phase portrait, if I may say so. So, the mere determination that an aperiodic behavior is some simple form of chaos, rather than random noise is a significant observation about the data. The second difference is that stable periodic behavior by itself gives little information about the underlying dynamics.

A simple periodic orbit can sample only a very tiny portion of the phase space. On the other hand, a chaotic (or strange nonchaotic) orbit can fill a dimensionally nontrivial fractal set. It explores possibly several independent directions,⁹ and in so doing gives evidence of a better bound, a lower bound, on the number of first order equations, m , needed for a model. The latter is synonymous with system size, *i.e.*, with the number of degrees of freedom needed to specify the underlying dynamics. The number m , which, of course, is always an integer, is called the topological dimension for the dynamics.

It would be nice if we reliably could measure m directly from the data!

Yes. It would.

Quest for the Grail

OK, so if I find that a noisy aperiodic quantity is actually the output of a chaotically behaving system I can infer the likely existence of a simple set of rules producing that behavior. And if I know that much, I might be led to search for a simple model despite the extremely complicated behavior in the data, and where otherwise I might have missed this opportunity. I'd just have to put physics in to get physics out, which is fine with me.

Exactly. Chaos is not a Silver Bullet. Still, you are a gentle scholar to be so agreeable, and to convert so readily. But you have just fingered one of the essential points. It naturally follows that a *simple* model could require that m not be too large! So one of the First Holy Grails of a chaos hunt in experimental time series is the uncovering of good clean evidence of the likely presence of low dimensional dynamics.

That's nice. Can you tell me another First Holy Grail story about chaos?

The "science of chaos" is a House with many Mansions, which space again does not permit me to go into, if I may say so. One example is the control of chaos, achieved by a loop generating a sequence of successive small parameter tweakings-*cf.* Equation (1), in which α , β and ρ are system parameters. It is sometimes possible to exchange chaotic behavior for a stably periodic behavior, which is what is meant by control here. Of course, any system parameters varied to achieve control, must be experimentally accessible.^{10,11} The method exploits the typical circumstance that a chaotic time series cycles in and out of unstable periodic segments, by locking in the periodic behavior one seeks to control for.

It seems to me that if I want to control chaos in this way, the first thing I'd like to have is that my system be deterministic as opposed to random. An m -dimensional governing dynamical system, a system of first-order ordinary differential equations, gives me this. I don't need m to be small to have determinism. So maybe I can control high dimensional chaos?

I'm glad to see you've recovered your critical faculty, gentle scholar. You are probably right, but the experimental situation is not yet entirely clear; see Reference.¹² For a

theoretical treatment see Reference.¹³ But now that I have told you more or less what chaos is, *exactly* where it comes from, pretty much where the physics is, if I may say so, and a couple of the motivational so what's for studying it, it is time to add the noise. For that, I should best turn to experiment, which indeed some say makes Truth appear!

Does experiment make technology applications appear?

Revelations II-through a glass darkly

Experimentalists operate from the back of the Book. What they measure are answers, and we'd like to learn as much about the questions as we can. These reflect the underlying laws and principles of system behavior. Experimental results always will be contaminated by some noisy background interference, however, so one really measures distorted versions of the answers, seeing through a glass darkly, as it were. From a theoretical point of view, there are two principal kinds of noise, additive noise and dynamical noise. The former models experimental effects of instrument noise, such as photomultiplier tube shot noise, while the latter models more subtle effects persistently present in the actual behavior of the experimental system we are trying to observe. The fluctuating effects of background or reservoir coupling to a system are an example of dynamical noise. Inevitably, both kinds are present, and while the presence of such noise may not be effective in masking a periodic system oscillation, it can create significant problems in an experimentalist's ability to recognize chaotic behavior. Noise contamination can also limit the efficacy of experimental control of chaos.¹⁴

I can't resist observing the irony present in the fact that a discipline as intricate and abstract as topology should lie bed fellow in the back of the Book with something as down to earth as experimental science. But I'll be quiet for a while now and let you finish your story of noisy chaos.

Thank you, you've been a wonderful foil. I'll make two final remarks, then describe some actual recent noise reduction work. The idea that extremely complicated data may sometimes reflect an underlying, albeit nonlinear, simplicity is an exciting and seductive one. It has been confirmed in numerous experiments. It has also been over-indulged, unfortunately, and some of the claims for observation of chaos are surely erroneous. The idea that chaotic dynamics might be uncovered after removal of a contaminating noisy background is an easy extension, and possibly even more exciting and seductive. It is probably not useless to note that this too is likely to be abused. If we are lucky, we will nonetheless find some successes here, too.

I don't think I can stress enough the fact that "chaos" is not a Silver Bullet. If I examine data from some complex system and find chaos, all I have found is a proper characterization of the content of an answer in the back of the Book. I may deduce the presence of some simple governing equations, but I don't get those equations nor more importantly the

principles of the underlying model they express. This last takes extra work. As the old saying in the physics community goes, I do have to put physics in, to get physics out. Or, physiology in, to get physiology out – medical researchers, please take note.

Wait. Please don't go on without telling me about technology applications potential.

I will. I'll tell you at the end.

II. Approaches to the problem

A complicating feature to the chaotic noise-reduction problem is that the chaotic signal will have a broadband spectrum just like the noisy broadband contaminant we wish to remove. Traditional band-passing techniques might result in unacceptable signal distortions, or simply be ineffective. A general approach to the problem may be seen to proceed through four steps, which exploit the phase space picture of a dynamical system. The local geometric projection (LGP) method introduced by the authors^{15,16} was the first to employ all four steps in a reasonably systematic way. Previous methods have employed variously one, two or three of these four elements. Given a scalar time series which may be a voltage output of a sensor through which an experimental observation is being made, for example, the four steps are:

- (1) represent $v(t)$ through some kind of embedding construction, as a "data-state" vector time series $p(t)$ in a d -dimensional Euclidean space R^d ;
- (2) through some protocol of "noise reduction," construct from $p(t)$ a "cleaner," candidate data-state vector replacement time series, $\hat{p}(t)$;
- (3) disembed, *i.e.*, construct from $\hat{p}(t)$ a suitable scalar replacement time series $\hat{v}(t)$; and
- (4) iterate the procedure until maximum, or at least satisfactory, noise reduction has been achieved.

The mathematical idea behind this approach is that, under the dynamical hypothesis on the data, there is a true noise-free orbit contained in an m -dimensional manifold M . The embedding construction, under which $M \rightarrow M' \subset R^d$ should give a complete representation of the orbit dynamics provided d is large enough that an embedding actually results (see Figure 2). The observed quantity $v(t)$ is then a noisy version of one of the coordinates for the system, such as the coordinate $x(t)$ in the Lorenz system, for example.

Most noise reduction methods so far have centered around various approaches to step (2). It has also been recognized that some noise reduction may be achieved right away in step (1) by what amounts to suitable coordinate choices. Except for the authors, scant notice has been paid the role of step (3); and likewise step (4) has received little or no attention except as an end in itself.

Original from

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN

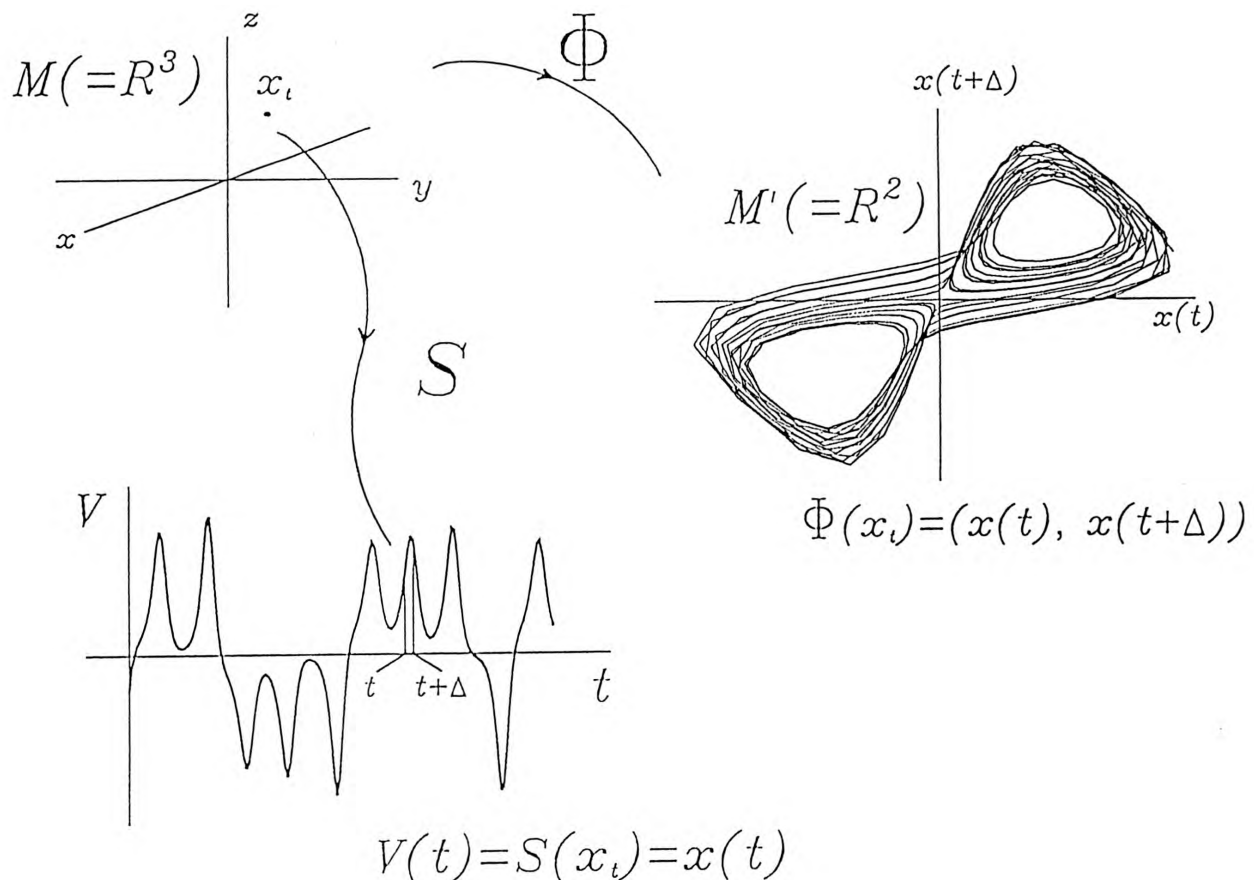
March 1993

An imaginative early approach based on linear signal processing ideas and constituting the first method based on step (1) was Broomhead and King's singular value truncation.¹⁷ Strictly, their "trajectory matrix" truncation procedure was not propounded as a noise reduction method, but rather an approximation approach based on efficient choice of coordinates to reduce the relative contribution of the noise. The method can be profitably applied only to moderately well sampled flow data (*i.e.*, analog data). An extensive investigation into questions of optimal choice of coordinates was given by Casdagli, *et al.*,¹⁸ In these approaches, steps (2)-(4) do not appear.

The most generally useful of the early noise reduction efforts is probably that of Kostelich and Yorke¹⁹. This approach, which employs steps (1), (2) and (4), never disembedding, makes use of dynamical information implicit in the data by means of a learning technique. Farmer and Sidorowich²⁰ also employ a learning technique and exploit technical properties of chaos, *viz.* expanding and contracting behavior along unstable and stable directions, to achieve noise reduction. These authors use steps (1), (2) and (4), also. The methods of References^{19,20} are applicable to both map and flow data although the latter reported application to only very

Figure 2.

The Ruelle-Takens embedding construction shown for the Lorenz system, where $M \subset \mathbb{R}^3$. x_i represents a point from the curve shown in Figure 1. S , in general, is the connection between the system under observation, which mathematically lies in M , and the experimental observation itself, $V(t)$. The construction is applied here to the case $d=2$, so it cannot be an embedding. This is reflected in the evident occasional self-intersections ("imperfections") of the fractal curve shown. When d is large enough the self-intersections go away, giving then an embedding, a "perfect" presentation. The $d=2$ picture can also be considered as a projection into the first two components of an embedding with sufficiently high d . Like the phase portrait in Figure 1, the embedded trajectory cannot have self-intersections.



low levels of noise. Shadowing methods,^{21,22} which have employed steps (1), (2) and (4), are somewhat related to those of Reference,¹⁹ but they have only been applied to two-dimensional maps. In an unusual approach Marteau and Arbarbanel²³ employ steps (1), (2) and (4), but assume *a priori* knowledge of a sample clean reference orbit. Noise reduction is achieved by a conditional probability technique. Substantial improvements can result from other noise reduction methods also when a sample reference orbit is available. This is true for the LGP method²⁴ and it is true for a prediction-based method (evolved trajectory minimization) due to Kadtko and Brush²⁵. The latter authors report use of only step (2) in Reference²⁵, but also have had good results using steps (1) and (2)²⁶. Grassberger and Schreiber²⁷ implement a modification of a conventional local smoothing procedure on the given scalar time series $v(t)$, with the smoothing coefficients being determined by the local distribution of data-state vector values $p(t)$ rather than just those of $v(t)$ local in time. Like our own, this method is actually a local geometric one, but where step (3) is unnecessary since step (2) is implemented only on $v(t)$ itself; the method leads to good results. Finally, and most recently, Sauer²⁸ has implemented a modification of our scheme, which he also applies to very high levels of noise, and with good results. He does not report results for application to maps or to coarsely sampled flows.

III. LGP method for noise reduction-a case study

To describe the LGP method, we describe what happens in each of the four steps listed in the previous section. Full details are in Reference¹⁶.

Step 1. Embedding construction. We consider first the noise-free chaotic case where the observation $v(t)=V(t)$. To have a complete representation of the chaotic dynamics producing $V(t)$ it is not necessary to measure all of the variables of the system. For example, if $V(t)=x(t)$ for the Lorenz system, eqs. (1), we do not somehow have to get $y(t)$ and $z(t)$ also. One possibility may be to choose as coordinates for the phase space, $(V(t), \dot{V}(t), \ddot{V}(t))$, say, where we have assumed three dimensions-about which more in a moment. This was the state space reconstruction studied originally by Packard, *et al.*²⁹ However, the numerical differentiation of observational data required for this method amplifies high frequency noise that inevitably must contaminate $V(t)$. Another approach, not suffering from this problem, appears first to have been suggested by D. Ruelle (see Ref.²⁹), namely, the delay coordinate construction (DCC). Here the coordinate choice for three dimensions is $(V(t), v(t+\Delta), v(t+2\Delta))$, where $\Delta \neq 0$ is called the delay. In each of these examples the data scalar $V(t)$ has been converted to a three-dimensional data-state vector.

Mathematical justification for each of these procedures was provided by Takens³⁰. Takens' Theorem says that the map

Φ from the m -dimensional manifold M for the dynamics reflected in $V(t)$, into the d -dimensional Euclidean Space R^d specified by the coordinate choices using $V(t)$ along with its derivatives ($d-1$ of them) or the delays (up to $(d-1)\Delta$), is very generally going to be an embedding (*i.e.* "perfect") as long as $d > 2m$. The data-state vector orbit in R^d will thus be contained wholly in the embedded image $M' = \Phi(M)$. Since we may not know m experimentally we do not generally know whether an embedding will result for a given choice of d (see Figure 2). However, more recent mathematical work has provided a considerable softening of the restriction on d to $d > 2d_\wedge$ where $d_\wedge$ is the fractal dimension of the attractor $\wedge$ holding the data-state vector orbit.³¹

Thus, while in fact an experimental observation is always somewhat noisy, *i.e.*, so that

$$v(t) = V(t) + \varepsilon \eta(t), \quad \varepsilon > 0, \quad (2)$$

where $\eta(t)$ is a normalized random process representing the noise, the DCC can at least provide an embedding for the deterministic part. The main features of the Takens construction are depicted in Figure 2.

Step 2. "Official" noise reduction. It is here that most methods differ from one another. But it should be stressed that significant noise reduction can also occur in Step 3 (see below)!

The idea of the LGP method is very simple. For sufficiently large d , all the dynamics implicit in $V(t)$ lies in M' , but the noise contribution to the data-state vector for $v(t)$, *i.e.* to

$$p(t) = (v(t), v(t+\Delta), \dots, v(t+(d-1)\Delta)), \quad (3)$$

explores every possible direction in R^d including those lying off of M' . So, we just estimate where M' is from the noisy data, $p(t)$, and move the points of $p(t)$ towards our best guess for M' ! Thus $p(t) \rightarrow \hat{p}(t)$, the "cleaned up" vector time series.

The problem of locating M' , whether it is some sort of distorted donut or hyper-torus, a garbly-looking multi-dimensional ellipsoid, or a funny saddle, in all cases if we look only at a small cloud of neighboring points, this small part will look flat. That is exactly what it means for M' to be a (differentiable) manifold, namely to be locally flat. So, our strategy is to estimate local hyperplanes tangent to M' . These are the so-called tangent spaces. Since there will also be some small non-flatness (see Figure 3), we don't specify $\hat{p}(t)$ by moving the noisy points of $p(t)$ into these estimated tangent spaces, but only halfway towards them.

Step 3. Disembedding. Although $p(t)$ is a delay coordinate construction by definition (eq. (3)), the cleaned up version is almost surely not. But the right answer, *viz.*

$$P(t) = (V(t), V(t+\Delta), V(t+\Delta), \dots, V(t+(d-1)\Delta)), \quad (4)$$

certainly is. Thus we seek a best scalar time series $\hat{v}(t)$ that will produce a data-state vector orbit as close as possible to

$\hat{p}(t)$. This is our first guess at $V(t)$. To do this we minimize the error,

$$\varepsilon = \sum_t \sum_{j=1}^d [\hat{p}_j(t) - \hat{v}(t + (j-1)\Delta)]^2 \quad (5)$$

Except for t -values near the end-points of the sum, the minimizing $\hat{v}(t)$ is given by

$$\hat{v}(t) = \frac{1}{d} \sum_{j=1}^d \hat{p}_j(t - (j-1)\Delta) \quad (6)$$

Similar expressions result for end-point range t -values.¹⁶

Notice if $\hat{p}(t)$ were actually $P(t)$ as shown in eq.(4), then $\hat{v}(t)$ in eq.(6) would become $V(t)$ as we should require. Since there will surely be some noise still present in $\hat{p}(t)$, we can imagine that the averaging that occurs in eq.(6) will provide further noise reduction. This expectation is borne out in numerical computations.

Summarizing the effects of steps (1)-(3), we have that $v(t) \rightarrow p(t) \rightarrow \hat{p}(t) \rightarrow \hat{v}(t)$, that is, the LGP algorithm A has mapped the scalar time series $v(t)$ to a cleaned-up scalar time series $\hat{v}(t)$, i.e. $A: v(t) \rightarrow \hat{v}(t)$. We are ready now for

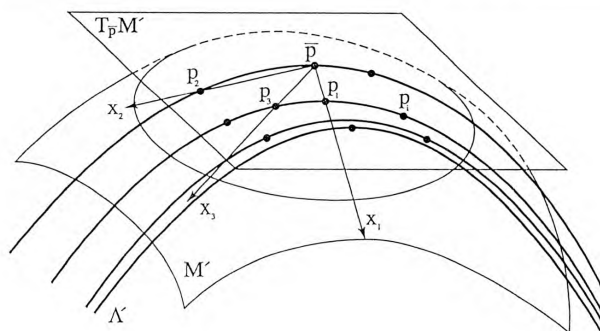
Step 4. Iteration to maximum improvement. We define the improvement in terms of signal-to-noise ratio (SNR). The initial SNR in dB is

$$S_i = 20 \log \frac{\|V(t)\|}{\varepsilon \|\eta(t)\|} \quad (7)$$

where $\|\bullet\|$ denotes root-mean square power. Following eq. (2), we write the cleaned-up time series as

Figure 3.

Tangent plane $T_{\bar{p}} M'$ to M' at a point $\bar{p}$. Λ' refers to the image of the strange attractor Λ on which the orbit under observation is assumed to lie.



$$\hat{v}(t) = V(t) + \varepsilon \hat{\eta}(t), \quad (8)$$

so the output SNR in dB is

$$S = 20 \log \frac{\|V(t)\|}{\varepsilon \|\hat{\eta}(t)\|} \quad (9)$$

Since ε can measure the strength of the noise, we simply normalize $\eta(t)$ to suit our convenience. We take $\|\eta(t)\| = 1$. Then the SNR improvement is

$$\delta = S - S_i = 20 \log \frac{1}{\|\hat{\eta}(t)\|} \quad (10)$$

which depends only on the relative output noise power. Under n iterates, $v(t) \rightarrow \hat{v}_n(t)$, and $\delta \rightarrow \delta_n$. In Figure 4 (a) we show plots of SNR vs. n for some model calculations. In all cases, if the iterations are carried far enough, the output SNR peaks for some $n=n_M$, and falls under further iteration.

IV. Performance of LGP method

We denote the improvement after n_M iterates by δ_M . The values achieved vary with the embedding trial dimension d , the dimension k of the tangent spaces used to estimate M' and the number of points $v+1$ chosen to model the local neighborhoods of M' . δ_M also depends on the delay Δ chosen for the DCC, the sampling time ΔT for the case of flow data, the initial noise fraction $\mathcal{N}_0$ whether it is additive or dynamical noise, the time series length, and the system itself. We have performed extensive systematic investigations of almost all these dependencies, much of which is reported in References.^{15,16} We offer here a representative sample of performance data reported in Reference³², which includes both additive and dynamical noise cases, in Figure 4. In Figure 5 we show phase portrait results for a 100% initial noise Lorenz System example. The initial noise levels in the examples shown in Figures 4 and 5 range from 1% to 100%, and the peak values of δ_M are typically from 6-12db. Longer time series result in higher improvements, with peak improvements rising roughly linearly with $\log N$, where N is the number of points in the time series.¹⁵

We remarked earlier that steps (1) and (2) have received the most attention in noise reduction efforts, but with little attention paid to steps (3) and (4). We addressed the role of step (3) in the previous section; here we offer a couple of remarks about step (4).

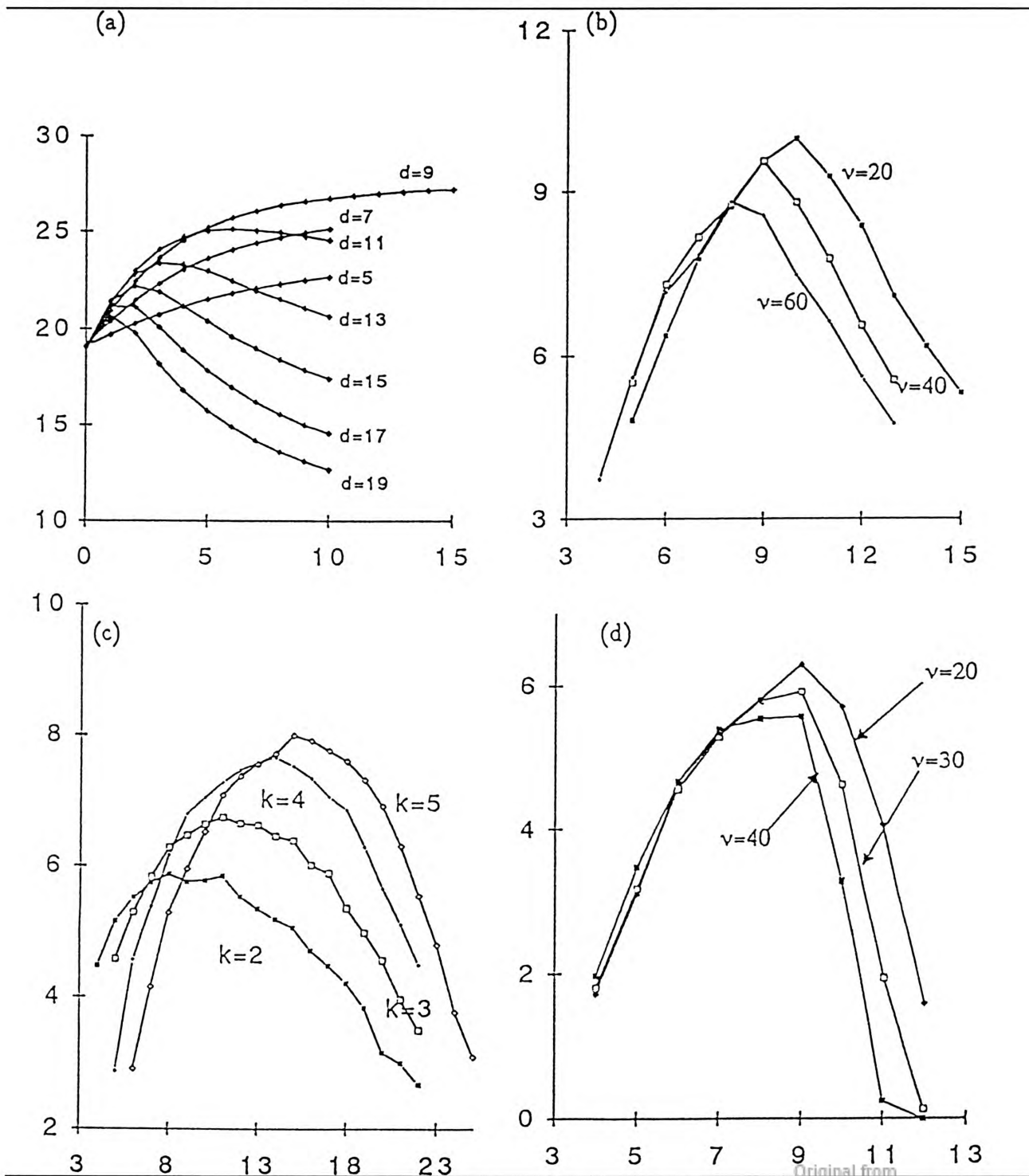
Consider the quantity

$$\sum_n = 20 \log \frac{\|v(t)\|}{\|v(t) - \hat{v}(t)\|} \quad (11)$$

and substitute eqs. (2) and (8) to get, for small ε

Figure 4.

Algorithm performance: (a) SNR(db) vs. iteration number n for Hénon map with $N=3,000$, $N = 10\%$, $k = 3$ and $v = 20$; (b) peak improvement δ_M vs. d for Hénon map with $N = 10,000$, $\eta = 10\%$ and $k = 3$; (c) peak improvement δ_M vs. d for Lorenz system with $\Delta T = 0.25$ (3.5 points per typical oscillation). The reduction parameters here are $\eta = 10,000$, $N = 10\%$ and $v = 40$; (d) 1% uniform dynamical noise for Hénon map with $N = 10,000$ and $k = 3$.



$$\Sigma_n = S_i + 20 \log \frac{1}{\|\eta(t) - \hat{\eta}(t)\|} \quad (12)$$

As n rises towards n_M , Σ_n falls, and approaches S_i itself. This assumes $\hat{\eta}(t)$ can be neglected in the second term of eq. (12). Note $\|\eta(t)\| = 1$. In fact, the theoretical situation is more complicated, but it is possible in some cases to estimate S_i from the behavior of Σ_n (see Reference ¹⁶). This sort of thing is particularly important for analysis of experimental data, where Σ_n is directly measurable from the raw data by simply applying the algorithm.

In fact, we believe that systematics of performance of a general chaotic noise reduction algorithm, $A: v(t) \rightarrow \hat{v}(t)$, should contain a good deal of useful information. This is an area where only the surface has been scratched. For example, in addition to S_i , we'd like to know quantitatively what algorithm parameters to choose for best results, *e.g.*, what values for (d, k, v) in the case of the LGP algorithm and how many times to iterate; we'd like to know what improvement we actually achieved. And it would be nice if there were a way to estimate an effective lower bound to m . These sorts of things probably do lurk in the n -dependence of functions like Σ_n , and we have found broad hints to this effect in our simulation studies.^{15,32}

V. Conclusion

I am not really the gentle scholar you take me for. On the contrary, I fancy myself a raging Philistine. Tell me now about applications.

Chaotic dynamics have been found and precisely analyzed in an impressively wide range of experimental systems. Among these are closed flow fluid systems, Taylor-Couette flow and Rayleigh-Benard convection, coupled mechanical fluid systems in open flow geometry, surface wave patterns in a vertically oscillating "coffee-cup" (Faraday experiment), in the very irregular beating of chick heart cell aggregates under suitable periodic electrical stimulation, numerous solid state systems, such as di-electric breakdown in Germanium, continuously stirred chemical reactor experiments (Belousov-Zhabotinsky reaction), chemical oscillations catalyzed on solid surfaces, numerous nonlinear electronic circuits, Josephson junction arrays, laser operation, and numerous quantum optical systems and processes, probably including multiphoton excitation, nonlinear mechanical and magnetoelastic systems, and even the heavens. Documenting the ubiquitous character of deterministic chaos, Marek and Schreiber³³ cite more than three hundred references in a chapter on chaotic dynamics in experiments. The list is dominated by accounts of hard experimental documentation of the observation of low dimensional chaotic dynamics.

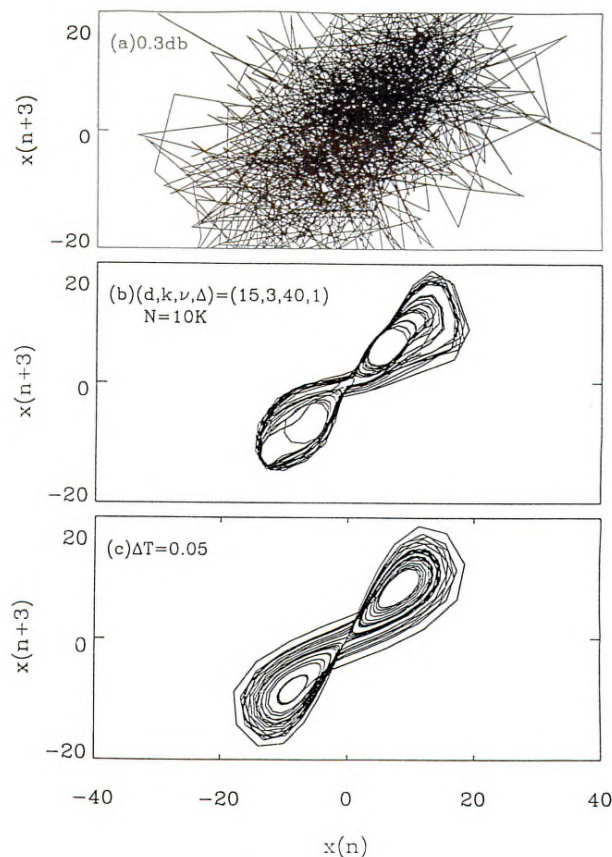
By far the greatest activity in the field of nonlinear dynamics has been among scientists, and much less activity on

the part of engineers and technologists despite the pervasiveness of nonlinear effects in real systems. That such a remarkable list as that above should fail to have important implications for technology applications, however, is unthinkable. A story making the rounds ten years ago concerned a sensor driven by a noisy electronic circuit that couldn't be "cooled". At liquid Nitrogen temperature the noise temperature was still 1000K. Three theses are said to have been written on this "remarkable" system. This was before chaos had become a household word. On the other hand, dynamics have been exploited more recently in the orbit design for an interplanetary probe, to the Jacobini-Zinner comet. It helps to be aware of the governing dynamical laws to exploit the corresponding phenomena, as well as to avoid mistakes!

There's an obvious moral to your story, and I think you are setting me up. But go ahead, let me have it!

Figure 5.

Lorenz phase portrait (DCC, $\Delta = 3$ first 1,000 points): (a) unsoiled; (b) white noise added; and (c) cleaned up, $n=50$, $\delta = 10.4\text{db}$.



We expect technology applications for chaotic noise reduction to occur in settings of the sort where an underlying and possibly unsuspected dynamics can be exposed. It is in the nature of the beast that prediction of just exactly where this will occur cannot be made with confidence. There are some places we ought to look, however, with medical applications at the top of the list. Methods of control of chaos that depend upon prior dynamical measurements could be applied more reliably if these could be performed on less noisy versions of system behavior, and noise reduction might help here. Biological processes, such as heart rhythms, should surely be examined from this point of view.

What else?

How much of the clattering of various kinds of machinery reflects low dimensional dynamical behavior? Could this be silenced by control, or by redesign? Radar performance is limited by system noise. Could some of this be dynamical, and perhaps be predictable (possibly after noise reduction) over short times, as chaotic processes typically are? The noise might then be subtracted on the fly, thereby improving performance! *Even where underlying dynamics is not present*, phase space methods such as those described herein appear to specify a novel signal separation into "smooth" or dynamics-like ($\hat{v}(t)$), and "rough" or noise-like ($v(t) - \hat{v}(t)$), parts. Thus these methods may be regarded as constituting a novel approach to time series modeling. Nothing of this approach as yet appears to have found its way into the mathematical statistics literature. Nonetheless these methods may have broad-based signal-processing uses. It may be that the prediction-subtraction idea could be applied to suppression of surface and volume reverberation, or of white-capping noise for sonar applications, or to radar sea clutter suppression for surface navy ship defense applications.

It sounds as though we need to have more engineers involved in nonlinear dynamics.

Yes, we do. But it won't likely be easy, the scientific ideas are still unfamiliar to most people. We may have to wait for a new generation of young technologists with more modern technical education.

Has anyone ever used noise reduction to improve the analysis of chaotic data? Or to uncover hitherto unsuspected low-dimensional dynamics, as you are suggesting here?

You are not at all the raging Philistine you fancy yourself. Your response and your questions are pertinent and fair. If the science of nonlinear dynamics, enabling as it is, is yet young, the consequent technology realizations and awareness potentialities are younger still. Much opportunity awaits the adventurous in this field. But to answer the first of your two questions: except for a couple of very simple examples of before and after measurements of Lyapunov exponents for known case systems¹⁶ (Henon and Lorenz) the answer is no – not as far as I know. As to your last question, again the answer is no, not to my knowledge.³⁴

Acknowledgements

This work was supported by the Office of Naval Research and the Independent Research Program of Naval Surface Warfare Center, Dahlgren Division.

Biographies

Dr. Cawley received a B.S. in Physics from the Massachusetts Institute of Technology in 1958, and an M.S. and Ph.D. in physics, with minor in astronomy, from the University of Illinois, Urbana, in 1960 and 1965, respectively. He taught physics at Clarkson University from 1965 to 1967, and came to Dahlgren Division (NSWCDD), White Oak, in 1969. In 1983, he started the NSWC Navy dynamics institute Program at White Oak with the help of ONR funding. He is currently a Principal Research Scientist at NSWCDD and Senior Staff Scientist to the Mathematics and Computations Branch. He was given a Navy Meritorious Civilian Service Award for his nonlinear dynamics work in 1991. His current scientific interest is chaotic noise reduction and its application to data analysis. Dr. Cawley is a member of the American Physical Society, the American Mathematical Society, and the Society for Industrial and Applied Mathematics.

Guan-Hsong Hsu received his B.S. in mathematics from the National Taiwan University in Taipei, Taiwan, Republic of China in 1978, and his M.S. and Ph.D. in applied mathematics from Courant Institute, New York University, in 1983 and 1986, respectively. He was a post doctoral associate shared between the University of Maryland and NSWCDD during 1986 to 1989. He taught mathematics at University of Missouri, Columbia, from 1989 to 1992. He is currently an ONR Post Doctoral Fellow at NSWCDD. His scientific interest is numerical methods in chaotic dynamics and their applications, especially to data analysis. Dr. Hsu is a member of the American Mathematical Society and the Society for Industrial and Applied Mathematics.

Address: Naval Surface Warfare Center, Mathematics and Computations Branch, Dahlgren Division, White Oak, Silver Spring, MD 20903, e-mail: rcawley@nswc-wo.navy.mil.

References

1. Newton, *Philosophiae Naturalis Principia Mathematica*, E. Halley, pub., 1686.
2. H. Poincaré, *Les Methodes nouvelles de la Mecanique celeste*, Gauthier-Villars, Paris, 1892 (t.1), 1893 (t.2) and 1899 (t.3). For a recent translation see H. Poincaré, *New Methods of celestial mechanics*, vols 1-3, Daniel L. Goroff, ed., AIP, 1993.
3. G.D. Birkhoff, *Dynamical systems*, American Mathematical Society Colloquium Publications, Vol. IX, 1927.

4. D.V. Anosov, Proc. Steklov Inst. Math. 90, 1(trans. from Trudy Mat. Inst. Steklov).
5. S. Smale, Bull. Amer. Math. Soc. **73**, 747(1967).
6. Externally driven systems, whose corresponding vector fields cannot vanish, are an exception.
7. A well-known exception is E.N. Lorenz, who did not file his ugly duckling in a bottom drawer. E.N. Lorenz, J. Atm. Sci. **20**, 130(1963).
8. B.B. Mandelbrot, The Fractal Geometry of Nature, H. Freeman & Co., New York, 1982.
9. The number of independent directions coincides with the number of non-negative Lyapunov exponents.
10. C. Grebogi, E. Ott and J.A. Yorke, Phys. Rev. Letters **64**, 1196(1990).
11. W.L. Ditto, S.N. Rauseo and M.L. Spano, Phys. Rev. Letters **65**, 3211 (1990).
12. A. Gills, C., Iwata and R. Roy, Ira B. Schwartz and I. Triandaf, Phys. Rev. Letters **69**, 3169(1992). These authors report experimental control of a laser system in "a higher dimensional chaotic mode," but do not report what that dimension value is. A qualitative theoretical basis for their assertion is in C. Bracikowski, and R. Roy, Chaos **1**, 49(1991).
13. Ditza Auerbach, Celso Grebogi, Edward Ott and James A. Yorke, Phys. Rev. Letters **69**, 3479 (1992).
14. M.L. Spano and W.L. Ditto, Taming chaos experimentally: a primer, in Procs. 1st Experimental Chaos Conf., Oct. 1-3, 1991, Arlington VA, Sandeep Vohra, Mark Spano, Michael Schlesinger, Lou Precora and William Ditto, eds., (World Scientific, Singapore, 1992), pp. 107-120.
15. Robert Cawley and Guan-Hsong Hsu, Phys. Letters **A166**, 188 (1992).
16. Robert Cawley and Guan-Hsong Hsu, Phys. Rev. **A46**, 3057 (1992).
17. D.S. Broomhead and G. King, Physica **D20**, 217 (1986).
18. M. Casdagli, S. Eubank, J.D. Farmer and J. Gibson, Physica **D51**, 52 (1991).
19. Eric Kostelich and James A. Yorke, Phys. Rev. **A38**, 1649 (1988).
20. J.D. Farmer and J. Sidorowich, Exploiting chaos to predict the future and reduce noise, in *Evolution, Learning, and Cognition*, edited by Y.C. Lee (World Scientific, Singapore, 1988) pp.277-325.
21. S.M. Hammel, Phys. Lett. **A148**, 421 (1990).
22. J.D. Farmer and J. Sidorowich, Physica **D47**, 372 (1991).
23. P.F. Marteau and H.D.I. Abarbanel, J. Nonlin. Sci. **1**, 313, (1991).
24. Robert Cawley and Guan-Hsong Hsu (unpublished).
25. James B. Kadtko and Jeffrey Brusch, Noise reduction methods for chaotic signals using empirical equations of motion, in Proc. SPIE **1699**, 338 (1992).
26. James B. Kadtko (private communication).
27. Thomas Schreiber and Peter Grassberger, Phys. Letters **A160**, 411 (1991).
28. Tim Sauer, Physica **D58**, 193 (1992).
29. N.H. Packard, J.P. Crutchfield, J.D. Farmer and R.S. Shaw, Phys. Rev. Letters **45**, 712 (1980).
30. F. Takens, Detecting strange attractors in turbulence, in *Dynamical systems and turbulence*, Warwick 1980, D.A. Rand and L.S. Young, eds., L.N. Mathematics, No. 898 (Springer-Verlag, Berlin, 1981), pp.366-381.
31. Time Sauer, James A. Yorke and Martin Casdagli, J. Stat. Phys. **65**, 579 (1991).
32. Robert Cawley and Guan-Hsong Hsu, Noise reduction for chaotic data by geometric projection, in *Applications of digital image processing XV*, Andrew G. Tescher ed., Proc. SPIE **1771**, 368-376 (1993).
33. M. Marek and I. Schreiber, *Chaotic Behavior of Deterministic Dissipative Systems*, Cambridge University Press, 1991 (Chapter 5).
34. One so far inconclusive attempt to recover chaos on geological time scales by examining sea floor topography data is S.A. Littel, D.K. Smith, P.R. Shaw, R. Cawley and G-H. Hsu (Abstract) EOS, Oct. 29, 1991, p.465.

Attractors of Iterated Systems and Their Application to Image Compression

*E.W. Jacobs and R.D. Boss
Naval Command, Control and Ocean Surveillance Center*

Introduction

Because of the increasing use of digital imagery, there is currently considerable interest in the image compression problem. In particular, image compression is a current and growing necessity for Navy applications including storage and transmission of maps, intelligence photographs, weather information, etc. General interest in image compression has led to the establishment by the Joint Photographic Experts Group (JPEG) of a standard based on discrete cosine transforms (ADCT).¹ There is also an ongoing effort in the research community to design improved vector quantization^{2,3} methods, and to develop schemes that utilize wavelet transformations.^{4,5}

A novel approach to the image compression problem, iterated transformations, has been presented by Jacquin.⁶⁻⁸ The iterated transform algorithm has received particular interest because of its relationship to simple fractal generating algorithms, and in a more general sense, dynamical systems which possess strange attracting sets. The method has its foundation in the theory of iterated function systems (IFSs), developed by Hutchinson⁹ and Barnsley,¹⁰ and recurrent iterated function systems.¹¹ The image is encoded as a set of transformations

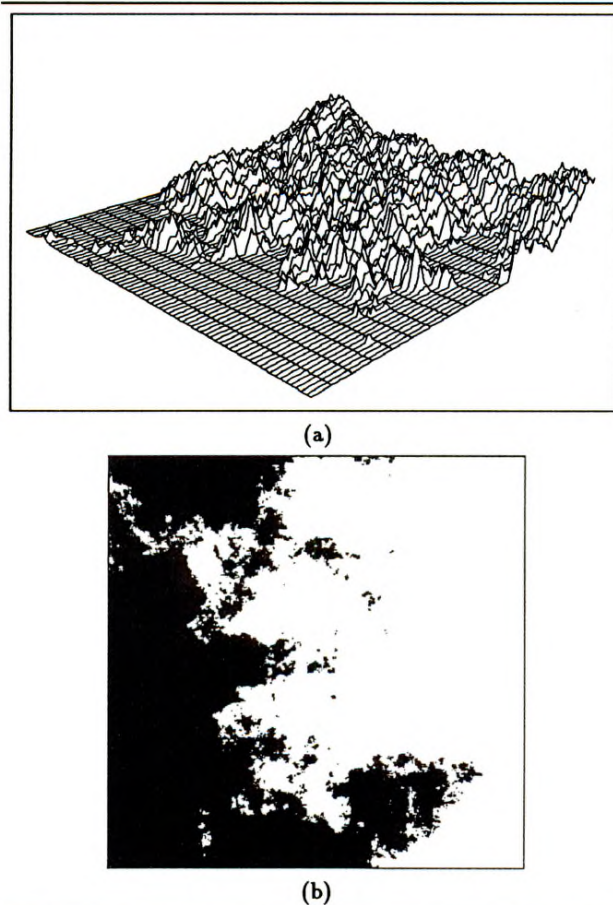
which, upon iteration, possess a fixed point that approximates the target image. The resulting decoded image can have fractal characteristics, i.e., by "zooming" into the image, features can be seen at finer and finer resolution. In this paper, an overview of some of the ideas leading to the development of the iterated transformation technique is presented. A brief description of the technique, some examples, and data that characterizes its performance relative to other methods follows.

Background

The work of Mandelbrot on fractal geometry has led to an avalanche of studies in a wide array of disciplines.¹² The driving motivation for this work is that nature often possesses fractal geometry over a wide range of scales. This fact is no more clearly evident than in natural imagery, which explains why some of the most widely disseminated work on fractal geometry has been that of application to generation of natural appearing images. Some of the most well known of these images are those made by R.F. Voss as part of reference 12. These images are visually stunning, but unfortunately, complex to construct. To better illustrate how complex images having fractal geometry can be generated from simple algo-

Figure 1.

a) A Brownian mountainous island-like surface, and b) the coastline for the island.



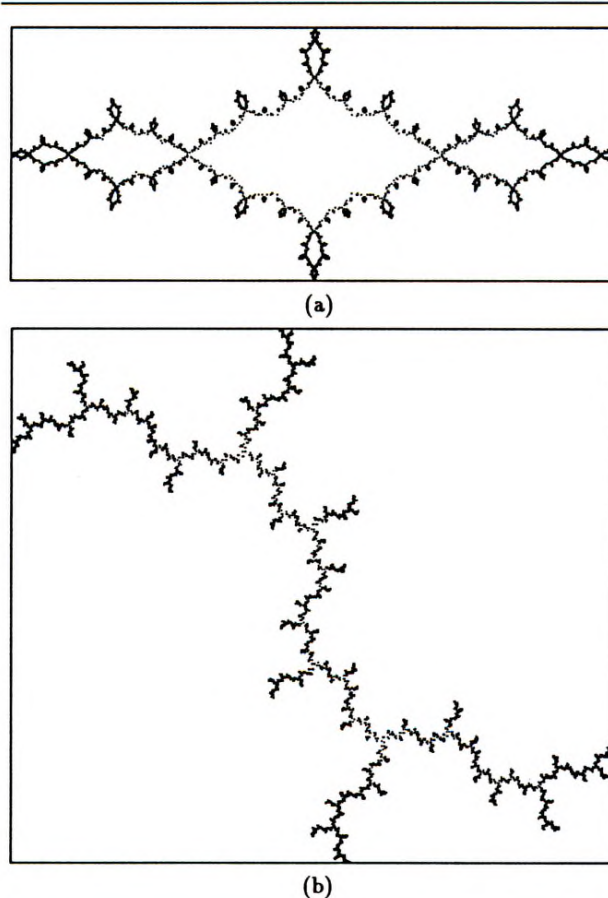
rithms, an example of a mountain (and coastline) generating algorithm, again due to Mandelbrot,¹³ is presented in the following paragraph.

Beginning with a flat surface (the xy plane), arbitrarily choose a line in the plane, and displace the part of the plane on one side of the line one unit in the z direction (thus generating a "fault"). Continue by repeating the process, where for the n th fault in the sequence, the height of the displacement in the z direction is scaled by $\frac{1}{\sqrt{n}}$. After choosing a sea level elevation, a surface generated by this algorithm is shown in figure 1a, and the corresponding coastline as viewed from above in figure 1b. Both the *Brownian surface* and the *Brownian islands* in figure 1 have fractal dimension, and they reproduce the qualitative features of real mountains and coastlines.

It is also well known that simple *iterative* systems can produce fractal images. Perhaps the most well known of these are the quadratic Julia sets (and the associated Mandelbrot set).

Figure 2.

a) The Julia set for $c = 1.2$, and b) the Julia set for $c = 1.0i$.



These Julia sets are the boundary between the basins of attractive fixed points for the iterative equation $z_{n+1} = z_n^2 + c$ (for a given value of c). For $c=0$, this Julia set is the unit circle (i.e., points in the complex plane inside the unit circle are attracted to the fixed point at zero, and points outside to infinity). For other values of c , the boundary is not so simple. A means to obtain a quick picture of this boundary for a given value of c is to "reverse" the iteration, thereby making the boundary an attracting set. That is, the map becomes $z_{n+1} = \pm\sqrt{z_n - c}$, where each root can be thought of as being a separate map;¹⁰

$$z_{n+1} = \omega_1(z_n) = \sqrt{z_n - c} \quad (1)$$

$$z_{n+1} = \omega_2(z_n) = -\sqrt{z_n - c}.$$

By starting with an arbitrary initial point z_1 , and at each iteration, arbitrarily choosing one of the two mappings, the attractor can be generated. Similarly, a deterministic method for generating the attractor can be performed by applying both

mappings to every point in the attractor at each iteration (thereby doubling the number of points in the attractor at each step). The resulting sets for $c = 1.2$ and $c = 1.0i$ are shown in figure 2a and b, respectively.

In the context of images, it is clear that to directly describe an image such as one of the attractors in figure 2 would require a great deal of information. In fact, for fractal images an infinite amount of information would be required to store the image. (In practice, images are digitalized, and the amount of information required to describe them is limited by the resolution of the display.) But it is seen in the previous example that the only information necessary to produce the images are the iterative equations 1, and the value of c . In essence, the iterative equations 1 are the *decoding* algorithm, and the *encoded* images are represented by the single number c . Consequently, an image, which requires a large amount of information to describe directly, can be described simply by specifying the associated value of c . This clearly represents an extremely effective *compression* for the image.

These examples serve to illustrate the power of simple fractal generating algorithms, and the concept of achieving compression using iterative equations with a limited number of parameters. But they do not solve the image compression problem. The first example in which a Brownian surface was constructed is a random process. The algorithm always generates mountain like images, but each time the algorithm is performed (with a different random sequence), a different image results. The Julia set example demonstrates how an image can be encoded in the form of parameters of an iterative equation, but it does not answer the question relevant to the image compression problem: How does one find an iterative system and its associated parameters which will encode a *particular* image?

Barnsley first suggested that, given a target image, iterative algorithms that possess (fractal) fixed points approximating the target image, should be sought.¹⁰ Barnsley called these algorithms iterated function systems (IFSs). He showed that for the purpose of image compression, affine transformations

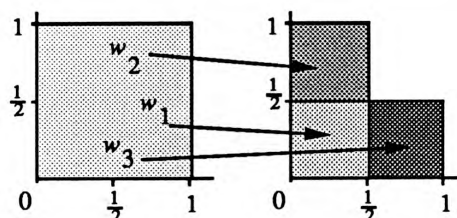
$$\omega_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix} \quad (2)$$

are useful for encoding images. The following IFS example serves as a simple illustration of some concepts involved in the iterated transform image-encoding scheme. The main concept is that the image of a set (a Sierpinski gasket, in this case) can be reconstructed from a set of affine transformations which may take less memory to store than the original image.

Consider the map which is made of the three transforms illustrated in figure 3. (The coefficients of the transformations are given in box 1.) Note that each of these transforms is *contractive*. That is, for any two points, they will be closer together after transformation than they were before the transformation. Transformations which fulfill such a contractivity

Figure 3.

Three affine transformations in the plane.



Box 1.

The transforms shown in figure 3 are:

$$w_1 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$w_2 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix},$$

and

$$w_3 \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}.$$

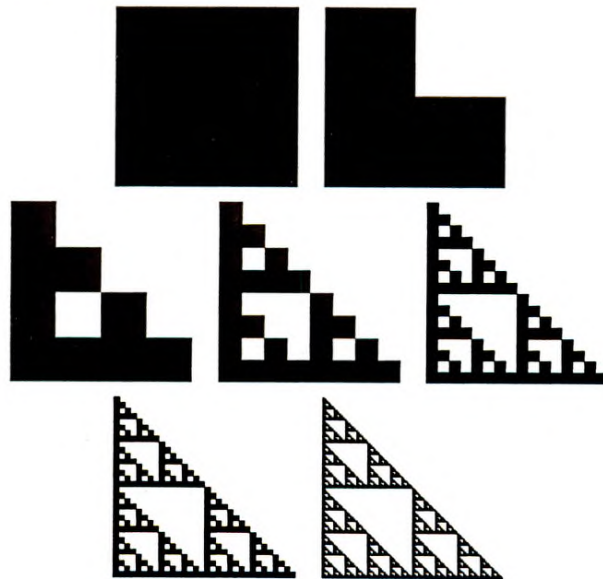
requirement will have an attractive fixed point which is independent of the initial starting point. This fact is stated by the contractive mapping fixed point theorem. That all compact initial sets converge under iteration to the same attractor is important—it means that the attractor is defined by the transforms only. Conceptually, this can be understood by realizing that, because the transformations are contractive, *all* detail of the initial image shrinks away as the mapping is iterated.

The first few iterations of the transforms of figure 3 (beginning with an *arbitrary* initial image) and the final attractor for these transformations are shown in figure 4. Each transform is determined by 6 real values, so that for this example 18 floating point numbers are required. In single precision, this requires 72 bytes. The memory required to store an image depends on the resolution; the image for the example above (the last image in figure 4) requires $256 \times 256 \times 1 \text{ bit} = 8192$ bytes of memory. Therefore, the resulting compression ratio in this simple example is 113.8:1.

The question that still must be answered is, given an image, what is the method for finding transformations that encode it? The contractive mapping fixed point theorem suggests how to answer this question. The contractive mapping fixed point theorem guarantees that, (under certain mathemat-

Figure 4.

An initial image, the first five iterates, and the fixed point for the three transforms shown in figure 3.



ical constraints) for a *contractive* mapping, when the mapping is iterated there will exist a unique fixed point. It is clear that the mapping, when applied to the fixed point, will result in the fixed point. Therefore, the transforms which constitute the mapping should be chosen so that they are contractive, and so that when the mapping is applied to the target image, the resulting image (which is called, after Barnsley, the collage) should be equivalent to the target image. It then follows that the fixed point of the mapping will approximate the target image. Referring back to the Sierpinski triangle example, it is seen that by satisfying this condition, the transformations encoding the image could be found.

In the Sierpinski triangle example, the collage of the image is perfect, resulting in an exact reproduction of the image. Given an arbitrary image, it is not possible in general to have a mapping which exactly reproduces the image with a small number of transforms. The obvious question is then: What happens if the reproduction is approximate? A corollary of the contractive mapping fixed point theorem, which Barnsley calls the collage theorem,¹⁰ puts a bound on the error between the final image (the fixed point) and the initial image, when the collage is not exact. The theorem says that the closer the collage is to the initial image, the closer the fixed point will be to the initial image, and that this is especially true if the transforms composing the mapping are very contractive.

The Sierpinski triangle, because of its inherent self similarity, is a particularly simple image to encode using the IFS technique. Writing a computer code to aid in encoding more

general images is straight forward. Such a code has been used to encode, and subsequently decode a digitized map of Point Loma (a peninsula at the opening of San Diego Bay on which the NCCOSC RDT&E Division is located).¹⁴ The digitized map of Point Loma, is shown in figure 5a. The computer program supplied an interactive graphics interface to aid the operator in choosing a set of transforms such that the collage approximated the target image. In figure 5b, the collage for the 48 transformations making up the mapping are shown (for clarity each transformation was applied only to the outline of the original image). When decoded, these transforms result in the image shown in Figure 5c. This encoding resulted in good fidelity with moderate compression.

It is important to note that the encoding program in the previous example did not find the transformations that encode the image automatically. Furthermore, in this example and the example of the Sierpinski gasket, the images are described by a set of pixels, each being either black or white. The problem of more interest for image compression applications is the encoding of gray scale images (i.e., an image in which each pixel has many possible gray levels, not just black or white). Two generalizations needed to make automated encoding of gray scale images feasible are discussed in the next section.

Encoding and Decoding Gray Scale Images Using Iterated Transformations

In the previous examples, each transformation operated on the entire image. The first generalization required to encode gray scale images is to restrict each transformation to operate on only a section of the image. The theory of IFS's has been extended by Barnsley and Jacquin to allow transforms to operate on only parts of the image rather than the entire image, in a method they call recurrent iterated function systems.¹¹ The particular section, or *domain*, upon which each transform acts must be stored as part of the encoded image. A key benefit of this *partitioning* of the image is that the encoding problem is broken down into a series of simple tasks which a computer can be programmed to perform.

The second generalization is that the transforms have to be generalized to three dimensions. A gray scale image can be thought of as a three-dimensional image, each pixel having a horizontal and vertical position, and an intensity value (x , y , and z , respectively). A convenient form of the transform for encoding gray scale images is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix} \quad (3)$$

Note that in this form the coefficient s_i acts to scale gray values, o_i acts to offset gray values, and the remaining coefficients

Figure 5.

The original Point Loma image, the collage, and the decoded image.

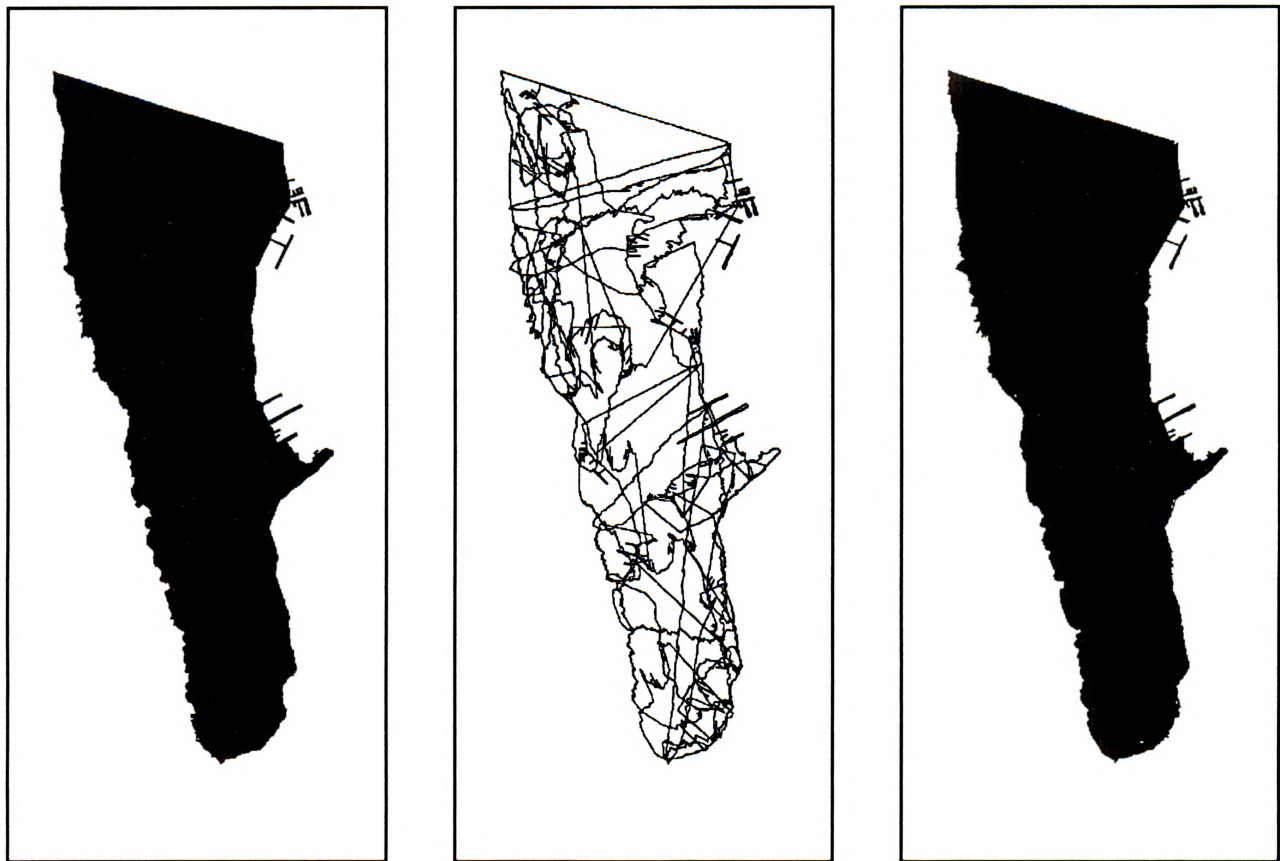


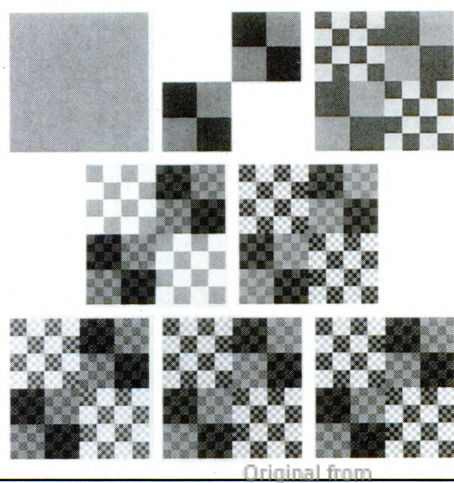
Table 1.

The 16 transforms which encode the fractal checkerboard (the attractor in figure6).

a	b	c	d	s	e	f	o
0.5	0.0	0.0	0.5	2.0	0.75	0.0	0.0
0.5	0.0	0.0	0.5	2.0	0.75	0.25	0.0
0.5	0.0	0.0	0.5	2.0	0.50	0.0	0.0
0.5	0.0	0.0	0.5	2.0	0.50	0.25	0.0
0.5	0.0	0.0	0.5	2.0	0.0	0.5	0.0
0.5	0.0	0.0	0.5	2.0	0.0	0.75	0.0
0.5	0.0	0.0	0.5	2.0	0.25	0.5	0.0
0.5	0.0	0.0	0.5	2.0	0.25	0.75	0.0
0.0	-0.5	0.5	0.0	0.25	0.25	-0.25	0.0
0.0	-0.5	0.5	0.0	0.25	0.25	0.0	0.25
0.0	-0.5	0.5	0.0	0.25	0.5	-0.25	0.25
0.0	-0.5	0.5	0.0	0.25	0.5	0.0	0.0
0.0	-0.5	0.5	0.0	0.25	0.75	0.25	0.0
0.0	-0.5	0.5	0.0	0.25	0.75	0.5	0.25
0.0	-0.5	0.5	0.0	0.25	1.0	0.25	0.25
0.0	-0.5	0.5	0.0	0.25	1.0	0.5	0.0

Figure 6.

An initial image, the first six iterates and the fixed point for the 16 transformations listed in table 1.



Box 2.

The image is encoded in the form of an iterative system (a space and a map from the space to itself) $W : F \rightarrow F$. The space F is a complete metric space of images, and the mapping W (or some iterate of W) is contractive. The contractive mapping fixed point theorem ensures convergence to a fixed point upon iteration of W . The goal is to construct the mapping W with fixed point "close" (based on a properly chosen metric $\delta(f, g)$) to a given image that is to be encoded, and such that W can be stored compactly. The collage theorem provides motivation that a good mapping can be found.¹⁰ Decoding then consists of iterating the mapping W from any initial image until the iterates converge to the fixed point.

Let $I = [0, 1]$ and I^m be the m -fold Cartesian product of I with itself. Let F be the space consisting of all graphs of real Lebesgue measurable functions $z = f(x, y)$ with $(x, y, f(x, y)) \in I^3$. Let $D_1, \dots, D_n$ and $R_1, \dots, R_n$ be subsets of I^2 and $v_1, \dots, v_n : I^3 \rightarrow I^3$ be some collection of maps. Define w_i as the restriction

$$w_i = v_i|_{D_i \times I}.$$

The maps $w_1, \dots, w_n$ are said to *tile* I^2 if for all $f \in F$, $\bigcup_{i=1}^n w_i(f) \in F$. This means the following: for any image $f \in F$, each D_i defines a part of the image $f \cap (D_i \times I)$ to which w_i is restricted. When w_i is applied to this part, the result must be a graph of a function over R_i , and $I^2 = \bigcup_{i=1}^n R_i$. This means that the union $\bigcup_{i=1}^n w_i(f)$ yields a graph of a function over I^2 , and that the R_i 's are disjoint. The map W is defined as

$$W = \bigcup_{i=1}^n w_i. \quad (1)$$

Thus the encoding process is: Partition I^2 into a set of ranges R_i . Since the goal is to limit the memory required to specify W , the set of R_i should be geometrically simple. For each R_i , a $D_i \subset I^2$ and $w_i : D_i \times I \rightarrow I^3$ is sought such that $w_i(f)$ is as δ close to $f \cap (R_i \times I)$ as possible; that is,

$$\delta(f \cap (R_i \times I), w_i(f)) \quad (2)$$

is minimized. The map W is determined by specifying the maps w_i . The w_i 's must be chosen such that W is contractive, or some iterate of W is contractive (i.e., W is eventually contractive).

A brief explanation of how a transformation $W : F \rightarrow F$ can be eventually contractive but not contractive is in order. The map W is composed of a union of maps w_i operating on disjoint parts of an image. If any of the w_i are not contractive, then W will also not be contractive. The iterated transform $W^{\circ m}$ is composed of a union of compositions of the form

$$w_{i_1} \circ w_{i_2} \circ \dots \circ w_{i_m}.$$

Since the product of the contractivities bounds the contractivity of the compositions, the compositions may be contractive if each contains sufficiently contractive w_{i_j} . Thus W will be eventually contractive if it contains sufficient "mixing" so that the contractive w_i eventually dominate the expansive ones.

represent the component of the transformation in the xy plane only. The following example illustrates how these two generalizations can result in a method for encoding gray scale images.

Consider the map consisting of the sixteen transformations given in table 1. The first eight transformations are restricted to act on the lower left quarter of the region (i.e., (x, y) such that $0 \leq x \leq 1/2$ and $0 \leq y \leq 1/2$), and the second eight transformations are restricted to act on the lower right quarter of the region (i.e., (x, y) such that $1/2 \leq x \leq 1$ and $0 \leq y \leq 1/2$). Let values of $z = 0$ be represented as black, and $z = 1$ as white, with intermediate values as shades of gray. The initial image is arbitrarily chosen as a completely flat gray image which is shown in figure 6a. The first six iterates, and the fixed point are shown in figure 6. In practice, the image is not represented at infinite resolution, but rather is discretized at some small number of intervals in each direction. When the image in this example is discretized as 128×128 pixels, with 8-bits per pixel (i.e., 256 gray levels), the encoder used herein automatically encodes this image (using an equivalent set of 16 transformations) with the resulting compression equal to 356:1.

A mathematical description of the encoding process is presented in box 2. Two pseudo codes for encoding images are presented in box 3. The interested reader should study boxes 2 and 3 for a more thorough understanding of the encoding algorithm. In the following paragraph, a simpler, albeit less complete explanation of the encoding process is presented.

In figure 7 part of the encoding process for this image is illustrated. The figure demonstrates how one section of the image, called a range, is covered as closely as possible by applying a transform (of the form of equation 3) to a section of the image called a domain. (In the figure, the domain, transformed domain, and range squares are expanded for clarity. In actuality, the transformation maps the section of the image defined by the domain square onto the section of the image defined by the range square.) To complete the encoding process, a transform and domain must be found to cover each range, and the ranges must completely *tile* the image. To facilitate compact specification of the transformations, the sets from which domains and ranges are chosen are restricted to be geometrically simple, and limited in number. As shown in figure 7, all the examples in this paper will use square ranges

Box 3.

Two pseudo-codes for an adaptive encoding algorithm.

- Choose a tolerance level e_c .
- Set $R_1 = I^2$ and mark it uncovered.
- While there are uncovered ranges R_i do {
 - Out of the possible domains $\mathbf{D}$, find the domain D_i and the corresponding w_i which best covers R_i (i.e., which minimizes expression 2).
 - If $\delta(f \cap (R_i \times I), w_i(f)) < e_c$ or $\text{size_of}(R_i) \leq r_{\min}$ then
 - Mark R_i as covered, and write out the transformation w_i ;
 - else
 - Partition R_i into smaller ranges which are marked as uncovered, and remove R_i from the list of uncovered ranges.

a. Pseudo-code targeting a fidelity e_c .

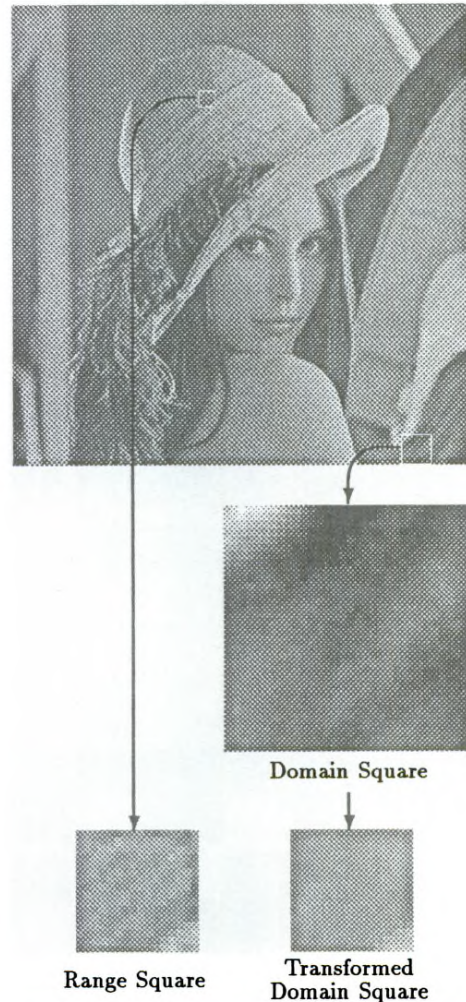
- Choose a target number of ranges N_r .
- Set a list to contain $R_1 = I^2$ and mark it covered.
- While there are less than N_r ranges in the list do {
 - Out of the list of ranges, find the range R_j with $\text{size_of}(R_j) > r_{\min}$ which has the largest $\delta(f \cap (R_j \times I), w_j(f))$ (i.e., which is covered worst).
 - Partition R_j into smaller ranges which are added to the list and marked as uncovered.
 - Remove R_j, w_j and D_j from the list.
 - For each uncovered range in the list, find and store the domain $D_i \in \mathbf{D}$ and map w_i which covers it best.
- Write out all the w_i in the list.

b. Pseudo-code targeting a compression having N_r transformations.

and domains. The transforms must be chosen such that upon iteration, a fixed point is reached. In light of the collage theorem which requires contractivity of the transformation and says that (all other things being equal) the more contractive the mapping the better, it may be surprising that when the mapping is constructed, it is not necessary to impose any contractivity conditions on the individual transforms. The only necessary contractivity requirement is that the mapping be eventually contractive.^{15,16} A description of what it means to be eventually contractive is given in box 2. Note that in the fractal checkerboard example above, half of the transforms are not contractive in the z direction. (i.e., $s_i > 1$), but the total transformation is eventually contractive, and thus a fixed point exists.

Figure 7.

Part of the encoding process.



As shown for the simple examples in the previous section (figures 4 and 6), decoding an image is performed by starting with an arbitrary initial image, and iterating the transformations until the fixed point is reached. This process is shown for an encoding of "tank farm" in figure 8. The compression of this 256 x 256, 8-bit per pixel image was 8.66:1, and the peak signal to noise ratio ($PSNR = -20 \log \left(\frac{\text{rms error}}{255} \right)$) was 33.6dB.

Typical images are not homogeneous. In some parts of an image there may be lots of detail, and in other areas, the image may appear flat. Therefore, it is advantageous to have an algorithm which uses ranges that can vary in size depending on the local image complexity. Allowing for ranges that can vary in size leads to a trade-off between compression and fidelity. For a given image, more transformations lead to better

Figure 8.

Initial image, first iterate, second iterate, and tenth iterate for an encoding of the "tank farm" image.

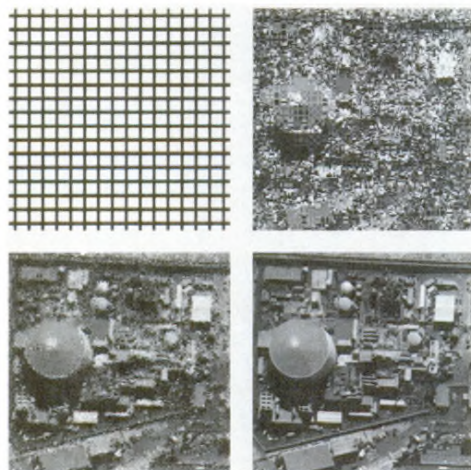
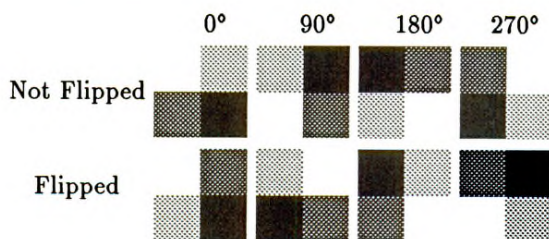


Figure 9.

The eight possible orientations for a square.



fidelity, but worse compression. This trade-off between compression and fidelity leads to two approaches to encoding an image, one targeting fidelity and one targeting compression. These approaches are outlined in the pseudo-codes in box 3.

Clearly, the selection of the domains, ranges, and transforms is one of the most important parts of an automated fractal encoding scheme. The notation D and R will be used to denote the set of all *potential* domains and ranges, respectively. Since the goal of the encoding is compression, compact specification of the transforms is a primary concern. To limit memory, only transforms of the form given by equation 3 are used. Furthermore, as stated above, D and R must be restricted to be geometrically simple. Restricting D and R to be collections of squares greatly reduces the information needed to specify the coefficients a_i , b_i , c_i , and d_i in equation 3. This can be seen in figure 9, where the eight possible orientations for mapping one

square onto another are shown. The coefficients e_i and f_i , are determined by the relative position of the range and associated domain. The values of the intensity transforming coefficients (s_i and o_i) are computed by least squares regression to minimize the collage error using the rms metric (i.e., expression 2 in box 2 is minimized), and are stored using a fixed number of bits. One could compute the optimal coefficients and then adjust them to the nearest allowed value for storage. However, a significant improvement in fidelity can be obtained if only storable values are used when computing the error during encoding.

Another concern is encoding time, which can be significantly reduced by employing a classification scheme on the ranges and domains. Both ranges and domains are classified using some criteria such as their edge-like nature, or the orientation of bright spots, etc. Considerable time savings result from only using domains in the same (or similar) class as a given range when seeking a cover, the rationale being that domains in the same class as a range should cover it best.

While varying the number of ranges directly affects the compression, varying the number of possible domains has a more complicated effect. It may seem that having a larger set D should decrease the compression since more information is required to specify the particular domain used. Likewise, it may seem that the encoding time must be increased since more domains must be searched. But with more domains, a range which would otherwise be partitioned may be covered, so that fewer ranges need to be covered, resulting in time and memory savings. Similarly, using more domains does not guarantee a better fidelity, since an other wise partitioned range may not be partitioned and thus covered more poorly.

Examples of Compression of Gray Scale Images

The image in figure 10a has been encoded in a simple way. The image is a 256 x 256 pixel image with 8-bits per pixel. The image was subdivided into 1024 non-overlapping squares (each square is 8 pixels by 8 pixels in size). These squares were used as the ranges for the encoding process. The set D was chosen to be all possible 16 pixel squares (there are 241 x 241=58081 of these squares). The encoding process consisted of searching through all domains to find the domain and corresponding optimal transformation which best covered each range. Each domain was tested in all eight possible orientations (see figure 9). The encoding process was performed using four values for the maximum allowed intensity scaling factor. The results of these encodings are given in table 2.

As stated above, the collage theorem suggests that it should be disadvantageous to allow for less contractive mappings. However, the results given in table 2 show that (for the rms metric), even though the covering accuracy improves only slightly, the resulting fixed point improves while the con-

Figure 10.

a) Original Mara, and b) decoded Mara with $s_{\max}=1.5$

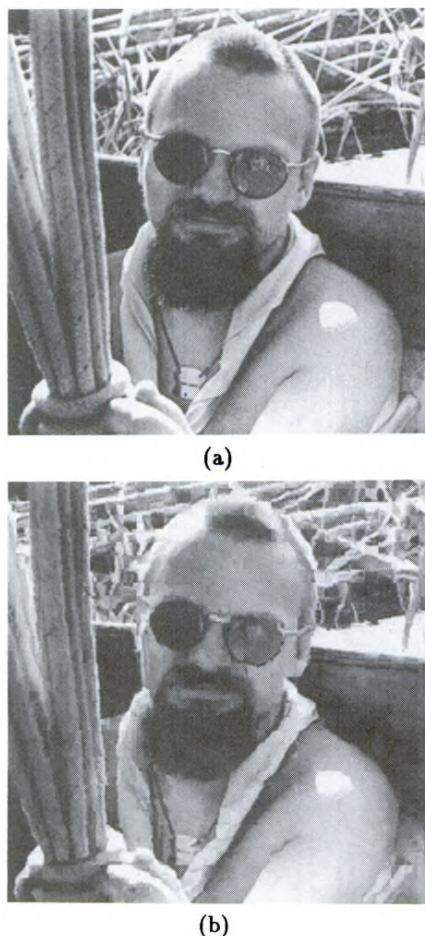


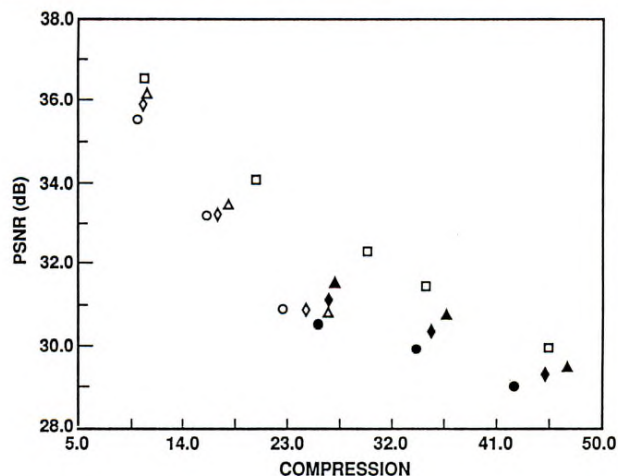
Table 2.

Root mean square errors for the collage and final attractors for a few simple encodings of Mara (see figure 10a). In each case the same ranges were used to encode the image with the best possible transform and domain used for each.

$s_{\max}$	collage error	attractor error
0.7	20.04	21.48
0.9	19.77	21.16
1.2	19.42	21.16
1.5	19.42	20.97

Figure 11.

PSNR versus compression as a function of e_c , $r_{\min}$, and n_c for 512×512 Lena. Open symbols represent $r_{\min}=4$ and solid symbols $r_{\min}=8$. Data for $e_c=5.0$, 8.0 , and 11.0 appear in clusters with increasing compression. Within each cluster, $n_c=1(\circ)$, $4(\diamond)$, and $72(\Delta)$. ADCT data ($\square$) is shown for comparison.



tractivity of the mapping decreases. Indeed, the last two results given in table 2 are for mappings which are not even contractive (although they are eventually contractive). These results show that the bounds given by the collage theorem should be viewed as motivation rather than as a real constraint.

A generalization of the previous example is to select ranges from a quad-tree partitioning of an image. A quad-tree partition is one in which a square may be subdivided into four smaller squares, each with one fourth the area of the original square. In reference 15, results for such an implementation with an assortment of adjustable parameters have been presented. A brief description of the encoder, and some results are presented below.

The encoder essentially followed the pseudo-code of box 3a. The domain squares had variable size and were restricted to have corners on a lattice with a fixed vertical and horizontal spacing. The domain set also included domains with sides slanted at 45 degree angles from the edges of the image. A classification scheme was used to classify all possible domains into classes. The classes were defined by the brightness and edge-like character of the quadrants of each square. Choosing a classification scheme that ordered the quadrants of a domain by their brightness determined a symmetry operation which limited the search for a good cover, since only one of the eight possible orientations was checked. When covering a range, domains (with side length twice that of the range square) from a fixed number of classes (n_c) were considered; those classes

Figure 12.

The original 512 x 512 8-bit per pixel image of Lena, and images generated using three different compression methods: a) Original Lena, b) Iterated Transformations, compression=32.05:1, PSNR = 31.16dB, c) ADCT, compression=31.71:1, PSNR = 31.98dB, and d) MRVQ, compression=32:1, PSNR = 28.38.



a) Original Lena.



b) Iterated Transformations, compression=32.05:1, PSNR=31.16dB.



c) ADCT, compression=31.72:1, PSNR=31.98dB.



d) MRVQ, compression=32:1, PSNR=28.38.

Table 3.

Relative time to encode Lena for various parameters. See text for details concerning the times.

s_{max}	n_c	r_{min}	e_c	size	time	Comp.	PSNR
1.2	1	4	8	512	1.0	15.95:1	33.13
1.2	4	4	8	512	3.1	17.04:1	33.19
1.2	72	4	8	512	35.5	17.87:1	33.40
1.0	4	4	8	512	3.1	16.74:1	33.30
1.2	4	4	5	512	5.3	10.49:1	35.92
1.2	4	4	11	512	2.0	24.62:1	30.85
1.2	72	8	8	512	7.5	36.78:1	30.71
1.2	1	4	8	256	0.14	9.09:1	30.63
1.2	72	4	8	256	4.5	9.97:1	31.53
1.2	72	4	10	256	3.7	11.85:1	30.58

were chosen to be the “most” similar to the classification of the range.

Other adjustable parameters were the target collage error (e_c), the minimum allowable range size r_{min} , and the maximum allowable scale factor s_{max} . When a range square of minimum size is covered, the best possible transform is stored, regardless of whether or not the error is less than e_c . For data presented here, s_i was stored with five bits, and o_i was stored with seven bits.

Figure 11 presents PSNR versus compression data of the standard 512 x 512 Lena (shown in figure 7). The compression was varied by adjusting e_c and r_{min} . Data is also shown for different values of n_c . The value of e_c separates the data into three widely spaced clusters, with smaller e_c leading to higher PSNR. For comparison, results are also shown for an ADCT algorithm similar to that described in reference 17. The ADCT data is comparable to the JPEG standard, being slightly better at higher compression.¹

In table 3, the relative encoding times for several encodings are presented. The data indicate the relative encoding time as a function of n_c , e_c , r_{min} , and image size. On an HP-Apollo 400t workstation, the relative time units in the table are approximately 1170 cpu seconds. The programs have not been optimized for speed.

In figure 12, the original Lena image, along with versions compressed using iterated transformation, ADCT, and mean-residual vector quantization (MRVQ), are shown. The iterated transformation encoding was performed with $r_{min} = 8$, $s_{max} = 1.2$, $n_c = 72$, and $e_c = 6.5$. The ADCT implementation was the same as that used to produce the data of figure 11. The only significant difference between this algorithm and the JPEG standard is that the algorithm used here segments the image into 16x16 squares, where as the JPEG standard uses 8 x 8 squares. Therefore the box-like artifact pattern resulting from the segmentation is different than that obtained with a JPEG

implementation. The MRVQ image was generated using the basic algorithm of reference 2. The code book (256 8 x 8 code vectors) for the encoding was generated from a set of five images, including faces, a crowd scene, and a couple standing in a room. So that the artifacts characteristic of each method can be observed, no post-processing was performed on any of the encodings that are shown. The presentation of these images is not for the purpose of a critical quantitative test between iterated transformations, ADCT, and VQ methods. Considering the vast array of the VQ algorithms, and the wide range of variations of the iterated transform algorithm that are easily implemented, such a critical evaluation would be difficult. Furthermore, when evaluating which image compression method to use, the requirements of a particular application may well dictate the choice. Nevertheless, these images do illustrate the differences in the images generated using the various algorithms, and indicate the relative level of compression/fidelity performance.

Conclusion

Some of the basic motivations and concepts leading to the development of an image compression method in which the decoded image is the fixed point of a set of iterated transformations have been presented.

Results of the implementations described above for iterated transformation encoders illustrate several interesting points. By adjusting appropriate parameters, encodings with reasonable compression/fidelity performance can be generated over a large range of compressions. Allowing non-contractive transformations can result in improved encodings even though this is contrary to what the collage theorem suggests. Therefore, the collage theorem provides motivation that a good encoding can be found, not a mathematically useful bound on the error. Also, based strictly on compression versus PSNR performance, current implementations do not outperform an ADCT method similar to the JPEG standard.

Although iterated transform image encoding is computationally intensive, the decoding process is computationally simple compared to ADCT methods. Because the iterated transform scheme is based on a self referential coding of an image as opposed to a fixed code book approach, it is well suited for the encoding of a more general class of images than vector quantization algorithms. Therefore, iterated transforms may be best applied in applications for which a large variety of pre-encoded images (possibly supplied by a central data base) need to be decoded quickly with software on off-the-shelf computer platforms. When confronted with a problem that requires image compression, there are several methods (some of them brushed upon in this paper) from which to choose. Depending on the application, iterated transformations may be a viable image compression option. Because the

method is relatively young, variations or improvements of the method may lead to more general application.¹⁸

In implementing the iterated transform method, the selection of classification method and scale and offset parameters are important, but it is the choices of the partitions for **R** and **D** which are least obvious. Very simple schemes work well for the same reason as vector quantization methods, i.e., the information in each transformation is relatively small, so that a large number of transformations can be used while maintaining reasonable compression. More elaborate partitions which might better exploit the self similar properties of iterated transform encodings must be stored as part of the encoding, and therefore become burdened by their own complexity. Improvements attained from generalizing a simple approach (such as the first example in the previous section) to a more elaborate approach (such as the quadtree method and an even more general partition utilizing general rectangles¹⁶) are not negligible, but small. When devising an iterated transform encoder, the issue of optimal algorithmic complexity may be the most interesting problem.

Acknowledgements

The authors wish to acknowledge the significant contributions to this work made by Dr. Y. Fisher while working at NOSC as an ONT postdoctoral fellow. The authors also wish to acknowledge the support of Drs. Ken Campbell, John Silva, and Randy Moore at NCCOSC, and Jim Smith at ONR.

Biography

Bill Jacobs received his Bachelors degree in Physics and Masters degree in Applied Physics from U.C. San Diego in 1981 and 1986 respectively. He has worked in the Materials Research Branch of the NCCOSC RDT&E Division since 1981, in which time he has studied a variety of research topics. Some of these included: properties of piezo-electric polymers; properties of high temperature superconducting ceramics; chaotic and stochastic effects in non-linear dynamical systems; and fractal based image compression.

Roger Boss received his B.S. from Kent State University and his Ph.D. in Analytical Chemistry from Michigan State University in 1980 and 1985, respectively. He has worked in the Materials Research Branch of the NCCOSC RDT&E Division since 1985. His past research interest have included: non-aqueous solution chemistry; spectroelectrochemistry of electron transfer; conducting polymers; high temperature superconducting ceramics; chaotic and stochastic effects in neurons; and fractal based image compression. His current research involves macromolecular solid state chemistry.

They have received the Office of Naval Technology award for best FY91 Independent Exploratory Development project for their work on iterated transformations.

Address: Naval Command, Control & Ocean Surveillance Center, RDT&E Division, San Diego, CA 92152-5000, e-mail: jacobs@gandalf.nosc.mil.

References

1. M. Rabbani and P.W. Jones, *Digital Image Compression Techniques*, SPIE Optical Engineering Press, Bellingham, Washington, 1991, Ch. 10, p. 121.
2. Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. on Comm.*, vol. COM-28, no. 1, pp. 84-95, (1980).
3. R.M. Gray, P.C. Cosman, and E.A. Riskin, "Image Compression and Tree-Structured Vector Quantization," chapter 1 in *Image and Text Compression*, J.A. Storer editor, Kluwer Academic Publishers, (1992).
4. S.G. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models," *IEEE Trans. on ASSP*, Vol. 37, p 2091, (1989).
5. R.A. DeVore, B. Jawerth, and B.J. Lucier, "Data Compression using Wavelets: Error, Smoothness, and Quantization," *Proceedings of Data Compression Conference '91*, J.A. Storer and R.H. Reif editors, IEEE Computer Society Press, (1991), p. 186.
6. A.E. Jacquin, "A Fractal Theory of Iterated Markov Operators, with Applications to Digital Image Coding," Ph.D. Thesis, Department of Mathematics, Georgia Institute of Technology, (1989).
7. A.E. Jacquin, "Fractal Image Coding Based on a Theory of Iterated Contractive Image Transformations," *SPIE Visual Comm. and Image Processing '90*, Vol. 1360, pp. 227-239, (1990).
8. A.E. Jacquin, "A Novel Fractal Block-Coding Technique for Digital Images," *1990 IEEE ICASSP*, Vol. 4, pp. 2225-2228, (1990).
9. J.E. Hutchinson, "Fractals and Self-Similarity," *Indiana University Mathematics Journal*, vol. 35, p. 5, (1981).
10. M.F. Barnsley, *Fractals Everywhere*, Academic Press, Inc., San Diego, CA, (1988).
11. M.F. Barnsley and A.E. Jacquin, "Application of Recurrent Iterated Function Systems to Images," *SPIE vol. 1001, Visual Comm. and Image Processing*, p. 122, (1988).
12. B.B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman and Co., San Francisco, CA, (1982).
13. B.B. Mandelbrot, *Stochastic Models for the Earth's Relief, the Shape and the Fractal Dimension of the Coastlines, and the Number-area Rule for Islands*, Proc. Nat. Acad. Sci. USA, Vol. 72, p. 3825-3828, (1975).
14. R.D. Boss and E.W. Jacobs, "Fractal-Based Image Compression," NOSC TR 1315, September 1989.
15. E.W. Jacobs, Y. Fisher, and R.D. Boss, "Image Compression: A Study of the Iterated Transform Method," to be

published in *Signal Processing*, vol. 29 no. 3, pp. 251-263 (1992).

16. Y. Fisher, E.W. Jacobs, and R.D. Boss, "Fractal Image Compression Using Iterated Transforms," chapter 2 in *Image and Text Compression*, J.A. Storer editor, Kluwer Academic Publishers, Norwell, MA, (1992).
17. W. Chen and W.K. Pratt, "Scene Adaptive Coder," *IEEE Transactions on Communications*, vol. 32, pp. 225-232, (1984).
18. *PC Magazine*, April 28, 1992, p. 337, and November 24, 1992, p. 42, describe commercially available proprietary fractal based image compression tools.

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
ARLINGTON, VA 22217

OFFICIAL BUSINESS