SIGNAL MODELLING: A VERSATILE APPROACH FOR THE AUTOMATIC ANALYSIS OF THE ELECTROENCEPHALOGRAM

by

Socrates Andreou Mylonas

A Thesis Submitted for the Degree of
Doctor of Philosophy

THE CITY UNIVERSITY
Centre for Information Engineering
July, 1995

Contents

List of Figures

List of Tables

Abstract

Despite recent advances in brain monitoring techniques, the electroencephalogram (EEG) is still widely used in the diagnosis and monitoring of epilepsy. To increase its effectiveness, long-term monitoring of patients was proposed but the large volume of recorded EEG signals produced, made their traditional interpretation by human experts difficult and automatic EEG analysis was proposed as an alternative.

This Thesis is concerned, primarily, with the on-line detection of epileptic transients (spikes) in the interictal EEG signals of patients. A review of previous methods, revealed that the limited success of automatic analysis systems was linked to the vagueness of neurophysiological definitions and the subjectiveness of human interpretation, which is based on experience.

To address these issues, it was realized that a common point of reference is required for the integration of medical and signal processing expertise, which could be provided by a model of the signal. Early attempts to develop such a model are described. These led to the development of spike detectors based on the derivatives of the EEG.

Later, by describing medical definitions with signal processing terminology, a comprehensive model of the signal was constructed. This was based on its decomposition into background activity, spikes, transients and noise and describing each one of them in terms of simple, random signals and quasi-linear systems.

This suggested a method of analysis based on inverse modelling for the decomposition of the EEG. The model for transients was estimated off-line. An on-line system, consisting of adaptive prediction error systems, constrained all-pole adaptive systems and a basic signal detection procedure was implemented. Several alternative adaptive realizations were investigated.

The spike detection procedure was generalized for the detection of other transients. Finally this procedure was replaced by a Multi-Layer Perceptron neural network, whose inherent ability to learn by example is important, as it provides the means to incorporate medical experience without requiring its explicit quantification. The system is flexible and its extension to detect any number of transients is demonstrated. The method may be applied to other signals and improved by new developments in signal processing.

Acknowledgements

Starting and completing my research and writing this Thesis required commitment, patience and hard work. Maintaining the necessary enthusiasm and persistence through the years, from its beginning to its completion, would have been impossible without support from the following people to which I wish to express my gratitude:

My family, in general, for their continuing support, especially my father, Andreas, for encouraging me and my sister to study and for undertaking most of the cost and my mother, Ariadne, for her invaluable advice and for giving me courage and patience to overcome the obstacles I encountered.

I would also like to thank my supervisor, Dr. Richard Comley, for his support during my doctorate and for encouraging me to follow my own direction into research. Also for his comments on the manuscript of the Thesis.

During my studies, numerous friends and colleagues, especially those in the Machine Vision Group have been available for discussions of new ideas. Frequent lively debates, not restricted to technical matters, served both as a source of inspiration and distraction from a long, monotonous task.

I would also like to express my gratitude to the technical staff, especially to Brian Burns for his prompt response to any arising technical problem and for maintaining computers and other equipment in working order, despite the odds. I am also indebted to Evelyn Cueto, for her support and encouragement during the last stages of my studies and for proof reading this report.

Finally, I would like to thank the committee of Vice Chancellors and Principals for providing me with an ORS award during the first three years of my studies.

Declaration

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Symbols and abbreviations

∇	Gradient operator for functions of a vector
$\mathcal{Z}$	the z-transform operator
$\cdot\mathbf{x}$	Gradient of vector $\mathbf{x}$
$\mathbb{N}$	The set of natural numbers $\{1, 2, \ldots\}$
$\mathbb{R}$	The set of real numbers
$\mathbb{Z}$	The set of integer numbers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$u_0(t)$, $\delta(t)$	Dirac (impulse) function (also unit sample for discrete-time signals)
$\mathbf{x}$	Vector $(\ldots, x_0, x_1, x_2, \ldots)^T$
A	Matrix A
A^T	Transpose of A
$r_{xx}(kT)$	Auto-correlation function of $x(nT)$
$S_{xx}(\omega)$	Power Spectral Density of $x(nT)$
R_{xx}	Auto-correlation matrix of $x(nT)$
$\{x_i\}$	Sequence, $\ldots, x_0, x_1, x_2, \ldots$
γ	Partial correlation (reflection) coefficient
$e_p^f(n)$	forward prediction error output of the pth stage of a lattice filter
$e_p^b(n)$	Backward prediction error output of the pth stage of a lattice filter
x'	Derivative of x
$E[\cdot]$	Expectation operator
σ^2, σ	Variance, Standard Deviation
$H_i(z)$	Transfer function for background activity model[1]
$G(z)$	Transfer function of spike generating system[1]
$\mathbf{F}(z)$	Transfer function vector for transient generating systems[1]
$e(n)$	EEG signal[1]
$u(n)$	Background activity generating sequence[1]
$b(n)$	Background activity signal[1]
$d(n)$	Epileptic spike generating sequence[1]
$s(n)$	Spike signal[1]
$\mathbf{c}(n)$	Vector transient generating sequence[1]
$t(n)$	Composite transient signal[1]
$v(n)$	Noise signal[1]
$r(n)$	Output of $H_i^{-1}(z)$[1]

$y(n)$	Output of $G^{-1}(z)$[1]
$y_i(n)$	Output of $F_i^{-1}(z)$[1]
$z_t(n)$	Transient detection signal (binary)[1]
$z_s(n)$	Spike detection signal (binary)[1]
α	Alpha brain activity (8–13 Hz)
β	Beta brain activity ($>$ 13 Hz)
θ	Theta brain activity (3–8 Hz)
δ	Delta brain activity ($<$ 3 Hz)
AC	Alternating Current (also signifies the varying part of a signal)
ADC	Analogue to Digital Converter/Conversion
ANN	Artificial Neural Network
AR	Autoregressive
ARMA	Autoregressive Moving-Average
CAT	Computed Axial Tomography
DC	Direct Current (also signifies the constant part of a signal)
DFT	Discrete Fourier Transform
ECG	Electrocardiogram (also EEG artifacts related to cardiac potentials)
ECoG	Electrocorticogram
EEG	Electroencephalogram, Electroencephalography
EMG	Electromyogram (also muscle artifact)
EOG	Electro-oculogram (also ocular artifact)
EPSP	Excitatory Postsynaptic Potential
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Frequency Modulation
IFSECN	International Federation of Societies for Electroencephalography and Clinical Neurophysiology
IIR	Infinite Impulse Response
IPSP	Inhibitory Postsynaptic Potential
MA	Moving-Average
MAE	Maximum Absolute Error
MLP	Multi Layer Perceptron
MRI	Magnetic Resonance Imaging

[1]Used extensively in Chapters 7 to 11

MSE	Mean Squared Error
PSD	Power Spectral Density
RAS	Reticular Activation System
REM	Rapid Eye Movement (sleep stage)
RMS	Root Mean Square
RLS	Recursive Least Squares (optimization)
QRS	Pattern caused by heartbeat in the ECG

Introduction

The brain consists of a vast network of highly interconnected functional elements. It is the centre of the nervous system, which spans the whole of the body controlling its actions. As a consequence, deviations from normal brain behaviour often disrupt other functions of the body, leading to complex conditions with variable symptoms.

One of the early methods of brain monitoring was the Electroencephalogram (EEG), a record of the evolution of the electrical activity of the brain taken by an inexpensive, non-invasive procedure. It has been known for almost a century and several characteristic EEG patterns have been identified and correlated with specific brain conditions. Unlike modern brain imaging methods which show the structure of the brain, the EEG is a record of its operation and is hence more suitable for investigating abnormalities in its *function*. Despite great advances in medicine and diagnostic techniques some of these conditions are among the least well understood and most difficult to treat.

One such disorder is *epilepsy*. Although known since ancient times, some of its aspects are still uncertain. In a number of cases, it is a symptom of another condition, like a brain tumour, and seizures are discontinued once the real cause is diagnosed and cured. In many other patients, seizures are the only manifestation of a transient abnormal brain behaviour. *Constitutional* (idiopathic) epilepsy is possibly the result of a variety of interacting and ambiguous factors and the effectiveness of its treatment may not always be satisfactory.

Perhaps the main obstacle in the investigation of idiopathic epilepsy is the transient nature of the brain disruption, which appears to be spontaneous and last for short periods of time, in some cases only for seconds. In a large number of sufferers, there is no residual evidence of an attack and even diagnosis is difficult. Medical experts agree, however, that an early diagnosis, especially in childhood, is important because if treated at an early stage, epilepsy may be easier to control or even eliminate. Several anticonvulsant drugs have been developed that can reduce the number of seizures, but most of them have serious side effects and the required dosage must be adjusted on an individual basis.

It is therefore important to have an effective technique to assist both in the initial diag-

nosis and in the subsequent monitoring, especially during drug therapy. It has been discovered that epileptic seizures are generally accompanied by abnormal changes in the electrical potentials of the brain, which become evident in the EEG. Their identification is a good indication of whether a person suffers from epilepsy or not. Some of these patterns appear spontaneously between seizures and may be used to monitor the effectiveness of medication.

The usefulness of clinical EEG in epilepsy is sometimes questioned. It has been argued that a traditional recording, which often lasts for less than half an hour, is too brief for the effective screening of patients. Although it is possible to provoke abnormal patterns, the results are not always conclusive and further recordings are sometimes necessary for confirmation. One cannot exclude the possibility that seizures may be provoked by factors not found in the EEG laboratory, but present in the patient's normal working environment. It is suspected, for example, that minor epileptic attacks in children may be associated with boredom, which is unlikely to occur during a routine recording.

The proposition to monitor patients over longer periods of time, even round the clock, in some more 'normal' environment, using telemetry, found limited use, as it revealed other problems. Traditionally, the EEG is analyzed by an experienced specialist who inspects its recording on paper and produces a report of his/her findings. Long-term recordings produce a large volume of data, which require long hours of inspection to be assessed. This is a costly procedure, demanding the constant attention of the expert over long periods of time. As a consequence fewer patients may be screened in a given time interval and the monotonous task is likely to cause fatigue, deeming the expert less reliable as his/her concentration is gradually reduced. With the increase of the work load in many hospitals, resulting in the use of inexperienced personnel to carry out some of the analysis tasks, using an expert analyst's time in such a way is highly undesirable. Another problem is the presence of artifacts. These are interfering signals that are picked-up by the recording equipment and make interpretation difficult, sometimes impossible. Artifacts are sometimes present in clinical EEGs, but it is anticipated that their numbers would increase as a result of a patient's movements and usual activities during ambulatory recording.

To overcome these problems it was suggested that some aspects of EEG analysis should be automated. Initial automatic analysis systems were based on analogue techniques, but the development of computer technology permitted more elaborate methods to be applied. Automatic EEG analysis became popular during the 1970's and 1980's, when computers became affordable bringing together expertise from the medical, the computing and the signal processing fields.

It was soon realized that a solution to the problem was far from trivial. Medical experts,

who have exercised for many years the analysis of the EEG by visual inspection, could not describe their task in detail, since experience, not accurate definitions, was the basis of their success. On the other hand, signal processing experts are accustomed to accurate descriptions that can be quantified, even if those are disconnected from the underlying causes and generating processes. Without accurate and explicit information, though, a successful automatic system could not be constructed. One could even question the reliability of traditional analysis practices, which is based on acquired, subjective criteria, affected by fatigue and depending so heavily on the unquatifiable factor 'experience'. Despite these, traditional human EEG interpretation is still considered the most successful procedure available, although there is no independent assessment method to quantify any means of analysis.

It may appear that automatic EEG analysis is a utopian dream. Indeed, it has not been possible to construct a system to replicate the human expert's task. Throughout the years, those involved in automatic EEG analysis have considered several methods (see Chapter 4), but none of them prevailed. Some were simple, lacking theoretical basis and worked to an extent. Others were well-founded on principles, but in practice they did not work satisfactorily. In many cases there was a trade-off between simplicity and accuracy, which was dictated by the capabilities of the hardware at the time they were developed. Good approaches were often overshadowed by crude, practical systems, which offered the advantage of an increased speed of operation. The latter were usually difficult to amend and their modification required the introduction of additional elements deeming the system a complicated assortment of unrelated elements, whose only purpose was to hide imperfections and overcome shortcomings of an initially modest design.

As technology advanced and computers became faster and more sophisticated, the emphasis moved in favour of good design practices and although most of the methods suggested lie between the extreme theoretical and crude practical approaches, the trend is still towards the latter. There is also an inclination to using novel approaches, dictated by trends in technology rather than their suitability and abandoning older ones when they have just started bearing useful results.

Perhaps the greatest obstacle in devising a reliable algorithm is still the poor understanding between medical experts and system designers. Medical and signal processing people see the problems and solutions from different view points. The advice of the first is often welcome by the second at the early stages of algorithm development, but the difference in the way of thinking makes medical experts unable to contribute during and after the design. After all, they are unable to express their expertise in a precise way. It appears that what is required is a common framework, in which experts from both fields are encouraged to

contribute and exchange ideas to achieve the ambitious targets of such a project.

The original objective of this work was the establishment of a reliable method for the automatic analysis of the interictal EEG of persons who are suspected of suffering from epileptic attacks, suitable for implementation as a portable device for long-term monitoring in real-time. But it was soon realized that there was little point in concentrating on the real-time aspect of any method. With the power of microcomputers continuously increasing, and their size and cost decreasing constantly, real-time operation is a volatile and unreliable guideline. Besides, the issues raised in the preceding paragraphs are more critical and ought to be resolved first.

Authors appear to agree that a common language must be defined, by quantifying the terms traditionally used to describe the EEG with accurate, definitions. But this has to be done gradually, with the contribution of medical experts so that the final definitions are accurate, but not unnecessarily strict.

Irrespective of the specific objectives, a general framework is necessary to bring together the worlds of medicine and signal processing without compromising the strengths of either. It is important to have the ability to integrate medical expertise at all stages of the design and be able to convey to them some impression of the system and to receive and interpret their comments. Only through this constructive interaction would a system developed make the most of their hard-learned experience.

While reading electroencephalography and neurophysiology, trying to understand and appreciate conventional EEG analysis, a number of bewildering issues came to light: There are standard definitions for EEG phenomena, but they are inaccurate and should be used only as guidelines; experience is difficult to attain, but it is important; EEG analysis is a rather subjective procedure. Attempting to understand what really took place during analysis, the process was interpreted as a recurring sequence of reading the EEG, looking for patterns conforming to the definitions and identifying those by their names and other attributes. After a while there was no need to look at the definitions and identification came almost spontaneously. Learning had taken place and an intuitive representation replaced the verbal definition. In other words, a *conceptual model* of what constitutes an EEG had been formed in the brain. The purpose of the initial textbook definitions was to form a framework on which subsequent analysis could be based. Experience may be regarded as the refinement of the model. As different experts have had different experiences, it is reasonable to expect small deviations in their response to a given record, especially when containing ambiguous patterns. This would explain why an experienced analyst is, in general, more accurate than an inexperienced one.

Developing a model for the EEG signal on which to base an EEG analysis method is important. Although this is not likely to be similar to the conceptual model used by human analysts, it bears its basic property, the potential for refinement with the assistance of experienced analysts. Secondly, it may make the development of an analysis procedure more methodical, orderly and not dependent on specific signal characteristics, which often come to light during testing. It may also be used for the production of simulated EEG records, to test the validity of a developed algorithm, before it is applied to real EEG signals. Finally, a model can convey a lot of information and would reflect the assumptions made during the design of the analysis algorithm, so that most limitations of the method may become evident at an early stage. As a result, the reliability of an algorithm may be assessed even before it is put to the test. Afterwards, any discrepancies from the expected response may be evaluated and used for its improvement.

To formulate such a model it was necessary to study and appreciate the EEG from the neurophysiological perspective. This preliminary material is covered in Chapters 1, 2 and 3.

Chapter 1 is an introduction to the physiology of the nervous system. It was considered necessary to acquire some understanding of the cellular mechanisms, processes and generating structures of the electrical potentials in the brain, which form the basis of the EEG. The identification of some of the sources of the recorded activity and the general nature of signals in the brain proved useful for determining whether a modelling assumption was reasonable or not.

Chapter 2 is a general introduction to electroencephalography. Being a continuation of Chapter 1 it gives an explanation of what signals may be recorded from the scalp that constitutes the EEG. This is followed by a brief description of traditional electroencephalography to appreciate what takes place during a routine recording as well as a description of the findings in a normal and an abnormal EEG, from the neurophysiological perspective. These proved necessary for modelling the signal.

Epilepsy, the brain disorder that provided the motivation for this work, is presented in Chapter 3. Its causes, symptoms and especially its relation to the EEG are emphasized. It will be appreciated that epilepsy is a rather varied condition and that it must be identified reliably and subsequent medical treatment must be udertaken carefully.

Following the neurophysiological background, a review of automatic EEG analysis is presented in Chapter 4. Although a wide range of methods is covered, the emphasis is on the detection of epileptic transients and on good or relevant approaches.

After this introductory material, the modelling approach to EEG analysis is presented in detail. It was decided to maintain enough detail for the work to be reproduced on the

basis of the descriptions given here. It has been the sad experience of the author that lack of detail often did not permit an in-depth study of some of the techniques proposed by other workers.

Based on the descriptions of Chapters 1, 2 and 3, methods for modelling the EEG signal were considered. Finding a suitable representation for the signal was not an easy task. Several signal processing tools were employed in turn. It was found that not all models could form the basis of an effective analysis procedure. To demonstrate, however the point made earlier that a model is useful in determining the shortcomings of a method, some of the early attempts for modelling have been included in Chapter 5. The experience gained from these unsuccessful models was valuable in the development of working systems later.

Some of the early working models (Chapter 6) were not particularly detailed, as they were formulated with specific epileptic patterns in mind. Two such methods are presented, demonstrating that by studying the signal and defining even a simple model, it is possible to construct successful analysis procedures. Both were based on the gradients of the EEG signal to detect epileptic transients. The first shows how a system was developed by considering the model, rather than individual EEG signals. The second followed the same guidelines but was more reliable, as it was more detailed and used information extracted entirely from the model.

A major advance in the modelling approach is presented in Chapter 7. Here, a complete model for the EEG signal was proposed. The EEG was analyzed for modelling purposes in a number of constituent signals, which were given neurophysiological significance and names. These were quantified in signal processing terms, but the definitions were maintained on a general level, free from details. The first half of this Chapter gives a rigorous description of the signal model, which was built entirely on the definitions of the first three chapters and is therefore independent of any test signal. The second half is a demonstration of how the model could be employed to develop a modular analysis system, still on a general level and implementation-independent.

Further refinement, additional assumptions and constraints on the general model were then imposed, leading to a specific implementation of the analysis system (Chapter 8). This has been tested with both artificial and real EEG records and proved satisfactory. To demonstrate the modularity and versatility of the approach, some alternative implementations of certain elements of the system were considered in Chapter 9. Not all of those worked well, but the general conclusion was that replacing one element of the system by another equivalent was possible without altering neither the other elements nor the concepts on which the system was based.

It is then shown that the system may be extended to detect other signals as well as epileptic spikes. The inclusion of those by the initial model made this task relatively simple, proving that the approach is general enough to encompass many aspects of computerized EEG analysis, not only related to epilepsy (Chapter 10). Finally, an artificial neural network was introduced, replacing some elements of the analysis procedure. This is capable of changing its behaviour with the guidance of medical expertise, but it raised other issues. A thorough investigation of proposals on how to optimize the performance of the network are discussed extensively in Chapter 11. It will be demonstrated that such a device need not necessarily be treated as a 'black box'.

Secondary material, like theoretical aspects and details have been included in a number of Appendices. It was decided not to include computer programs, as these were lengthy and would add little to the understanding of the approach, while increasing the size of an already long report. Although there is enough detail in this Thesis to reproduce them, the system implementation and other auxiliary programs, are available on a computer disk.

As stated in the conclusions of this thesis a general approach has been proposed, that may be useful both in future system development, but also as a point of reference for signal processing and medical experts alike. Perhaps the most important contribution of this work is the proposition not only of a working system, but of a general and flexible framework that may be utilized for the development of better techniques in the future. The system realized may be understood, used and perfected by medical experts, gathering their experience. Apart from accomplishing its primary function, the suggested realization produces a number of secondary results that may be useful in further analysis. The general approach may be extended to other signals, not necessarily of biological origin.

Chapter 1

The nervous system

1.1 Introduction

The nervous system is described here briefly, with emphasis on the structure and function of neurons, its basic cellular elements and anatomical regions of the brain that are linked to the EEG, like the cerebral cortex, the thalamus and the reticular formation. A brief word on the protection of the nervous system concludes this Chapter, which contains founding material to understand the nature, significance and problems associated with the electroencephalogram and to appreciate some of the aspects and the foundation of the methods presented later in Chapters 5 to 11.

This Chapter is also an introduction for readers without previous knowledge in the subject. It is by no means complete, but more details may be found in an internal report[132], which follows the same organization as the present Chapter.

1.2 Cellular composition of the nervous system

Despite its complexity, the tissue of the nervous system consists of a remarkably small number of cell types. From these only one, the neuron, is actively involved in its function, whereas the others have a supporting role[118, p. 334].

1.2.1 Structure and morphology of neurons

Neurons are highly specialized cells, unable to reproduce but with a long time span and high metabolic rate[118, pp. 336–337].

Although they differ from one another (see Figure 1.1), they have two identifiable parts; the cell body, also known as perikaryon or soma, and the processes (*dendrites* and *axons*),

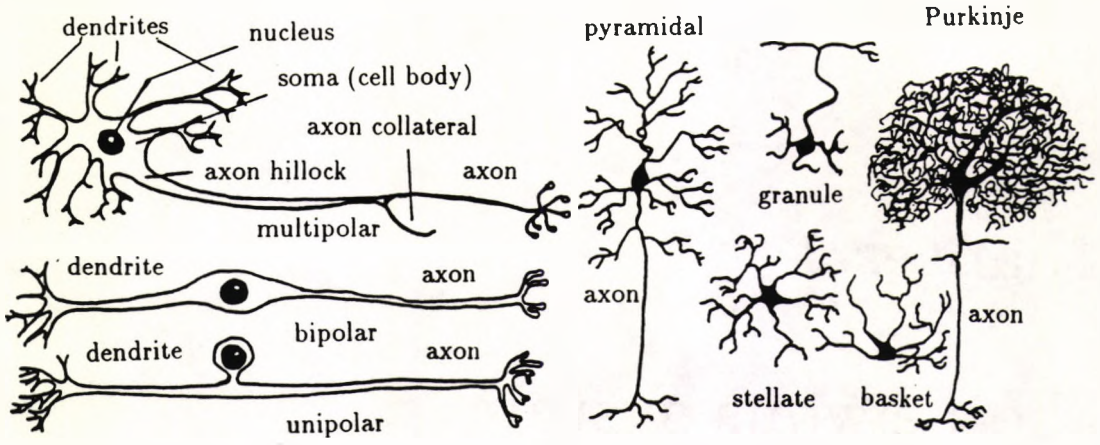


Figure 1.1: An assortment of neurons and their basic elements

which are cytoplasmic 'branches', sprouting from the cell body. Dendrites receive information from other neurons or body tissue. The unique axon of the neuron is attached to the cell body by the axon hillock. It is its transmitting end and, being considerably longer than the dendrites, it may stimulate many other neurons or muscle fibres[63, pp. 21–23].

Neurons may be classified according to their morphology and purpose[118, pp. 338–341]. *Unipolar* neurons have one dendrite linked to the axon, *bipolar* have a distinct dendrite whereas *multipolar*, the commonest type of neurons, have many dendrites. *Afferent neurons* conduct information from the sensory organs to the central nervous system whereas *efferent neurons* activate muscle fibres and *association neurons*, which consist the vast majority process information provided by other neurons. Association neurons are subdivided into various categories, depending on their shape and purpose (pyramidal, stellate, granule, basket, Purkinje etc., see Figure 1.1).

1.2.2 Supporting cells

There are six categories of supporting cells sometimes called *neuroglia*[118, pp. 334–336]. Some form an electrically insulating myelin sheath by coiling around the processes of the neurons in the central nervous system (*oligodendrocytes*) and the bodies and processes of neurons in the peripheral nervous system (*satellite* and *Schwann cells*, respectively). In the axons the myelin sheath has gaps, the *nodes of Ranvier*, from where *axon collaterals* sprout. Other neuroglia, *astrocytes*, hold neurons close to blood capillaries, mediating the exchange of substances and controlling their chemical environment, whereas *microglia* protect and clean the nervous system from invading substances. Finally, *ependymal cells* are involved in the production and circulation of the cerebrospinal fluid.

1.3 Neurons and electrical potentials

The communication of information in the nervous system is effected in electrical form[118, pp. 347–365]. Ironically this is true for many modern man-made communication systems, although the mechanisms involved in the production of neural signals are rather different and are briefly outlined here. A more complete explanation may be found in [132].

1.3.1 Cellular structures and mechanisms

The *cytoplasm* of neurons and the *extracellular fluid* that surrounds them have similar composition, consisting mainly of a water solution where several organic and inorganic substances, many in ionic form, float. The *plasma membrane* contains many types of protein molecules each one facilitating the exchange of a particular type of molecules between the cytoplasm and the extracellular fluid. Owing to the selective permeability of these proteins, the concentrations of the various elements in the cytoplasm differ from those in the extracellular fluid[118, pp. 62–74].

1.3.2 Neurons at rest, resting potential

All electrical phenomena in the nervous system are caused by ions, which are the only electrically charged elements in this environment. Sodium (Na^+), chlorine (Cl^-) and potassium (K^+) ions are in abundance in both the cytoplasm and the extracellular fluid. In the extracellular fluid the concentration of Na^+ and Cl^- is high and that of K^+ low. The opposite happens in the cytoplasm, where the concentration of K^+ is high.

Chloride ions are freely permeable, across the membrane, since the chlorine protein channels allow the bidirectional diffusion of Cl^-. Similar mechanisms (sodium and potassium ion channels) permit the inflow and outflow of Na^+ and K^+, attempting to bring their concentrations and the electrical condition across the membrane to an equilibrium[118, pp. 62–74]. When this occurs, the potential difference between the cytoplasm and the extracellular fluid is approximately $-70\ mV$[118, pp. 62–74][63, pp. 12–17][103, p. 3]. The high concentration of K^+ inside the cell encourages these ions to leave the cytoplasm, but the negative electrical potential reduces their outflow. The inflow of Na^+ is encouraged by both their low concentration and the negative electrical gradient, but is slow, because of the low permeability of the plasma membrane to these ions.

The *electrical polarization* across the membrane to the aforementioned voltage[118, pp. 62–74] is maintained by an active mechanism, the *sodium-potassium ion pump*, a special

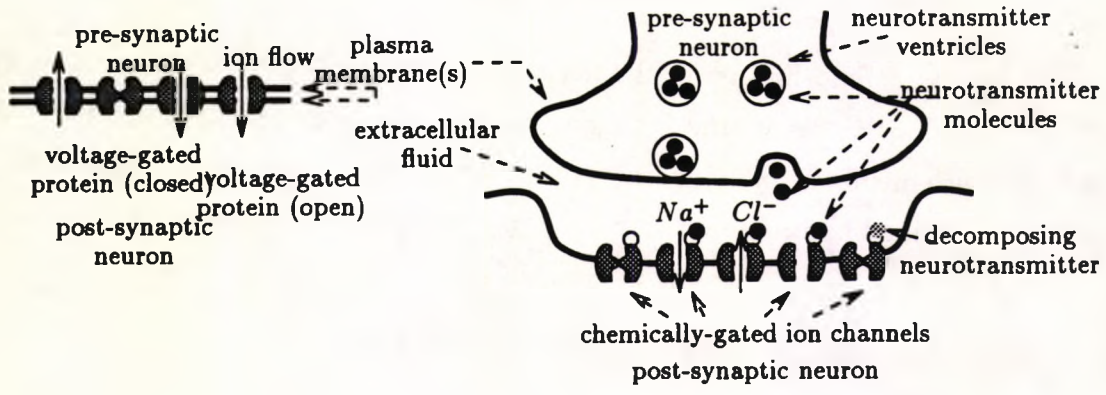


Figure 1.2: Types of synapses

protein channel, working against the concentration gradients, pumping K^+ in and Na^+ out of the cytoplasm[63, pp 12–17].

Some protein molecule ion channels, called *gated channels*, do not operate continuously. They are activated by electrical stimuli (voltage-gated channels) or to specific chemicals called *neurotransmitters* (chemically gated channels)[63, pp. 52–63]. Gated channels are involved in the production of neural signals, which is explained in the next two sections.

1.3.3 Stimuli, synapses and synaptic potentials

Neurons receive and process electrical stimuli that normally originate at the axons of other neurons (see later section 1.3.4) with the exception of sensory inputs, which generate in sensory organs and are an integral part of sensory neurons[118, pp. 344–347][63, pp. 64–69].

Stimulation of a neuron is effected through small 'swellings' on the membrane of its body and the dendrites (section 1.2.1), where axons of other neurons are attached to form electrical and chemical *synapses.*

Electrical synapses (Figure 1.2(a)) are direct connections of the cytoplasm of two neurons via voltage-gated channels. An electrical stimulus in the first neuron opens the channels transferring ions between the cytoplasms of the attached neurons, and hence the stimulus to the second.

Chemical synapses (Figure 1.2(b)) are by far the commonest. No physical contact between the neurons exists. An action potential in the transmitting neuron triggers the release of special chemicals, *neurotransmitters.* These bind on receptor molecules on the chemically-gated channels in a recess on the membrane of the receiving neuron, causing them to open. Ions flow between the extracellular fluid and the cytoplasm, causing a local change, of the resting potential, called a *postsynaptic potential.* The neurotransmitter is decomposed by

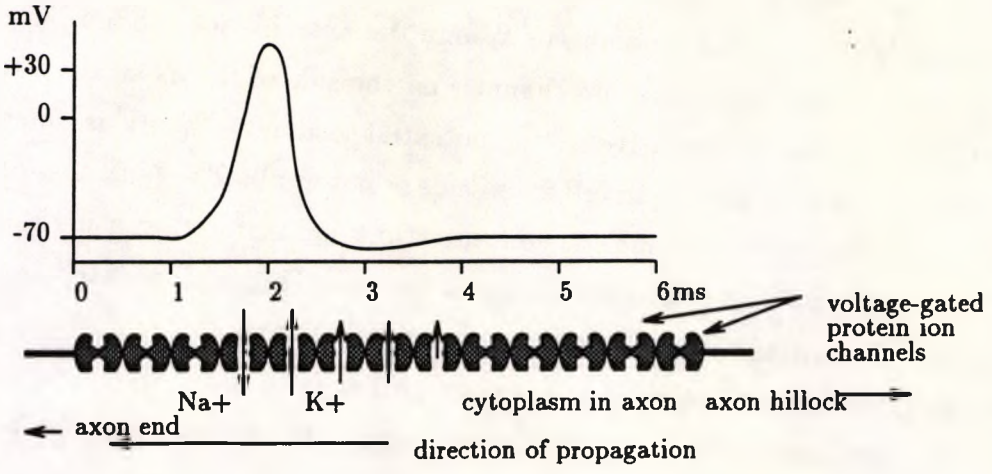


Figure 1.3: Generation of an action potential

special enzymes in a few milliseconds, closing the channel, and the ion concentrations and resting potential are eventually restored by the action of the passive ion gates and the sodium-potassium pump[63, pp. 52–63].

The postsynaptic potential caused by a synapse always has the same direction. If it is less negative than the resting potential it is called *excitatory postsynaptic potential* (EPSP), because it encourages the generation of an action potential (section 1.3.4), whereas if it is more negative it discourages it and is called *inhibitory* (IPSP).

A neurotransmitter may cause the generation of either an EPSP or an IPSP depending on which type of channel (K^+, Na^+ or Cl^-) it gets attached to. The most well-known neurotransmitters are *acetylcholine* (*ACh*), usually excitatory and *γ-aminobutyric acid* (GABA), usually inhibitory. A more comprehensive list may be found in [132]. The action of neurotransmitters is affected by drugs, like cocaine, LSD, heroine and amphetamines and is often linked with some forms of epilepsy and other disorders of the nervous system (Chapter 3).

1.3.4 Neuron responses, action potentials

The function of a neuron comprises the processing of postsynaptic potentials. The outcome of this is a signal, called an *action potential*, which is propagated along its axon to stimulate other neurons, activate muscle fibres, etc.

Neurons have a large number of synapses (typically over a thousand) from which they receive stimuli from other neurons. The resulting postsynaptic potentials, though prominent near the synapses, spread through the cytoplasm and reach the axon hillock with much reduced amplitudes, depending on their strength and their proximity.

Their spatial and temporal integration may give rise to a potential, sufficiently large to cause the depolarization of the membrane around the axon hillock[118, pp. 350–356][63, pp. 25–31], opening voltage-gated sodium channels on the side of the axon. As a result, a large positive transition from the negative resting potential to about $+30\ mV$ is generated. This is a binary phenomenon, occurring in full amplitude or not at all. The Na^+ channels eventually close and the resting potential and ion concentrations are restored, causing a *nerve impulse* of amplitude around $100\ mV$ and duration of a few ms. This is propagated (regenerated) in a repetitive cycle until the potential reaches the end of the axon[63, pp. 25–31]. The velocity of the propagation of this potential varies between 1 m/s (thin, unmyelinated axons) to 130 m/s (thick, myelinated axons). In myelinated axons propagation occurs in 'leaps' between the nodes of Ranvier (section 1.2.1).

The processes of generation and propagation of an action potential, called *firing* of a neuron, are demonstrated in Figure 1.3. More details may be found in an internal report[132].

1.4 Anatomy of the brain

The rather simple operation of individual neurons could not account for the complex processes of thought or the initiation and execution of movements. Much of the complexity of the brain function may be attributed to the great number of neural interconnections. These are not random, as thought initially, but recur with surprising regularity. Disruptions in their formation may lead to the death of the neurons involved[118, pp. 342–343, 370]. Moreover neurons never act in isolation. Groups of neighbouring neurons often exhibit similar behaviour. It is the common potentials in these regions that give rise to significant extracellular potentials that may be measured by placing electrodes on the surface of the scalp (see Chapter 2).

The main elements of the nervous system are the *brain* and the *spinal cord*, which is the mediator of neural signals from the sensory organs to the brain and from the brain to the muscles. It is also an important reflex centre, generating reflex actions without the intervention of the brain. A summary of its function may be found in [118, pp. 404–413] and in [132].

The human brain is subdivided into several regions with distinct morphological as well as functional features[118, pp. 376–404], the *cerebral hemispheres*, the *diencephalon*, the *brain stem* and the *cerebellum* (Figure 1.4). Some other structures consist of elements dispersed in the brain, but functionally related, like the *limbic system* and the *reticular formation*. There are also four hollow spaces, embedded deep in the brain, the *brain ventricles*, where

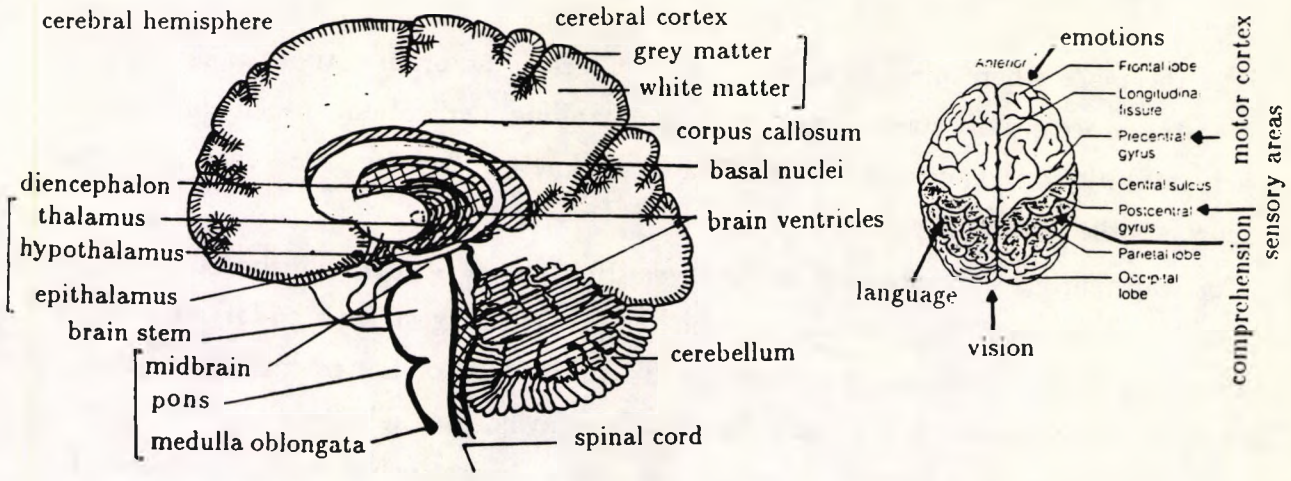


Figure 1.4: The human brain

the cerebrospinal fluid (see section 1.4.6) is produced.

Although the brain function, as a whole, is the result of their complex interactions, the brief description that follows, based on [118, pp. 376–404], [103, pp. 10–20] and [132], is biased towards the ones responsible for recognized EEG phenomena.

1.4.1 The cerebral hemispheres

The two (left and right) cerebral hemispheres are the most prominent regions of the brain, engulfing most of the other structures and occupying most of the space in the head. They are joined only at a narrow region, situated deep in the brain called *corpus callosum* (see section on white matter).

Their crinkled surface has numerous grooves, called *fissures*, if they are deep and *sulci*, if they are shallow (Figure 1.4). The windings formed between the sulci are called *gyri* and although individual brains differ in their details, certain landmarks are always identifiable. The central, parieto-occipital and lateral sulcus separate the surface of each hemisphere into four *lobes*, the *frontal*, *parietal*, *occipital* and *temporal* lobes, named by their location. The *precentral* and the *postcentral gyri* are also easily identified (Figure 1.4).

Anatomically the surface of the hemispheres, the *cerebral cortex* is distinguished by its grey appearance from their interior *white matter* where isolated grey regions, the *basal nuclei* are embedded.

The cerebral cortex

This is the place where most of what is often perceived as brain activity takes place, including the processes of thought, memory, understanding, perception, speech and voluntary movement initiation. The cerebral cortex has a thickness of 2–4 *mm* and consists primarily of neuron bodies and glial cells.

Cortical neurons are organized in six layers[103, pp. 6–8], distinguished by the types of association neurons they contain (section 1.2.1). The largest (pyramidal) neurons, are organized in columns running perpendicular to the surface of the cortex[103, pp. 9–10]. Their axons project inwards to the centre of the brain, whereas their larger dendrite projects towards the surface of the cortex. Other dendrites are comparatively short. This organization of cortical neurons leads to the formation of groups that are stimulated by deeper structures in the brain like the thalamus (section 1.4.2). As a result, it was initially thought that separate cortical areas specialize in the performance of different brain functions (*regional specialization theory*). Contrarily, the *aggregate field view* supported the distribution of brain functions over the whole of the cortex. Later studies have demonstrated that although areas of specialization do exist (shown in Figure 1.4) these do not act in isolation[118, pp. 376–404], whereas higher mental functions, such as logic, memory, abstraction and understanding cannot be localized.

Sensory areas, receiving and processing responses from sensory neurons, and motor areas, issuing instructions to the muscles to move, lie respectively in the postcentral and the precentral gyri. Other sense organs, like the ears and the eyes have their own separate specialization areas (auditory and visual cortex respectively).

A cerebral hemisphere is mainly associated with the opposite part of the body. In the vast majority of people, there is *cortical lateralization*, prominent in the non-specialized areas, where one hemisphere dominates the other.

The white matter

White matter consists, primarily, of myelinated fibres (axons and dendrites). These are processes of cortical neurons as well as neurons from the deeper brain structures to which they are connected.

Fibres are usually encountered in bundles, linking neurons from one localized region to another. They are distinguished in *association fibres*, connecting neurons of the same hemisphere, *commissural fibres* connecting the two hemispheres and forming the corpus callosum and *projection fibres*, which link to areas outside the brain.

The basal nuclei

These are groups of neurons that are located symmetrically in the interior of the cerebral hemispheres (Figure 1.4). There are three basal nuclei in every hemisphere, the *putamen*, *globus pallidus* and *caudate nucleus*, collectively called *corpus striatum*. Although their function is still a matter of debate, they appear to communicate with the motor cortex for the initiation of movements, especially stereotype ones.

1.4.2 The diencephalon

The diencephalon is located under the corpus callosum and forms the walls of one of the brain ventricles (the third ventricle). It consists primarily of the *thalamus* and two smaller structures, the *hypothalamus* and the *epithalamus*.

The thalamus

This consists of neuron bodies, located ipsilaterally on the walls of the third ventricle, linked together by a bundle of fibres, the *intermediate mass*.

The thalamus is located in the path of all the sensory fibres arriving from the body (Figure 1.4) and the the sense organs on the head (apart from the nose). Sensory inputs enter the thalamus and are preprocessed by at least one of its nuclei. It appears that the primary function of the thalamus sorting and grouping related sensory responses and routing them to the appropriate parts of the cerebral cortex for analysis. This is evident from the projecting connections between some thalamic nuclei, (*specific nuclei*)[103, pp. 18–19] and the narrow specialized regions of the cortex (see section 1.4.1). Other nuclei (*non-specific*) stimulate a much wider area of the cortex. Effectively there exist thalamic connections to every site in the cortex[63, pp. 124–133]. Kiloh et al.[103, pp. 22–28] explain that there is experimental evidence that the specific and possibly the non-specific thalamic nuclei play the role of the pacemaker of brain activity, by controlling the level of excitability of cortical neurons.

The hypothalamus and the epithalamus

The hypothalamus, located under the thalamus (see Figure 1.4), is responsible for the balanced function of the body. Among its nuclei are the regulators of the heart rate, blood pressure, body temperature, respiration, food and water intake and digestion, sleep cycles and wakefulness[63, pp. 165–181]. It also controls the *hypophysis cerebri*, whose hormones regulate all other glands.

The epithalamus, situated on the top of the thalamus, contains the biological clock, regulating body rhythms.

1.4.3 The brain stem

Projecting from the spinal cord to the diencephalon, the *brain stem* is the centre of vital autonomic functions and the path of most somatosensory signals. It has three constituent parts, the *midbrain*, on top, the *pons*, in the middle, and the *medulla oblongata*, at the bottom.

The midbrain groups the cortical motor fibres, connecting some of these to the cerebellum and relaying the rest towards the spinal cord. The auditory and visual reflex centres are situated here.

The pons (bridge) consists mainly of fibres connecting the midbrain to the medulla oblongata and the cerebellum. One of its nuclei is the respiratory control centre.

The medulla oblongata fuses in the spinal cord and contains the bundles of motor fibres, the *pyramids*, that originated in the cerebral cortex. At the back it connects to the cerebellum. Important reflex nuclei of the medulla include the *cardiac centre*, the *respiratory centre*, and the *vasomotor centre*.

1.4.4 The cerebellum

The *cerebellum* is a large growth at the back of the brain stem, under the cerebral hemispheres. It has two cerebellar hemispheres, linked by the vermis. Their 'pleated' surface consists of grey matter. Its interior consists of fibres and nuclei of neurons. Like the cerebrum, it has fissures and sulci that separate its surface into lobes and gyri. The inputs to the cerebellum are sensory and motor signals. Its outputs to the motor areas of the cortex provide feedback for the proper coordination and smooth execution of voluntary movements[63, pp. 147–151].

1.4.5 Functional groups in the brain

Not all functions of the brain are localized. Some functionally related parts exist in the brain that do not have clear physical boundaries. The *reticular formation* and the *limbic system* are such systems[118, pp. 396–398].

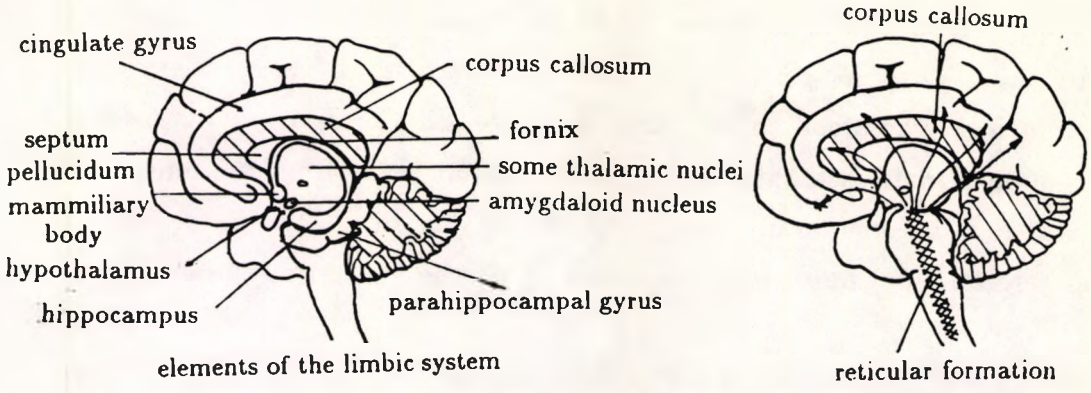


Figure 1.5: The limbic system and the reticular formation

The reticular formation

The *reticular formation* (Figure 1.5) consists of fibres, small nuclei and individual neurons located in the interior of the brain stem. Their processes have connections in all regions from the spinal cord to the cerebral cortex. Perhaps the most important function of reticular neurons is the implementation of the *reticular activation system* (RAS)[63, pp. 123–134], which sends impulses to the brain structures continuously, maintaining cortical and other structures 'awake'. The RAS is inhibited during sleep, anæsthesia, etc. Because they are connected to the thalamus (section 1.4.2), they affect the operation of the pacemaker, responsible for cortical rhythms. The reticular formation is also responsible for weakening repetitive, expected stimuli and highlighting unexpected, unusual and strong ones that may cause the arousal of the cortex during sleep[103, pp. 22–28].

The limbic system

This is more dispersed than the reticlular formation[63, pp. 184–198]. Its various parts encircle the thalamus and the brain stem (Figure 1.5) and constitute the emotional centre, responding with general changes to stimuli, causing the physical manifestations of emotions. Several personality and emotional disorders arise from damage to elements of the limbic system.

1.4.6 Protection of the nervous system

The EEG is certainly linked to the electrical activity of the brain, but this is distorted by the layers of tissue which surround and protect it from external hazards, viruses and chemical substances existing in the body[118, pp. 399–402][63, pp. 208–220][132].

From the exterior, these are the *hair and skin*, the leathery *periosteum*, that covers the bones of the skull, and the three *meninges*. The outermost *dura mater* is a leathery tissue. Between its two layers, the *periosteal* and the *meningeal dura mater*, there exist the *dural sinuses*, which convey blood to the veins of the neck. The *arachnoid* is a loose cover underneath the dura mater separated by the *subarachnoid space* from the innermost meninge, the *pia mater*, a fine membrane lining even the smaller sulci. The subarachnoid space is filled with cerebrospinal fluid, which is absorbed by the sinuses at the *arachnoid villi*.

The nervous system is suspended in the buoyant *cerebrospinal fluid*. This derives from blood plasma, to which it has similar composition but different molecule concentrations and is also the mediator of the hormones produced by the hypophysis cerebri.

1.5 Comments and observations

The nervous system consists of an astronomical number of neurons, estimated as several thousand million in the cortex alone[132], which are highly interconnected and act, to a great extent, independently one from another. The evident dissimilarities in the fine details of brains of different individuals make the study of the brain on a neuron by neuron basis rather pointless.

Therefore, studies are focussed either on the details of the function individual neurons or on the groups of neurons, that are known to exhibit similar behaviour. The first approach aims to understand the underlying biochemical phenomena, whereas the second aims to interpret localized or generalized neurophysiological observations[28]. Responses that are common to groups of neighbouring neurons and their effects on other groups have been recognized many years ago in electroencephalographic recordings.

Chapter 2

The electroencephalogram

2.1 Introduction

The electrical potentials generated by neurons and their large accumulations on the cerebral cortex (Chapter 1) give rise to extracellular potentials. These were first recorded in 1875 and with the technological advances of our century their recording was refined and became a non-invasive and standardized procedure. The recording of the varying electrical potentials from the surface of the scalp, termed the *electroencephalogram* (EEG), from the Greek words electro=electrical enkephalos=brain and graphō=write, is common practice in many hospitals. Early expectations on the significance of the EEG were abandoned when later studies revealed that it constitutes only a distorted shadow of the underlying brain activity[137].

Traditional EEG analysis is performed by visual inspection and the diagnostic value of the brief routine EEG has been questioned, as it depends on the experience of the analyst and can be contaminated with non-cerebral signals (artifacts).

Studies of the EEG have revealed similarities between different individuals and recurring variations with age and with the state of consciousness. Because there is correlation between abnormal EEG patterns and a variety of brain disorders the former has been used extensively for diagnostic purposes. Traditional EEG analysis does not usually take into account the generating mechanisms of the signal, but is based largely on observations from its recording on paper.

2.2 Origin of scalp potentials

The electrical activity inside individual neurons (section 1.3) gives rise to extracellular electrical potentials. Despite their small amplitudes, these may be recorded with sensitive in-

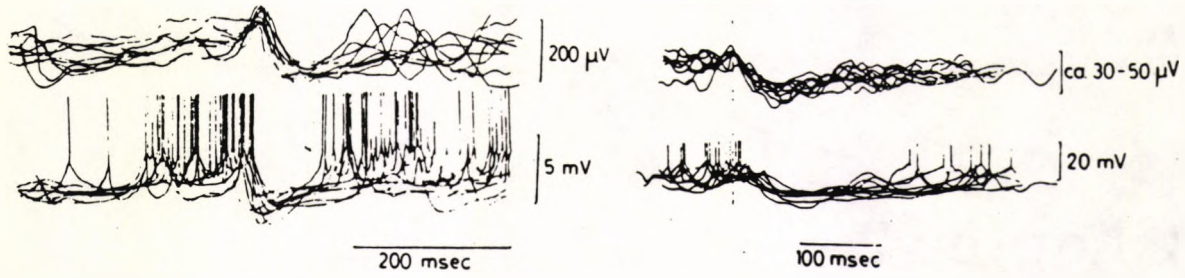


Figure 2.1: Correspondence of EEG (top) and underlying neuron activity (bottom)
(Reprinted from Kiloh et al.[103])

struments from the surface of the cortex or even from the scalp.

2.2.1 Neural activity and extracellular potentials

From the electrical phenomena in a neuron (see section 1.3), action potentials and cytoplasmic currents either have brief duration or large electrical resistances, leaving excitatory and inhibitory postsynaptic potentials in the the largest dendrites and the soma as the most likely sources of significant extracellular potentials[103, p. 22–23]. This was demonstrated by simultaneous recordings from the surface of the cortex and underlying individual neurons (Figure 2.1).

The rhythmic and often synchronized nature of recordings from various points of the cortical surface suggested the existence of a subcortical pacemaker, later localized to the *non-specific thalamic nuclei*[103, pp. 22–28], which project to all cortical regions (section 1.4.2). Andersen and Sears[103, p. 25] proposed a model of the thalamus capable of producing synchronous, periodic discharges in the form of excitation-inhibition cycles[133]. These cause large EPSPs on the long apical dendrites of the pyramidal neurons in the cortex, which, in turn discharge and activate neighbouring stellate neurons causing feedback IPSPs. Rhythmic thalamic discharges, however sustain and synchronize pyramidal neurons, causing periodic phenomena in the cortex, (section 2.5.1), characteristic of sleep and anæsthesia[103, p. 26].

During wakefulness rhythms are more irregular and faster due to the desynchronization of cortical neurons, attributed to the action (RAS) of the reticular formation (section 1.4.5), which causes EPSPs in the non-specific thalamic nuclei, counteracting their internal inhibitory mechanisms[103, pp. 22–29].

2.2.2 The generation of cortical potentials

According to a simple model of a pyramidal neuron[103, pp. 29–33], a PSP on its apical dendrite or the soma causes a potential difference, V_{DS}, across these points, since the other

is still at the resting potential (sections 1.3.1 and 1.3.2). The neuron then behaves like an *electrical dipole*[103, p. 29] causing an electric field through the extracellular fluid.

In the cortex, most neurons are organized in columns, a structure which is more prominent in large pyramidal neurons (section 1.4.1). Many neighbouring neurons have similar thalamic connections, receiving almost simultaneous stimuli. These represent electrical dipoles with the same orientation producing an aggregate electric field strong enough to be be recorded by remote electrodes[103, pp. 7–8,21]

Despite criticism[47, p. 11], this model has been used for the interpretation of many EEG potentials[103, pp. 30–32], like the epileptic spike-and-wave precursor (section 2.5.2), which is believed to be the result of a spike, caused by thalamic EPSPs on cortical neurons, followed by a slow wave, the result of the IPSP on their soma by inhibitory stellate neurons [133].

2.2.3 Scalp potentials

The evolution of the electric field produced by cortical neurons may be recorded from the cortical surface to produce an *electrocorticogram* (ECoG). ECoGs taken from electrodes located only millimetres apart are generally different and any similarities in the activity of disjoint regions suggest either their direct or subcortical interconnection[47, p. 11]. An ECoG is only taken in exceptional cases, since electrode placement requires neurosurgery.

For routine electroencephalograms (EEGs), the electrodes are placed on the surface of the head, a simpler and non-invasive procedure, but the presence of layers of protective tissue (section 1.4.6) in the path of cortical potentials, introduces significant distortion in the recorded signals[133][47, p. 13][103, p. 40][177, pp. 188–189].

The electrical resistance of the layers of tissue introduces a five-fold *attenuation* to cortical signals, which is variable over the scalp; their capacitance is responsible for the attenuation of high-frequencies, phase distortion and temporal averaging.

Spatial averaging of cortical potentials is the result of large surface of scalp electrodes and the diffusion of the field of neurons within a large radius around it through the intervening tissue.

The convoluted nature of the cortical surface (section 1.4.1) results in *variations in contributions* since neurons on the tips of the gyri, due to their orientation, contribute more to the recording than those in the fissures and sulci which are obscured by scalp electrodes. Similarly *non-uniformities in the conductivity* of the tissues cause the recording of distant interfering cerebral potentials[47, p. 13][103, p. 40]

Finally the EEG is very susceptible to non-cerebral potentials, caused by external signals

or the muscles the eyes and the heart (section 2.5.4), because these have comparable (in some cases larger) amplitudes to the attenuated cerebral potentials.

The poor localization, distortion and interference introduced deem the EEG suitable only for general observations[47, p. 11–12][103, p. 40].

2.3 Brief history of electroencephalography

Electroencephalography took many years to become established. In the early days, progress was hindered by a sceptical society and the lack of appropriate facilities and suitable instrumentation. Only the most important landmarks can be mentioned in this brief account. A more complete description may be found in a book by Brazier[27] (see also [133]).

The first report on the electrical activity of the brain was produced in 1875 by a British physiologist, Richard Caton. He observed alternating currents and a *negative variation* in the cortical potentials following stimulation.

Between 1887 and 1890 many workers, including Caton and Adolph Beck in Poland discovered the *spontaneous activity* of the cortex and the controversy that followed the claim for originality brought many workers together, drawing the attention of others to their work. The investigation of the brain potentials was the subject of intensive research in Russia, with Sechenov (1882), Kaufman (1910) and others, who recorded induced epileptic seizures.

Perhaps *the* most important development of electroencephalography was the first book on the human electroencephalogram, *"Über das Elektrenkephalogramm des Menschen"*, by Hans Berger in 1929. This contained fourteen reports on his observations on the normal EEG and abnormalities related to pathological conditions, like epilepsy, using a cathode ray tube[69]. Many of his observations, speculations and terms (*electrenkephalogramm*, α and β rhythms, etc.) have been very influential in clinical neurophysiology.

Subsequent publications, by Adrian and Matthews and others, followed after 1930 confirmed Berger's findings. More recently many researchers, like Beaudoin and Fessard, Brémier, Walter, Gibbs, Lennox and Jasper have contributed to the field. The need for standardization led to the foundation of the *International Federation of Societies for Electroencephalography and Clinical Neurophysiology (IFSECN).*

Historically, advances in electroencephalography were related to advances in technology[27][47, pp. 1–2]. The sensitive galvanometers and elaborate photographic processes used by the first workers were abandoned, as inventions, like the valve amplifier and the oscilloscope in the 1930's, pen recorders and the differential amplifier in the 1940's and the transistor in the 1950's improved the quality and the cost of EEG recordings.

Developments in electromagnetic recording provided a compact means of storing many hours of EEG on a single magnetic tape using FM techniques to store many channels simultaneously[47, pp. 172–174]. Tapes are suitable for automatic assessment, as they can be played back later for off-line analysis. The transmission of EEG signals via *telephone lines* and the use of *telemetry* for prolonged monitoring of ambulatory patients[47, pp. 175–180][177, p. 196] was also used experimentally[133].

Automatic EEG analysis was the subject of research in the 1970's and early 1980's because of the wide use of computers. Targets varied, from alternative visualization of the EEG to full automation of the analysis. Among the methods applied were ad hoc wave measures, frequency and correlation analysis, parametric and statistical methods (see Chapter 4 and references [7], [47, pp. 231–314], [133] and [134] for details).

2.4 Technical aspects of electroencephalography

In order to appreciate some aspects of the electroencephalogram, the apparatus and the basic procedures followed for recording and analyzing the signal are described.

Electrodes: the interface to the scalp

Types of electrodes

Since the early days of electroencephalography, electrodes were a major problem during a recording. Bad contact with the scalp is reflected as a poor, often illegible record.

The two commonest types of scalp electrodes are the *pad* and the cup (*stick-on*) electrodes (Figure 2.2 (a) and (b))[47, pp. 15–16][177, pp. 8–9]. Both use saline as their conducting medium and have similar electrical properties, but pad electrodes are easier to attach, whereas cup electrodes are more stable[133].

Other electrodes like *metal disk*, *plastic cup*, *platinum needle*, *Sphenoidal*, *nasopharyngeal*, *electrocorticographic* and *intracerebral* electrodes (Figure 2.2) are used in special cases[133][177, pp. 71–72][47, p. 16–17].

An electrical model for an electrode[47, pp. 24–25] is shown in Figure 2.2. R_s is the resistance of the saline jelly, R_c the resistance and C_c the capacitance of the boundary of the metal electrode and the electrolyte, C_t the stray capacitance and Z_w, the *Warburg impedance*, is a frequency-dependent resistance and capacitance. The values of these components define the properties of the electrode[133][47, p. 24].

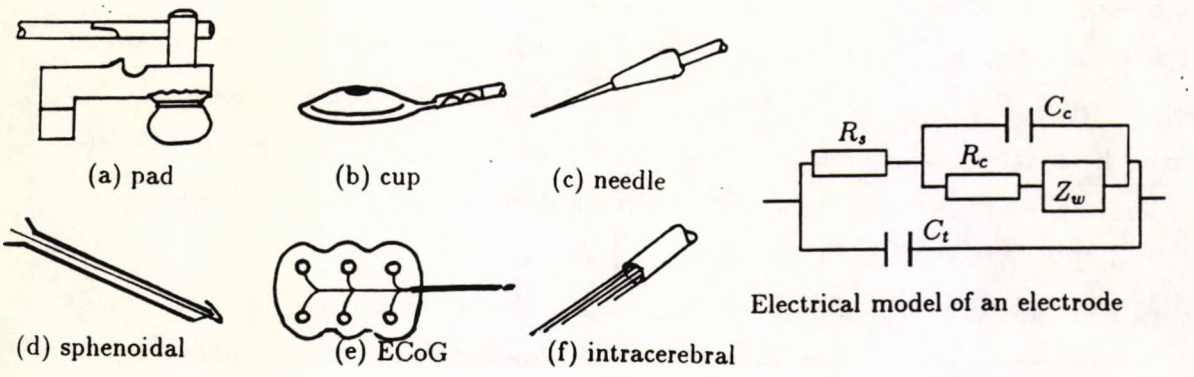


Figure 2.2: Electrodes used for recording brain potentials

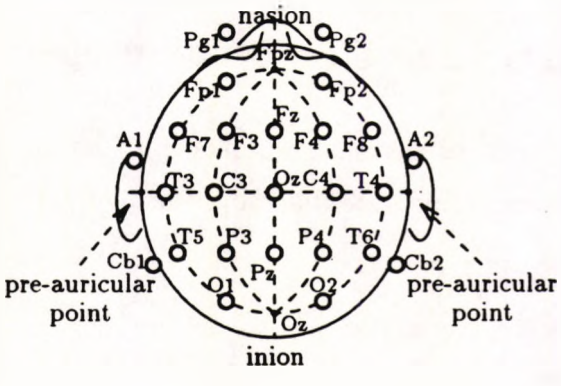


Figure 2.3: Electrode placement according to the 10–20 system

Electrode placement: the 10–20 system

Irrespective of the type of electrodes used, their placement on the scalp is quite standardized. The most widely-accepted standard, called the 10–20 system, was recommended by the IFSECN in 1958, after suggestions by Jasper, but not without criticism of its deficiencies[47, p. 92]. Electrodes are placed equidistantly in a mesh-like structure, covering the scalp[47, p. 94].

Electrode placement is based on four bony landmarks on the scalp, the *nasion*, the *inion*, and the *preauricular points*, (Figure 2.3). A systematic procedure for identifying the locations of the electrodes, using these points, was described by Cooper et al.[47, pp. 93–94] (also listed in [133]).

By convention, electrodes on the left side of the scalp have odd numbers, whereas those on the right have even numbers. The letters indicate the location, F for *frontal*, C for *central*, T for *temporal*, P for *parietal*, O for *occipital*. These are shown in Figure 2.3 which

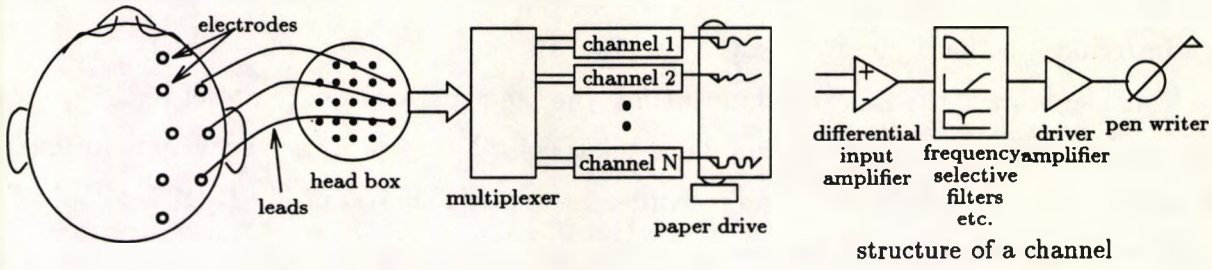


Figure 2.4: Block diagram of an EEG recorder

also depicts the cerebellar (C_{b1} and C_{b2}), nasopharyngeal (P_{g1} and P_{g2}) and *ear lobe* (A_1 and A_2) electrodes[47, p. 94].

EEG recorders usually have a limited number of channels and, during the course of a recording, they are switched between electrodes.

2.4.1 The EEG recorder

A block diagram of an EEG recorder is shown in Figure 2.4. At the input there is a 'head box', a diagram of the head with sockets at the electrode positions for the systematic, error-free connection of electrodes to the apparatus[177, p. 11]. A multiplexer, controls the connection electrodes to available channels. The rest of the recorder consists of identical channels, comprising amplifiers, filters and writing pens.

Each channel has a bipolar preamplifier to which two electrodes are connected. This has a high gain (several thousand) and input impedance (several $M\Omega$) and a high common mode rejection ratio[47, p. 57–67]. It should also have a linear frequency response and good noise and distortion properties in the range 0–100 Hz[177, p. 15]. A driver amplifier provides the necessary current to operate the writing pen (Figure 2.4). The gain is calibrated and adjusted using internal calibration signals and gain controls[47, p. 63].

Three frequency selective filters may be introduced if necessary at the output of the preamplifier to suppress unwanted frequency components. The *low frequency filter* is a high-pass filter whose time-constant may be chosen from a standard set[47, p. 72]. It is used to remove very low frequencies that make visual inspection of the EEG difficult. The *high frequency* (low-pass) filter, is utilized for the reduction of noise and other interference at the high frequencies of the EEG[47, p. 72]. Finally, the *notch filter* may eliminate narrow-band signals, like 50 or 60 Hz mains interference. Cooper et al.[47, p. 73] recommend the use of filters with caution, because of the phase distortion they introduce in the recorded signals.

The writing pens are usually responsible for high amounts of distortion because of the

slow response of the mechanical system they represent and the time and amplitude errors they introduce on the printed record[47, pp. 73–75].

Other elements of the EEG machine include the the paper drive with preset paper speeds, anti-blocking devices, modulators and demodulators for DC recordings, event markers, channels with lower gain and magnetic tape recorder interfaces[133][47, pp. 79–81][103, p. 36][177, p. 15].

2.4.2 Auxiliary equipment

It is customary not to have many pieces of equipment in the EEG laboratory, to minimize electrical interference. Some devices are often used as part of the EEG procedure, like the *stroboscope*, with variable flashing rate and devices producing auditory somatosensory and visual stimuli[177, p. 16]. An FM recorder is sometimes available too.

2.4.3 Monitoring techniques

The differential amplifiers of the recording apparatus permit the free selection of the reference to which the potential at an electrode may be measured. Different schemes, called *derivations*, have been implemented.

Bipolar derivations measure potential differences between consecutive electrodes[103, p. 42] (Figure 2.5), reducing the amplitude of common potentials and enhancing different ones. Graded potentials exhibit polarity reversals at the point of their maximum[47, p. 37].

In *common reference derivations* the same reference point is used for all channels (Figure 2.5). Electrically inactive reference points, like the ear lobes, the nose and the chin, are used[47, p. 38], although these are susceptible to some cerebral activity and interference[177, p. 14]. The *average reference* and *source* derivations[47, pp. 40–44][103, p. 43] (Figure 2.5) are variants suggesting alternative means of defining the reference point.

Montages (configurations) are guidelines for connecting sets of electrodes to the recorder[47, p. 96][103, p. 40]. According to these, negative potential differences should produce upward deflection of the recording pen. Channels should be connected sequentially in rows or columns, running anterioposteriorly or transversely, ordered from front to back and from right to left, as shown in the examples of Figure 2.6.

2.4.4 Electroencephalography as a medical diagnostic procedure

Nowadays EEG departments exist in many hospitals and are used by many doctors like neurologists, psychiatrists, general physicians, pædiatricians, surgeons and neurosurgeons for

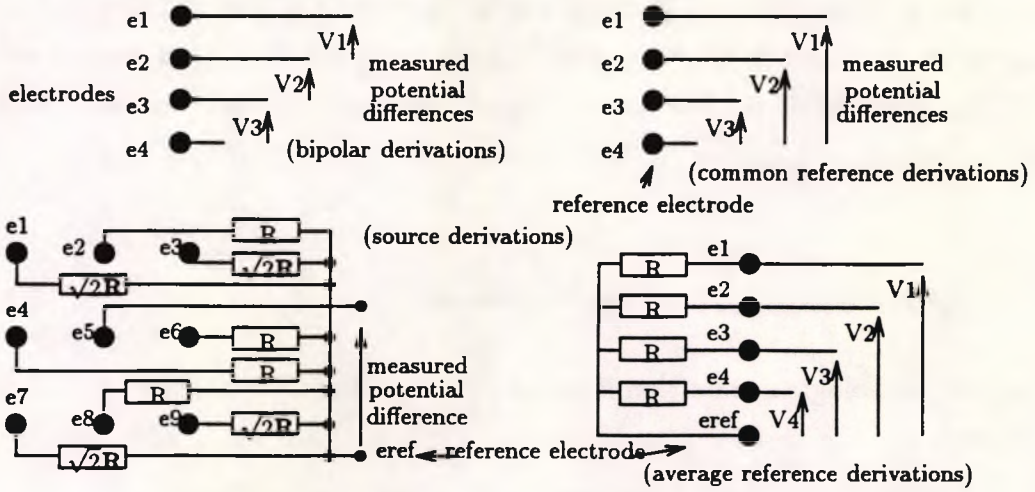


Figure 2.5: Standard schemes (derivations) used for EEG recording

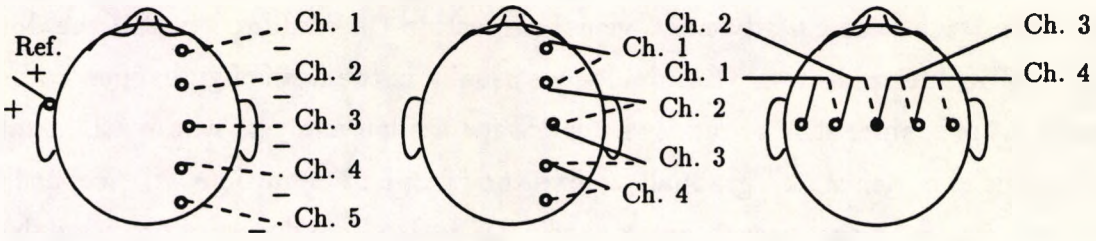


Figure 2.6: An assortment of electrode montages

diagnostic and preventive purposes. The preparation of the patient, placement of electrodes, calibration of the recorder and recording and annotating the EEG, marking abnormalities, state of consciousness, and artifacts (see sections 2.5 and 2.5.4) is the responsibility of an *EEG technician.* The EEG is then received and analysed by an expert *electroencephalographer*, who prepares a report containing a general description of the EEG, observations of significant EEG phenomena and abnormalities, possible clinical correlates and other comments[177, pp. 2–3, 173–178]. Sometimes, *the technicians have to produce the report without supervision*, due to restrictions of staff[177, p. 16,191].

2.4.5 Special monitoring techniques

When a specific pathological condition is suspected, techniques for provoking EEG phenomena may be employed to provide more evidence to support or reject a suspected diagnosis.

Hyperventilation (slow, deep breathing) induces slow-waves (section 2.5), characteristic of some pathological conditions[47, p. 102][177, p. 25]. *Photic stimulation* using a stroboscope[47, p. 103], provokes epileptiform discharges (section 2.5.2) in photosensitive epilepsy and other patterns in anxiety and migraine[103, pp. 177–179]. The *administration of specific drugs* is used for the diagnosis and localization of epileptic attacks (section 3.6). Finally, *sleep recordings* (spontaneous, or drug induced) may contain phenomena related to epilepsy[47, p. 105][103, pp. 64–68].

2.5 Visual analysis of the EEG

Despite recent technological advances, visual inspection ('eyeballing') is still the dominant method for EEG interpretation. Terminology is usually in the form of guidelines, rather than precise definitions, since it is primarily intended for human analysts who resolve ambiguity using previous experience and gradually refine their implicit definitions. Hence, although a clinical electroencephalographer interprets the EEG with a high degree of precision, he or she is often unable to explain why visually similar patterns have different significance, stating that it is "a matter of experience". In the brief account of EEG patterns that follows, the reader should note the qualitative nature of the 'definitions' and should treat any quantitative descriptions as guidelines, Texts on EEG analysis give more comprehensive descriptions[47, pp. 123–146, 187–227][103, pp. 52–68,201–221][177, pp. 18–32,107–118, 164–172] (see also [133]).

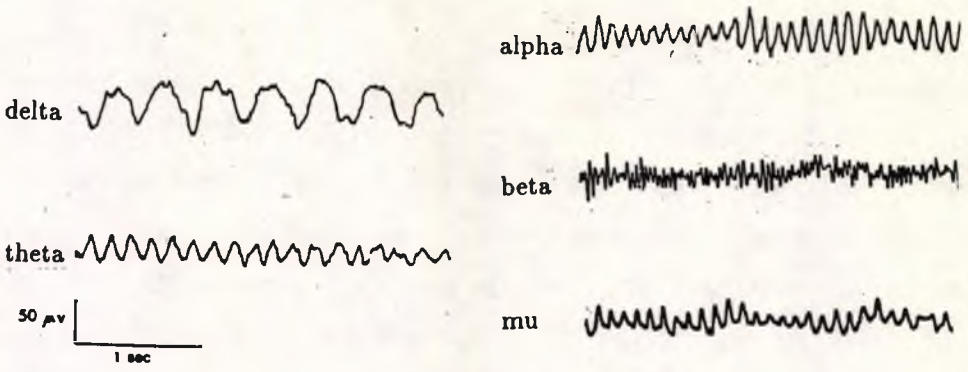


Figure 2.7: Normal EEG rhythms

(Reprinted from Scott[177])

2.5.1 Normal findings

The EEG of alert adults is characterized by *alpha (α) rhythm* (originally observed by Berger). It consists of rhythmic, quasi-periodic, almost sinusoidal oscillations of frequency between 8 and 13 Hz, occurring in both hemispheres and more prominent in the occipital regions, when the eyes are closed, disappearing with the opening of the eyes or mental concentration[177, p. 22] (Figure 2.7). Phase differences may exist between the two hemispheres[103, p. 52]. The frequency of the α rhythm in most people is almost constant having a mean deviation of about 0.5 Hz. In some individuals there is no α rhythm even with eyes closed whereas in some the rhythm re-appears shortly after eye opening. Complete unresponsiveness is an abnormal finding[103, pp. 52–54].

Beta (β) activity has frequency above 13 Hz and amplitude much less than that of α. It is observed in alert adults, especially in the frontal and central regions (Figure 2.7). It is not affected by eye opening and closure but may be altered by the abuse of some drugs[103, p. 55].

The other two frequency bands of activity in normal subjects are *theta* (θ) and *delta* (δ), having frequencies from 4 to 8 Hz, and below 4 Hz, respectively[47, p. 126] (Figure 2.7). θ waves are sometimes observed on the vertex and the temporal regions especially during drowsiness[103, p. 54]. Delta activity is generally an abnormal characteristic in adults, but quite normal in children (see later in this section). It is difficult to determine limits on amplitudes and frequencies beyond which θ and δ activity may be considered abnormal[103, pp. 25,55].

Finally, *mu (μ) rhythm*, a sharp, comb-like signal with frequency between 7 and 11 Hz, is relatively rare in the EEG, but prominent in the ECoG of the central regions (Figure 2.7). The μ rhythm is abolished by movements, or the intention of moving[103, p. 56][177, p. 23].

Age dependency

Changes are observed in the EEG with ageing, especially noted in the weeks following birth. Studies of prematurely born babies show discontinuities in the EEG. The EEG of neonates consists of continuous irregular and desynchronized delta activity, gradually becoming more regular. Theta activity slowly appears and by the age of 12 months it dominates the EEG. Alpha rhythms are also observed, but these are not yet localized to the posterior regions, until the age of 5. With increasing age, the other rhythms disappear in favour of the alpha activity whose amplitude decreases from 50–100 μV, during adolescence, to 20–40 μV in adult life[177, pp. 108–119].

The EEG and the level of consciousness

During drowsiness and sleep the monotonous α rhythm of an awake adult gives way to a variety of of EEG patterns[177, p. 134].

Sleep is separated into various phases, or stages, beginning with drowsiness. This is characterized by alternating periods of α waves and low-amplitude ragged slow-wave activity (section 2.5.2). Eventually, alpha activity disappears and theta activity with some beta components appear in sleep *stage 1*, which lasts for a few minutes and is followed by *stage 2*. Generalized, bilaterally synchronous *k-complexes*[177, p. 137], transient δ waves, *sleep spindles* (often associated with *sharp waves*) and vertex sharp transients[47, pp. 131–132] are observed. *Sleep spindles* dominate the frontal and *lambdoid waves* (resembling λ waves) the occipital regions. Definitions of these patterns may be found in section 2.5.2.

Stage 3 occurs approximately 45 minutes after the onset of sleep and contains mainly delta activity with some of the phenomena of stage 2. In *stage 4*, high-amplitude, irregular delta activity is the main characteristic.

Stage 4 is succeeded by *rapid eye movement (REM)* sleep, which lasts for about 20 minutes[177, p. 139]. The EEG is similar to that of wakefulness, the eyes move rapidly in clusters and dreaming occurs.

Patterns characteristic of each sleep stage are shown in Figure 2.8. During a night's sleep, the cycle of the five sleep stages is repeated four to five times.

2.5.2 Transient patterns in the EEG

Most of the EEG signals described so far appear, persist and characterize whole stretches of the EEG (cf. alpha activity). These are often called *rhythms* to emphasize their repetitive nature[41]. The rhythm may change from time to time, but their sequence constitutes the

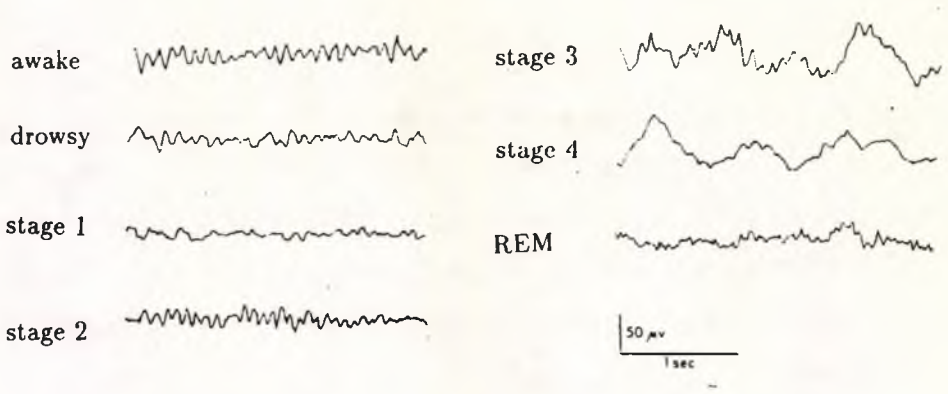


Figure 2.8: The EEG during sleep

(Reprinted from Scott[177])

(quasi-periodic) *background activity*, providing the setting where other brief *transient* signals may occur. These are isolated waveforms, or complexes[1], which are distinguished from the background activity. These are described in a glossary of terms used by clinical electroencephalographers, compiled by Chatrian et al.[41] on which the list of common transients that follows was based.

K-complex: It consists of a slow wave followed by a sequence of waves with a frequency of approximately 14 Hz. The amplitude is maximum at the vertex (C_z) and is variable but approximately 200 μV. It appears spontaneously, or in response to a sudden stimulus during sleep.

Lambda (λ) waves: These are sharp transients, generally positive with respect to other areas, occurring in the occipital regions during visual exploration, in awake subjects. They are related to eye movements and have variable amplitude, usually below 50 μV.

Kappa (κ) waves: Bursts of alpha or theta waves, recorded from the temporal regions during intense mental activity. Their cerebral origin has not yet been confirmed.

Mu (μ) waves: Stretches of arch or comb-shaped waveforms with frequency between 7 and 11 Hz, most prominent in the central and parietal regions, during wakefulness. They have an amplitude of less than 50 μV, which is severely attenuated, or blocked, by readiness to move, or thought to move, or actual movement or tactile stimulation. The blockage refers to contralateral movements.

Sleep spindle (s-wave): A succession of waves at a frequency of approximately 14 Hz prominent over the frontocental regions maximum at the vertex (C_z) and amplitude up to

[1] a complex is a succession of more than one wave which appear consistently as a group, distinguished from the background activity

50 μV. It appears during some sleep stages.

Sharp wave: A transient with a pointed peak (at conventional paper speeds), of duration between 70 and 200 ms and very variable amplitude, which is negative with respect to other areas. Its should appear distinct from background activity and should not be confused with lambda waves. Its morphology is similar to the spike (whose description follows); the distinction between spikes and sharp waves is rather arbitrary and serves descriptive purposes only.

Slow wave: Is any single wave with duration longer than alpha waves ($\frac{1}{8}$ s).

Sharp-and-slow-wave complex: This consists of a sharp wave followed by a slow wave.

Spike: A transient with a pointed peak (at conventional paper speeds), duration between 20 and 70 ms and negative polarity compared to other areas and variable amplitude. It should be distinguished from the on-going (background) activity and should not be confused with spikes recorded from individual neurons using microelectrodes. Spikes are similar to sharp waves, described earlier.

Spike-and-slow-wave complex: A pattern consisting of a spike, followed by a slow wave. The term is often used to describe a rhythm of 3 per second spike-and-slow-waves which is described below.

3 Hz Spike-and-slow-waves: It is a paroxysm consisting of regular sequences of a spike and a slow wave with a period of repetition of 2.5 and 3.5 Hz. The paroxysms occur bilaterally and synchronously and are generalized, although their amplitude, which is variable, is often larger in the frontal regions where it may reach 1000 μV.

Multiple spike-and-slow-wave complex: It resembles the-spike-and-slow-wave complex, but the slow wave is preceded by two or more spikes. It is also called a *polyspike-and-slow-wave-complex*.

Vertex sharp transient (V-wave): A sharp potential of duration 70–200 ms maximal at the vertex (C_z), negative in relation to other areas. It may appear spontaneously during sleep or during stimulation at sleep or wakefulness. Its amplitude does not exceed 300 μV. It is more prominent in children during sleep.

From the above transients, spikes, sharp waves, sharp-and-slow-waves, spike-and-slow-waves and multiple spike-and-slow-waves (Figure 2.9) are characteristic patterns of some forms of *epilepsy* (see Chapter 3).

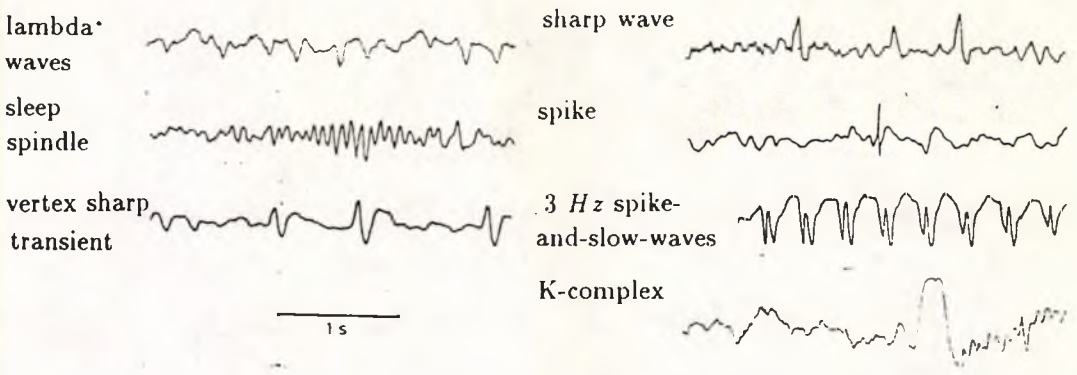


Figure 2.9: Examples of some EEG transients

(Reprinted from Scott[177])

Non-spintaneous transients, like *event-related potentials*, which are linked to stimuli, are beyond the scope of this report. They are discussed extensively in some of the references[47, pp. 222-225][177, pp. 168–171] (see also [133]).

2.5.3 The abnormal EEG

It is sometimes difficult to qualify an EEG as abnormal, since the abnormality and the interpretation depend on many physiological and psychological factors, as explained in texts on clinical electroencephalography[47, 103, 177]. For example an EEG with slow components is abnormal for an adult, but perfectly normal for a child (section 2.5.1). Sleep records contain components that are pathological indications if observed during wakefulness (section 2.5.1). Changes in the sequence and content of sleep stages is sometimes considered abnormal. Transients, like spikes and slow waves are often related to conditions like epilepsy, which is implicitly linked with this thesis, and is discussed separately in the next Chapter.

2.5.4 Artifacts

According to Scott[177, p. 18], potentials of non-cerebral origin, called *artifacts* in EEG nomenclature, are one of the most impressive feature of a record, due to their large amplitudes. They are a permanent headache in EEG interpretation, which is sometimes deemed impossible[47, p. 111]. They arise from a variety of sources, the main ones outlined here.

External, non-biological interference

Electrical induction between external electrical sources and the recording electrodes or unscreened parts of the recorder[47, p. 111] causes these potentials. Mains interference appears in the form of sinusoidal oscillations of frequency 50 or $60Hz$[103, p. 47] (Figure

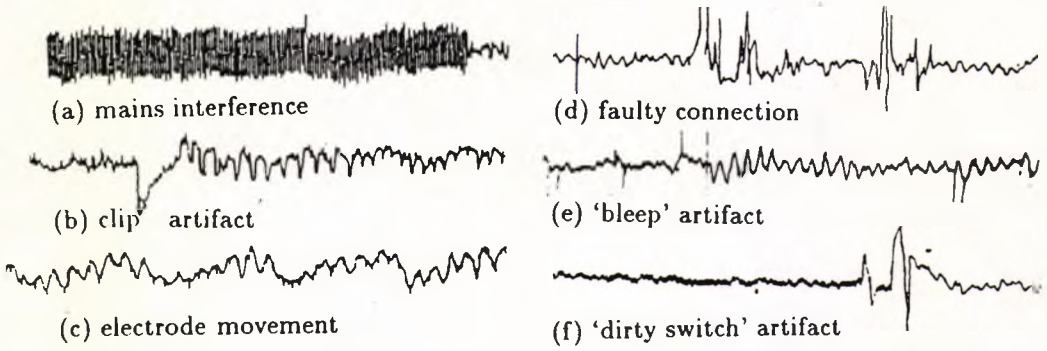


Figure 2.10: Non-biological artifacts

(Reprinted from Scott[177])

2.10 (a)). Other sources include electrostatic charge on clothing, unscreened transformers, telephones[177, p. 23] and, less frequently, radio interference[47, p. 113].

Ageing equipment is likely to cause increased noise and non-linearity as well as gain variations. These artifacts are not easily identified, because they occur progressively[103, pp. 46–47].

The disturbance of the electrical contact between an electrode and the scalp or the head box, defective internal connections and dirty switches result in changes in impedance and, hence, the recorded potentials (Figure 2.10). Technicians and electroencephalographers need to resort to 'experience' for their rejection because they sometimes resemble cortical potentials[47, p. 113]. Hence their prevention is preferable[47, p. 114][103, p. 47].

Artifacts of biological origin

Various functions of the body generate, directly or indirectly, electrical potentials that are picked-up by the EEG electrodes.

The potential difference between the cornea and the retina is the cause of *ocular artifacts*[47][p. 115], which are quite common. Eye closure causes positive and eye opening negative electrical disturbances on the anterior electrodes (Figure 2.11(a),(b)). Similarly, eye movements, blinking, drowsiness etc. cause a variety of artifacts, especially in the frontal electrodes[47, p. 116] (Figure 2.11 (c),(d)).

Uncomfortable and anxious patients have tense muscles, which give rise to spiky potentials, especially in the temporal regions[103, p. 48]. These *muscle potential artifacts* may run continuously, in bursts or rhythmically (Figure 2.11 (e),(f)). Although filtering could reduce their amplitudes, the residue may be indistinguishable from genuine cerebral rhythms[47,

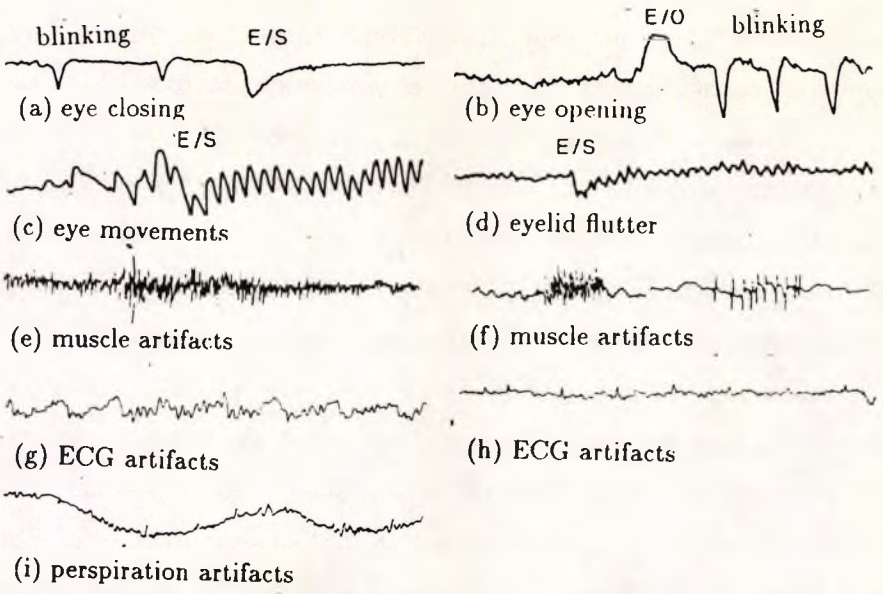


Figure 2.11: Examples of biological artifacts

(Reprinted from Kiloh et al.[103] and Scott[177])

p. 117]. Head and body movements in children and unconscious patients also cause muscle artifacts[177, p. 23].

Heart activity is a source of artifacts also. These are either induced by the pulse, to electrodes placed close to arteries or by the direct recording of the cardiac electric field due to the QRS (heartbeat) complex. The first has the form of a saw-tooth waveform, whereas the second has the shape of QRS complexes, with the R-wave being more prominent (Figure 2.11 (g),(h))[47, p. 118]. This artifact is normally absent in bipolar recordings[103, p. 48].

Perspiration is the cause of low-frequency waves because of the disturbance of the skin and electrode contact resistances[103, p. 49] (Figure 2.11 (i)). The high-frequency filter is sometimes useful in their elimination[47, p. 119].

2.6 Uses and limitations of electroencephalography

Speaking in general terms, EEG recording is a non-invasive and relatively inexpensive procedure. It can be repeated frequently (if necessary) and has the ability to show both functional and structural abnormalities in the brain. The EEG is often an early prognostic tool for tumours, and hæmatomas[103, pp. 222–228] and some psychiatric conditions. It finds uses in the monitoring of head injury, cerebral infections, anæsthesia and coma[103, pp. 222–228] and in the exclusion of neurological factors as the cause of other symptoms[177, p. 169].

It is rather unfortunate that EEG observations are non-conclusive since the number

of neurons represented is small[177, p. 189], the analysis subjective, affected by artifacts (section 2.5.4), the level of consciousness and other physiological factors[103, pp. 222–228] (section 2.5.1). This was used by some to discredit its diagnostic value[103, pp. 222-228][120, 178] although these problems are also encountered with other clinical tests. The complexity of EEG interpretation puts the reliability of its assessors in doubt. In a reliability study carried out by Woody[103, p. 223], low intra- and inter-reader reliability was indicated.

The recording of the magnetic, instead of the electric field was suggested by the transparency of the brain and scalp to magnetic fields to avoid the problems listed in section 2.2.3. Recording equipment, however, is bulky and shielding stray magnetic fields difficult.

Nowadays the EEG is not the only method for investigating brain disorders. Brain imaging devices, are generally available in large hospitals[48, 58]. *Computed Axial Tomography* (CAT) and *Magnetic Resonance Imaging* (MRI) produce three-dimensional images of the brain and are better than the EEG in showing and localizing structural abnormalities in the tissue, like brain tumours.

The EEG is of great value in the diagnosis and treatment of constitutional epilepsy[19] (see next Chapter), for example petit mal in children[103, pp. 222–228][177, p. 193], which is normally a transient functional disorder occurring in the absence of structural brain abnormalities.

Chapter 3

Epilepsy and the EEG

3.1 Introduction

The nervous system (Chapter 1) is immensely complex and spans large areas. It is, therefore, unlikely that a stimulus or a malfunction will remain localized. Defects, however, are usually detected and confined by the brain itself and pass almost unnoticed. The fault tolerance of the brain is often capable of maintaining most of its functionality even when sections of it get damaged.

Nevertheless, damaged neurons are irreplaceable, the tissue delicate and sensitive to chemical imbalances, which disturb the function of the nervous system. Head injuries, viruses, intoxications and chemical variations in the body sometimes lead to symptoms like headache, delirium, paræsthesia, seizures and behavioural changes[165, pp. 3–58]. Distinguishing causes and effects is sometimes impossible and the name *disorders* has been adopted to collectively describe both symptoms and diseases. Their persistence may be transient or permanent and their severity variable. Texts on neurology[165, 193] give a comprehensive description of many such disorders.

One of these, *epilepsy*, has peculiar, often ambiguous causes. It is generally associated with abnormal EEG characteristics, more prominent during an epileptic attack (seizure), a transient phenomenon. Despite the strong relation between specific EEG abnormalities and epilepsy, the evidence in the latter is inconclusive, because the EEG of some people suffering from epilepsy is normal and vice-versa[22].

In this Chapter the causes, mechanisms, classification, clinical symptoms and treatment of epilepsy are briefly outlined. The emphasis is on the EEG patterns characterizing each type of epilepsy. More details may be found in texts on neurology[165, 193] and in an internal report[134].

3.2 What is epilepsy?

Epilepsy is a condition that cannot be defined unambiguously. Gibbs (1937) defined epilepsy as "paroxysmal cerebral dysrhythmia". Walton[193, p. 1093] described it as

> *... a paroxysmal and transitory disturbance of the functions of the brain which develops suddenly, ceases spontaneously and exhibits a conspicuous tendency to recurrence. Though in its most typical forms it is characterized by the sudden onset of loss of consciousness, which may or may not be associated with tonic spasm and clonic contractions of the muscles, many varieties of epileptic attacks occur, their distinctive features depending upon differences in the site of origin, extent of spread and nature of the disturbance of function ...*

According to McLeod and Lance[122] it is

> *... a recurring disturbance of cerebral function caused by paroxysmal neuron discharges, usually (but not always) accompanied by loss of consciousness ...*

and Goldensohn et al. avoid to define it explicitly, suggesting that it is *a group of symptom complexes that have many causes ... causing cerebral neurons to become excessively excited*[165, p. 780]. Definitions have been subject to debate, branded by their critics as imprecise, if not meaningless[103, p. 70]. Descriptions often clash with social and legal implications and thus many clearly epileptic attacks have been given different names or considered a symptom of other disorders, when these have been diagnosed. In conclusion, calling epilepsy either a symptom or a disease is neither entirely true nor false. As explained later, there is evidence that all individuals may suffer from seizures, given the necessary provocation[73]. According to Brown, little is really known about epilepsy and there may be a relation in its mechanisms and those of other non-epileptic disorders[30].

3.3 Causes of epilepsy

Epilepsy is caused by various local factors or widespread abnormalities or both. The type of seizures occuring are associated with these causes and their interactions. Minor brain abnormalities, sometimes observed, may be the consequence as well as the cause of the disorder[193, p. 1100].

3.3.1 Local factors

When the seizures mostly affect one brain region (section 3.6.2), their causes are usually localized also. The indications in the EEG reflect their origin and, sometimes, the causes responsible[193, pp. 1095–1096][122] (see also [134]). Common local factors include:

- *Intracranial lesions* (like tumours)
- *Inflammatory and demyelinating conditions*
- *Brain injuries*
- *Congenital abnormalities* (like tuberous sclerosis)
- *Inborn errors of metabolism and degenerations*
- *Vascular disorders* (like cerebral hæmorrhage)
- *Reflex causes* (stimuli)

3.3.2 General factors

When the onset of the seizures cannot be localized, factors that affect the whole of the brain may be responsible. According to Walton[193, pp. 1095–1096] the main general causes are the following:

- *Exogenous substances*, like drugs and poisons
- *Anoxia*, caused by asphyxiation etc.
- *Disordered metabolism*, when it affects neurons and the concentrations of ions and neurotransmitters[122, p. 233]
- *Endocrine disorders*
- *Infections during childhood*[122]

The rôle of *hereditary factors* in epilepsy is a controversial issue[63, pp. 783–785], but there is evidence that a predisposition to generalized seizures is inherited[122, p. 233]. *Psychological factors* may provoke a seizure, although they are not the primary cause of epilepsy[193, p. 1096].

3.3.3 Constitutional epilepsy

The exclusion of known factors leaves a considerable number of cases in which the cause of epilepsy cannot be identified. The name *constitutional (idiopathic) epilepsy* has been adopted to describe these cases, which are possibly the result of a predisposition to seizures or the complex interaction of factors too diverse or too specific to a patient[193, p. 1097]. The epileptic threshold theory (section 3.4.3) is a possible explanation.

3.4 Neurophysiology of epilepsy

The study of the mechanisms of epilepsy is not easy, as neural interactions cannot be studied in detail and seizures are transient and sometimes violent. Nevertheless, they appear to result from the diffusion of transient abnormal discharges of neurons to the cortex[70, 193].

3.4.1 Neural mechanisms of epilepsy

Several intrinsic and extraneous factors control the function of a neuron (see section 1.3) and their deviation from normal levels may lead to extraordinary responses and the concept of the *epileptic neuron* (see also [134]).

The mechanisms that maintain the resting membrane potential (sections 1.3.1 and 1.3.2) may deteriorate. If the membrane potential becomes less negative than it should, the neuron may be excited more easily and vice versa resulting in a predisposition to abnormal discharge patterns.

Abnormal stimulation (section 1.3.3), assisted by disturbances of the resting potential, may cause an extraordinary neuron response. Neurons are generally tolerant to small timing discrepancies and the unexpected absence or presence of a stimulus, but many simultaneous abnormal or strong PSPs may cause erratic discharges. Deviations in the strength of postsynaptic potentials caused, for instance, by imbalances in the concentration of neurotransmitters (section 1.3.3), alter the firing pattern of neurons[193, p. 1095].

Small operational discrepancies of a few neurons are counteracted by self-regulating mechanisms. These may not suffice when the discrepancies are large or generalized and there is a point, the *epileptic threshold*, beyond which the neural function may break down. According to this theory, all individuals are liable to seizures, given the necessary provocation[73][193, p. 1095] and those suffering from epilepsy have neurons that normally operate near their threshold, where small provocations may cause a seizure.

3.4.2 Effects on groups of neurons

Chemical imbalances and other factors (section 3.3) usually affect *groups* of neurons making them liable to discharge abnormally, spontaneously, or after stimulation. Consequently, other neurons become depolarized and discharge rhythmically in rapid succession causing an avalanche of discharges, encouraged by two factors. The insufficiency of the IPSPs caused by inhibitory stellate neurons (section 2.2) and neural 'kindling' (the increased probability of a neuron discharging following a series of periodic discharges)[122].

3.4.3 Disturbances on the cortical scale

Cortical regions communicate with each other (section 1.4.1) and with subcortical structures like the thalamus and the reticular formation (sections 1.4.2, and 1.4.5). These connections are evident in the similarities of the corresponding electroencephalograms and their relative delays show whether communication is direct or mediated by subcortical structures[70][103, p. 8] (section 2.2).

Gastaut assumed that the brain consists of interacting groups of cortical and subcortical clusters of neurons[103, p. 92]. Therefore, a malfunction in either the cortical *or* the subcortical cluster could affect the other element of the group. Partial seizures are confined to a few clusters and usually originate in one of them, called the *epileptic focus*. Generalized attacks affect the whole of the cortex and either originate in subcortical structures, which make connections to large numbers of cortical neurons (section 3.6.1), or in a cortical cluster, which is relayed to a subcortical structure (section 3.6.3).

3.5 Pathology–classification of epileptic seizures

Epileptic attacks have diverse symptoms, varying origins and debatable causes. It is not surprising that there are many ways to classify them.

3.5.1 International classification

The universally agreed classification, accepted by the International League Against Epilepsy, the World Federation of Neurological Societies and the IFSECN[103, p. 72] is based on the physical manifestation of seizures and the extent of the disturbance of consciousness. It is based on a suggestion by Janz (1969) amended by Sutherland et al. (1974) and is shown in Table 3.1.

1. Generalized seizures (Bilaterally symmetrical without local onset)
   (a) Absences (petit mal)
   (b) Bilateral myoclonous
   (c) Infantile spasms
   (d) Clonic seizures
   (e) Tonic seizures
   (f) Tonic-clonic seizures
   (g) Akinetic seizures
2. Partial seizures (Seizures beginning locally)
   (a) Elementary symptomatology
      i. Motor[1]
      ii. Sensory[1]
      iii. Autonomic
   (b) Complex symptomatology
      i. Impaired consciousness[2]
      ii. Complex hallucinations[2]
      iii. Affective symptoms[2]
      iv. Automatism[2]
   (c) Partial seizures becoming generalized tonic-clonic seizures
3. Unclassified seizures (Seizures that cannot be classified due to incompleteness of data)

[1] Jacksonian Epilepsy

[2] Temporal Lobe Epilepsy

Table 3.1: International classification of epileptic seizures

This classification provides no information on the the mechanisms and origins of seizures and is ambiguous on the classification of partial seizures that later become generalized. Alternative classifications that rely on neurophysiological findings rather than the symptomatology of seizures do not suffer from these problems.

3.5.2 An alternative classification

Jasper and Kershman in 1949 proposed a classification of epileptic seizures based on the EEG and other neurophysiological findings. This was modified by Hill in 1958 and distinguishes seizures into four groups[103, p. 73],

1. Primary subcortical epilepsy
2. Primary cortical epilepsy
3. Secondary subcortical epilepsy
4. Epilepsy arising from numerous areas both in the cerebral cortex and subcortical structures.

These categories are related to the international classification; subcortical epileptic seizures are usually generalized, whereas cortical ones are of local nature. This classification was adopted in the next sections, because of its relation to the EEG.

3.6 Symptoms of epileptic attacks

The emphasis of the brief description of the symptomatology of epileptic attacks that follows is on EEG manifestations. More complete descriptions of epilepsy and other disorders of the nervous system may be found in Walton[193] and Kiloh et al.[103, pp. 71–104] (see also [134]).

3.6.1 Primary subcortical epilepsy

There is strong evidence that subcortical structures, like the thalamus and the reticular formation (sections 1.4.2 and 1.4.5), are actively involved in the onset of generalized epileptic seizures[103, p. 73], which are associated with synchronous, widespread patterns in the EEG. The common characteristic of primary subcortical epilepsy is the transient loss of consciousness. Symptoms depend on the type of the attack, whether major, minor or myoclonic.

Major epilepsy (grand mal)

The extensive muscle activity during grand mal seizures limits the clarity of EEG recordings, which usually contain muscle artifacts[103, p. 74] (section 2.5.4). Walton [193], divides a seizure into the aura, the convulsion stage and the post-convulsive phase.

The aura, a warning that may precede the forthcoming seizure characterizes attacks that begin locally (section 3.6.3). Feelings of familiarity (*déjà vu*), unreality (*jamais vu*), disembodiment and inexplicable fear are common. Visual or auditory hallucinations, vertigo, clonic movements or spasms may be present.

The convulsion sometimes begins with the epileptic cry followed by loss of consciousness, a symmetrical *tonic* spasm on both sides of the body, and arresting of respiration for 15 to 30 seconds. The EEG contains fast activity, followed by generalized rhythmic high-amplitude waves with frequency 8–12 Hz, which decreases towards the end of the phase. The *clonic phase* that follows, is characterized by bursts of spikes, occasionally followed by slow waves, coinciding with sharp, interrupted muscle movements. Biting of the tongue, foaming in the mouth, urine incontinence and pupil dilation are possible.

In the *post-convulsive phase*, spikes become infrequent and slow waves appear. Their frequency and amplitude decreases[103, p. 75] leading to unconsciousness and then to normal sleep, as the EEG and the mental state returns to normal.

Attacks of major epilepsy in young children are called *febrile convulsions* and may lead to the development of an epileptic focus.

Minor epilepsy (petit mal)

Petit mal epilepsy constitutes of mild idiopathic attacks, characterized by transient impairment of consciousness[103, p. 75] that begin in childhood[193, pp. 1103–1104]. These last for a few seconds and are characterized by a blank facial expression, rolling of eyes in the upper lids, momentary inhibition of activity, muscle relaxation and, sometimes, a short myoclonic twitching. Several attacks per day may be present[193, p. 1104]. These may continue during adolescence and adulthood, occasionally developing into major attacks.

The EEG during minor epileptic attacks contains sequences of spike-and-wave complexes (section 2.5.2) occurring simultaneously over the whole of the scalp, but more profound in the frontal regions[103, p. 75]. Such paroxysms repeat at an average rate of 3 per second, faster at the beginning and slower at the end of the seizure. Sleep EEG recordings often contain polyspikes, polyspikes and waves, as well as spike-and-wave discharges during REM sleep, becoming more irregular during stage 4 sleep[103, p. 76]. According to Gastaut and Fischer-Williams (1960) petit-mal may characterize individuals with a very strong inhibitory

mechanism in their reticular formation where the attacks are believed to originate.

Other forms of epilepsy with local onset may be clinically indistinguishable from true petit mal.

Myoclonic attacks

Attacks of myoclonic epilepsy resemble grand mal seizures (section 3.6.1) in miniature and may be provoked by sudden, unexpected noises and photic stimulation. Bilaterally symmetrical jerking of the face and the upper limbs is the main visible symptom of this type of epilepsy. Loss of consciousness is extremely brief[103, p. 77]. The EEG may contain transient discharges, resembling spike-and-wave complexes (section 3.6.1) or polyspike-and-wave patterns (section 2.5.2). Myoclonic attacks are related to centrencephalic structures, like the reticular formation and the subcortical motor system.

Interseizure patterns

These are 'larval forms' of epileptic discharges appearing in a clinical EEG recording of a large proportion of sufferers. They are significant for clinical assessment, although their presence or absence is sometimes inconclusive as Kiloh et al.[103, p. 79] emphasize.

A typical interseizure pattern of primary subcortical epilepsy is the symmetrical, generalized, bilaterally synchronous spike-and-wave complex (epileptiform) characterizing 50% of grand mal and 80% of petit mal cases[103, p. 78]. In some cases, spikes appear with very low amplitude and the pattern resembles δ activity.

These patterns or even epileptic attacks may be induced using hyperventilation, photic stimulation and the administration of drugs like Chloropromazine (see section 2.4.5).

3.6.2 Primary cortical epilepsy

When local factors (section 3.3.1) affect a group of cortical neurons, called an *epileptic focus*, they may exhibit spontaneous abnormal rhythmic discharges. These may progressively increase in both amplitude and frequency, affecting neurons located in the vicinity of the focus, causing a seizure[103, p. 85].

The local character of the abnormality is indicated in the EEG, which may contain spikes, sharp waves or spike-and-wave complexes and, occasionally, α, δ or θ waves, primarily in the area around the epileptic focus[103, p. 92]. Their frequency of occurrence decreases from 12 to 2 per second during the end of the seizure.

A mirror focus may appear in the other hemisphere and, although its activity is initially synchronized with the primary focus, due to the intervention of subcortical structures, it later becomes autonomous. If many independent foci exist, epileptic activity may shift between them. In some cases, the foci appear to move[103, pp. 89–91]. The migration of a focal discharge to a subcortical structure, may precipitate a generalized attack (section 3.6.3). Occasionally the epileptic focus is in a subcortical structure but projects only to a small area of the cortex.

Psychomotor attacks, a common form of primary cortical epilepsy, arise from foci in the temporal or the frontal lobes[103, p. 85]. Auras, when present, are characteristic of the location of the focus. Seizures are accompanied by confusion, clouding of consciousness, memory and personality disturbances, hallucinations, fear, anxiety, depression and other syndromes[193, p. 1105]. The EEG sometimes contains spikes and sharp waves which quickly diminish in amplitude and frequency, becoming synchronous discharges of frequency 2–10 Hz, occasionally interrupted by sharp waves[134]. These are asymmetrical to one of the hemispheres.

Other forms of this type of epilepsy, include *Jacksonian epilepsy*, originating in one of the motor areas[193, p. 1105] and *sensory epilepsy* is related to foci in the sensory areas of the cortex[134].

Common interseizure patterns include surface-negative spikes of duration 20–60 ms having amplitudes of 25–500 μV. Interseizure patterns and seizures may be provoked by hyperventilation, photic stimulation, sleep and drugs, such as Chloropromazine. They are used to detect and localize the focus and its nature.

3.6.3 Secondary subcortical epilepsy

The name is used to describe several varieties of cortical epilepsy that exhibit secondary bilateral synchrony. This occurs when a focal discharge is transmitted to subcortical structures and then relayed back to the whole of the cortex, to cause spike-and-wave discharges. Unlike minor epileptic attacks (section 3.6.1) these are irregular, variable and often asymmetrical with amplitudes higher in the hemisphere containing the focus[103, p. 85].

Primary and secondary subcortical epilepsy are often discriminated by the injection of *sodium amylobarbitone*, which suppresses focal but not generalized abnormalities[103, p. 97] (see also [134]).

3.6.4 Epilepsy arising from numerous cortical and subcortical areas and structures

In this type of epilepsy the EEG is abnormal even between seizures, characterized by widespread bilateral irregular spike-and-slow wave discharges of frequency 2–4 Hz superimposed on a background of diffuse slow activity. Frequently there exist multiple independent focal sources of polyspikes, sharp waves and spike-and-wave complexes[103, p. 98]. Causes include diffused or scattered brain lesions, trauma at birth, various infections and intoxications and degenerative cerebral disorders.

Common forms include *hypsarrhythmia* and *phenylketonuria* in infants and *myoclonous epilepsy*[103, pp. 99–100] (see also [134]). The first is characterized by high-amplitude slow activity, and the second by the rapid development of high-amplitude sharp or slow-wave epileptic foci. In myoclonous epilepsy the EEG is grossly abnormal, containing θ, δ and β components and multiple spikes, of subcortical origin.

3.6.5 Other epileptic and similar conditions

Status epilepticus

Prolonged or repetitive seizures without relapse of consciousness result in a continuous state of epilepsy that is evident in both the electroencephalogram and the clinical findings[103, pp. 101–102][193, p. 1108] (see also [134] for a brief description). The term *convulsive status* is often used to describe sequences of major epileptic seizures without intermediate recovery periods[3]. Petit mal status, psychomotor status and epilepsia partialis continua[103, pp. 101–102] are prolonged attacks of other types of epilepsy, lasting for hours or even days.

Epilepsy with normal EEG

If a focus is too discrete, abnormal activity may be too localized or not hyperpolarized to cause a significant scalp potential (see section 2.2.3). EEG abnormalities may also be absent if the focus is located a few millimetres below the cortical surface or in a deep subcortical structure, as the discharges may not reach the surface of the brain.

Other EEG-related brain disorders

Not all transient attacks are epileptic in nature. *Tonic fits*, caused by tumours, high intracranial pressure, and damage of the brain stem[103, p. 104], *syncopal attacks* following

cardiac arrest or acute cerebral anoxia and *breath-holding spells* caused by cerebral anoxia may be discriminated from epilepsy using the EEG.

3.7 Management and treatment of epilepsy

3.7.1 Medical treatment

Once the cause of the seizures has been identified, medical treatment may begin. The earlier the treatment begins the greater the probability of controlling the seizures effectively, although the possibility of an initial misdiagnosis is higher[39, 160]. Richens[160] gives comprehensive guidelines concerning the choice of drug, monitoring and discontinuing the treatment. Anticonvulsant medication is usually administered, usually in the form of a single drug (monotherapy). The dosage is gradually increased until a therapeutic level is reached[193, p. 1120][122, p. 236], as an overdose may affect the alertness and intellectual ability of the patient[13][122, p. 236][193, p. 1120]. The administration of more than one type of anticonvulsant is normally avoided, because side effects are additive[13][193, pp. 1120–1121]. For partial and tonic-clonic seizures *phenytoin*, *carbamazepine*, *sodium valproate*, *barbiturates* and *clonazepam* are used. For minor epilepsy and myoclonic attacks *sodium valproate* and *ethosuximide* are more common. These are described in [122][pp. 237–240] (see also [134]). Usual side effects include nausea, sedation, vertigo, drowsiness, irritability, weight gain, hepatic damage, coarsening of features and others.

In some cases where seizure control was not effective, change or increase of types of drugs or even their decrease may be useful[2, 73]. If anticonvulsant medication does not control seizures effectively, surgical intervention, like anterior temporal lobectomy, may be suggested for some types of focal epilepsy[122, p. 240]. Surgical treatment is becoming more rare, as a result of the increased effectiveness of anticonvulsants[193, p. 1118].

3.7.2 Social issues

The personal and social stigma following the diagnosis of epilepsy may be more devastating that the seizures themselves[172, p. 81]. This mystery of epilepsy made the sufferers subject to ritualistic abuse, as it was attributed to 'possession'[172, p. 40]. School children should be encouraged to continue their usual activities[122, p. 235] to acquire social skills and confidence needed for a normal adult life. Some types of occupation are unsuitable for individuals suffering from epilepsy[122, p. 82], but some gainful employment is recommended as a source of self esteem, satisfaction, companionship, status and independence[172, p. 94].

Driving regulations are quite restrictive in the case of epilepsy[172, p. 52]. Marriage and having children is another serious consideration[122, p. 236]. Managing epilepsy is a more general issue than the control of seizures.

3.8 Comments

Although the investigation and treatment of symptomatic epilepsy, due to tumours, lesions and other brain abnormalities is now performed by brain imaging techniques[48, 58], traditional electroencephalography is still a valuable tool, especially in the diagnosis and treatment of minor (petit mal) and other forms of constitutional epilepsy.

The social stigma[172, p. 49] of people suffering from epilepsy is an important issue and the sooner the seizures are controlled the better. It is therefore important to diagnose the condition early and tailor medication quickly, two factors where the EEG may prove invaluable. Because seizures are the result of transient abnormal discharges of neurons, not the primary deterioration of the function of the brain, the intellectual ability is not impaired and once the seizures are controlled, no sign of the disorder remains. It is interesting to note that although the boundary between people suffering from epilepsy and those that do not is not clear at all, in contrast with other handicaps, epilepsy is still not accepted, even by modern societies.

Chapter 4

Review of automatic EEG analysis

4.1 Introduction

In the early days of electroencephalography the collaboration of neurologists with engineers and physicists was necessary for the development of instrumentation, monitoring procedures and for the interpretation of results. When the EEG became a medical diagnostic procedure this interaction was weakened, as equipment was standardized and analysis followed those of other medical procedures, consisting primarily of the visual inspection of the paper EEG records.

With the increasing use of the EEG, its visual analysis became a tedious, monotonous and time-consuming task and some form of automation was sought, particularly for prolonged and sleep recordings (sections 2.5.1 and 3.6), event detection and the presentation of the information in alternative, more compact and informative ways. The relation with engineering was renewed to take advantage of technological advances, like microcomputers, and the analysis tools offered by digital technology (signal processing, pattern recognition, expert systems and other). Inter-disciplinary expertise became a necessity for the more effective use of the available resources[5, 176].

The application of technological advances for the analysis of the EEG and other biological signals[32, 187] is far from trivial. The diversity of the techniques used and the absence of widely acceptable trends in the methodology is characteristic of the problems encountered, primarily due to the difference between the informal, often imprecise medical definitions and the strict engineering and scientific formalism[174] and also the inability to quantify medical experience.

In this Chapter a number of methods of automatic EEG analysis will be outlined. Some bear a relation to those investigated in later Chapters, whereas others were included for

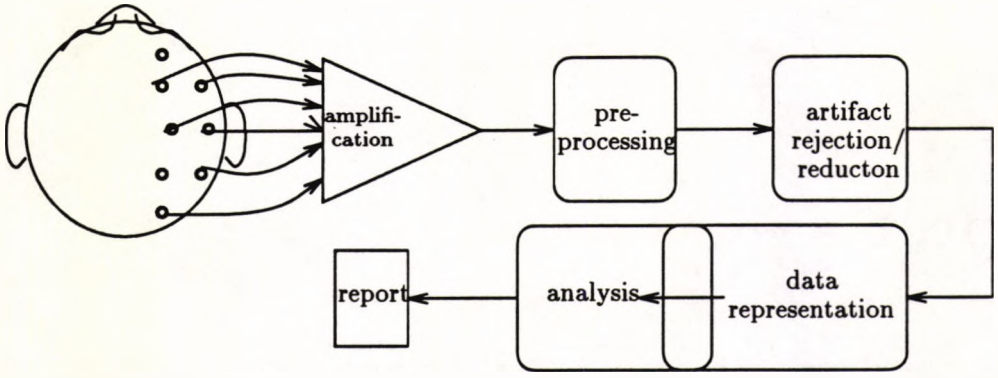


Figure 4.1: Schematic representation of the EEG analysis procedure

comparative purposes or as an overview of the field. Only a small proportion of the available literature has been considered. More details on these and other methods may be found in review articles published occasionally[7, 43, 47, 65, 68, 79] (see also [135]).

4.2 Objectives and methods

Figure 4.1 attempts to highlight the main stages of traditional EEG analysis (section 2.4.4). Preprocessing involves any type of preliminary analogue operations performed on the signal, mainly by the recording equipment (section 2.4.1). Artifact reduction and rejection is performed by the EEG technician during the the recording (sections 2.4.4 and 2.5.4) and by the electroencephalographer's skilled eye during analysis. Sometimes alternative means of signal representation (power spectral density, stages of sleep, spatiotemporal maps, averaging) is performed. These transformations are often inseparable from the analysis procedure, which is a complicated task performed by human expert analysts, resulting in the compilation of a report, summarizing the characteristics of the signal, using relatively standard terminology.

The automatic realization of one or more of these tasks normally attempts to achieve some of the following aims:

- Relief of human experts from the repetitive, tedious and monotonous task of routine assessment, so that they can concentrate on special or difficult cases more effectively
- Purification of the EEG by reduction of extraneous signals to make visual inspection easier/possible
- Provision of more informative representations of the available information for the benefit of human analysts

- Opening new prospects in clinical neurophysiology, by allowing long-term monitoring of patients, which is unsuitable for visual analysis, due to the large volume of data.

The methods employed to realize these targets are rather varied.

4.3 In the beginning it was analogue

Many years ago, when computers were a scarce luxury, techniques were mainly analogue. Despite their simplicity and limitations, their objectives included artifact rejection, sleep stage monitoring and automatic detection of EEG phenomena. A small sample of methods will be presented here for comparison with those that made use of microprocessor technology.

4.3.1 Artifact rejection

Automatic artifact reduction and rejection (Figure 4.1), are important, whether subsequent analysis is performed manually or automatically. Because of the variable morphology of these signals (section 2.5.4), techniques are usually developed to deal with a limited range of artifacts.

Barlow[8] proposed a method to remove of isolated muscle spike artifacts by observing that these are characterized by a high gradient. An analogue differentiator, whose output was full-wave rectified and smoothed indicated the amplitude of the gradient of the EEG. This was compared to a set value to mark the presence of an artifact, and to activate a sample-and-hold to remove it from the EEG signal. Distortion of EEG patterns with high gradient (such as spikes) was evident. Differential sample-and-hold units, were also used by Peper and Grimbergen[146] to remove artifacts arising from current pulses used in cortical stimulation.

Sudden changes in the DC potential at an electrode (section 2.5.4), cause 'electrode pop artifacts' having the form of an exponential decay, after AC coupling by the amplifiers. Barlow[11] employed analogue differentiation, rectification and thresholding to detect their leading edge and C-R networks for their estimation and cancellation. False detections, caused 'ghost' artifacts in the output signal.

The same author suggested the removal of continuous muscle artifact activity by analogue filtering[9]. Butterworth low-pass filters with a cut-off frequency at 12.5 or 15 Hz were built in a multi-channel device. The method compared favourably with a digital filter of Gotman et al.[78], described later. The success of the method was annihilated by the distortion, or even elimination of fast spikes, whose frequency spectrum overlapped with that of the artifacts.

4.3.2 Sleep stage monitoring

The interpretation of sleep recordings is based on the recognition of the various sleep stages, their duration and alternation (section 2.5.1). Manual sleep staging for one night's sleep is difficult, because of the large volume of data produced and automatic systems were employed to produce a profile (parametrization) based on the sleep stages, for the human analyst to read, instead of the EEG (Figure 4.1).

A portable(!) system, suggested by Frost[62], discriminated between wakefulness, sleep stages 1, 2, 3, 4, REM and coma using light indicators and a graphic output. The EEG was fed to a system of comparators with adjustable references and logic elements. These triggered a positive and a negative pulse generator, representing the frequency and amplitude of the signal, which were added and integrated to producing a continuous indication that was high for wakefulness and low for deep sleep. Quantization produced the various sleep stages. REM sleep was distinguished from stage 1 using an EOG channel, thresholded and gated with the other indicators. There was good agreement of this system and human analysts, although the latter were sometimes better, but less consistent. Sleep stage scoring is popular in automatic analysis (see section 4.6.3).

4.3.3 Automatic detection of EEG phenomena

The automatic identification of EEG transients, especially those associated with epilepsy has always been of interest, especially for long-term monitoring purposes.

Walter et al. proposed a semi-automatic quantification method of sharp transients, based on the large gradient changes around their peak[192]. Double analogue differentiation of the signal followed by full-wave rectification produced an indication, enhanced by the suppression of small derivatives by the dead-band of the rectifier. Sharp muscle artifacts, however caused false detections.

A different approach was adopted by Fitch and Willison[57] for the detection of spike-and-slow-wave complexes 60 times faster than real-time. Spikes were detected by a band-pass filter in the range 12–18 Hz followed by a spike rhythm filter (2–4 Hz) and slow waves with a 2–4 Hz band-pass filter. The two outputs, after rectification and analogue integration were thresholded and ANDed to produce a single indication. A computer calculated the number of such events and plotted it with respect to time. Detection of individual spikes was not possible and false detections due to muscle artifacts and eye blinks were evident.

The same method but with slightly different pass bands was used by Besag et al.[16], as the basis of a pocket-sized spike-and-wave monitor. The resulting indication activated

audible and visible indicators and was registered in solid-state memory. Results were more consistent than manual analysis.

Band-pass filtering was the basis of the long-term on-line monitor of De Vries, Wisman and Binnie[191]. Detection of individual events was possible using Schmitt triggers with preset threshold levels. The succession of spikes by slow waves was tested using a counter. Genuine epileptic events were detected using a monostable and a counter, triggered by the initial spike to detect sequences of three events in a period of 1.5 seconds. False detections were mainly due to chewing artifacts.

4.4 The impact of computer technology

Analogue methods were cheap and fast, but were limited by the analogue hardware they employed. The complexity of the required logic functions and the implementation of some fundamental mathematical operations are a nightmare in analogue design. The function of the prototype system cannot be reproduced exactly in replicas of the design, due to tolerances in analogue components. Many of these limitations were overcome by the use of digital techniques and particularly microcomputers, especially after the 1970's. The ability to program the analysis task opened up new prospects, as difficult analogue tasks became trivial programming exercises. Complicated operations could be performed and replicated exactly. Digital signal processing offered a greater repertoire of methods than its analogue counterpart and new concepts, numerical and symbolic, were made available. The rapid decrease in size and price and increase in sophistication and speed made computers a powerful, cost-effective tool.

Despite the multitude of methods published, computers are no panacea and none of those matched the performance of the human analyst. Some of the available literature is briefly reviewed here, with emphasis on methodology (see also [135]).

4.5 Artifact reduction: Preprocessing or target?

Artifact rejection (Figure 4.1) could make the task of the human analyst easier and more effective and would prevent their false interpretation by a subsequent automatic system, particularly important for long-term unsupervised analysis. Effective artifact reduction could give value to stretches of otherwise unusable EEG. Automatic techniques have targeted, primarily, ocular (EOG), muscle (EMG) and cardiac (ECG) artifacts.

4.5.1 Dealing with ocular artifacts

These are caused by potentials arising from eye movements and blinks (section 2.5.4) and are a serious problem due to their large amplitudes, especially in the study of event-related potentials. Fortgens and De Bruin[59] suggested their automatic removal off-line, using electrodes placed around the eyes to form four ocular potential channels. Assuming independence, the mean squared error (MSE) of the difference between the EEG and a weighted sum of the measured EOG signals,

$$M\bar{\epsilon^2} = \sum_{j=1}^{M}\{EEG(j) - \sum_{i=1}^{4} A_i EOG_i(j)\}^2 \tag{4.1}$$

was minimized, leading to a system of equations, solved for the unknown weights, $\{A_i\}$. The difference, $\epsilon(j) = EEG(j) - \sum_{i=1}^{4} A_i EOG_i(j)$ was an approximation of the artifact-free cerebral activity. ECG artifacts could be removed in a similar manner by placing electrodes on the torso. According to the authors this method worked better than other alternatives at the time of publication.

Based on a similar model, Gratton et al.[80] developed a multi-channel scheme, catering for differences in the mechanisms of eye blinks and eye movements by using distinct weighting factors. Records of EEG and EOG for N trials were used off-line, after the subtraction of the average event-related activity. Blinks, EB were identified by their high gradient and along with other eye movements, EM they were subtracted from the EEG of individual trials using linear weighting.

$$EEG_{corrected}(j) = EEG_{raw}(j) - aEM(j) - bEB(j) \tag{4.2}$$

The weights, a and b were estimated by MSE minimization, as in the previous method.

Ifeachor et al.[90] recognized that the elegance of these methods was offset by the large number of computations required for the estimation of the weight vector, θ_m, by solving the normal equations:

$$\mathbf{R}_{xx}\theta_m = E[X_m^T Y_m] \tag{4.3}$$

involving the EOG and the EEG vectors, X_m and Y_m, respectively. The inversion of the autocorrelation matrix, $\mathbf{R}_{xx}$ is not suitable for on-line implementation and the authors proposed the method of Recursive Least Squares, RLS, to compute $\mathbf{P} = \mathbf{R}_{xx}^{-1}$ iteratively with every new sample:

$$\begin{aligned} \theta(m+1) &= \theta(m) + \mathbf{P}(m+1)X(m+1)E(m+1) \\ \mathbf{P}(m+1) &= \frac{1}{\gamma}\left[\mathbf{P}(m) - \frac{\mathbf{P}(m)X(m+1)X^T(m+1)\mathbf{P}(m)}{\gamma + X^T(m+1)\mathbf{P}(m)X(m+1)}\right] \end{aligned} \tag{4.4}$$

where $E(m+1) = Y(m+1) - \theta^T(m)X(m+1)$. Due to potential instabilities of this algorithm, the U-D factorization $\mathbf{P}(m+1) = \mathbf{U}(m+1)\mathbf{D}(m+1)\mathbf{U}^T(m+1)$, where $\mathbf{U}$ is a triangular and $\mathbf{D}$ a diagonal matrix was employed. A later more detailed study[91] took into account serial correlations of the background EEG investigating three methods for estimating θ: Ordinary Least Squares (OLS), the iterative method (assuming the EEG is serially correlated) and OLS on the difference of the data (implying that the EEG is a random walk). The iterative and the difference methods were considered closer to reality and were favoured by this study. Ifeachor et al.[92] noted that small numbers of carefully selected EOG channels was preferred, as large numbers of them caused overcorrelation with the EEG.

In a review of methods for removing ocular artifacts, Jervis et al.[96] suggested that only a few of the proposed electrode arrangements gave satisfactory results and from the methodologies those of Least Squares were preferred. They recommended the difference or iterative methods for decorrelation and the RLS technique for on-line processing.

These recommendations influenced the work of Van Den Berg-Lenssen et al.[189], who produced a study of an ocular artifact-removing system by modelling the cerebral component, $\underline{e}(n)$, of the recorded EEG signal, $\underline{y}(n)$, by an autoregressive process, excited by white noise, $\underline{x}(n)$ with additive artifacts:

$$\underline{y}(n) = \sum_{j=0}^{m-1} M(j)\underline{u}(n-j) + \underline{x}(n) - \sum_{j=1}^{r} A(j)\underline{e}(n-j) \tag{4.5}$$

where $\underline{u}(n)$ are the EOG signals. The order of the models, m and r, were determined by experiment to produce decorrelation without overparametrization. Experiments demonstrated the requirement for three or four EOG channels for effective artifact removal. Coefficients calculated for eye blinks were successfully applied for the removal of other eye movements and were effective for a short period of time after the data used for their computation.

4.5.2 The troublesome muscle artifacts

Compared to the rejection of ocular artifacts that of muscle (EMG) artifacts is a near impossible task, owing to the absence of a reference signal (unlike the EOG), or even a well-defined morphology (section 2.5.4). Muscle potentials mainly contaminate frequencies above 15 Hz[142] and often have morphological similarities to epileptic transients causing problems in the detection of the latter.

Gevins et al.[66] used artifact rejection in their multi-channel spectral analysis system, based on the power and cross-spectral density of epochs of about 1 s, evaluated using the FFT. The mean and variance of the PSD of artifact-free epochs was computed for the

frequency ranges 0–1 Hz, 0–4 Hz and 34–44 Hz, the respective regions for perspiration, ocular and muscle artifacts, according to the authors. These artifacts were detected (but not corrected), when the PSD in their frequency range was higher than the mean by an integral multiple of the standard deviation in the respective range of the artifact-free record. The accuracy of the method (65%) permitted its use under human supervision, although the consistency of human analysts was also low.

Gotman et al.[78] attempted to remove EMG artifacts from records containing epileptic seizures, using digital filtering. These authors noticed that the PSD of the artifact-free EEG did not have significant components above 12 to 28 Hz, whereas EMG records had their main components above 15 Hz. They proposed an FIR filter with a cut-off frequency at 25 Hz to eliminate artifacts without significant distortion of the cerebral activity. According to the authors, activity masked by artifacts and revealed after filtering should be interpreted with care, since residual muscle activity sometimes resembled cerebral potentials.

Two types of digital filters were developed and applied by Panych et al.[142] with encouraging results. The first, was a non-linear low-pass IIR filter with a parameter varying with the signal gradient, lowering its cut-off frequency in the presence of signals of high gradient (like muscle artifacts), thus clipping them. The second was a linear FIR digital filter like that of Gotman et al. (see above), but with softer cut-off characteristics to preserve the the signature of the artifact.

A different approach was adopted by Johnson et al.[98]. The EEG was modelled as the linear summation of four lightly damped oscillators, α, β, δ and θ, driven by white Gaussian processes and additive Gaussian, zero mean, white observation noise. Muscle artifacts were modelled by three second-order systems, driven by independent impulse Poisson processes, to represent spikes of duration 10, 15 and 20 ms (Figure 4.2) which were equally likely. Their parameters and impulse responses were estimated using the PSD of real muscle spikes falling in the three categories.

Muscle spikes were detected using three matched filters and a maximum likelihood criterion to determine which one, if any, had occurred. These were eliminated by exciting the corresponding system model by the output of the matched filter, subtracting its output from the EEG signal and applying a Kalman filter[31][101, pp. 254–289] to eliminate noise due to mismatch of the simulated and the actual spike. Only normal EEG records were considered, but, according to the authors, the detection of abnormal transients was a trivial extension of the proposed system.

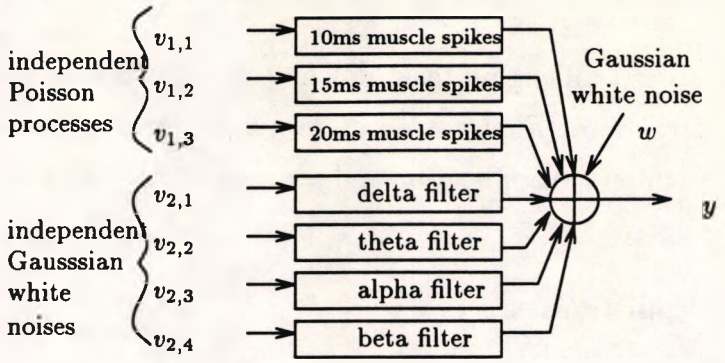


Figure 4.2: The EEG signal model used by Johnson et al.

4.5.3 Cardiac potentials and other sources of interference

A wide range of signals of biological and non-biological origin interfere with the EEG. Cardiac potentials (ECG) are another source of artifacts (section 2.5.4). These have similar properties to ocular artifacts, and some authors, like Fortgens and De Bruin[59], have applied common methods for their removal (section 4.5.1).

Witte et al.[203] suggested the detection and elimination of the cardiac activity from the EEG using spectral analysis. Since the ECG is periodic, its spectrum in the contaminated EEG, would contain discrete equidistant lines, corresponding to harmonics of its frequency of repetition. The PSD of the EEG, estimated by the FFT, and differenced to reduce the smoother cerebral activity was tested for ECG periodicities using its discrete spectral properties. Confirmation by transforming to the time-domain and comparing with the recorded ECG led to the subtraction of an average ECG from the EEG signal. The correction of the artifact was tested by repeating the test in the PSD. Ocular artifacts were eliminated by testing for coherence in the low-frequency range of the EEG and the EOG using the FFT. If present, the EEG was corrected by subtracting the scaled EOG signal and repeating the coherence test for confirmation.

Although drowsiness is a cerebral phenomenon, Gevins et al.[67] treated it as an artifact, because it introduces bias in the establishment of wakefulness. Using the PSDs of a number of EEG channels (see method by the same author in section 4.5.2) the amounts of alpha (8–13 Hz), theta (4–7 Hz) and delta (1–3 Hz) activity present were estimated. The ratios of every one with the two others were computed and compared to individual drowsiness thresholds, different for every channel, to determine an increase in the slow activity, characterizing drowsiness. Bilateral indications were accumulated and their summation over a number of epochs determined the state of consciousness (awake or drowsy) at the end of every epoch.

4.6 Computerized analysis techniques

Artifact rejection a rather challenging task, did not inhibit workers to attempt the to automate the analysis procedure (Figure 4.1), as motivation was quite strong (section 4.2). Targets varied with the intended application and some were more general than others. These fall in one of these categories:

- Detection of epileptic transients
- Detection of event-related potentials
- Sleep stage analysis
- Overall EEG assessment
- Data Compression
- Integrated systems

The methods applied to realize these targets were quite varied and classification is difficult. In the review that follows the emphasis on the detection of epileptic transients is evident and there is a bias towards methods bearing resemblance to those employed in later Chapters.

4.6.1 Detection of epileptic transients

The importance of the EEG in the diagnosis and treatment of epilepsy was made apparent in the past two Chapters (sections 2.5, 3.6 and 3.8). Most epileptic patterns in the EEG are associated with spikes and possibly slow waves, whose appearance is often spontaneous. Methods for their automatic detection exploit some of their many properties, although most use some form of pre-processing (feature extraction) followed by detection (classification). Since preprocessing, is their most prominent part, it was used for their discrimination in this review.

Methods based on gradients

Epileptic transients, like spikes and sharp waves are characterized by abrupt, short-lived changes in the EEG, having amplitudes typically larger than the rest of the record and rapid onset. These give rise to large gradients (derivatives) with respect to time, a fact that may be used for their detection.

Carrie[35, 36], used the EEG and its second derivative, evaluated by an analogue computer for the detection of sharp transients. These were sampled and fed into a computer which marked spikes by comparing the second derivative of individual samples to a measure based on its average, computed from a number of samples and a preset relative value. This was confirmed if followed by a slow wave of duration between preset limits, determined from the original EEG samples. In a later development[37], false detections due to muscle artifacts were prevented by filtering out the low frequencies (0–26 Hz) and detecting bursts of waves in the residue to indicate artifacts. Spikes detected during these periods were rejected.

Based on similar principles, Smith[181] modelled spikes as triangles with unequal sides and a base-line of duration 20–70 ms and proposed the first derivative as an indication of the steepness of the two sides. These were detected by their abnormally high gradient using a preset threshold level. The sharpness of the apex of the spike was calculated as the difference in the average slopes of the detected edges and not as the second derivative, which is affected by noise. Differentiation was performed by an impulse invariant transformation of the analogue pseudo-differentiator:

$$G(s) = \frac{s}{s+b} \tag{4.6}$$

The problem of muscle artifacts with their high derivatives was recognized but not solved.

Digital differentiation was the basis of spike quantification proposed by Ktonas and Smith[107]. A digital low-pass filter derived by an impulse-invariant transformation of an analogue filter reduced high-frequency noise before applying a 'successive difference' digital differentiator. Quantification utilized five slope and time parameters. Their values were consistent for spikes extracted from three patients. The authors pointed out that this method of quantification offered a more detailed and precise spike description than the one used during visual inspection, which was subject to errors. At a later development of the method[108] the definition of the parameters was changed a little and their number increased to ten, to include sharpness and amplitude measures. Derivatives were estimated using digital differentiators, derived from analogue transfer functions via the bilinear transform and included filtering. Differences in the measures between subjects and between spikes and sharp waves were observed.

Based on this method, Glover et al.[71] later developed a system capable of analyzing 16 channels at speeds of 8 times that of real-time. One slave processor was allocated for every channel, programmed individually, but controlled by a master processor. A multi-channel processor was connected with a terminal for downloading commands and reporting results. It also correlated spike detections from individual channels to reduce artifacts and classify 'borderline' cases.

A prototype microprocessor system was designed and tested by Comley and Brignell[45]. It was based on the concept of the 'roving slave processor', a stand-alone system, connected to a larger computer only for downloading programs and extracting results. Differentiation was performed digitally, using an impulse-invariant transformation of the analogue differentiator, truncated to a simple difference operation. The gradient was filtered by a low-pass digital Butterworth filter with a cut off frequency of 50 Hz to reduce the high level of noise. Consecutive EEG segments having gradients above a positive and below a negative threshold and the time required for the transition between them were utilized for spike detection. Muscle artifacts were the main source of false detections.

Instead of a differentiator, Quian et al.[157], used a difference filter in the form

$$d(n) = x(n) - x(n-k) \tag{4.7}$$

emphasizing specific frequency components (multiples of 16 Hz) which were in the range of the fundamental and harmonics of a spike. If the product of $d(n)$ and its past value $d(n-K)$, remained above a threshold (20.5 times its mean deviation) for at least 15 ms (chosen experimentally), a spike was detected. The product operator was found to reduce, but not eliminate, high-frequency artifacts and noise.

Finally the real-time spike and wave detector, developed by Stelle and Comley[183, 185] had similarities to that of Comley and Brignell (described earlier). Stelle developed the detector entirely in the discrete z-domain. To eliminate the need for two level-detectors, due to the bipolar gradient of epileptic spikes, a Hilbert transformer was cascaded with the differentiator to make the output monopolar. The transfer function of the ensemble system

$$h(n) = A\frac{\sin^2(n/2)}{n/2}\cos(\pi n) \tag{4.8}$$

found by the inverse Fourier transform, was truncated to form an FIR system, whose output was raised to the power 3, enhancing large gradients but maintaining signal polarity. Subsequent slow waves were detected by a Butterworth digital low-pass filter with cut-off frequency at 7 Hz and a level detector. False detections, mainly due to muscle artifacts were reduced by defining a window in which the occurrence of more than one spike was rejected.

The use of domain transformations

Time-to-frequency transformations cause the spread of brief phenomena, such as spikes, making their detection more difficult. They have been used, however, in some special cases.

Spectral analysis was used by Gotman et al.[74] for the detection of prolonged abnormal slow wave activity. A representative 40 s segment was chosen interactively and then

automatically subdivided into 5 s segments. The FFT was applied to each segment and the resulting PSDs were averaged for every EEG channel and displayed along with the relevant EEG segment. The weighted ratio of slow to fast activity

$$\frac{(a \cdot \delta) + (b \cdot \theta)}{(c \cdot \alpha) + (d \cdot \beta)} \tag{4.9}$$

where the weights a, b, c and d depend on the location of the channel, was computed and plotted graphically, localizing excessive slow wave activity and asymmetries between the two hemispheres.

Darcey and Williamson[51] used relative measures derived from power spectra for the automatic quantification of seizures. Recordings from implanted depth electrodes, were manually divided in the pre-ictal and ictal phases which were separately processed by the FFT. The mean, $M(i, f_a, f_b)$ and standard deviation, $S(i, f_a, f_b)$, of the PSD of the pre-ictal record were evaluated for the frequency range $f_a = 8\ Hz$ to $f_b = 30\ Hz$ (8–30 Hz) for every channel, i. During a seizure, the integral of the PSD in the same frequency range, $BIP(i, j, f_a, f_b)$, was computed for each epoch j and was used to calculate two relative measures,

$$BIPZ(i, j, f_a, f_b) = \frac{|BIP(i, j, f_a, f_b) - M(i, f_a, f_b)|}{S(i, f_a, f_b)} \tag{4.10}$$

$$BIPR(i, j, f_a, f_b) = \frac{BIP(i, j, f_a, f_b)}{M(i, f_a, f_b)} \tag{4.11}$$

which had large values during a seizure and were used to locate the onset of focal attacks. The relative measures proved superior to $BIP(\cdot)$ in the localization of the epileptic focus.

Saltzberg[168] employed the FFT to reveal the presence of subcortical spikes in the EEG. In the analysis provided, it was demonstrated that the presence of non-periodic subcortical spikes in the EEG would cause ripple frequencies in its PSD, corresponding to the reciprocals of the time intervals between them. According to the author, the presence of these spikes was sometimes more evident in the PSD than in the time domain.

To utilise both time and frequency information for the more accurate detection of the spike-and-wave complex, Stelle and Comley[184] used the Wigner-Ville distribution, a transformation converting a time function, $x(t)$, into a time-frequency distribution:

$$W_x(t, \omega) = \int_{-\infty}^{+\infty} x(t + \frac{1}{2}\tau) \cdot x^*(t - \frac{1}{2}\tau) e^{-j\omega\tau} d\tau \tag{4.12}$$

where $x^*(t)$ denotes the conjugate of $x(t)$. For practical evaluation, a discrete form, incorporating a window, $w(n)$ was employed:

$$W_x(m, k) = \sum_{n=0}^{N-1} w(n)x(m + n) \cdot w^*(n)x^*(m - n)e^{-j\pi nk/N} \tag{4.13}$$

After smoothing by a two-dimensional Gaussian filter, the distribution for spike-and-wave complexes appeared to have a characteristic 'T' shape, which was visible even in the presence of artifacts.

Four regions in the distribution were identified, the spike, A_s, the slow wave, A_w, the high frequencies of the spike, A_{hf} and an area of inspection, A_i. Variations in A_{hf} were enhanced by differentiation and the ratio, $K1$, of the standard deviation and mean of the region was computed. The maximum, $K2$, of the A_s and A_w regions was then determined. The area A_i is almost clear for spike-and-wave complexes and the normalized function

$$y(m,k) = \left| \frac{A_i(m,k)}{K2} \right| \cdot K1 \cdot k \tag{4.14}$$

was a measure of muscle artifacts, which have significant components in A_i. These were detected by counting the points in this region whose values were higher than a preset level, achieving the rejection of artifacts and the detection of contaminated spikes[185](pp. 191–192).

Parametric techniques

Formal modelling the EEG signal is not a trivial task, due to its changing and unpredictable nature. Nonetheless, under certain assumptions and allowing for discrepancies, stochastic modelling proved a powerful analysis tool.

Lopes Da Silva et al.[179], one of the pioneers in the application of this approach in the detection of epileptic transients, modelled cortical signals as the output of an autoregressive filter, $H(z)$, excited by white Gaussian noise, $\epsilon(n)$, with additive spikes (Figure 4.3). For analysis the inverse of system, $A(z) = H^{-1}(z)$ was applied to the EEG signal and the output contained the generating sequence, $\epsilon(n)$. The (15) coefficients of $A(z)$ were estimated, by the method of least squares based on the autocorrelation function of a window of the EEG signal and its output was squared, averaged, and tested for spikes as outliers in the assumed χ^2 probability distribution function (Figure 4.3). The number of detections in a number of channels was used for the localization of an underlying epileptic focus and agreed with other clinical findings. This method did not discriminate between spikes and other transients, but demonstrated the presence of spikes with characteristics outside the internationally agreed limits.

Autoregressive modelling was also the basis of the work of Bodenstein and Prætorius[25, 153], who proposed that the EEG should be regarded as a sequence of quasi-stationary segments, each characterized by its length and a (constant) power spectral estimate. Under additional constraints for real-time operation and data reduction, linear prediction was employed

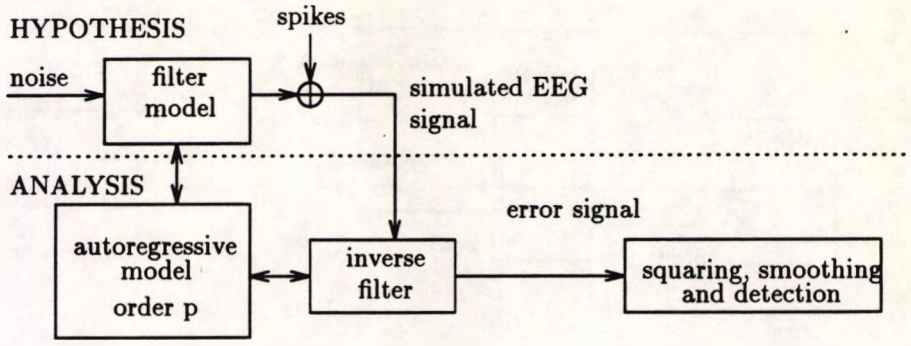


Figure 4.3: The autoregressive model of Da Silva et al.

for spectral parameter estimation. Their algorithm began with the computation of the auto-correlation function of a window of the signal to construct the auto-correlation matrix and solve the normal equations for the optimal autoregressive filter coefficients, $\{a_i : i = 1 \ldots p\}$ (see section 8.1.1). The inverse system was applied to subsequent samples and the estimated auto-correlation sequence, of the output error function, $r(n;k)$, was the basis of the spectral error measure:

$$SEM(n) = \left(\frac{r(0;0)}{r(n,0)} - 1\right)^2 + 2\sum_{k=1}^{N}\left(\frac{r(n;k)}{r(n;0)}\right)^2 \tag{4.15}$$

whose value was compared to a preset upper limit to determine the end of a segment and the beginning of the next. This procedure was repeated for every new segment. To prevent the premature termination of segments due to the large transient SEM, caused by spikes and other sharp phenomena, the average squared error for consecutive samples was used as a measure of spikeness. If above a given threshold, the corresponding error values were clipped to reduce the SEM. Segmentation led to data reduction, since each segment was described by its length, the power of the error and the prediction coefficients. From these the time behaviour and PSD of the signal could be reconstructed. The authors proposed the possible description of the EEG as a Markov process.

Clustering of similar EEG patterns was investigated by Birkemeier et al.[17]. Double differentiation, estimated by fitting a parabola on consecutive segments of the EEG improved clustering of normal activity and abnormal transients. Better clustering was achieved by linear prediction of five samples in the future since it decorrelated the background activity and made transients more prominent. The prediction coefficients were estimated using window-based correlation estimates and direct matrix inversion. Further improvements in the discrimination between normal and abnormal patterns were possible by cascading linear prediction and double differentiation.

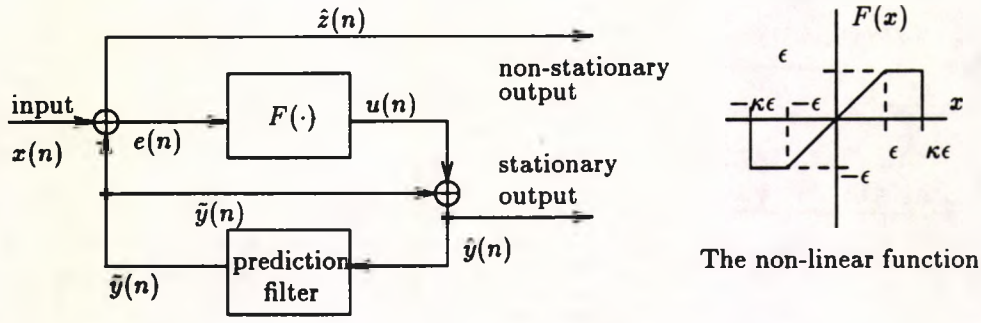


Figure 4.4: The non-linear filter suggested by Arakawa et al.

The merits of autoregressive filtering for the segmentation of the EEG signal were exploited by Penczek et al.[144], who considered both stochastic and quasi-periodic repetitive abnormalities during a seizure. In order to follow rapid changes in the signal, a (multichannel) Kalman filter predictor with sparsely updated parameters was employed (see [140, pp. 173–188], [31] for a description of the Kalman filter). The resulting matrix of prediction coefficients indicated intra- and inter-channel relations. A linguistic (syntactic) approach was adopted to classify segments. Similar matrices were grouped to form 11 labelled classes and the signal was treated as a string of labels. Sub-strings were found to recur during seizures. The authors highlighted the similarity of this approach to visual analysis which is based on general characteristics, rather than individual detections.

Arakawa et al.[4] proposed a non-linear system for separating stationary and non-stationary EEG activity. The system consisted of a prediction filter and a non-linear function, $F(\cdot)$ (Figure 4.4). The stationary part of the signal, $\hat{y}(n)$ was predicted by an autoregressive filter, implemented as an LMS adaptive filter (described in Chapter 7 and in many texts and articles[14, 15, 140, 200, 201, 202]). The error of the prediction, $e(n) = x(n) - \tilde{y}(n)$, was applied to the non-linear system, $F[e(n)]$, which was linear for $|e(n)| \leq \varepsilon$, constant at ε for $\varepsilon < |e(n)| \leq \kappa\varepsilon$ and zero for $|e(n)| > \kappa\varepsilon$, which were assumed to be caused by non-stationarities. The stationary output, $\hat{y}(n) = \tilde{y}(n) + u(n)$, was thus separated from the non-stationary, $\hat{z}(n) = e(n) - u(n)$. The parameters ε and $\kappa\varepsilon$ were estimated adaptively, based on fixed relative probabilistic measures and experimentation and on the running variance estimate $V(n)^2 = (1-\alpha)V(n-1) + \alpha e(n)^2$. This system did not distinguish between spikes and other non-stationarities.

Inverse and matched filtering

Inverse and matched filtering are normally used to detect patterns of brief duration and well-defined shape, typically buried in noise. Unfortunately, variations between individual transients and the correlation of the background activity limited their success, as explained by Comley and Brignell[45] who had used matched filters for spike-and-wave detection.

Weinberg and Cooper[196, 197] pointed out that matched filters are based on energy properties and hence the output of a spike-and-wave filter would be more dependent on higher-energy slow waves, although the presence of a spike is equally important and good overall matching is deemed impossible. The authors suggested separate matched filters $\{x_m\}$, for every sub-pattern, m whose outputs, $r_m = \frac{\sum x_m y}{\sqrt{\sum x_m^2 \sum y^2}}$, were combined to form the recognition index, defined as their product if all are positive and zero otherwise. This measure is amplitude-insensitive and could reveal the presence of the pattern by thresholding.

Inverse and matched filtering were combined for spike detection by Pfurtscheller and Fischer[148] after recognizing that failure of matched filtering was caused by the high auto-correlation of the EEG. This was removed by inverse autoregressive modelling, (see Lopes Da Silva et al. in previous section). A number of different spike 'templates' were processed by the inverse filter to produce matched filters in the output domain. Spikes were detected by thresholding the outputs of these filters, applied to the output of the inverse filter. The effectiveness of the combined technique was significantly improved, compared to direct matched filtering.

Wave parameter analysis

Many of the developed methods were based on heuristic and intuitive criteria, using parameters measured from individual waves. These required only moderate computing power for their implementation.

A real-time system based on quantification of single waves was developed by Leader et al.[109]. 'Relative' maxima and minima were detected as zero crossings of the derivative of the EEG (estimated by a difference operator) preceded and followed by four samples of the same polarity. 'Absolute' maxima were those whose amplitude from the nearest absolute minimum was above a threshold. Absolute minima were defined in a similar manner. These were used to define a peak-to-peak amplitude and a period measure for every half wave, which were averaged over a time interval and used to classify the EEG segment in one of 13 normal and abnormal classes using a nearest neighbour algorithm.

Carrie[34] analyzed transient EEG abnormalities using the peak, duration and amplitudes from the baseline at specific fractions of the duration of the negative half-waves of the signal.

For every new wave, each measure was divided by the corresponding running mean and compared to a preset threshold to detect abnormal levels.

Hill and Townsend[86] used the same technique to select features from individual waves. The peak angle of a half-wave, considered an important feature and was evaluated by fitting straight line segments on both sides of the peak. Peak angles, normalized by an amplitude measure, were found to follow different distributions for 'spikes' and for 'waves' and their discrimination threshold was estimated. The mean frequency of spike occurrence was subsequently evaluated as a useful diagnostic parameter.

Following a review of methods, Gotman and Gloor[75] concluded that a working real-time system had to be a compromise between a formal method and a simple, heuristic one. The time differences and amplitudes between local maxima and minima were measured. Successive extrema formed basic segments, which were re-grouped using amplitude and duration criteria, to eliminate short intermediate segments, caused by noise. Waves were then defined as sets of two consecutive segments. The average amplitude of all segments in a 5 s interval before a segment was used for the normalization of its amplitude. The sharpness of its peak (measured from the estimated second derivative) and two duration measures were evaluated. Epileptic spikes were detected as pairs of consecutive segments satisfying predefined amplitude, duration and sharpness criteria. Inter-channel correlation of detections revealed their time relation and improved the reliability of the system. A graphic and an interactive user interface were incorporated during a subsequent development by Gotman et al.[76]. This system was modified later to include a derivative-based wave asymmetry measure[77], defined as the slope of the line joining the apex to the point of middle-duration of a half wave.

Segmentation using regrouping of elementary segments was utilised by Koffler and Gotman[105] to detect bursts of epileptiform activity during long-term monitoring. A half-wave was detected as a spike if its relative amplitude compared to the background activity exceeded a threshold. The threshold was lowered if a slow wave was subsequently detected. Spike bursts were identified by counting the number of spikes in a given period for each channel and between channels. Their amplitudes, durations and mean frequency were used to discriminate genuine spikes from muscle artifacts.

Pola and Romagnoli[149] employed differencing between consecutive samples to eliminate slow waves. The difference of the differences, introduced as a corrective factor, was claimed to reduce sampling rate fluctuations. A typical spike was used as a model to derive spike margins, which were compared to the differenced waveform.

Relative measures were used by De Oliviera et al.[139]. The peak amplitudes of the

signal, and its derivatives (computed by analogue differentiators) were normalized by the corresponding variances and combined with various duration measures to a single feature vector. The boundary between normal and spike vectors was determined by maximizing the sum of specificity and sensitivity (the inverse of false and missed detections).

Davey et al.[54] combined elements of the previous methods. The extrema of the waveform were detected by a difference operator using the previous and the next sample. Three parameters (duration, amplitude and peak sharpness), were extracted from the set of ten suggested by Ktonas et al. Sharpness was computed according to Hill and Townsend and the definition of background activity was borrowed from Gotman and Gloor. These features were applied to a rule-based system, developed with the assistance of a human analyst. Classification of the record could distinguish focal and generalized activity.

Other methods

Finally, there are methods that could not be classified under any of the previous headings, either because they used too specialized or too varied techniques.

Ehrenberg and Penry[56], for instance, detected spike-and-wave discharges using the duration between baseline crossings of the signal to estimate wave periods. These were accumulated over an epoch and compared to a running mean to determine three-epoch periods where a threshold was exceeded.

A microcomputer system constructed by Principe and Smith[155] could detect petit mal paroxysms and its variants by separating slow components from the fundamental of the spike using band-pass filtering. Spikes and waves were detected separately using the duration and peak amplitude of half waves. These were compared to preset limits and a seizure was marked as a succession of three slow waves containing at least one spike.

Panych et al.[141] constructed a seizure detection unit based on digital band-pass filtering, and thresholding of the estimated RMS value of its output. The system was a basic one, but its many false detections were acceptable, since its purpose was the unsupervised control of a mechanism for video, audio and EEG recording of seizures.

Grochulski et al.[81] followed the syntactic approach for real-time analysis of the epileptic EEG. Peak-to-valley amplitudes and durations were used as parameters. All possible conditions were assigned labels in a mutually exclusive and exhaustive manner based on the similarity of their parameter values. The EEG was then described as a string conforming to a regular grammar implemented as a finite-state automaton with nine terminal conditions recognizing sub-strings representing significant EEG phenomena. Later, Penczek and Grochulski[145] simplified the initial automaton via state reduction methods. A machine,

constructed using the theory of Markov chains defined states and transition probabilities between them, was applied to a multi-channel system to model interactions between brain structures after direct stimulation. Eleven states and their transition probabilities were estimated by clustering spectral parameters, extracted from an autoregressive filter constructed via the FFT.

Finally, Glover et al.[72] recommended the use of a knowledge-based system to incorporate spatiotemporal context information for the prevention of false detections in long-term seizure monitoring. Various time and amplitude features were extracted from the EEG, ECG and EOG channels and fed into a rule-based system to detect and group possible epileptic transients. This was followed by multi-channel analysis with rules to correlate the various findings in the EEG, EOG and ECG channels to confirm detections and remove false events, due to artifacts.

4.6.2 Detection of event-related potentials

Event Related potentials (ERPs) are minute fluctuations in the EEG signal that correlate with external stimuli, often presented to the subject many times. Traditional averaging of many responses to reveal the ERP is not always satisfactory and experimental techniques for its extraction from the much larger ongoing activity have always been valued. The methods summarized here relate to the the work presented in later Chapters.

To decrease the time variability in the responses that could invalidate averaging, Vidal[190] suggested the estimation of a Wiener filter based on the pre-stimulus to reduce the on-going EEG signal, followed by an F-test to estimate a suitable weight vector to enhance clustering and a Bayes posterior probability rule for classification.

Wiener filtering was the core of the work of Cerutti et al.[38]. The filter was developed in the frequency domain and assumed that the signal (ERP) and noise (ongoing EEG) were uncorrelated.

$$H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{nn}(\omega)} \tag{4.16}$$

The power spectral densities of the signal, $\Phi_{ss}(\omega)$, and the noise, $\Phi_{nn}(\omega)$, were estimated from the pre- and the post-stimulus waveforms, respectively. The filter was applied to the post-stimulus in the frequency domain and the signal was revealed by transforming back to the time domain. This method decreased the necessary number of experiments, compared to averaging, by a factor of 10.

A different approach was adopted by Fridman et al.[61], who used the amplitude and phase of the average FFT-based spectra of individual responses to synchronize them. Fil-

tering and differentiation in the frequency domain were applied to enhance the signal before the inverse FFT was applied.

The classification of ERPs using the FFT was attempted by Moser and Aunon[125]. Separability of the transformed pre- and post-stimulus data and their synchronization was assisted by a cosine transformation of amplitude and phase. Their most significant features were then used to construct FIR bandpass filters which were applied to the EEG signal and their outputs were used as features for the classification based on a Log-likelihood ratio.

Spectral analysis was the basis of a technique suggested by Jervis et al.[97], who assumed an EEG model, consisting of the superposition of the spectra of the ongoing activity and the ERP. Statistic measures for specific spectral components were extracted and used for the classification of patients suffering from Huntington's disease and normal subjects.

4.6.3 Monitoring stages of sleep

Overnight monitoring of the EEG during sleep has always been a challenging classification problem. A lengthy EEG record is to be segmented and annotated into states of consciousness (section 2.5.1). Work in this domain dated back to the early days of computers. Cox et al.[49] produced a review of early methods in 1972. This included amplitude and frequency analysis and the simpler, but equivalent, period analysis, based on zero crossings and variances of the signal and its derivatives. The sophistication of more recent developments was inherently linked to technological advances.

A real-time system was implemented by Lim and Winters[110] for the analysis of the EEG of cats. The peaks and zero crossings of the signal were used, since they signify amplitudes and frequencies. Artifacts were identified by their atypically large amplitudes. Waves were classified into δ, θ, α, σ, γ and β and their relative ratios determined the state of consciousness and the classification in one of seven primary states, whose limits were determined experimentally, to resemble the performance of the human analyst.

Principe et al.[154] transformed Chebyshev analogue filters to discrete-time using the impulse invariant transformation and zero placement at $z = \pm 1$ to construct five band pass filters, α (7–17 Hz), β (13–35 Hz), σ (11–29 Hz), θ (1–10 Hz) and Δ_w (0.5–3 Hz). These separated the various frequencies in the input EEG signal for subsequent qualification in sleep stages. Different filter realizations and low-pass to band-pass transformations were compared for narrow- and broad-band filters, but no explicit method for EEG quantification was proposed.

The automatic sleep staging method of Haustein et al.[84] used one EEG, one EMG channel and FIR filters to subdivide the EEG in the frequency bands, δ, θ, α and β. The

mean peak amplitude in each filter output were estimated from the amplitudes between consecutive local maxima and minima and combined in a vector. The EMG was similarly separated into amplitude measures to quantify continuous, low-amplitude and discontinuous high-amplitude muscle vector activity. Using the inner product of the mean amplitude vector with weight vectors, three orthogonal vector functions were formed. Sleep quantification was a vector pointing from the average for an awake person to that of deep sleep (stage 4) and was quantized to indicate discrete sleep stages. The other two were a calibration and an artifact detection vector. The authors reported good correspondence with human scoring, but supported the view that separation of sleep into stages was arbitrary and the continuous parameter was preferable.

Sleep stage analysis was one application of a programmable medical instrument, developed by Papp et al.[143] as an integrated, real-time, multi-tasking software system. A signal processing library and a command language, supported both general-purpose and knowledge-based system programming as well as user interface design. Sleep staging was based on monitoring amplitude trends in every EEG channel, computing power spectra and evaluating the activity in the conventional EEG frequency bands, used by a rule-based inference engine to detect sleep stages. This system was a general one, not bound to a specific method or signal.

One of the more recent developments in sleep stage analysis was the approach of Roberts and Tarassenko[161, 162]. The EEG sequence, was parametrized using a linear prediction Kalman-Bucy filter:

$$x(t) = \sum_{i=1}^{p} \theta_i \, x(t-i) + \epsilon(t) \tag{4.17}$$

where $\theta = \{\theta_i; i = 1, \ldots, p\}$ are the (time-varying) Kalman coefficients and $x(t)$ and $\epsilon(t)$ are the EEG and a white noise sequence, respectively. The order of the filter, $p = 10$ was chosen by experiment to reduce the mean squared error without over-parametrization. The extracted θ was applied to a self-organizing map, a neural network that enhanced clustering. It consisted of a 10×10 lattice of processing elements, each having an internal characteristic vector, $\mathbf{m}_{jk}$. Its function consists of computing the Euclidean distance of an input vector $\|\theta(t) - \mathbf{m}_{jk}\|$ for all its elements, and suppressing all outputs but the one with the minimum distance, whose output is typically set to 1. Training is unsupervised, presenting to the network the input vectors in random order, after initializing the $\mathbf{m}_{jk}$ to small random values. Each time, the distances are computed, an area, N around the node with the smallest distance is determined in which the $\mathbf{m}_{jk}$ are modified, according to the learning rule:

$$\mathbf{m}_{jk}(t+1) = \mathbf{m}_{jk}(t) + \alpha(t)\beta(n,t)[\mathbf{x}(t) - \mathbf{m}_{jk}(t)] \tag{4.18}$$

where $\alpha(t)$ is the learning rate and $\beta(d,t)$ is a function of the distance in output space of the element from the one with the smallest distance. These were made functions of 'time', because both have to be reduced as learning progressed for successful operation of the map. During the course of sleep, the output of the map was found to 'migrate' from one area to another following characteristic paths and exhibiting stationary 'states'. Their correspondence with conventional sleep stages, was determined using a linear classifier on the outputs of the map and the likelihoods for every condition, which were computed with supplementary information from an EMG and an EOG channel to discriminate awake, light sleep and REM sleep.

4.6.4 Overall EEG assessment

Some methods were developed without any particular EEG features in mind, but applied certain procedures for the general quantification or analysis of the signal. The articles reviewed here are only a small sample of the vast literature available and are characteristic of the variability in the approaches.

Domain transformations

Excluding transients, the EEG signal has relatively constant statistical properties. The signal maintains its properties for a long period of time (a few seconds or longer). Domain transformations, especially the Discrete Fourier Transform (DFT), have been employed by different groups of workers.

Matoušek and Petersén[119] employed the FFT on 60 s EEG epochs to detect immaturity patterns, associated with a variety of brain conditions. The power in the usual frequency bands of activity (δ, θ, α and β) was evaluated. Their ratios and ratios of linear combinations were used in a formula to estimate the age and compared to the actual age of the subject.

The average PSD of several 2 s epochs evaluated using the FFT was used by Brenner et al.[29] for the discrimination of demented, depressed and normal subjects. The logarithms of the total power density in the usual EEG activity bands and the average frequency were subjected to covariance analysis, which revealed differences in their distributions for the three groups.

Oken and Chiappa[138] evaluated the PSD by applying the FFT on 4 s windowed epochs to investigate the short-term variability in the EEG. The mean, variance, skewness, kurtosis, maximum, minimum and the ratio of the variance and the mean for the usual frequency bands were evaluated. Some were found to vary from epoch to epoch, whereas others remained fairly constant.

For real-time analysis, Weide et al.[195] employed the orthonormal Discrete Walsh transform (DWT), because it is considerably faster to compute than the FFT. Walsh 'spectra' contain marked peaks in the presence of sinusoids, at the same 'frequency' (called sequency). Compared to the Fourier, the Walsh domain is not shift invariant, there is a greater spread of components and an enhancement of those with high-sequency.

Parametric methods

Almost three decades ago, Wennberg and Zetterberg[198] suggested a model for the EEG signal consisting of a near-stationary stochastic process with additive transients (e.g. spikes). The autocorrelation of the stochastic component was described by the superposition of three families of functions,

$$\text{Type } 0: \quad r_0(\tau) = E \tag{4.19}$$

$$\text{Type I}: \quad r_I(\tau) = Ge^{-2\pi\sigma|\tau|} \tag{4.20}$$

$$\text{Type II}: \quad r_{II}(\tau) = e^{-2\pi\sigma|\tau|}[G\cos(2\pi f_0\tau) - H\sin(2\pi f_0|\tau|)] \tag{4.21}$$

The coefficients, $\{a_i\}$ and $\{b_j\}$ of an ARMA system excited by white noise, $e(n)$,

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + e(n) + \sum_{j=1}^{q} b_j e(n-j) \tag{4.22}$$

were computed from samples of the auto-correlation function. Separating the system into first and second order processes, the parameters (E, G, H, σ and f_0) were evaluated. These were more descriptive than the $\{a_i\}$ and $\{b_j\}$, because they are interpreted as amplitudes and resonant and cut-off frequencies, readily providing a parametrization of the PSD.

Isaksson and the aforementioned authors[93] used parametric models to overcome the requirement for long EEG segments necessary to achieve adequate frequency resolution when applying the FFT, coming to conflict with the non-stationarity of the signal. The parametrization of the PSD was considered preferable to the arbitrarily chosen δ, θ, α and β ranges. An ARMA model was preferred over the simpler AR, because of its lower order. It was implemented as a Kalman Filter, for the recursive computation of the coefficient vector, θ, which may be time-varying, following a method similar to equation (4.4).

One year later, Pfurtscheller and Haring[147] reported the fast computation of the PSD using an auto-regressive filter. The p filter coefficients, $a_1, \cdots, a_p$ were evaluated from the first Np auto-correlation values and recursively applied to them to estimate the remaining auto-correlation values, a faster procedure than their direct computation, especially if $M \gg p$. The PSD was then evaluated using the FFT. Kleiner[104] explained that the PSD could have

been computed by squaring the FFT of the EEG samples, a faster method, or by evaluating the transfer function of the filter at N discrete points $z = e^{j\pi i/N}$, even faster.

Crowell et al.[50] used an autoregressive model to summarize spectral information in a small number of parameters. The model order, $p = 6$, was selected by minimizing the Akaike Information Criterion and its parameters, were subjected to statistical analysis which revealed differences between age groups and abnormalities in the maturation process.

The AR (autocorrelation) and ARMA (Yule-Walker) models were employed by Smith and Lager[182] to compute amplitudes and frequencies of complex exponential functions, relating to pairs of poles of the system. Problems were reported in selecting the optimal model order and the duration of signal window for coefficient evaluation. The ARMA model produced less biased estimates of the correct parameters than the AR model for simulated data, but the variance of the former was higher. On real records, the AR system proved more versatile, even in the presence of non-stationarities and artifacts, and was preferred as it was simpler. Bias decreased with increasing filter order.

Blinowska et al.[21] fitted an autoregressive model to the EEG using numerical optimization and converted to the continuous transfer function form:

$$H(s) = \sum_{i=1}^{p} \frac{C_i}{s - \alpha_i} \tag{4.23}$$

The C_i and α_i parameters related to peaks in the signal spectrum and to the superposition of parallel processes. The PSD was evaluated from these parameters for the benefit of human analysts.

Hjörth analysis

One of the early methods of EEG quantification, Hjörth analysis, called after the person who suggested it, was so well-formulated, simple and mathematically sound that it caught the attention of many workers[47].

In 1970, Bo Hjörth[87] defined three parameters based on time-domain properties to summarize EEG signal behaviour:

- *Activity*, A, the variance (power), σ_0^2 of the signal
- *Mobility*, M, the ratio of the standard deviations of the derivative of the signal, σ_1 to that of the signal itself:

$$M = \frac{\sigma_1}{\sigma_0} \tag{4.24}$$

expressed the standard deviation of the power spectrum along the frequency axis

- *Complexity*, C, the ratio of the activities of the derivative and the signal,

$$C = \frac{\sigma_2/\sigma_1}{\sigma_1/\sigma_0} \tag{4.25}$$

is a measure of fine detail, deviating from unity (for sinusoidal signals) with frequency spread.

Although these parameters were defined in the time-domain, they are versatile, relating to the moments, μ_k, of the power spectral density since $\mu_0 = \sigma_0^2$, $\mu_1 = 0$, $\mu_2 = \sigma_1^2$, $\mu_3 = 0$, $\mu_4 = \sigma_2^2$, etc. Hjörth reported that his parameters were sensitive to changes of consciousness. He demonstrated that the activity, mobility and a generalized complexity could describe the auto-correlation function, whose diminution with increased lag justified the use of a few parameters for its description[88]. First and second order systems could be quantified exactly by the suggested descriptors (A, M and C).

Saltzberg and Burch[167] proposed a fast procedure for the computation of Hjörth parameters using the zero crossings, N_k, of the kth derivative of the auto-correlation function $R(\tau)$ and the relation $\frac{d^k R(\tau)}{d\tau^k}\big]_{\tau=0} = (-1)^{(k/2)}\mu_k$ for $k = 0, 2, 4, \ldots$ because under some conditions on which the EEG signal generally abides, $\mu_k = (\pi N_{k-2})^2$.

Binnie et al.[18] augmented Hjörth descriptors by an additional parameter, the form factor (similar to complexity) defined by:

$$F = \left(\frac{\sigma_2^2}{\sigma_0^2} - \frac{\sigma_1^2}{\sigma_0^2}\right)^{\frac{1}{2}} \tag{4.26}$$

and used them in the quantification of normal and abnormal EEGs.

Techniques based on correlations

Saltzberg[169, 170] used the sample auto-correlation function, $R(n\Delta t)$ for the fast computation of moments, μ_k of the power spectral density, $P(f)$ related to its mean, variance, skewness and kurtosis:

$$\mu_k = \frac{\int_0^F f^k P(f)df}{\int_0^F P(f)df} = \frac{F^k}{k+1} + \frac{1}{R(0)} \sum_{n=1}^{\infty} A_{n,k} R(n\Delta t) \tag{4.27}$$

where $A_{n,k} = \frac{2}{F} \int_0^F f^k \cos\left(\frac{n\pi f}{F}\right) df$ diminish with increasing n making necessary only the first few terms of $R(n\Delta t)$.

Ricci et al.[159] developed a cross-correlation routine in assembly language taking into consideration the fact that the neural signals to be correlated usually had a small time lag between them, hence it would have be unnecessary to compute correlations using the FFT.

Other methods

The focus of the preceding sections was on trends followed in computerized EEG analysis. Some interesting attempts constituted isolated or too varied methods to be uniquely classified.

To overcome problems from windowing data for the computation of power spectra using the FFT, Daskalova[52] suggested an alternative fast technique, based on the quantification of wave extrema, by their amplitude and time difference. Insignificant extrema were eliminated by employing time and amplitude criteria and the remaining were grouped to form waves, described by their amplitude and frequency (inverse of duration). These were sorted in frequency bins and smoothed to form an estimate of the PSD.

Barlow[10] compared three techniques for monitoring EEG changes with carotid clamping affecting low and high frequencies. These were analogue band-pass filtering to extract the ranges 0–1 Hz and 4–20 Hz, followed by rectification and smoothing, inverse filtering, based on the EEG before clamping, to enhance changes and correlation of a window before clamping. The last method was reported as the most successful, whereas inverse filtering had the worst results.

In 1987 Katz[102] quantified states of consciousness in the EEG using the fractal dimension, defined by Mandelbrot[117] as the ratio of the length of a curve and the longest distance between any two points (its diameter). This was used as a measure of similarity and randomness, although it is sensitive to the units of measurement.

Jansen[95] used chaos theory for the description of EEG signals. EEG rhythms, too regular to be random but not exactly periodic, having bounded amplitudes were treated as the limit cycles out of relatively simple non-linear systems, described by state-space concepts. The 'fractal dimension' of these spaces, computed experimentally, was used to quantify states of consciousness. A system of non-linear differential equations, exhibiting chaos, was proposed for the generation of EEG rhythms.

A different objective, especially important in long-term monitoring, is the compression of EEG signals. McLochlin et al.[123] looked for a compression technique to preserve shape and time relations. The signal was segmented by consecutive fitting of one of three functions (hold, ramp, cosine) using the mean absolute error as a trade-off between compression and fidelity of reconstruction. Using the parameters of these functions (amplitude, offset, slope and frequency) and a two-bit code, compression rates of 3 to 5 were achieved for sleep recordings, sampled at 200 Hz with 12 *bit* accuracy. Data compression was one of the objectives of a 32-channel system developed by Wheeler and Valesano[199]. This was based on storing peaks and times of occurrences instead of the complete signal. The compression

rates and the fidelity of reconstruction were not reported.

4.7 Comments

It is evident that there is no unique or generally acceptable method for automatic EEG quantification, even for specific objectives (e.g. detection of epileptic spikes). Trends in methodology are often associated more with developments in signal processing and other related fields, rather than with their suitability, a clear fact taking into account not *which* method was applied but *how* and in what *context.* All methods were successful to some extent, but none of them was outstanding. Those based on heuristics, are often founded on implicit assumptions that may not be generally valid. On the other hand, rigorous methods with good theoretical foundations are often more robust, but require moderate to high computational powers for their realization. Versatility in the presence of artifacts and noise may be quite low in general, since artifact rejection is a difficult problem requiring special attention.

The designer of an EEG analysis system, has a difficult task at hand. The signal is variable, poorly described and rather unpredictable. Despite the inability to quantify or replicate the highly acclaimed 'experience' of the analyst, automatic techniques have revealed the subjectiveness of human EEG interpretation and the presence of inconsistencies between experts. Could it be that the apparent poor performance of automatic analysis methods is the result of poor definition of the objectives and the design criteria and by comparison with non-ideal human analysts?

Chapter 5

Early models of the EEG signal

5.1 Introduction

A research project often begins with a vague description of its objectives, which is refined later. For this project the main objective was the development of a method for automatic analysis of the interictal EEG of persons who are suspected of suffering from epileptic attacks by indicating transient patterns related to this pathological condition, preferably in real-time (see Introduction). The available EEG records contained primarily one type of transient, the spike-and-slow-wave complex (section 2.5.2), which is associated with minor epileptic attacks (section 3.6.1). This limited the scope of analysis, but not the range of possible methods. A multitude of hypothetical schemes come to mind, only to be rejected on grounds of unsuitability, after refinement or because of previous evaluation by one or more articles in the vast literature on the subject. In EEG analysis the discouragement was most profound, since it appeared that all techniques one could think of had been considered before (see Chapter 4).

At this stage the only element that seemed clear was the need of some pre-processing or parametrization of the EEG record, as the direct classification of the 'raw' signal is not generally recommended. The possibility of applying the developed method to the detection of other types of epileptic events, or even to other signals containing transients was considered important. Hence any development was visualized in the context of its future generalization.

The need for generalization was shadowed by the prospective limitations in its success. As stated in the Introduction and demonstrated by numerous methods in Chapter 4, a sound theoretical method may not be as successful as a crude, practical one, when applied to real signals. In contrast, practical methods are often difficult to amend and modifications often result in a complicated assortment of dissimilar methods to hide the many imperfections

and overcome the shortcomings of an initially not well-thought out design. This is not a general problem with the theoretically-based systems, whose development is often a matter of refinement of the theoretical basis, where the general ideas remain unchanged. The contrast between these two approaches was obvious from the previous Chapter (particularly from section 4.7).

This Chapter is dedicated to the description of some of the early attempts to give a mathematical description (model) of the EEG signal, with emphasis on spikes and similar events. Although these were not particularly successful and were eventually rejected, they highlight some of the problems associated with EEG signal analysis, which were taken into consideration for the methods presented in later Chapters.

5.2 Some notes on EEG analysis practices

In the introductory Chapters on epilepsy and the EEG, the various epileptic patterns, mainly spikes, slow waves and the spike-and-wave complex were described (see sections 2.5.2 and 3.6). For the computer recognition of any pattern, it is necessary to have some quantitative definition that could be formulated into an algorithm. According to the definitions of section 2.5.2, a spike is a transient with a pointed peak (at conventional paper speeds), negative polarity and duration between 20 and 70 ms, clearly distinguished from the on-going activity. A slow wave was described as a single wave with frequency below that of α activity (< 8 Hz) and the spike-and-wave complex was defined in a pattern consisting of a spike followed by a slow wave. Examples of spike-and-wave complexes are shown in Figure 5.1.

Reading the definition given above the electroencephalographer has a fairly clear view of what it describes. The designer of an automatic detection system, on the other hand, is deeply concerned with the qualitative, impressionistic nature of the definitions, the lack of detail and the remembrance of articles stating that even when quantitative, numerical attributes have been used, they should not be treated strictly, but more as guidelines. He/she would probably then recall with appreciation the accurate and detailed definitions and statements of mathematics, physics and engineering and wonder why are neurophysiological definitions not defined in a similar way.

One is inclined to answer that mathematical precision in medical definitions was not necessary. Until recently, medical diagnosis was strictly performed by human experts, like doctors and experienced medical staff. Hence the descriptions of the various phenomena did not need to be too specific, nor quantitative. Humans are more used to qualitative terminology; one tends to use characterizations rather than numbers. And perhaps one of

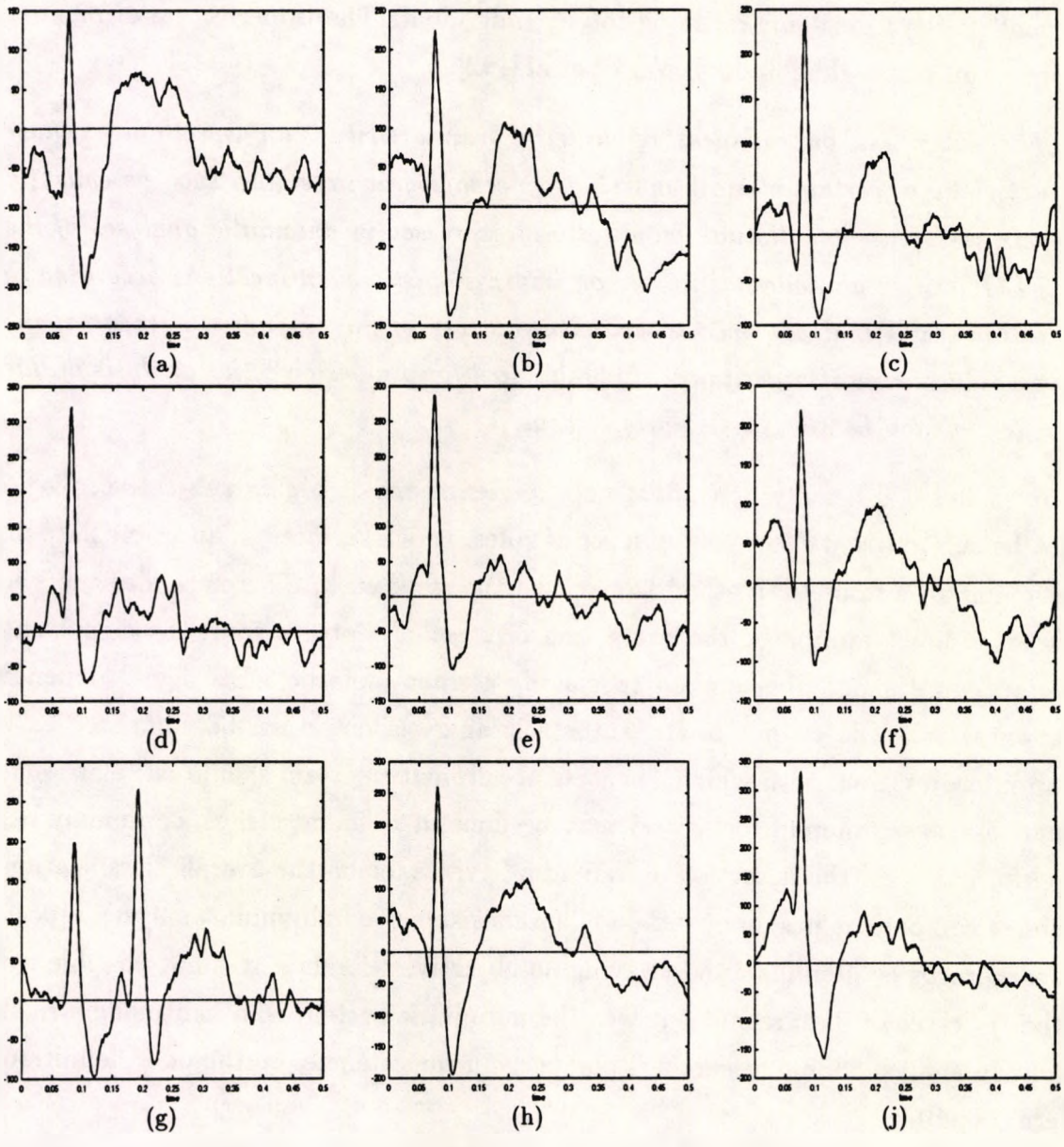


Figure 5.1: Examples of the spike-and-wave complex pattern

the greatest assets of human decision making and work is the ability to confront and overcome problems whose solution has not been prescribed, starting from a general description of the task and 'filling-in' the gaps in the knowledge and improving their skills with experience. With a certain degree of cynicism one could argue that precise methodology was originally created as a means of expanding human ability and removing subjectivity, ensuring that a statement conveys the same meaning to any individual. The imperfection of manual EEG interpretation was highlighted by Walter et al.[192]:

> *The 'sharpness' or 'spikiness' of electrical brain activity is an aspect that is characteristic of certain clinical and electroencephalographic states, such as epileptic seizures. It is actually not quantitatively expressed in automatic analyses of the EEG activity and one has to rely on impressionistic quantifications according to rating scales that are open to criticism because of the subjective element of the procedure. In a recent clinical study the need for a more objective analysis of this phenomenon became particularly evident...*

Human analysis is heavily dependent on experience and is highly subjective. Experience cannot be summarized effectively in a set of rules, which surfaces as an immediate problem in designing an automatic method to perform the same task. Therefore, designers have to follow an indirect approach; the extraction of a set of features from the signal on which some procedure should be applied to classify elements of the EEG signal depending on the application. The output of the automatic analysis may then be compared to that of human interpretation, with which a successful automatic system should correlate well. This comparison takes different forms and may be done at different stages, commonly based on the performance on the detection of individual events, or on the overall classification. This method of comparison has many pitfalls. The consistency of individual analysts, especially on atypical records is often disputed. Having no objective reference, it is not possible to decide whether in cases of disagreement it was the automatic system who had made a mistake or the human analyst. Some researchers blame the inconsistencies on the poor definition of the patterns[16, 107]:

> *The potentialities of clinical and research electroencephalography cannot be realized until such impressions [sharp, distinguished] are replaced by numbers and subjective descriptions are replaced by mathematically derived characterizations.*[107]

Regarding the definition of the spike-and-wave complex, Principe and Smith[155] reported a disagreement of instances of the paroxysm in records with the definition, despite its generality:

> *...results show that the requirement of spikes followed by slow waves was too restrictive for most petit mal paroxysms ...*

These authors do have faith in human decision making and attribute misinterpretations to the small, often improper, set of features used in automatic systems. Such a statement is true for any pattern recognition mechanism, as a set of improperly chosen features could easily ruin the performance of the best classifier.

> *In general, attempts at definition are either qualitative in nature or, even when quantitative, lacking in parameters which describe the phenomenon exactly.*[108]

Observations such as these are crucial in medical diagnostic practice, as they indicate that human decision-making may not be trusted fully, especially when the task involves such a strong subjective element and imprecise criteria. On the other hand, even if an automatic method outperformed human analysts, there is always the possibility of it failing to perform correctly in a small number of cases. On such occasions the question of who should take the responsibility for the misdiagnosis might emerge. These are highly sensitive issues and despite their importance, are open to debate. Besides, should a successful automatic EEG analyzer ever become widespread, it would be the users, not the designers who should decide how it is used and to what extent it could be trusted.

It is evident that the issue should not be that of competition between humans and machines in the task of medical diagnosis. In the interest of the patients both should be used to complement each others imperfections. Perhaps some form of feedback, or interaction, between automatic systems and human electroencephalographers would improve the consistency and hence the objectivity of both.

5.3 The problem of data representation and modelling

Defining a pattern accurately is crucial for its identification, detection recognition or classification. For small sets of patterns the definition may be stated explicitly as the collection of all their instances, but for most real-life applications, this is not practical, since the number of elements in the set is large, if not infinite. An approximation to this explicit definition is a set of examples of what might constitute the pattern, like the collection of spike-and-wave complexes shown in Figure 5.1. It is intuitive that any such set cannot define completely all possible spike-and-wave complexes, in the same way as any set of examples of a circle cannot define all possible circles. Besides, such definitions cannot be related and studied in a systematic way and the direct use of these pattern representations is generally avoided.

A better alternative is the extraction of a set of attributes (features) that characterize the pattern in question for further treatment. In other words the practically encountered patterns are related to a theoretical *model*. This is similar to defining a circle as the set of points on a plane that are located at a specific distance r from the centre point, c. The modelling approach uses a mathematical representation of the pattern and quite often compresses the pattern in a small number of meaningful parameters, much like a circle is defined by r and c. In practice, instances of the pattern may deviate from the model and may be contaminated by noise. It is often possible, however, to approximate individual examples of the pattern by changing the value of the parameters, to achieve the best fit. Usually, there is a region of values of the parameters corresponding to these examples, which also represent valid instances of the pattern. Such regions are seldom isolated and the boundaries of the formed clusters are a good criterion for what constitutes the pattern. This may be assimilated to the human cognitive system, where examples of a pattern are believed to be stored in terms of subjective internal representations and new instances are classified using some form of interpolation and extrapolation from a priori information. Unfortunately, these representations and mechanisms are not used consciously and cannot be expressed verbally, which accounts for the imprecise way of definition of EEG patterns.

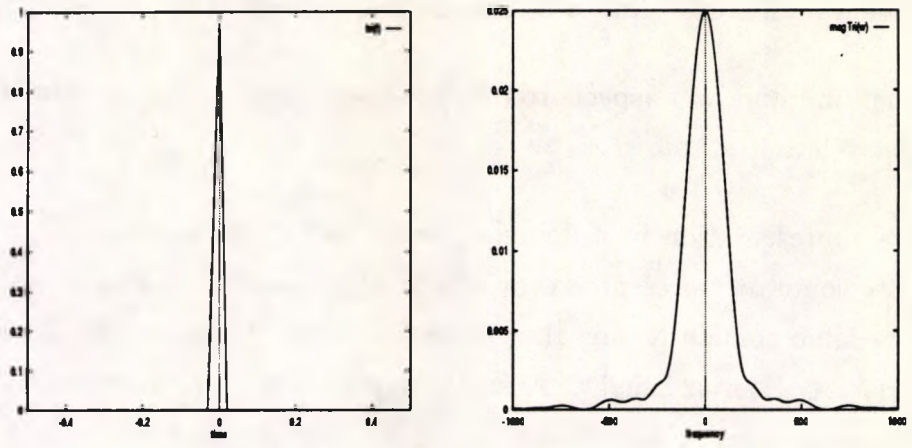
Within this framework, learning may be regarded as the transfer of mechanisms and representations from an experienced analyst to an inexperienced one through the rather limited channels of human communication. In the case of automatic identification, the task is even more difficult, as both mechanisms and representations are non-existent and there is no intuitive way to formulate them.

Signal modelling is an approach to form some internal pattern representation for the specification of the pattern of interest, in this case the spike-and-wave complex. A good model should

- describe consistently most, if not all, of the instances of the pattern
- preferably describe the signal in a unified manner
- have sound mathematical foundations and properties

Additionally, since it would form the basis of EEG analysis, the suggestion of a procedure for analysis by the model would be an additional advantage.

The models that follow in this Chapter constitute early attempts to devise a method for the description of the spike-and-wave pattern. Most of them did not conform to all the aforementioned criteria but their inclusion is believed necessary, as they have influenced the



(a) Triangular spike model (b) Its amplitude spectrum

Figure 5.2: Triangular model for spike in the time and frequency domain

work that followed and their properties, especially their limitations, may be considered by those following a similar line of research.

5.4 Study of triangular waveforms

One of the first attempts to model EEG spikes was directly linked to models used in spike detection by numerous authors (see section 4.6.1). This underlying model was implicitly or explicitly used by methods based on derivatives of the EEG, like those of Comley and Brignell[45] and Stelle and Comley[183]. It was based on the appearance of the spikes in the EEG record. According to this, spikes were described as the peak formed by joining two line segments, one of positive and the other of negative slope, which approximate the rising and falling edges of the spike, respectively. The magnitudes of the slopes could be different. An example of a spike so described and its Fourier Transform are shown in Figure 5.2.

The apparent spread of the frequency components led to the conclusion that the Fourier domain, despite its broad use in signal analysis, was probably not the most appropriate for the representation of spikes, as the broad-band nature of isolated spikes would necessiate the assembly of the relevant frequency components before detection, thus causing complications in their recognition. An alternative transformation was sought having the following properties:

- considerable data reduction, as groups of data values are represented with a few parameters

- enhancement of 'features' that are believed to be important
- depletion of unimportant aspects of the data to prevent them from affecting classification of the relevant patterns

Such a data representation in a domain (vector space) where the main morphological aspects of spikes would be represented concisely in a few parameters was sought. A transformation from the time domain to this alternative space, using a (non-sinusoidal) set of basis functions, $\{\phi_i(t); i \in \mathbb{N}\}$ was sought. A function $f(t)$ may be represented in the range of the transformation as a set of coefficients $\{a_i; i \in \mathbb{N}\}$, one for each of the $\phi_i(t)$ using a least mean squared error fit. The general formula for this computation is

$$\sum_{i=0}^{\infty} a_i \int_{t_1}^{t_2} \phi_i(t)\phi_j(t)dt = \int_{t_1}^{t_2} f(t)\phi_j(t)dt \tag{5.1}$$

where t_1 and t_2 determine the period where the transformation applies.

A natural choice for the $\{\phi_i(t)\}$ is a set of *triangular waveforms*, which was suggested by the definition of the spikes given earlier on (see Figure 5.2). Assuming a duration $T = t_2 - t_1$, two types of triangular functions, $\phi_{A_T}(t)$ and $\phi_{B_T}(t)$ might be defined:

$$\begin{array}{lcll}
\phi_{A_0}(t) & = & 1 & nT - \frac{T}{2} < t < nT + \frac{T}{2} \\
\phi_{B_0}(t) & = & 0 & nT - \frac{T}{2} < t < nT + \frac{T}{2} \\
\phi_{A_T}(t) & = & 1 + \frac{4}{T}(t - nT) & nT - \frac{T}{2} < t \geq nT \\
 & = & 1 - \frac{4}{T}(t - nT) & nT < t \geq nT + \frac{T}{2} \\
\phi_{B_T}(t) & = & -2 - \frac{4}{T}(t - nT) & nT - \frac{T}{2} < t \geq nT - \frac{T}{4} \\
 & = & \frac{4}{T}(t - nT) & nT - \frac{T}{4} < t \geq nT + \frac{T}{4} \\
 & = & 2 - \frac{4}{T}(t - nT) & nT + \frac{T}{4} < t \geq nT + \frac{T}{2}
\end{array} \tag{5.2}$$

where $n \in \mathbb{N}$. The $\phi_{A_T}(t)$ is an even function, equivalent to a cosine in the Fourier basis, whereas $\phi_{B_T}(t)$, being an odd function is equivalent to a sine. Examples of these functions are shown in Figure 5.3.

The abrupt gradient changes at their peaks were initially considered an advantage over the smooth sinusoidal basis of the Fourier domain. Discrete versions of these functions could also be defined, sampling both the time and the 'frequency' ($\frac{1}{T}$). as with the Discrete Fourier Transform.

Orthogonality of the basis functions

$$\int_{t_1}^{t_2} \phi_i(t)\phi_j(t)dt = 0 \qquad i \neq j \tag{5.3}$$

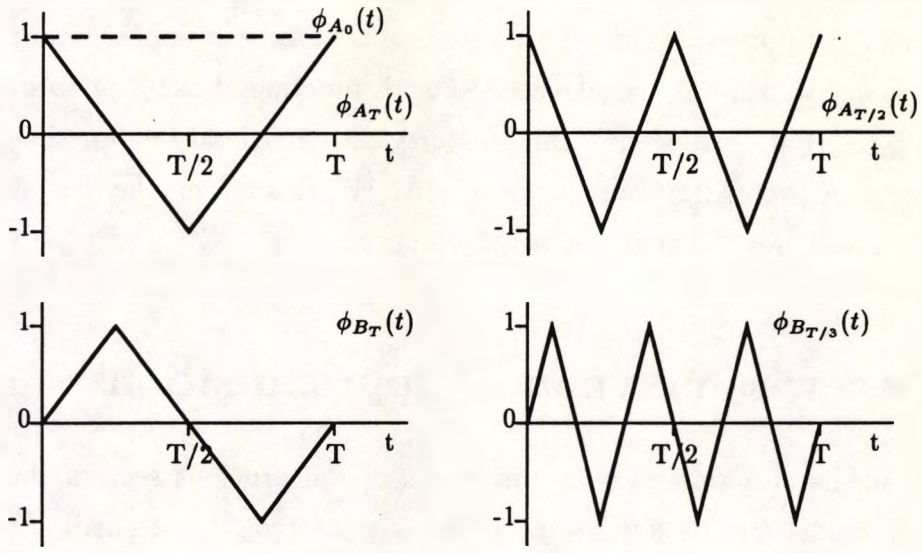


Figure 5.3: Examples of the proposed triangular basis functions

is a crucial property in their use. Not only does it simplify the computation of the transform, but it also makes their computation independent from each other:

$$a_j = \frac{\int_{t_1}^{t_2} f(t)\phi_j(t)dt}{\int_{t_1}^{t_2} \phi_j^2(t)dt} \tag{5.4}$$

Non-orthogonal functions use equation 5.1, which requires matrix operations for the solution of the resulting system of non-trivial simultaneous equations, necessary for the estimation of the $\{a_i\}$. This is a serious disadvantage as it leads to numerically unsound or computationally intensive techniques.

In terms of theory, some useful properties only apply to orthogonal transformations. For example, the independence of each coefficient from another often called the *finality of coefficients*[121, p. 28], is highly desirable, because the values of the subset of the transformation, $\{a_0, \ldots, a_N\}$, computed from the subset of the basis, $\{\phi_0(t), \ldots, \phi_N(t)\}$, will remain unchanged under the introduction of new function, ϕ_{N+1}, as this affects only the a_{N+1} coefficient. Hence an arbitrarily close approximation of $f(t)$ may be achieved in a systematic way by recursively introducing new functions belonging to the basis and computing the new coefficients.

The converse is even more useful for signal analysis, as coefficients with small values, hence the corresponding basis functions, may be ignored, without any changes in the remaining ones. This is a standard procedure in pattern recognition used for the reduction of the number features used as well as for the prevention of noise in these components affecting classification.

Undoubtedly orthogonality is perhaps the most important property that a good set of basis functions must possess. When the question of orthogonality of the triangular basis was raised, the answer was simply negative. Although functions belonging to different groups ($\phi_{A_T}(t)$ and $\phi_{B_T}(t)$) are orthogonal, this property does not hold for functions belonging to the same group. A proof is given in Appendix A. For this reason the use of the triangular waveforms as a basis for the analysis was abandoned.

5.5 Piecewise polynomial approximations

After the problems encountered with the model of the previous section the suitability of the triangular model for the description of EEG spikes was questioned. Are EEG spikes really composed of line segments forming a sharp peak? With the bandwidth of an EEG recorder limited to, say, 0–75 Hz and assuming a spike duration of 70 ms (corresponding to a fundamental frequency of 14.28 Hz) components of up to the fifth harmonic are recorded. Sharpness is a property attributed to high-frequency components. Therefore, band-limiting by the recording equipment, augmented by low-pass filtering before sampling, would make spike peaks smooth, assuming they were sharp in the first place. It would not be unreasonable to suggest that the 'sharpness' of spike phenomena is a rather subjective attribute, possibly related to the conventional recording paper speeds and their relation to the rest of the EEG record. As such, it is a 'visual' rather than an actual characteristic and it disappears with stretching of the time axis. The possibility of using smoother functions was considered as a better alternative. These usually have better mathematical properties and can be studied with more ease than those exhibiting abrupt transitions.

Polynomials (with finite order) are continuous mappings, but cannot approximate infinitely oscillating functions, like the EEG, since they may only have a finite number of local extrema (maxima and minima). Although this increases with the order of the polynomial, so does the computational load for fitting it to the data. The order of the approximating polynomial is also likely to increase even further in the presence of isolated transitions of large amplitude, like spikes. Besides, high-order polynomials, fitted to sets of data, tend to provide a better approximation in the middle of the range and often suffer from oscillations at the two extremes[89, p. 142], an undesirable feature as the error of approximation is expected to be uniformly distributed over the range.

These problems may be remedied using a piecewise approximation of the EEG signal with polynomials of fixed order. Each one is fitted to a small section of the signal. The whole signal is then represented by the concatenation of all these polynomial sections, or

implicitly by their coefficients and the range where each one applies.

The way these segments are fitted to the data points defines the properties of the approximation. It is possible to separate the EEG record into consecutive segments of equal duration and fit a polynomial to each one of these using an optimization criterion, like the minimization of the mean squared error function. This, method, however is not always satisfactory, as the resulting functional approximation will suffer from discontinuities at the boundary points (end-points) of each segment and the residual mean squared error may vary considerably from one segment to the next. A better technique is based on *polynomial splines.*

5.5.1 Traditional piecewise polynomial fit through a set of points

Traditionally spline functions find applications in interpolation. Polynomial splines consist of fitting polynomial sections through a discrete set of sample points to describe the unknown underlying function with a smooth curve passing through them. The smoothness of this piecewise approximation depends on the order, n, of the polynomials, which is fixed during the process. The nature of the approximation ensures not only the continuity of the fitted curve but also that of its first $n-1$ derivatives, which is achieved by setting the requirement that adjacent polynomial segments have equal value and derivatives up to the order $n-1$ at the end-points where they meet[89][pp. 144–149][6, pp. 141–150]. This ensures a smooth transition from one segment to the next, maintained even if the curve is differentiated $n-2$ times. The method of fitting a polynomial spline to a set of data points is quite systematic. The problem is, typically, to estimate the parameters of a polynomial segment in order to approximate an unknown function between two known points, $s_i = (x_i, y_i)$ and $s_{i+1} = (x_{i+1}, y_{i+1})$, where the stated rules of continuity of derivatives apply. The collection of these points is divided into consecutive pairs, $(s_0, s_1), (s_1, s_2), \ldots, (s_{N-1}, s_N)$, where the end-point of one pair is the starting point for the next.

For nth order polynomials, there are a total of $n+1$ coefficients to be computed for each, that is a total of $(n+1)N$. The continuity conditions at each one of the $N+1$ boundary points generate $2N$ equations and those of the derivatives that apply for the $N-1$ internal points generate another $(n-1)(N-1)$ equations, that is a total of $(n+1)N-(n-1)$ equations, requiring another $n-1$ equations to solve uniquely. These are selected by *choosing* the conditions at the end points, typically setting the high-order derivatives to zero. The system of equations so generated is then solved for the $n+1$ coefficients.

Usually, the order of the polynomials is maintained low. First-order polynomials only guarantee continuity, but not smoothness, since the approximation is piecewise linear. Sec-

ond order (quadratic) have the form $a_2x^2+a_1x+a_0$ and are not employed because they tend to cause oscillations between the pairs of boundary points they are attached to, owing to their parabolic shape. *Cubic splines* are extensively used, since they are continuous and have continuous first and second derivatives at the end-points. Their capability to have points of inflection as well as maxima and minima ensures a smooth fit with minimal oscillatory behaviour while computational requirements are not excessive, due to their low order. In fact, the necessary equations may be manipulated to form a tridiagonal matrix, which is easy to solve[89, pp. 144–149][6, pp. 141–150]. Higher-order polynomials, are often considered inferior, since they may exhibit oscillations, due to their many maxima and minima and they require more computations[6, pp. 144].

5.5.2 Piecewise polynomial approximation of infinite sequences

Spline functions of the class described cannot be applied directly for the purpose of EEG approximation owing to two reasons. Firstly, the boundary points, which are critical in the computational procedure are not known a priori; the problem here is not to interpolate between the few known key points, but to compress sequences of points into a few parameters. The second problem is the increasing number of polynomials that need to be computed as new data become available in the infinite evolution of the signal; it is understood that the change in the end conditions with the incorporation of every new end-point results in a new set of equations that need to be solved again. This is rather impractical, hence an alternative method, akin to splines, was considered. Remembering that the EEG is uniformly sampled at a fixed sampling rate, the problem may be reformulated by maintaining the same number of equations, but modifying the conditions for fitting each polynomial. Therefore what is sought is a *criterion* to select a subset of the available points to act as boundary points of polynomial segments and a *different* set of equations for computing their coefficients.

Formulation and solution of the estimation problem for one segment

The second problem will be considered first. The presence of points before and after a selected boundary point, $y(kT)$, allows the computation of estimates of the derivatives at these end points, using, for example, the simple approximation

$$y^{(n)}(kT) = \frac{y^{(n-1)}[(k+1)T] - y^{(n-1)}[(k-1)T]}{2T} \tag{5.5}$$

where $y^{(n)}(kT)$ denotes the nth time-derivative of the sampled signal $y(kT)$ at the point kT. This can be applied recursively to compute the first, second, or higher derivatives, as required.

Then, instead of the continuity of the first $n-1$ derivatives, the new condition is the *coincidence* of the values of the $\frac{n-1}{2}$ derivatives (estimated in the suggested way) at a specific boundary point, with the ones computed from the two polynomial segments meeting at that point. This implies that only the first $\frac{n-1}{2}$ derivatives will be continuous. A disadvantage of this is the reduced smoothness for a given order, compared to ordinary splines, and the inability to use even ordered polynomials, due to the existence of one equation more than the number of unknowns.

The first useful polynomial is the cubic one, which has a continuous first derivative at its boundary points. Fifth-order polynomials were also considered, since they have continuous first *and* second derivatives. In both cases the approximation consists of a collection of polynomials, $\{p_1(t), p_2(t), \ldots, p_i(t), \ldots\}$, each valid in their corresponding time interval $\{[t_1, t_2], [t_2, t_3], \ldots, [t_i, t_{i+1}], \ldots\}$. For a cubic spline these have the form

$$p_i(t) = a_3 t^3 + a_2 t^2 + a_1 t + a_0 \quad \text{for} \quad t \in [t_i, t_{i+1}] \tag{5.6}$$

For continuity and smoothness, the end points, (y_i, y_{i+1}), and their derivatives, (y'_i, y'_{i+1}) must be shared between the previous and the next polynomial segment, that is $p_i(t_i) = p_{i-1}(t_i) = y_i$, $p_i(t_{i+1}) = p_{i+1}(t_{i+1}) = y_{i+1}$, $p'_i(t_i) = p'_{i-1}(t_i) = y'_i$ and $p'_i(t_{i+1}) = p'_{i+1}(t_{i+1}) = y'_{i+1}$. These values together with t_i and t_{i+1} are used for the computation of the unknown coefficients, a_0, a_1, a_2 and a_3. To prevent the large increase in the values of t, as the signal progresses, the origin of the t-axis is relocated to coincide with the beginning of the segment, thus $t_i = 0$ at the beginning of the ith segment. Substituting into equation (5.6) and its derivative, $p'_i(x) = 3a_3 t^2 + 2a_2 t + a_1$, gives four equations

$$\begin{aligned} y_i &= a_0 \\ y'_i &= a_1 \\ y'_{i+1} &= 3a_3 x_{i+1}^2 + 2a_2 x_{i+1} + a_1 \\ y_{i+1} &= a_3 x_{i+1}^3 + a_2 x_{i+1}^2 + a_1 x_{i+1} + a_0 \end{aligned} \tag{5.7}$$

which may be solved uniquely for the unknown coefficients:

$$\begin{aligned} a_0 &= y_i \\ a_1 &= y'_i \\ a_2 &= 3\frac{y_{i+1} - y_i}{t_{i+1}^2} - \frac{y'_{i+1} - y'_i}{t_{i+1}} \\ a_3 &= \frac{y'_{i+1} + y'_i}{t_{i+1}^2} - 2\frac{y_{i+1} - y_i}{t_{i+1}^3} \end{aligned} \tag{5.8}$$

With similar considerations, it can be shown that the coefficients of a fifth-order polynomial segment can be expressed in terms of the values (y_i, y_{i+1}) and the first and second derivatives (y'_i, y'_{i+1}, y''_i,y''_{i+1}) at the two end points by the following set of relations:

$$\begin{aligned}
a_0 &= y_i \\
a_1 &= y'_i \\
a_2 &= \frac{y''_i}{2} \\
a_3 &= 10\frac{y_{i+1}-y_i}{t^3_{i+1}} - \frac{4y'_{i+1}+6y'_i}{t^2_{i+1}} + \frac{y''_{i+1}-3y''_i}{2t_{i+1}} \\
a_4 &= -15\frac{y_{i+1}-y_i}{t^4_{i+1}} + \frac{7y'_{i+1}+8y'_i}{t^3_{i+1}} - \frac{2y''_{i+1}-3y''_i}{2t^2_{i+1}} \\
a_5 &= 6\frac{y_{i+1}-y_i}{t^5_{i+1}} - 3\frac{y'_{i+1}+y'_i}{t^4_{i+1}} + \frac{y''_{i+1}-y''_i}{2t^3_{i+1}}
\end{aligned} \tag{5.9}$$

This procedure is an alternative to the one for splines, with the difference that the coefficients of a segment depend only on the conditions at its end-points and hence may be computed as soon as its second boundary point becomes available. This, however, does not suggest a way to select these points, and hence define a segment, which was the other half of the polynomial segment fitting problem.

Selection of boundary points for a piecewise polynomial approximation

The selection of the boundary points, defining the end of a segment and the beginning of a new one, was overcome by assuming that every new data point is potentially the end of the current segment, fitting the polynomial as described and measuring the error of the approximation, terminating the segment before a predefined error criterion failed. To demonstrate how this is achieved, one has to consider the following algorithm, applied to a continuously sampled signal, $y(0), y(T), y(2T), \ldots, y(nT), \ldots$:

1. Initially set the beginning of the current segment to one of the first available samples, $y(nT)$. This should be preceded by enough samples to compute its derivative(s) (see next step).

2. Compute the derivative of the EEG at the beginning of the segment, $y'(nT)$ (also $y''(nT)$ for fifth-order polynomials).

3. Set the length of the segment, $k = 2$, that is, use $y(nT)$ and $y[(n+2]T)$ as the boundary points of the segment, with $y[(n+1)T]$ as the only intermediate point.

4. Compute the derivative(s), $y'[(n+k)T]$ (also $y''[(n+k)T]$ for 5th order) of the end point $y[(n+k)T]$ as in step 2.

5. Use equations (5.8) or (5.9), whichever applicable to compute the coefficients of the polynomial segment passing through the end-points (notice that $t_{i+1} = kT$).

6. Compute an error function, ξ, between the actual intermediate points, $y[(n+1)T], \ldots, y[(n+k-1)T]$ and those computed from the fitted polynomial.

7. If the value of ξ is less than a preset upper error bound, ξ_{max}, increase the length of the segment, k, by 1 and repeat from step 4. Otherwise, signal the end of a segment by storing the coefficients computed when the segment length was $k-1$ and the length of the segment. Set $y[(n+k-1)T]$ as the beginning of the new segment and repeat from step 2.

It should be noted that the derivatives computed above are not exact. A first approximation was taken using the average of the forward and backward differences at the point of reference (see equation 5.5):

$$y'(mT) = \frac{y[(m+1)T] - y[(m-1)T]}{2T} \tag{5.10}$$

$$y''(mT) = \frac{y[(m+2)T] - 2y[mT] + y[(m-2)T]}{4T^2} \tag{5.11}$$

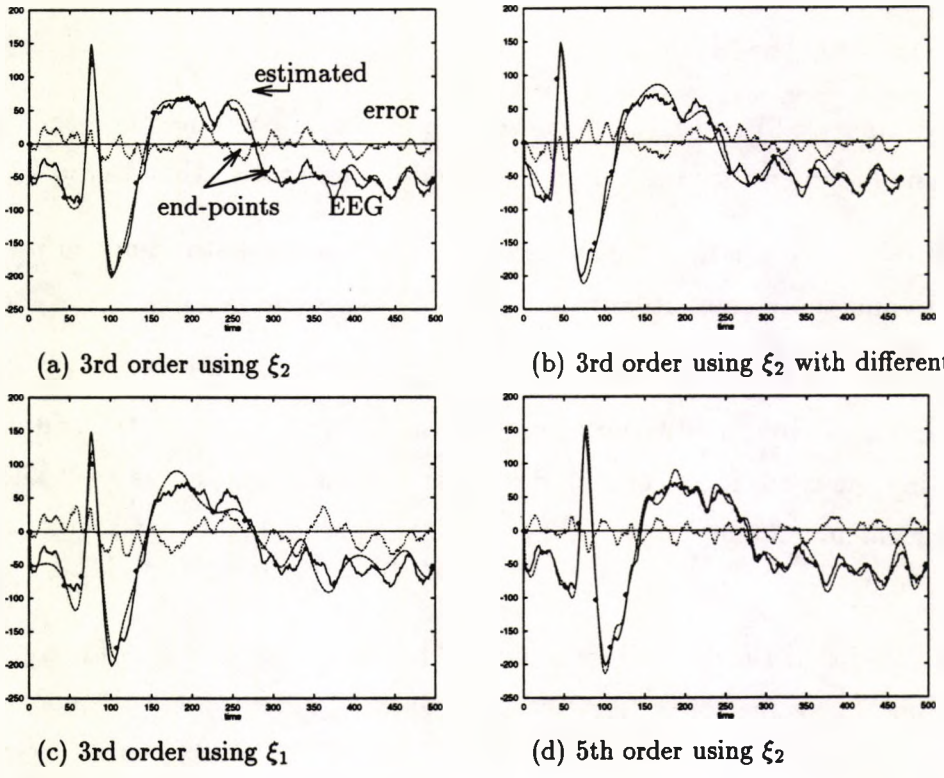
A similar procedure with a cubic polynomial has been used for data compression of ECG signals by Sandman and Sapir[171] who employed the maximum absolute error as the criterion for the termination of segments. For the EEG, both the *maximum absolute error* and the more traditional *mean squared error* were compared:

$$\xi_1 = \max_{i=1}^{k-1} \left\{ \left| y((n+i)T) - \sum_{j=0}^{3} a_j[(n+k)T]^j \right| \right\} \quad \text{maximum absolute error} \tag{5.12}$$

$$\xi_2 = \frac{1}{k-1} \sum_{i=1}^{k-1} \left\{ \left[y[(n+i)T] - \sum_{j=0}^{3} a_j[(n+k)T]^j \right]^2 \right\} \quad \text{mean squared error} \tag{5.13}$$

5.5.3 Results and discussion

Examples of polynomial section fitting to a spike-and-wave complex for both criteria are shown in Figure 5.4(a) and (c). The results of the fit for either criterion have been similar, although the comparison was not exact, because the criteria cannot be compared for the same value of ξ_{max}. Choosing 'equivalent' values of this limit for ξ_1 and ξ_2 was based on a simple technique, which was not always consistent. The amplitude of the error of the approximation



(a) 3rd order using ξ_2

(b) 3rd order using ξ_2 with different origin

(c) 3rd order using ξ_1

(d) 5th order using ξ_2

Figure 5.4: Fitting polynomial sections on a spike-and-wave complex

was assumed to follow a zero-mean Normal (Gaussian) probability distribution. The Mean Squared Error function (ξ_2) is related to the variance of the error, whereas the Maximum Absolute Error (ξ_1) to its modulus. With a 99% confidence, for a variance σ^2 the maximum squared modulus is $(2.575)^2\sigma^2$. Hence for a selected MSE limit, $\xi_{2_{max}}$, the 'equivalent' MAE limit would be $\xi_{1_{max}} = 2.575\sqrt{\xi_{2_{max}}}$. In the examples of Figure 5.4, whenever ξ_2 was used, $\xi_{2_{max}} = 300$ whereas whenever ξ_1, was employed, the error bound was $\xi_{1_{max}} = 45$.

There was no significant difference in the number of segments obtained between criteria for a given polynomial order, although the selection of end-points was (as expected) different, since the fitness criteria are dissimilar. The level of the residual error was comparable for the two error criteria. For some parts of the signal, ξ_1 yielded a lower residual error, for others ξ_2 was better.

For the fifth-order polynomial, the fit on the spike was slightly better than the cubic, although some small oscillations were observed after the slow-wave.

The results in the presence of noise were difficult to compare, as the errors increase rapidly with segment size and exceed the thresholds resulting in a large number of short segments.

It would appear that through this method the problems of feature extraction and segmentation of the EEG signal, which are critical in pattern recognition had been solved

simultaneously. The coefficients of the polynomial would be the features describing the segment, defined by the procedure that computes them. If additionally these features and the segmentation procedure were consistent, then the representation problem, stated at the beginning of this section, would have been solved. Could the segment-ends defined actually be the boundaries of important elements of the EEG which could not be defined intuitively? Unfortunately, this had not been the case. Consistency of the segmentation is still a problem and its consequences on data representation by polynomial coefficients are even more serious, as demonstrated in Figure 5.4(b). This shows the same signal, containing a spike-and-wave complex as in (a), (c) and (d) of the same Figure, with the origin of the first segment displaced in time. Since in practice the location of the origin of a segment cannot be predefined, as it lies at the end of the previous segment, this describes a situation commonly arising in practice. The presence of the spike with its large slopes resulted in a new segment defined near its summit, but the polynomials that describe it are different, owing to the dislocation of the starting point. For the segments including the rising edge of the spike in Figure 5.4, with $T = 1\ ms$ the polynomials in (a) and (b) were

$$\begin{aligned} s_a(t) &= 18831996.00t^3 - 370599.84t^2 - 15085.00t + 119.26 \\ s_b(t) &= 88244160.00t^3 - 3586203.25t^2 + 24940.00t + 94.82 \end{aligned} \tag{5.14}$$

respectively. The discrepancy is largely attributed to the difference in the starting point and the associated conditions between the segments. This indicated a serious limitation of the method when used for segmentation and parametrization, its *sensitivity to initial conditions.* It is not difficult to deduce that inconsistency is also present if noise is introduced in the data, which would affect segmentation even if the same initial point is chosen.

In an attempt to reduce this problem, the condition for termination of a segment (see step 7 of the procedure) was modified. Instead of using an error criterion, segments were defined as fragments of the signal between baseline (zero) crossings, but the results were disappointing as the elimination of error control in the approximation resulted in either too many segments, if many zero crossings were present, or a poor approximation, in cases of baseline drift. Noise and low-amplitude high-frequency oscillations discredited the definition of the boundaries of the segments causing excessive segmentation. Similarly, waves occurring away from the baseline were poorly approximated.

The inability of the spline process to segment the EEG signal in a meaningful way put their use in dispute. If an independent segmentation procedure is introduced, there is little point in restricting the parametrization method to cubic splines, as other methods may arise from that. Splines are an alternative method for representing the signal and may be used for its compression, for storing purposes, or for the estimation of its derivatives, as explained in

the next Chapter (section 6.4.2).

5.6 A recursive short-time frequency-sampling method

The failure of polynomial splines to represent the signal consistently was caused mainly by their inability to maintain the same set of parameters for similar signal patterns, and their dependence on relatively few points of the signal, which were not necessarily relevant in subsequent analysis. It was also impossible to represent features, like spikes, by a single polynomial segment contrarily to initial expectations. These issues arose from the method of parametrization, which was performed on a sparse sample basis, rather than on a continuous one, taking into consideration information contained in every individual sample. The piecewise approximation was therefore rejected and an alternative was sought.

5.6.1 Frequency analysis of time-varying signals

For an infinitely oscillating signal like the EEG the sinusoidal basis of Fourier was reconsidered. Sinusoids have infinite duration and can model the extrema of the EEG. They are, however, rather ill-equipped to describe signals of brief duration (transients), such as spikes. To maintain some emphasis on these signals, brief segments of time must be considered and their spectra evaluated separately. This again is contrary to the argument, as it represents a piecewise operation, which was a serious limitation in the case of polynomial segments. Of course the effects here would be less dramatic. Since the Fourier transform of signals shifted in time have the same amplitude, segments containing the same signal (a spike for instance) are expected to have similarities in their amplitude spectra although their phase characteristics may differ[121, pp. 132–133].

It is therefore sufficient to make consecutive segments long enough to enclose the spike pattern and introduce enough overlap between consecutive segments to ensure that spikes not fully contained in a particular segment would be contained in the next. This overlap should optimally be as long as the pattern itself.

Traditionally the Fourier Transform of a finite-length discrete signal segment, is computed by the Fast Fourier Transform (FFT)[46] algorithm, an efficient technique, requiring $\frac{N}{2}\log_2 N$ complex multiplications for the evaluation of the Transform of a window of N consecutive samples.

For an on-line system detecting transients buried in signals of long duration, the application of the FFT is quite inflexible. By the nature of the algorithm, computation cannot begin until all the N samples of the signal segment become available. Therefore it must

be computed either in the inter-sample period, or computation must be interrupted every time a new sample arrives. The first solution would impose a severe restriction on the bandwidth of the sampled signal, whereas the second, despite being more flexible, would still be subject to limitations caused by the nature of EEG transients and the requirement for overlapping windows. If the frequency resolution needs to be improved, the number of samples in the window must be increased, either by choosing a longer segment or by raising the sampling rate. In either case the computational load would increase according to the $\frac{N}{2}\log_2 N$ rule. Furthermore, the frequency components of the FFT always represent equal bandwidths and if only a few frequency components prove relevant by subsequent analysis, no increase in computational effectiveness would be gained, as all components are evaluated simultaneously.

From the theoretical point of view, if a long segment is chosen, to achieve high-frequency resolution, the signature of brief transients like the spike-and-slow-wave complex in the frequency domain would deteriorate.

Therefore an alternative, possibly equivalent, computational method was sought, which could perform the Fourier Transform of consecutive overlapping signal segments in a more gradual way, being flexible in the number of frequency components it could compute.

5.6.2 Frequency-sampling digital filters

An algorithm conforming to these specifications was formulated from a special class of *frequency sampling* digital filters, highlighted by Lynn[114, 115]. These are based on the so-called *comb filter*, which has transfer function

$$C(z) = 1 - z^{-M} = \frac{z^M - 1}{z^M} \tag{5.15}$$

and possesses M zeros, distributed at equal angles on the unit circle in the z-plane and an equal number of poles at the origin. Its time and frequency response are shown in Figure 5.5.

If a zero of the comb filter is annihilated by a coinciding pole, the complete system may be made to have low-pass, band-pass or high-pass characteristics. This is equivalent to cascading $C(z)$ with an all-pole first or second order system, $D(z)$, with a pole at $z = 1$, for low-pass, or at $z = -1$, for high-pass behaviour. The most interesting case arises when $D(z)$ has two complex conjugate poles at $z = \cos(\frac{2\pi}{M}k) \pm j\sin(\frac{2\pi}{M}k)$ for $k = 1, 2, \cdots, \frac{M}{2} - 1$. In this case, $G(z) = C(z)/D(z)$ has a pass-band centred around the frequency $\frac{2\pi}{M}k$. It should be noted that the resulting system is FIR, since it has zeros but no poles[114]. This is easy to observe by considering the impulse response of $G(z)$. The comb filter initially excites

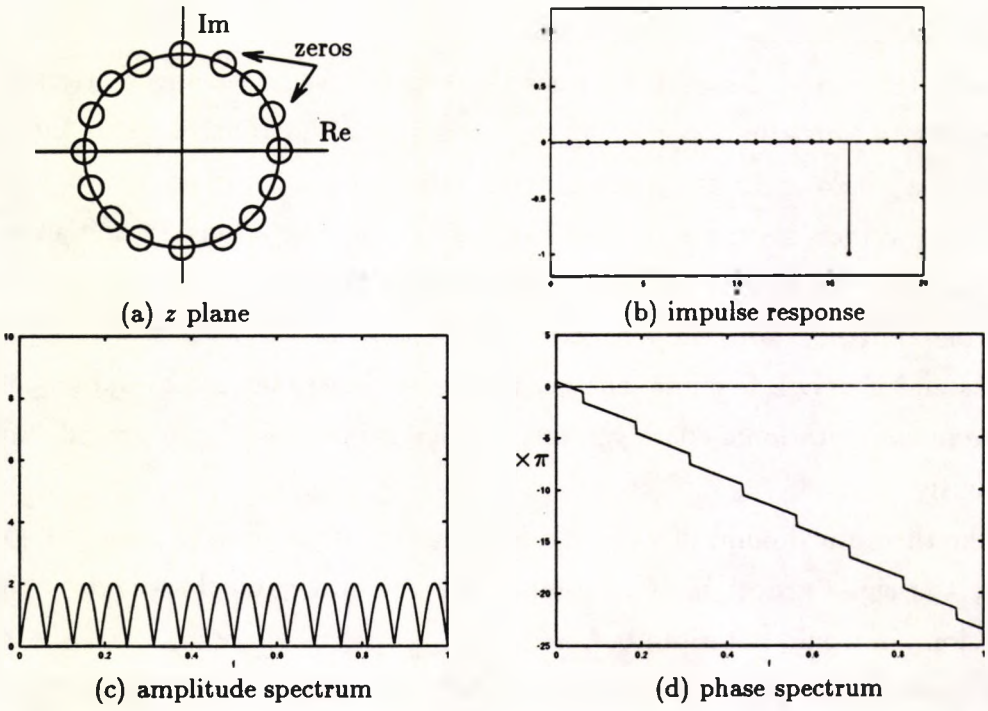


Figure 5.5: Properties of the comb filter

the resonator, which will oscillate at its centre frequency until the negative impulse of the response of the comb-filter inhibits its operation m samples later (Figure 5.6). This occurs because all such resonator have a frequency $\frac{k}{M}$, which is a multiple of $\frac{1}{M}$ and hence have a cycle beginning after M samples.

By cancelling each time a different zero (or conjugate pair) a bank of filters may be constructed, which may be adjacent if even-numbered zeros are cancelled, or partly overlapping if all zeros are used. The impulse response of these filters is equivalent to a sinusoid of frequency $\frac{k}{M}$, multiplied by a rectangular window of M samples. The comb filter was also employed by Principe and Smith[156], although these authors introduced additional zeros, instead of poles to achieve a pass-band.

5.6.3 A filter bank for implementing the DFT

The introduction of zeros in the basic resonator to modify its phase characteristics permitted the construction of a pair of systems having short-time sinusoidal and cosinusoidal responses respectively[121][55, p. 645]:

$$\begin{aligned} p(n) &= w(n)\cos(\tfrac{2\pi}{M}k) \\ q(n) &= w(n)\sin(\tfrac{2\pi}{M}k) \end{aligned} \tag{5.16}$$

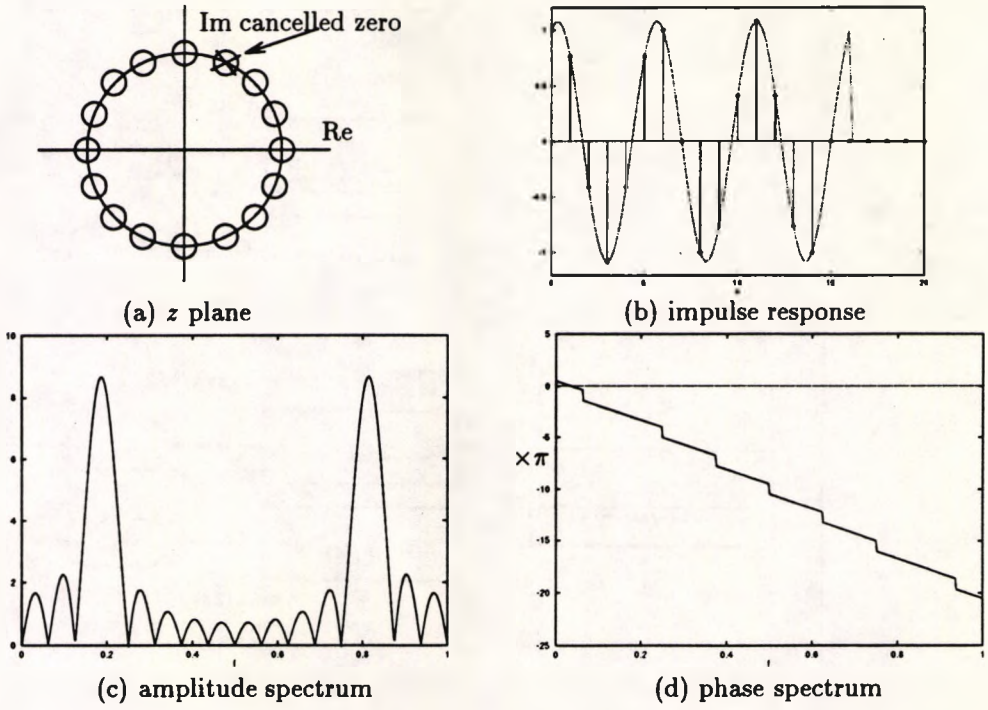


Figure 5.6: The basic element of the filter bank for continuous DFT

Filter type	Transfer function	
low-pass	$G_{0_{re}}(z) = \frac{1}{1-z^{-1}}$	$(G_{0_{im}}(z) = 0)$
band-pass	$G_{k_{re}}(z) = \frac{1-\cos(\frac{2\pi}{M}k)z^{-1}}{1-2\cos(\frac{2\pi}{M}k)z^{-1}+z^{-2}}$	(real output)
	$G_{k_{im}}(z) = (1-z^{-M})\frac{\sin(\frac{2\pi}{M}k)z^{-1}}{1-2\cos(\frac{2\pi}{M}k)z^{-1}+z^{-2}}$	(imaginary output)
high-pass	$G_{\frac{M}{2}_{re}}(z) = \frac{1}{1+z^{-1}}$	$(G_{\frac{M}{2}_{im}}(z) = 0)$

Table 5.1: Transfer functions of the filter bank for continuous DFT

where $w(n)$ is a rectangular window of magnitude 1 and duration from $n = 0$ to $n = M - 1$. The transfer functions for the filter bank (excluding the comb filter $1 - z^{-M}$) are tabulated in Table 5.1.

After every input sample is processed, the outputs of these filters, represent the corresponding components of the DFT of the input signal, based on a window of M past samples. The amplitude and phase response of the spectrum may be evaluated, as with the FFT by transforming the complex outputs of the filters, $y_k(n) = y_{k_{re}}(n) - jy_{k_{im}}(n)$, into polar form, $y_k(n) = r_k(n)e^{j\theta_k(n)}$, where $r_k(n) = \sqrt{y_{k_{re}}(n)^2 + y_{k_{im}}(n)^2}$ and $\theta_k(n) = \arctan\left[-\frac{y_{k_{im}}(n)}{y_{k_{re}}(n)}\right]$ (see Figure 5.7).

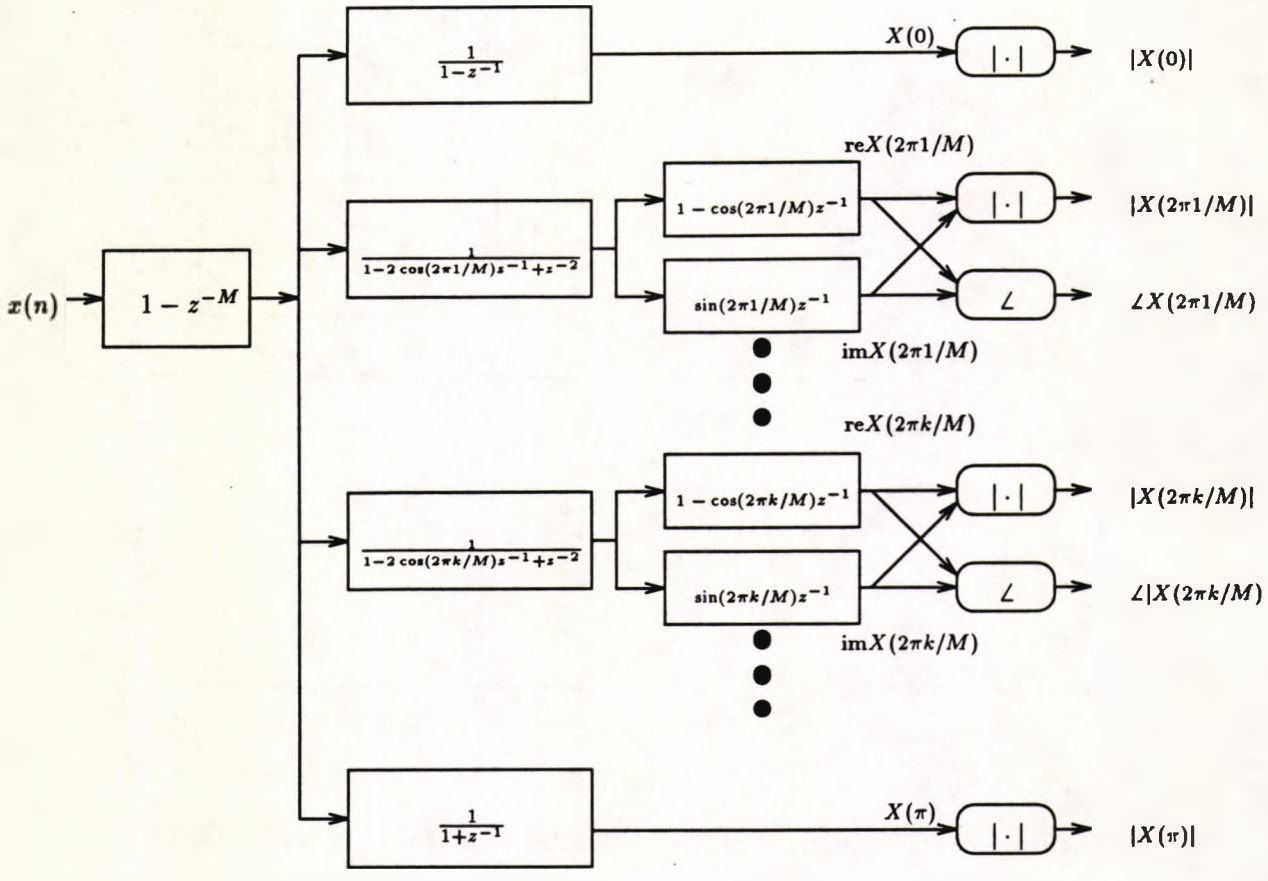


Figure 5.7: Diagram of the DFT filter bank with amplitude and phase outputs

Compared to the FFT algorithm the filter method presented here has advantages, when applied to EEG signals. First of all, the signal is treated on a sample-by-sample basis, equivalent to a sliding window by one sample every time. As a result, the processing is constant from one sample to the next and there is no need for interrupting the procedure when new samples arrive. This feature is naturally applicable to on-line processing. Moreover, since the transform is performed on every sample, patterns may be detected at the beginning of a window, without the need for phase normalization to compensate for its appearance elsewhere in the window. Then, the algorithm may use an arbitrary window size, M, whereas in the FFT this must be a power of 2.

Finally, the method is more efficient than the FFT for continuous sample-by-sample transformations, and can be made even faster since specific filters may be combined or eliminated if they prove unnecessary. The numerical efficiency of the algorithm is easily evaluated. The low-pass, high-pass and the comb filters require no multiplications. The band-pass filters require four real multiplications each, that is a total number of $4(\frac{M}{2} - 1) = 2M - 4$. An equivalent FFT requires $\frac{M}{2}\log_2 M$ complex multiplications (each representing 3 or 4 real operations, depending on the implementation), which is more expensive, even for

low values of M.

A disadvantage of the method, compared to the FFT, lies in the concept of zero cancellation by coinciding poles. Using low precision for the representation of the coefficients of the resonators may cause imperfections in the location of the poles, leading to a pole-zero function with IIR, instead of FIR characteristics, thus causing error in the computation of the spectra. The deterioration of the method is, however, gradual and can be minimized with the use of double precision arithmetic.

5.6.4 A modified filter bank

The proposed technique, however, is not without limitations. The band-pass filters have significant side-lobes with amplitudes at approximately 21% of the main lobe[114]. Moreover, the phase response of these filters is not linear, as Lynn suggested, but piecewise-linear, suffering discontinuities at the location of the zeros, as correctly pointed out by Kohn[106]. The phase discontinuities are introduced by the change in sign of the frequency response of the filter, representing a 'jump' of magnitude π.

Squaring the corresponding transfer functions as recommended by Kohn[106], eliminated this source of phase distortion and reduced the side-lobes to a maximum of 4.5%[114]. This operation is equivalent to cascading two comb filters and two resonators of every type. The impulse response of the new filters is the convolution of the response of each filter in the original bank by itself. For band-pass filters this is equivalent to a sinusoidal signal multiplied by a triangular (Bartlet) window of length $2M$. This cascade, however, affected the phase characteristics of the elements of the filter bank, since $G^2_{k_{im}}(z)$ is no longer in quadrature with $G^2_{k_{re}}(z)$. This is easily established by noting that as a consequence of squaring, the frequency characteristics of the new transfer functions has squared amplitude, compared to the original transfer function and twice the phase. Hence, since the phase of $G_{k_{re}}(z)$ at the 'resonant frequency' is zero, that of $G^2_{k_{re}}(z)$ is also zero. But for $G_{k_{im}}(z)$ this is $\frac{\pi}{2}$ and hence for $G^2_{k_{im}}(z)$ it is π. In conclusion, $G^2_{k_{im}}(z)$ is in antiphase with $G^2_{k_{re}}(z)$, instead of being in quadrature and their responses are not those of a sine-cosine any more. There are a number of ways to maintain the 'imaginary' part of the filter in quadrature with the 'real' part, without changing other important characteristics of the filter. One suggestion is to modify the phase response of the two cascaded resonators, constituting the imaginary output transfer function, to be at $\frac{\pi}{4}$, instead of $\frac{\pi}{2}$, compared with the resonators for the real output. Using the trigonometric identity $\cos(\frac{2\pi}{M}kn - \frac{\pi}{4}) = \cos(\frac{2\pi}{M}kn)\cos(\frac{\pi}{4}) + \sin(\frac{2\pi}{M}kn)\sin(\frac{\pi}{4})$

and transforming to the z domain gives the following transfer function:

$$G_{k_{im}}(z) = (1 - z^{-M}) \frac{1 - [\cos(\frac{2\pi}{M}k)\sin(\frac{2\pi}{M}k)]z^{-1}}{1 - 2\cos(\frac{2\pi}{M}k)z^{-1} + z^{-2}} \tag{5.17}$$

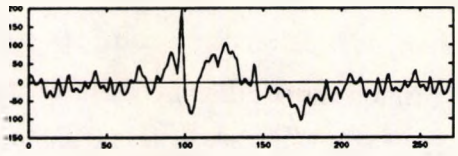
When two of these are cascaded, the overall system will have a phase difference of $-\frac{\pi}{2}$ at its resonant frequency and a 'sinusoidal' impulse response. Alternatively, one may operate on the cascaded structure. If the first resonator section for both the real and the imaginary parts is made the same, having the form of $G_{k_{re}}(z)$ (see Table 5.1) their output will be in phase. By introducing $G_{k_{re}}(z)$ as the second section in the real part and $G_{k_{im}}(z)$ in the imaginary, their final outputs will be in quadrature, as required. Because the first sections are only required for the annihilation of the double zeros of the cascade of the comb filters, they may be made the same for both the real and imaginary branches of the system. Hence, they may be replaced by a single system operating on the output of the comb filter to reduce the complexity and computational requirements of the system. This may have the form of, say, $G_{k_{im}}(z)$ or may even be an all-pole section with transfer function $G_k(z) = \frac{sin(\frac{2\pi}{M}k)}{1-2\cos(\frac{2\pi}{M}k)z^{-1}+z^{-2}}$. In these cases, the filter responses will not have the correct absolute phase (0 and $\frac{\pi}{2}$), but they will still be in quadrature. The use of $G_k(z)$ simplifies the structure of the filter and may be preferable. In this form of the system, the outputs of the band-pass filters have an amplitude that is half of that for the low-pass and the high-pass filters. A normalizing factor has been introduced to bring all outputs to the same level.

Although the new filter bank required approximately twice as many operations to implement, it is still more efficient compared to the FFT, especially when windowing is applied to the input of the latter.

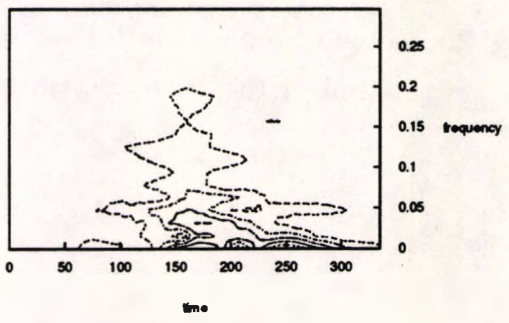
5.6.5 Results and discussion

The filter bank was applied to EEG containing instances of the epileptic precursor. An example of the evolution of the spectrum of an EEG segment containing a spike-and-wave complex, taken with the modified bank of filters described in section 5.6.4 is shown in Figure 5.8. The sampling rate was 200 Hz. The comb filter window length was $M = 64$ (Figure 5.8(b)) and $M = 200$ (Figure 5.8(c)) corresponding to a frequency resolution of 3.13 and 1 Hz respectively. Because the amplitudes of the frequency components decrease rapidly with increasing frequency, in the above Figure only those below 60 Hz (0.3 times the sampling rate) are shown. This demonstrates one of the advantages of the method, the ability to control the frequency resolution independently from the number of 'frequency samples'.

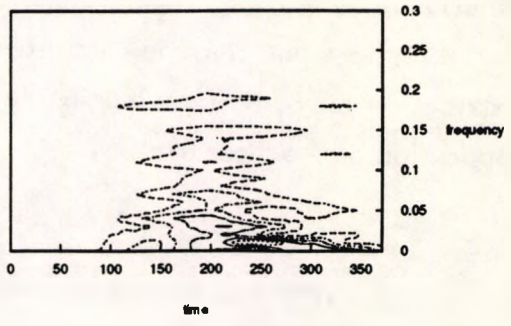
In general, the EEG spectrum is concentrated in the low frequency range. When the spike occurred (around 100 samples from the origin in Figure 5.8), there was a spread of



(a) time sequence



(b) Amplitude spectrum
small window ($M = 64$)



(c) Amplitude spectrum
large window ($M = 200$)

Figure 5.8: Examples of the evolution of the spectrum of a spike-and-wave complex

energy in the higher frequency components. The spike was more localized for $M = 64$, whereas for $M = 200$ it influenced the spectrum for a large number of consecutive samples (bad localization). The slow-wave was not as clearly indicated for $M = 64$, due to the low frequency resolution, but was more obvious for $M = 200$.

The results from frequency analysis were not particularly encouraging. The broad-band spectrum of spikes was the primary concern for the application of this technique as a pre-processing stage for subsequent detection. Because spikes cover a wide range of frequencies, an easy and reliable procedure for their detection was not suggested by the method. Moreover, their localization is dependent on the window size and it got worse as the window size increased to isolate the slow wave. These problems are not inherent to this method of computation of frequency spectra, but to the general concept of windowed time-to-frequency transformations, including the FFT. The susceptibility of the spectrum of the spike to other elements in the window containing it as well as morphological variations between spikes was considered as a potential hazard.

Despite these limitations, the efficient sample-by-sample analysis, is well-suited for the computation of the Wigner-Ville Distribution[26] (see equations (4.12) and (4.13)). This was applied for the off-line analysis of epileptic EEG signals by Stelle[184, 185] (see also section 4.6.1), who investigated its reliability as an analysis tool, rather than its efficient implementation. The latter issue was addressed by Boashash et al.[23] and Sun et al.[186],

who proposed systems based on the repetitive application of the FFT on adjacent windows of the signal multiplied by a delayed version of its complex conjugate. The recursive method can compute repetitive DFT's much more efficiently than the FFT, as already explained. It may also be used for the computation of the *analytic signal* associated with a real signal, which is usually employed to prevent aliasing in the computation of the Wigner-Ville distribution, arising from its periodicity at the Nyquist, instead of the sampling frequency[24]. The analytic signal may be computed using the method proposed in this section and the distribution may be evaluated in the frequency, instead of the time domain by a further application of the filter bank.

5.7 Comments and conclusions

In this Chapter three methods of EEG pre-processing were considered, a triangular basis of functions, piecewise polynomials and recursive frequency analysis. Although none of these formed the basis for subsequent analysis, some considerations and especially their limitations were used for the development of the techniques employed in the Chapters that follow.

The triangular basis of functions pointed out one problem of inadequacy associated with poor modelling. Triangular functions, which lack special properties, were considered because of their relation to a model of an epileptic spike that was empirical. In reality spikes are band-limited and hence their representation as a *pointed* peak is not as descriptive as initially expected.

Piecewise polynomial fitting proved that the EEG and epileptic spikes could be modelled with smooth functions, thus confirming the conclusions of the previous paragraph. Although their properties were better than triangular functions, these were limited by the inability to produce a reliable, meaningful representation suitable for analysis. The parametric method of approximation, however, led to an alternative to conventional differentiation, which has found some application in a later method (see section 6.4).

Finally, recursive frequency analysis highlighted some other difficulties linked with EEG signal analysis. The signal is non-stationary, consisting of a mixture of low frequency elements and broad-band signals, like spikes, causing problems when methods based on windowing of the data are applied. A short window to localize and enhance transient isolated spikes would cause loss of resolution for others. In this case all waves with period longer than that for the window would not be discriminated. A long window could cause the loss of the transient, whose spectral elements are averaged over the length of the window. This suggested that the treatment of brief isolated transients and of waveforms longer periods

should be performed using *separate methods*, for their effective discrimination.

Chapter 6

Spike detection and the derivatives of the EEG

6.1 Introduction

From the conclusions of Chapter 5 it is clear that the early models failed to describe EEG spikes adequately, not only due to their inherent limitations but also because of the nature of the signal. Spikes vary from one to another both in amplitude, duration and shape. The background activity has lower amplitude than epileptic transients, but it persists for longer duration and hence contaminates the time sequence before and after a spike. It is evident that a suitable transformation to separate spikes from background activity is still required. Frequency analysis proved inadequate, because the localization of spikes in the time-domain was depleted and the dependency on the size of the window was an important limitation.

Several methods were considered, including matched filtering, a combination of short and long-term frequency transforms, the use of moments of the Power Spectral Density and Hjörth analysis (see section 4.6.4). One general limitation of all these methods is their inability to describe (model) spikes with a consistent set of parameters which could be manipulated for their detection. It was realized that perhaps a very accurate spike model was probably not feasible, but perhaps unnecessary for the development of a successful method. Several simple mathematical descriptions were therefore constructed, based on general observations on real spikes and information from thir formal definition (section 2.5.2. Some of these are listed later in this Chapter. Their construction was geared towards the development of methods that utilized the abrupt amplitude transitions that often characterize spikes.

These characteristics were considered in the development of the first successful spike detection methods, which were based on the derivatives of the signal with respect to time. The

concept is not new, since a multitude of publications on related techniques exists (see section 4.6.1). The systems described in this Chapter are different not only in their construction details, but in the approach for their development also. There is an explicit definition of a signal model, on which the method was based, thus minimizing the number of experimentally deduced parameters. This approach makes a developed system more self-contained and independent from the test data. Its behaviour and response are therefore more predictable and, to some degree, easier to amend.

In this Chapter, a justification for using derivatives of the signal will be provided (section 6.2), followed by a presentation of methods for discrete-time signal 'differentiation' (section 6.3). Then two methods for spike detection, based on derivatives of the signal will be presented, the simpler (earlier) first. For each one of the two procedures simple formal models of a single wave, the basic element of signal analysis, are described briefly. This is followed by a description of the technique, with emphasis on the use of the model for its development, followed by the presentation of results and a brief discussion. The Chapter closes with some general comments on the methods covered and their limitations.

6.2 Why look at derivatives

Derivatives are useful for the detection of transients in the EEG and have been used extensively in the past for this purpose. They are also related to signal description and modelling in general. It is my opinion that the reasons behind these statements are more important than the statements themselves in the description and detection of signals.

The connection between derivatives and transients in the EEG is easily established. What defines the onset of some transient activity is a rapid change in the behaviour of the signal. Such a variation is associated with an increase or decrease in amplitude, phase reversal and frequency changes. When the transient ceases, these transitions are reversed, although they may be less abrupt than at its onset. These phenomena are associated with changes in the gradient (derivative) of the signal with respect to time. While the EEG signal progresses 'normally', behavioural variations are usually gradual as they require a long time interval to come into effect. The transition from one sleep stage to another, for instance, is not instantaneous and there is a transition period between them, where one 'gives way' to another (see sections 2.5.1 and 4.6.3). On the contrary, a spike begins rapidly and terminates equally abruptly in a few milliseconds, causing the derivative of the EEG signal to rise at the onset of the spike to an abnormal level ('normal' here refers to the values it acquired in the periods before and after the spike). Similarly, the second derivative has been a useful measure for

quantifying transients, such as spikes and sharp waves. Denoting the rate of change of the slope, it signifies the onset and offset of transients, which are related to large transitions of the derivative in opposite directions. In the case of a spike, for example, the second derivative acquires abnormally large values at its summit, where the slope changes from positive to negative in a short period of time. These characteristics have been exploited by many workers to detect spikes and sharp waves (see section 4.6.1).

The connection between the description (and modelling) of a signal and its derivatives is clearly stated in any text on linear signals and systems analysis[115, 121, 150, 151, 158]. Continuous-time signals and systems are usually described using differential equations, which utilize derivatives of the signal. The analysis of complicated signals is sometimes made easier if they are decomposed into simpler ones. This might be done by manipulating the differential equations that describe the signal to formulate it as the superposition of the outputs of one or more systems, when excited by inputs of simpler form, like steps, pulses, sinusoids and white noise.

For example a causal signal, $y(t)$, may be represented as the output of a linear time-invariant system with transfer function $H(s) = \frac{1}{1+\sum_{i=1}^{n} a_i s^i}$, excited by, say, white noise with zero mean, $n(t)$. One may write $Y(s) = H(s)N(s)$, where $Y(s)$ and $N(s)$ are the Laplace transforms of $y(t)$ and $n(t)$, respectively. In the time-domain, this translates to the differential equation

$$y(t) + \sum_{j=1}^{n} a_j y^{(i)}(t) = n(t) \tag{6.1}$$

The first n derivatives of $y(t)$, $\{y^{(i)}(t); i = 1, 2, \ldots, n\}$ may be estimated numerically from the signal itself and by employing a method of 'best fit' it is possible to compute a set of $\{a_i\}$, so that when applied to the appropriate derivatives will give a sequence $n(t)$, with the expected properties. A discrete-time equivalent of this signal representation method has been the basis of a method discussed in later Chapters.

6.3 Discrete-time differentiation

The derivative of a function of a continuous variable, such as time, $f(t)$, is defined by the limit

$$f'(t) = \lim_{\delta t \to 0} \frac{f(t + \delta t) - f(t)}{\delta t} \tag{6.2}$$

This is of course defined at points where $f(t)$ is continuous, where the limit is well-defined. Discrete-time functions are obtained by 'sampling' a continuous time function at points in time, by multiplying it by a sequence of equally-spaced impulse (Dirac) functions $f(t)u_0(t -$

$nT) = f(nT)$, which result in the definition of the discrete-time equivalent

$$f_d(t) = \begin{cases} f(t) & t = nT \\ 0 & t \neq nT \end{cases} \tag{6.3}$$

This function is obviously discontinuous and its derivative, $f'_d(t)$ is not defined at the points of interest, $f(nT)$. In discrete-time, the term differentiation does not refer to the definition of equation 6.2. It is rather a process that yields a discrete-time sequence of the derivative of the underlying continuous-time function, $f'(nT) = f'(t)\, u_0(t - nT)$. Notice that $f'(nT)$ exists in contrast with $f'_d(t)$, which does not.

The problem of evaluating $f'(nT)$ may be solved in many ways. Working directly from the definition, one may differentiate the continuous-time signal $f(t)$ and sample its derivative. This was the approach used in some early systems[36, 192]. The need for additional analogue hardware to implement differentiation as well as the limited precision and other limitations they impose on the process made 'differentiation' after sampling an elegant alternative, which has been used extensively by designers of digital (discrete-time) systems.

6.3.1 Discrete-time approximations of differentiation

Given the data sequence arising from sampling of a signal, the question is how to compute sequences corresponding to the derivatives of the underlying analogue signal at the sampling instants. Methods vary in terms of theory and approach. Some of these will be described shortly, with emphasis on the ones that have been used in the work presented in sections 6.4 and 6.5 that follow.

Approximations based on continuous-time pseudo-differentiators

Differentiation may be regarded as a linear operation, described by a transfer function. It is therefore possible to derive an equivalent discrete-time transfer function by means of a transformation from the continuous s-plane of the Laplace transform to the discrete z-plane.

The *Impulse invariant transform* ensures that the unit sample (pulse) response of the resulting transfer function is identical to the sampled impulse response of the continuous time system. It can be shown[151, p. 604] that the required transformation is $z = \mathrm{e}^{sT}$, where T is the sampling period.

This transform was used by Smith[181] to convert an 'analogue pseudo-differentiator' with transfer function $H(s) = \frac{s}{s+b}$ into a discrete-time equivalent. The impulse response of this function, $h(t) = u_0(t) - b\mathrm{e}^{-bt}\, u_1(t)$ (where $u_1(t)$ is the unit step function) was sampled and z-transformed to produce the required discrete-time transfer function.

Impulse-invariant transformations are usually avoided, since they may result in frequency aliasing, if the impulse response contains frequency components above $\frac{\pi}{T}$, as these are replicated every $\frac{2\pi}{T}$, due to the folding of the s domain so that strips of width $j\frac{2\pi}{T}$ map into the entire z-domain[151, p. 603–608]. These problems may be avoided by using a *bilinear transformation*, which converts directly from the s-plane to the z-plane by using a substitution of the following form:

$$s = \frac{2}{T}\frac{z-1}{z+1} \tag{6.4}$$

This maps stable continuous to stable discrete-time systems, the entire imaginary axis $s = j\omega$ onto the unit circle $|z| = 1$ and has the additional property that it compresses the entire s-domain into the strip $-j\frac{\pi}{T} < s < j\frac{\pi}{T}$ prior to mapping into the z-plane, thus preventing aliasing. This method was used by Ktonas et al.[108] to derive the discrete-time equivalent of an analogue pseudo-differentiator.

Simple discrete-time approximations

Analogue differentiators, however, are approximations to the actual process. There is little, if any, justification in using an approximation, imposed by physical limitations of analogue hardware. Restricting the more flexible digital system, where the discrete-time function is to be implemented, is pointless. The required approximation may be formulated directly in discrete-time, where simplicity and speed may be traded off for precision.

One of the early 'all-discrete' differentiators is the *successive difference*, which has its origins in the numerical solution of differential equations by approximating them by equivalent difference equations[151, p. 598–601]:

$$\hat{f}'(t) = \frac{f(nT) - f(nT - T)}{T} \tag{6.5}$$

Differentiation may also be represented by multiplying the Laplace transform of a signal by s. Since $z = \mathrm{e}^{sT}$ the required operation is

$$s = \frac{1}{T}\mathrm{Log}(z) \tag{6.6}$$

where the restriction $-\pi \leq \arg(z) < \pi$ of the domain of the mapping was introduced to ensure the uniqueness of the not well-behaving complex logarithm. By truncating the Taylor expansion of the logarithm around $z = 1$, $\mathrm{Log}(z) = \mathrm{Log}(1 + u) = u - \frac{1}{2}u^2 + \frac{1}{3}u^3 - \ldots$ to its first term and replacing $u = z - 1$, differentiation is approximated by the relation $s = \frac{1}{T}(z - 1)$, or the causal system $H(z) = \frac{1}{T}(1 - z^{-1})$, which represents the first difference equation (6.5). This procedure was used by Comley and Brignell[45].

The same approximation may be obtained alternatively from the equation (6.2), of definition of the derivative, if the limit is dropped and the infinitesimal step δt is replaced by a finite-size step, say Δt:

$$\hat{f}'(t) = \frac{f(t+\Delta t) - f(t)}{\Delta t} \tag{6.7}$$

Since $f(t)$ is known at specific points spaced T apart, a natural choice for the step is $\Delta t = T$. In practice, the backward difference is used, since it leads to a causal system and has better properties than the forward one. Hence the derivative may be approximated by equation (6.5). This approximation has been used for the detection of epileptic spikes by Ktonas and Smith[107]. Other authors suggested that the difference equation $\frac{1}{T}[f(nT) - f((n-m)T)]$ gave better results, as it suppressed certain frequencies (see Quian et al.[157]). Alternatively, the derivative may be estimated from the difference of the previous and the next sample. This, however, belongs to a broader category of differentiators derived by more formal criteria, whose description follows.

A systematic discrete-time procedure

Like many Finite Impulse Response (FIR) filters, digital differentiators may be designed using *frequency specifications* of the required transfer function and time-limiting *windows* to truncate the length of the filter, which is often large, or even infinite. This has become one of the standard procedures in signal processing. With the closely related *frequency sampling technique*, they are the dominant methods for the design of such filters[121, p. 321–323][151, p. 544–596]. Briefly, the design procedure is as follows:

1. Start from the Fourier Transform representation of the transfer function (filter), $H(\omega)$

2. Evaluate the unit sample response,

$$h(n) = \frac{1}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} H(\omega) e^{j\omega nT} d\omega \tag{6.8}$$

   (notice that T is the sampling period and $\frac{2\pi}{T}$ is the sampling frequency).

3. Truncate the resulting sequence $\{h(n);\ n \in (-\infty, +\infty)\}$ to make it practically realizable, by multiplication with a time-limited window function,

$$w_N(n) = \begin{cases} a_n & \text{for} \quad |n| \leq \frac{N}{2} \\ 0 & \text{otherwise} \end{cases} \tag{6.9}$$

   Window functions are usually symmetrical about the origin and have the property of smoothing the frequency response of the filter compared to the ideal, $H(\omega)$. Their use,

however is often recommended, if the truncation of a sequence (equivalent to multiplication by the Rectangular window, if $a_n = 1$ irrespective of n) causes unacceptable fluctuations and side-lobes in the frequency domain.

A differentiator is represented in the Laplace domain $H(s) = s$ and in the Fourier domain by $H(\omega) = j\omega$. Following the design procedure just described, the impulse response of this transfer function is given by

$$h(n) = \frac{cos\ \pi\ n}{nT} = \frac{(-1)^n}{nT} \tag{6.10}$$

This method was employed by Stelle et al.[183] to determine a differential operator, although for the application it was cascaded to a Hilbert transformer ($\frac{\pi}{2}$ phase shifter). Notice that the sequence is antisymmetric and has zero value for $n = 0$. If a window of length 3 is applied to $h(n)$, the truncated impulse response will have the form of the forward-backward difference operator, $\frac{y(n+1)-y(n-1)}{T}$ which was the approximation proposed at the end of the previous paragraph. Its unit sample (impulse) response is shown in 6.1(a) and its frequency response (purely imaginary) in Figure 6.1 (b). More terms may be taken using a window with longer duration[151, p. 544–596] (Rectangular,Hamming, Hanning, Tukey etc.). Differentiators of length 9 and 31 truncated using the Rectangular window and their corresponding Fourier transformations are also shown in Figure 6.1 (a)(b).

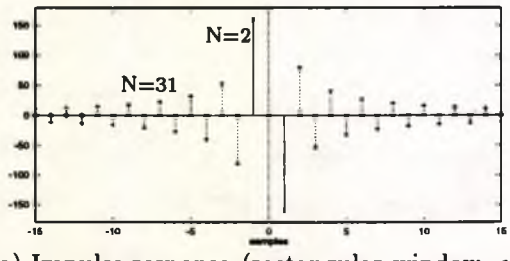
The same differentiators but with the application of a Hanning window instead, are shown in Figure 6.1(c) and (d). The frequency response of these is much smoother and at low frequencies it approximates the ideal differentiator quite well.

In the case of the EEG, a modified differentiator was also tested. This has the property of 'differentiating' up to a particular cut-off frequency, $\omega_c = \alpha\frac{\pi}{T}$, where $0 < \alpha \leq 1$, above which it produces zero output. The impulse response corresponding to this frequency behaviour may be determined in the same way as before:

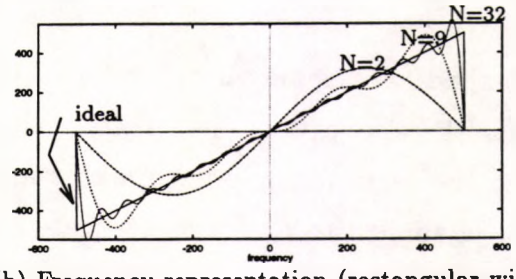
$$h(n) = \frac{\alpha\ \cos\ \alpha\pi\ n}{nT} + \frac{\sin\ \alpha\pi\ n}{\pi n^2 T} \tag{6.11}$$

This sequence, truncated by a Hanning window of length 31, and the corresponding spectrum are shown in Figure 6.1(d) and (e), respectively, for a cut-off frequency of 0.625π. The significance of this value will become apparent later in this Chapter.

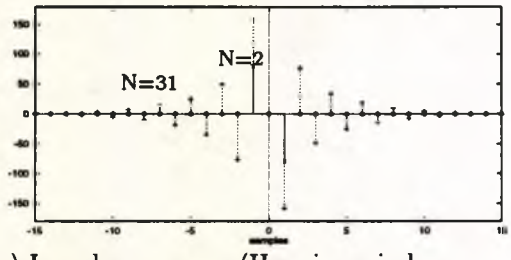
The cut-off frequency parameter, which is available in the differentiator described last may be considered advantageous in the presence of muscle artifacts. These contaminate the high-end of the frequency spectrum of the EEG and would be enhanced by a full-band differentiator. By selecting the cut-off frequency carefully it might be possible to prevent them from producing a high-level derivative in the absence of genuine spikes.



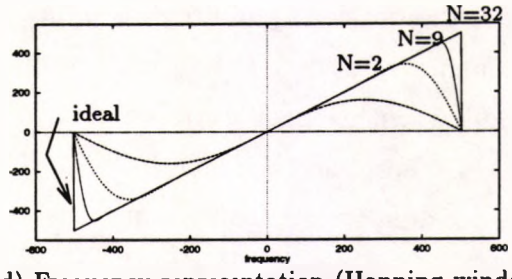
(a) Impulse response (rectangular window, $\alpha = 1$)



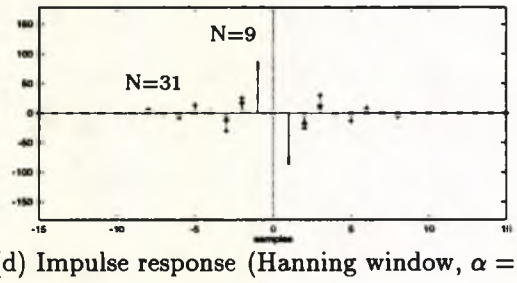
(b) Frequency representation (rectangular window, $\alpha = 1$)



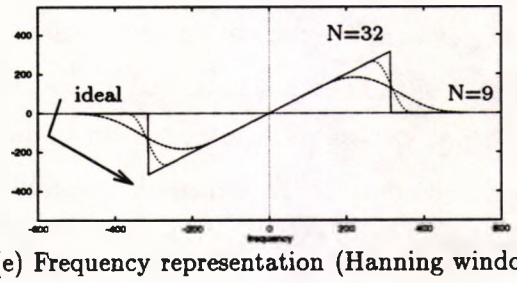
(c) Impulse response (Hanning window, $\alpha = 1$)



(d) Frequency representation (Hanning window, $\alpha = 1$)



(d) Impulse response (Hanning window, $\alpha = 0.67$)



(e) Frequency representation (Hanning window, $\alpha = 0.67$)

Figure 6.1: Discrete-time differentiators, unit sample sequences and frequency responses (in all cases $T = 6.25ms$)

It should be noted that the above procedure is equivalent to the implementation of $\mathrm{Log}(e^{jwT})$ as a Fourier series, and permits the evaluation of $\mathrm{Log}(z)$ around the unit circle in z with an arbitrary precision (depending on how many terms of equation (6.10) are selected). In fact the series can evaluate $\mathrm{Log}(z)$ in an annular region extending from zero to infinity, $0 < |z| < \infty$ and is also called the *Laurent series* of this function[85, pp. 223–226], which is valid in the above region. This is far more general than the Taylor series (used in the previous section) which only applies around the point of definition, $z = 1$ (low frequencies).

This design procedure is systematic and was formulated in a small program. This takes as input parameters the sampling frequency, f_s, the ratio of the cut-off to the sampling frequency, α, the length of the impulse response of the differentiator after truncation, N and the type of window (Rectangular or Hanning). Its output is a file containing the coefficients of the resulting differentiator, its length and other definitions in a form that may be directly incorporated as a module in a program written in the C language, making the rest of the system independent from the length and type of the differentiator. This permitted the evaluation of the behaviour of a number of differentiators with different construction parameters without modifications to the core program that used them.

A method based on piecewise polynomial approximations

Finally, derivatives may be estimated using curve fitting on the EEG data points, $\{f(nT)\}$. The *piecewise polynomial approximation* algorithm, described in section 5.5 provided a method for obtaining derivatives at the points of interest. As explained then, the curve fitted using 3rd or 5th order polynomial segments, $\hat{f}(t)$ is smooth. Its derivative is continuous and may be computed using the normal rules of differentiation for polynomials:

$$\begin{aligned} \hat{f}(t) &= \sum_{n=0}^{N} a_n t^n \\ f'(t) \approx \hat{f}'(t) &= \sum_{n=1}^{N} n\, a_n t^{n-1} \end{aligned} \tag{6.12}$$

where N is the order of the polynomial. For fifth order polynomial segments, both the first and the second derivatives are continuous.

The spike detection algorithms that follow were based on one of these methods to evaluate the derivatives of the EEG.

6.4 A basic spike detector using derivatives

6.4.1 Analysis

The apparent success of spike detectors based on the derivatives of the EEG (section 4.6.1) was rather influential at the early stages of this work. Although most of the reviewed methods had their basis on the first derivative of the EEG signal, the second derivative was also used by some workers. Each derivative contains different attributes of the spike. The first measures slope, hence it monitors the large changes in amplitude in a relatively short time period, that characterize the spike edges, compared to those of the 'normal' EEG activity, which are more gentle. The second derivative is a measure of the curvature at the summit of the spike, where it acquires a large negative value (indicating the rapid change of slope from large positive to large negative).

Although closely related, the information of the first derivative is different from that of the second. In fact, workers into the field of EEG modelling, like Ktonas and Smith[107], Ktonas et al.[108] and Glover et al.[71] had suggested that the description of spikes and other sharp transients ought to be based on more than one parameters, suggesting various amplitude, slope, sharpness and duration measures. These suggestions motivated the combination of the first and second derivative into a single measure of 'spikiness'. Instead of using the triangular approximation for a spike, to reflect its sharp transition from a large positive to a large negative slope, a smooth waveform was preferred. This is in fact more realistic, since band-limiting the signal to remove high-frequency interference tends to decrease the sharpness at the summit of spikes and other sharp transients. It was, therefore assumed that spikes are single sinusoidal oscillations with period considerably smaller and amplitudes larger than those of the rest of the signal. Spikes are superimposed on the ongoing EEG signal, so that their onset is gradual and causes no sharp transitions at their starting and ending points. For analysis purposes, a spike was approximated by a sinusoid, of frequency ω_0, equal to the fundamental frequency of a typical spike, time-limited by a cosine (Hanning) window of duration equal to the period of the transient, $t_0 = \frac{2\pi}{\omega_0}$. Without loss of generality, the waveform was considered centred at the origin. This function and its first two derivatives are given by

$$y = -\frac{A}{2}(1 + \cos\omega_0 t)\sin\omega_0 t \tag{6.13}$$

$$y' = -\frac{A\omega_0}{2}(\cos\omega_0 t + \cos 2\omega_0 t) \tag{6.14}$$

$$y'' = \frac{A\omega_0^2}{2}(\sin\omega_0 t + 2\sin 2\omega_0 t) \tag{6.15}$$

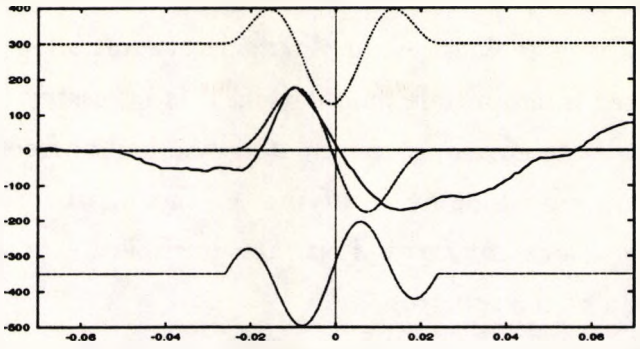


Figure 6.2: A simple spike model and its first two derivatives

which applies in the range $-\frac{t_0}{2} \leq t \leq \frac{t_0}{2}$. All three functions are zero everywhere outside this range. It is easy to show that they are also continuous, since $y = 0$, $y' = 0$ and $y'' = 0$ at the ends of the range $t = \pm\frac{t_0}{2}$. This is clearly demonstrated in Figure 6.2. An actual spike was superimposed on the model. Its derivative is shown on the top, scaled by a factor of $5T$ and its second derivative, shown underneath, is scaled by a factor of $25T^2$. Evidently, this simple model does not reflect accurately real spikes, but it contains two of its fundamental attributes, amplitude and oscillatory behaviour.

One would then look for a way to combine the derivatives into a single indication for the presence or absence of a spike. The following was assumed true

Proposition 6.1 *If both the amplitude, A and the frequency ω_0 of a spike are larger than the equivalent parameters for "normal" waves in the EEG, these are enhanced by the first and second derivatives of the EEG. The product of these derivatives is more indicative of the presence of the spike than either one on its own.*

This is easily verified, by forming the product in question

$$p = y' \cdot y'' = -\frac{A^2\omega_0^3}{8}(\sin\omega_0 t + \sin 2\omega_0 t + 3\sin 3\omega_0 t + 2\sin 4\omega_0 t) \tag{6.16}$$

This clearly shows that for a signal of given amplitude, p is enhanced by a factor of ω_0^3, compared to ω_0 and ω_0^2 for y' and y'', respectively. Similarly for a signal of fixed frequency, ω_0, the amplitude enhancement is proportional to A^2, rather than A, as in either of the derivatives. In other words, signals of large amplitude and frequency are enhanced more by p than by either y' or y'' alone. This property of p to enhance spikes and suppress normal EEG activity is clearly visible in Figure 6.3(c), where the difference between the large burst of mixed frequencies in p for the duration of the spike and the rest of the signal is quite profound.

There is a small limitation in practice, the absence of a constant term in p that makes the detection of its amplitude tedious. A simple practical solution, borrowed from communications, proved adequate. In amplitude modulation, it is necessary to detect the amplitude of a slowly varying signal (the envelope), modulated by a higher frequency one (the carrier). The separation may simply be done by rectifying the signal, which introduces a harmonics of the carrier, including a constant term. It is then possible to remove these harmonics by filtering-out the high end of the spectrum.

In the case of p, no precise reconstruction of the envelope is necessary. Hence, instead of rectification, squaring of p was used. This is somehow slower than 'rectification' (implemented as the absolute value on a computer), but it is simpler to analyze mathematically and provides further amplitude enhancement when a spike is present. Using basic trigonometric identities and after algebraic manipulation, it can be shown that

$$\begin{aligned} p^2 &= \frac{A^4\omega_0^6}{128}(15 + 19\cos\omega_0 t + 9\cos 2\omega_0 t 2\cos 3\omega_0 t - 12\cos 4\omega_0 t \\ &\quad -4\cos 5\omega_0 t - 13\cos 6\omega_0 t - 12\cos 7\omega_0 t - 4\cos 8\omega_0 t) \end{aligned} \tag{6.17}$$

This function is rich in higher-order harmonics, but contains a constant term, which may be isolated fairly easily using a low-pass filter with a cut-off frequency below ω_0 to remove any oscillations. The details of the implementation of a spike detector based on this idea follow.

6.4.2 Implementation

The formulation of this method for spike detection, given in the previous section points to a fixed procedure to follow. This is also shown as a block diagram in Figure 6.4:

1. Differentiation of the signal, twice
2. Multiplication and squaring of the derivatives
3. Filtering of higher frequencies
4. Thresholding of the filtered signal

Differentiation is the procedure that is more flexible in its implementation. Several methods have been presented in section 6.3.1 on how to 'differentiate' a discrete-time signal. Step 2 is rather straightforward to implement, whereas filtering requires some more consideration. Adjusting the threshold on the filtered signal is a simple operation, but it has serious implications on the accuracy, or even the validity of the method, which will be pointed out later.

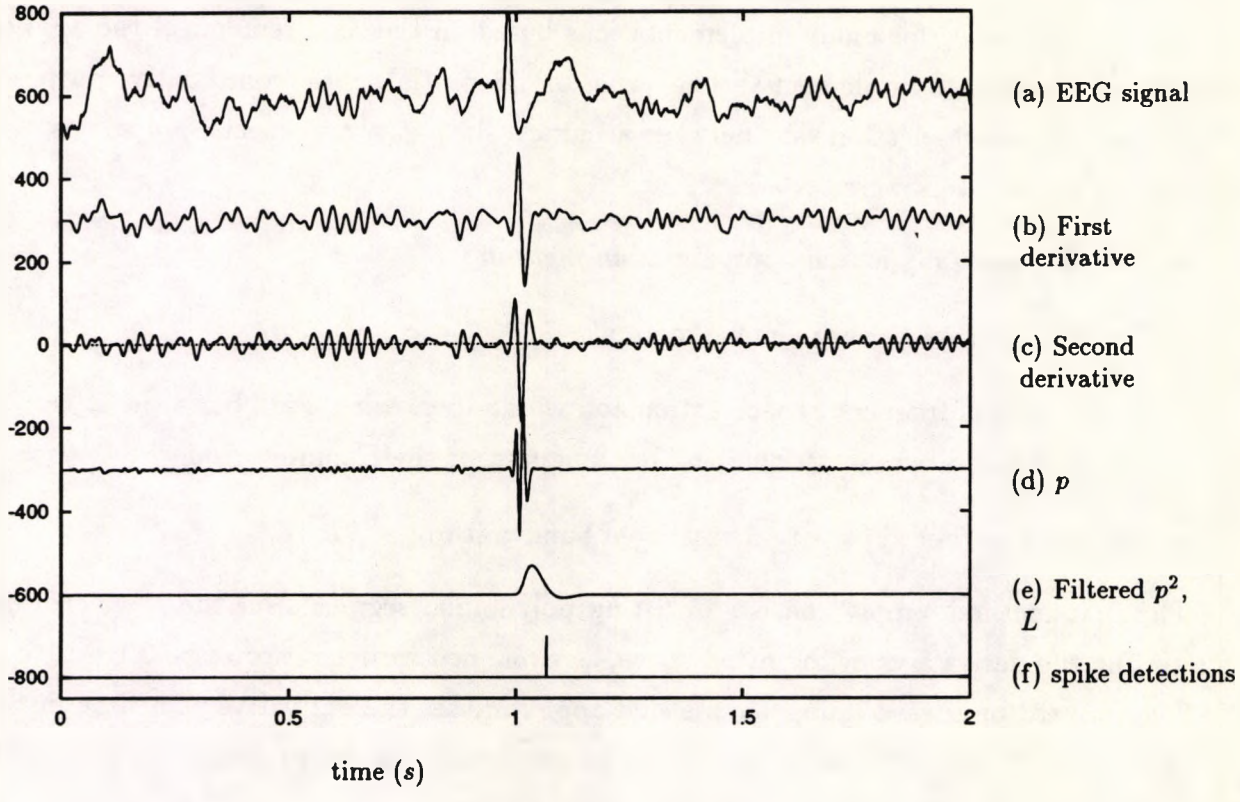


Figure 6.3: Typical waveforms of the multiplicative spike detector

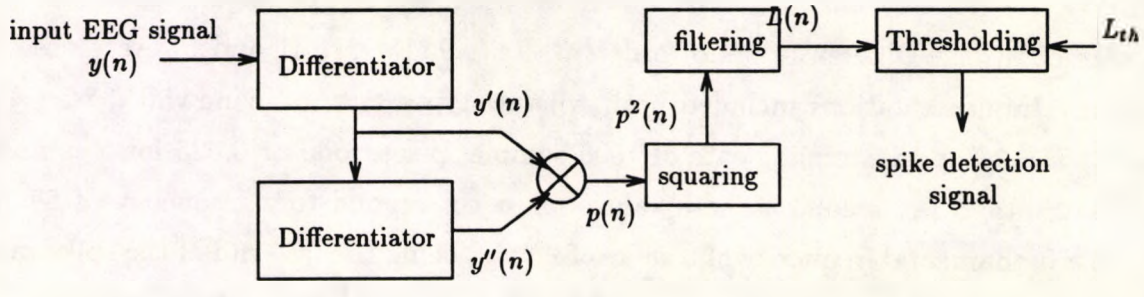


Figure 6.4: Block diagram of the basic spike detector

Differentiation

A number of the methods for discrete-time 'differentiation' proposed in section 6.3.1, have parameters that allow for many implementations based on the requirements of the application. It was decided to implement various types of differentiators for comparative purposes. The criteria for their selection vary between accuracy, simplicity and special properties. The following differentiators were tested:

- The piecewise polynomial approximation method
- The forward-backward difference method
- The method of frequency specification and windowing, using a full bandwidth ($\alpha = 1$), with different orders and either the Rectangular, or the Hanning window
- The same method, but with a restricted bandwidth ($\alpha < 1$).

The first method simply consists of fitting polynomial segments on the waveform and computing the derivatives of the fitted curve, as explained earlier in section 6.3.1.

The forward-backward difference method approximates the derivative at a time instant by taking the difference between the sample following and the one preceding, dividing it by the time interval between them:

$$y'(nT) = \frac{y[(n+1)T] - y[(n-1)T]}{2T} \tag{6.18}$$

This may be regarded as a specific case of the method for frequency specification with a Hanning window of length 3.

The frequency specification method with bandwidth of application has been extensively covered in section 6.3.1. This category has three parameters, the length of the truncated sequence, N, the cut-off point, α, and the window used, either Rectangular or Hanning in this implementation. Differentiators of orders 3, 5, 7, 9, 15, 21, 31 and 51 were used for comparison. Implementations included both window types and full-bandwidth, $\alpha = 1$ as well as for $\alpha = 0.1$ for a sampling rate of 1000 samples per second or 0.625 for a sampling rate of 160 samples per second. These values for α correspond to a frequency of 50 Hz, which is the fundamental frequency of a spike of 20 ms, being the fast-end of the spike range and providing the minimum bandwidth for which no spike will be suppressed.

Examples of these differentiators and their corresponding frequency responses were shown before in Figure 6.1.

In the actual realization, because the differentiation operators are non-causal, it was necessary to simulate 'future' samples before evaluating the gradient. Three data buffers

were employed, to store the signal, y, its derivative, y', its second derivative, y''. The differentiator was first applied to y, These were implemented as cyclic buffers with two pointers, one for the next available (future) sample and one for the sample on which the rest of the operations are performed. Hence the detection of spikes is indicated with $\frac{M}{2}$ samples delay, where M is the length of the buffer.

Spike detector formulation, squaring and filtering

The function, p was simply the product of the first and the second derivative at every sample, $p(nT) = y'(nT) \cdot y''(nT)$ and its square, p^2, to introduce a constant term a trivial operation. Removing the non-constant terms needed some more consideration. From equation (6.17) it is evident that for a spike, oscillations in p^2 are caused by frequency components at the fundamental frequency ω_0, and its harmonics. In fact, these are not discrete, but appear as peaks of a smoother function of frequency, owing to the windowing operation for time-limiting the sequence. Restricting the duration of a signal may be treated as the multiplication of the time sequence by a Rectangular window. Due to the duality properties of the time and frequency domains, this corresponds to convolution by the function $\frac{\sin \omega \frac{t_0}{2}}{\frac{\omega}{2}}$ in the frequency domain, which results in smoothing of the spectrum. Note that $\omega_0 = \frac{2\pi}{t_0}$, which converts the above equation into $\frac{\sin \pi \frac{\omega}{\omega_0}}{\frac{\omega}{2}}$. In discrete-time, assuming that $t_0 = (N_0 + 1)T$, this corresponds to

$$H(\omega) = \frac{\sin \pi \frac{\omega}{\omega_0}}{\sin \frac{\omega}{2}} \tag{6.19}$$

which is a periodic function. Since both functions have most of their energy in the same frequency band, $|\omega| < \omega_0$ in the range of frequencies $|\omega| < \frac{\pi}{T}$, it is unimportant whether the filter is developed in discrete-time directly, or in continuous-time and then converted. The selection of the direct approach here is somehow arbitrary.

The interest is drawn to the low-frequency elements of the spectrum, produced when convolving the function of equation (6.19) with that for p^2, derived from a discrete-time equivalent of equation (6.17). There is a large amplitude in the band $0 \leq \omega < \omega_0$, due to the constant term and another one, $0 < \omega < 2\omega_0$, due to the first harmonic. These bands overlap. Hence the cut-off frequency of the low-pass filter was chosen as their mid-point, $\omega_c = \frac{\omega_0}{2}$, which suppresses as much oscillations as possible while maintaining most of the useful low-frequency components needed for the detection.

Formally, spikes have durations between 20 and 70 ms, which correspond to fundamental frequencies, $\omega_0 \in [89.76, 314.16] c/s$, giving cut-off frequencies in the range $\omega_c \in [44.88, 157.08] c/s$. Selecting either extreme of the range has side effects. If the low-end is chosen, some useful signal power may be lost when encountering spikes with higher frequen-

cies. Selecting the upper bound may allow oscillations of significant amplitude to occur in the filtered output, causing multiple detections. The lower bound was chosen instead, as it would result in excessive smoothing of the raw detector output for some spikes, which is not too critical. It is possible to 'tune' ω_c for the spikes of each patient, assuming that their duration is reasonably constant but in practice the system functioned satisfactorily even for cut-off frequencies outside this range , indicating that despite these considerations, ω_c is not a very critical parameter. This, however, is not conclusive since the detector was tested on a relatively small set of EEG records.

Several low-order filters with a cut-off frequency $\omega_c = 44.88c/s$ were tested. Because there is significant overlap between the spectrum of the signal to detect and the one to suppress, the Chebyshev filter[150, p. 568–576] structure was chosen, which has fast transition characteristics. Ripples in the pass-band are acceptable, since no precise gain characteristics are required. A ripple factor of 0.5 dB was chosen, corresponding to the filter parameter $\epsilon = 0.3493$. The transfer functions for a first- and second-order filter for a sampling rate of 1000 s^{-1}:

$$H_1(z) = \frac{0.06038(1 + z^{-1})}{1 - 0.87925z^{-1}} \tag{6.20}$$

and (6.21)

$$H_2(z) = \frac{7.39539 \cdot 10^{-4}(1 + 2z^{-1} + z^{-2})}{1 - 1.93508z^{-1} + 0.93804z^{-2}} \tag{6.22}$$

and for the rate of 160 s^{-1}:

$$H_1(z) = \frac{0.28784(1 + z^{-1})}{1 - 0.42432z^{-1}} \tag{6.23}$$

and (6.24)

$$H_2(z) = \frac{2.45395 \cdot 10^{-2}(1 + 2z^{-1} + z^{-2})}{1 - 1.57497z^{-1} + 0.673129z^{-2}} \tag{6.25}$$

In both cases the bilinear transform $s = \frac{2(z-1)}{T(z+1)}$ was used for the conversion of the transfer functions from continuous to discrete time, according to the well-known procedure[151, p. 608–612]. Although the final results for the records tested were similar for either the first, or the second order filter, the computational complexity of the second over the first is marginal. Therefore this was the one incorporated in the end, since it provided a precautionary 'performance margin'.

Thresholding

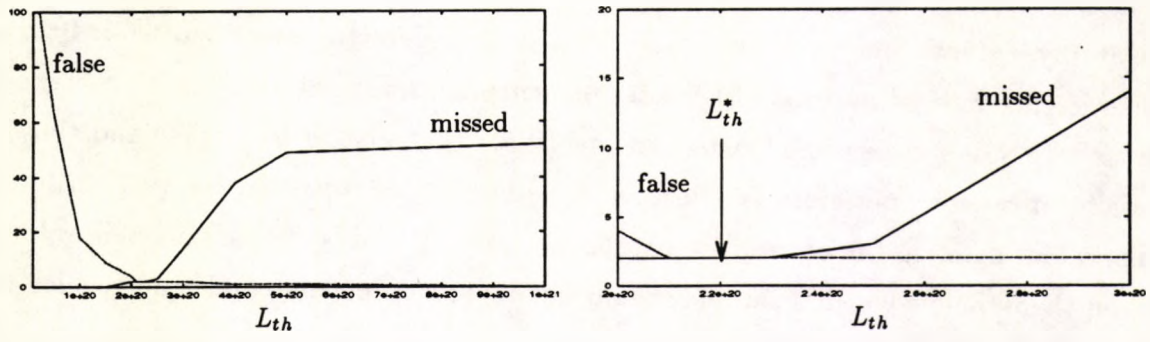
The output of the filter is smooth with a rather small amplitude, almost zero, maintained over the length of the EEG recording, but suffers transient increases during the occurrence

of spikes. What is actually required is a binary indication, of the occurrence or absence of a spike at any instant. Since the amplitude of the filtered detector output, $L(n)$, is high when a spike is encountered and low otherwise, one may assume that this function maps EEG samples into a one-dimensional space, so that those that belong to a spike and those that do not form two distinct clusters. Because $L(n)$ is a scalar function, the two clusters may be separated, defining the corresponding classes 'normal' and 'spike', by a single value, L_{th}, between them, below which a sample should belong to the 'normal' and above which to the 'spike' class.

$$\begin{aligned} L(n) \leq L_{th} \Rightarrow \text{class}(n) &= \text{'normal'} \\ > L_{th} \Rightarrow \text{class}(n) &= \text{'spike'} \end{aligned} \tag{6.26}$$

Using the constant term of equation (6.17), $\frac{15A^4\omega_0^6}{128}$, assuming $\omega_0 = 89.76$, gives a value of $L_{th} = 5.23 \times 10^{11} A^4$. From the available spikes, $A \approx 100$, hence $L_{th} = 5.23 \times 10^{19}$ for a 'slow' spike of low amplitude. This value was not satisfactory, as it resulted in a either a number of missed spikes or false detections, depending on the other parameters of the system. In practice, a different value was estimated, from experimental data. Since the clusters of spikes and non-spikes are not always distinct, there is a region of values of L corresponding to points belonging to both spike and non-spike samples. The optimal value,L^*_{th}, was chosen as the value of L_th which minimized false detections (i.e. the probability that a normal sample is incorrectly marked as a spike) as well as missed detections (i.e. the probability that a genuine spike is classified wrongly as a normal sample). In other words, L^*_{th} should be selected to minimize the probability of incorrect classification (the point where the probability of making a mistake is equal for both classes).

In order to do this, EEG records containing a known number of pre-identified spikes were repeatedly processed with the suggested method, each time increasing the value of L_{th} from a low value, which caused an infinite number false, but no missed detections to a level where the number of detected spikes approached zero. For each level, the number of false and missed detections were recorded and plotted on the same set of axes. Figure 6.5 shows these two functions for a record sampled at 160 s^{-1} and processed using a differentiator with $N = 9$ and $\alpha = 1$. It was then possible to determine the abscissa of the point of intersection of the two graphs, which gave the required value L^*_{th}. The small number of points in the region around L^*_{th} and the distinct peaks of the two curves are characteristic of good cluster separation. Therefore, by selecting any threshold level in this region the performance of the detector should remain largely unchanged. This procedure was automated using a variant of the spike detection program with multiple values of L_{th} and the results of spike detections were subsequently analyzed manually.



(a) Total view of the functions

(b) Detail, showing optimal threshold value, L^*_{th}

Figure 6.5: Missed and false detections as a function of L_{th}

Spike detection, however, requires the association of the binary indication after thresholding, which *persists* over a period of time, with a single spike *event*, which is registered *at* a particular time instant. It was decided to register a spike occurrence at the end of the indication period, provided that the period itself was less than about 0.1 s, measured by counting the duration in samples and comparing it to the number of samples corresponding to the mentioned duration. This was chosen to be a little above 70 ms, which represents the longest spike duration, so that filtering which may prolong the response of the multiplicative detector did not cause genuine spikes to be missed. This acted as a safeguard against the false detection of some artifact bursts, which often persist for longer durations than isolated spikes and resembles a similar concept used by Stelle and Comley[183].

6.4.3 Results and observations

Data preparation

The developed method was applied to a number of EEG records, which were saved on an FM magnetic tape as a continuous time signal and then digitized using a 12-bit data acquisition circuit (DT3100), based on 80286 computer. This was driven by a commercial digital signal processing package (ILS) to sample the EEG records played back from tape. The sampled signals were saved in separate files, one for each recording.

Four records containing mixtures of spikes and artifacts were sampled to produce an equivalent number of files, which were used off-line for testing different spike detectors. For comparative purposes, sampling was repeated twice for every record, with sampling rates of 1000 and 160 samples per second. Anti-aliasing filters were used in both cases (fourth-order, low-pass Butterworth with cut-off frequencies at 400 Hz and 70 Hz, for the 1000 and 160 s^{-1} sampling rates, respectively).

The sampled data contained a large constant offset, which is not necessarily a characteristic of the signal itself, but one of the recording procedure. This was removed after discretization using a digital second-order high-pass Butterworth discrete-time filter with a cut-off frequency of 0.1 Hz, to ensure a flat response in the pass-band. Conversion to discrete time was achieved by the bilinear transformation as before. Although this might appear unnecessary, since subsequent differentiation would remove any constant (DC) term, in practice it helped limiting the dynamic range of the products, during the application of the differentiator with the signal. This is a crucial aspect, if the use of fixed-point arithmetic proves necessary for the portable realization of the system on a small microcomputer.

Record contents

The four records, labelled A, B, C and D contained various types of EEG signals, mainly spikes and muscle artifacts (see Table 6.1).

Record A had a duration of 210 s and contained a total of 52 spike-and-wave complexes. Although it was fairly free from non-cerebral activity, at two points it contained low-amplitude high-frequency components which resembled muscle artifacts.

Records B and C, having durations of 130 and 123 seconds respectively, contained a few spikes each (8 and 10). Muscle potentials were more abundant in these records than in record A, but the contamination of the cerebral potentials was only partial.

Record D, the one with the longest duration (242 seconds) was the most severely contaminated with muscle and other artifacts. The general high-frequency spiky appearance, and the large amplitude muscle spike bursts which completely obscured the underlying cerebral activity were characteristic of this record. No spikes were present in this record or, to be precise, none could be detected.

Variants of the basic system

It must have become obvious from Section 6.4.2 that there are many possible combinations of parameters that define the overall multiplicative differential spike detector, most of which related to the construction of the differentiator. Depending on the sampling rate f_s, the cut-off frequency, $f_c = \alpha\frac{f_s}{2}$, the order, N and the window used, numerous differentiator realizations are possible. The results for various combinations of these parameters are recorded in Table 6.1. Tests were mainly carried out for the sampling rate of 1000 s^{-1}, since the results were more variable than those for 160 s^{-1}. Differentiators with orders from 3 to 51 and a bandwidth from 0 to either 50 Hz or the Nyquist frequency were tested. The windows used were either the Rectangular or the Hanning.

As mentioned earlier, filtering of the squared product of the derivatives was performed by a second order Chebyshev filter. No differences in the results were observed with the use of a first order filter.

Finally, perhaps the most crucial element of a successful detector is the optimal spike detection threshold, L^*_{th}, estimated as explained in section 6.4.2, based on the results of record A, since it is the only one containing a large number of spikes. This must have introduced some bias towards a lower value than required to reduce false detections in, say, the contaminated record D, but simulates better what may happen in practice, since relatively artifact-free recordings at an EEG laboratory could be used for setting up the system, when during its operation it may be subjected to artifacts, as there is no precise knowledge about the nature of artifacts that might occur. It may be argued that missing spikes is more crucial than falsely detecting some when there are none, since it is possible to record segments of the EEG signal around the detected spike for verification of detections by an expert human analyst later, whereas there would be no means of recovering genuine spike occurrences if they have been missed by the system.

6.4.4 Discussion

The results of table 6.1 are a good means of comparison of the various settings of the system, but the small availability of EEG records make the generalization of the conclusions drawn difficult.

Although the differentiator based on the piecewise polynomial approximation method had a number of variable parameters (sampling rate, order, type and value of the approximation criterion), only one example is given in Table 6.1. This had order 5 to guarantee the continuity of both the first and the second derivative, and hence prevent sudden 'jumps' in the value of the raw spike detector, The low sampling rate of 160 s^{-1} did not result in a good approximation of the EEG signal, hence the sampling rate of 1000 s^{-1} was considered only. Results were not found to depend on the criterion of fit, but the tested configuration used the MSE (ξ_2) and the value of ξ_{max} set to 300 as in section 5.5.3.

It is evident that from the presented results this type of differentiator was not satisfactory especially in the presence of artifacts. As there is no precise way of controlling its bandwidth, spiky artifacts appear to cause polynomial approximations consisting of a few samples each with large oscillations between their end points. These gave rise to large derivatives and hence excessive numbers of false alarms. Some spikes were also missed, because their small amplitude did not give rise to a large error in the approximation and hence were approximated by one or two smooth polynomial segments, which meant low derivatives and hence inability

Record							A		B		C		D	
Duration in seconds							210		130		123		242	
Number of spikes present							52		8		10		0	
Muscle artifacts present?							some		yes		yes		many	
System parameters							Detections							
#	Diff/tor Name	f_s (Hz)	f_c (Hz)	N	Win	L^*_{th}	√	×	√	×	√	×	√	×
1	pcw. poly†	1000	ξ_{max}=300			5.7×10^{22}	48	4	6	31	7	80	0	—
2	D03_H100	1000	500	3	Hann.	9.2×10^{21}	50	2	7	18	8	18	0	860
3	D05_R100	1000	500	5	Rect.	3.1×10^{15}	48	4	6	22	7	30	0	1231
4	D05AH100	1000	50	5	Hann.	2.6×10^{15}	50	2	8	12	10	15	0	636
5	D09_R100	1000	500	9	Rect.	6.3×10^{24}	46	6	6	28	6	56	0	—
6	D09AH100	1000	50	9	Hann.	2.5×10^{15}	50	2	8	6	10	12	0	186
7	D15_R100	1000	500	15	Rect.	1.8×10^{25}	44	8	5	29	4	60	0	—
8	D15AH100	1000	50	15	Hann.	2.5×10^{15}	50	2	8	2	10	7	0	154
9	D31_R100	1000	500	31	Rect.	2.9×10^{25}	41	11	5	49	4	78	0	—
10	D31AR100	1000	50	31	Rect.	2.5×10^{15}	50	2	8	2	10	7	0	153
11	D51AH100	1000	50	51	Hann.	2.5×10^{15}	50	2	8	2	10	7	0	88
12	D03_H016	160	80	3	Hann.	1.3×10^{20}	50	2	8	3	8	10	0	94
13	D09_R016	160	80	9	Rect.	2.2×10^{20}	50	2	7	3	8	14	0	123
14	D09AH016	160	50	9	Hann.	1.0×10^{20}	50	2	8	3	10	6	0	53
15	D31AH016	160	50	31	Hann.	1.0×10^{20}	50	2	8	2	10	6	0	41

†: Fifth order piecewise polynomial approximation method (see text)

Table 6.1: Results of multiplicative differential spike detector

to discriminate. Because of its rather unpredictable behaviour with artifacts, the problems with low-amplitude spikes and the requirement for more computations, compared to the other methods presented, this technique was considered no further.

The *sampling frequency* appeared to be a crucial element in the success of the spike detector. The results for 160 s^{-1} were generally better than those for 1000 samples per second. This, however, is only an indirect observation. The actual reason is the effect of the anti-aliasing filter of lower cut-off frequency in the first case, which considerably reduced signal components, especially artifacts, at frequencies above 70 Hz. Moreover, since differentiation, even at full bandwidth, could not enhance frequencies above 80 Hz, the effect of these on spike detection was much reduced (this is related to the bandwidth issue, which

is addressed later).

There has been some evidence that the reliability of the detector increased with the *order* of the differentiator. This observation only applies to the differentiators of restricted bandwidth. For the full bandwidth ones, the increase in the order led to the deterioration of the performance, which was expected since the increasing order perfected differentiation close to the Nyquist frequency, as is evident from Figure 6.1. The enhancement of high-frequencies, where artifacts are more prominent, resulted in false spike detections. It should be noted that the improvements in the performance of the system with the increase in order of the limited bandwidth differentiators is not so much associated with the perfection of the differentiation in the pass-band as with the *perfection of the suppression of frequencies in the stop-band.* This is demonstrated by the results of the system for the low sampling rate of 160 s^{-1}, which proved accurate even for low orders.

The issues already addressed appear related to a large extent on the *bandwidth* of the differentiator. This is defined by the cut-off frequency, f_c and the Nyquist frequency, $\frac{f_s}{2}$. The improved performance of the detector when the bandwidth of the differentiator was restricted to 0–50 Hz is clearly marked in Table 6.1. Full-bandwidth differentiators tend to enhance artifacts, which affect the higher frequencies and, since the amplification at these frequencies is considerably larger than those in the frequency band occupied by spikes, even low-amplitude artifacts gain substantial enhancement after being differentiated twice! This is, of course, observed much more for the sampling frequency of 1000 Hz, as the bandwidth of the EEG is restricted to a small fraction of the spectrum, whereas the region of the enhancement is what artifacts are likely to affect most. This is hardly observed if the sampling rate is 160 s^{-1}, since most of the bandwidth is now occupied by the EEG and the only interference from artifacts occurs in frequency components within the bandwidth of the EEG or slightly above. Therefore, the restrictions in the bandwidth here had been less beneficial over those on the 1000 s^{-1} rate, though the rejection of artifacts, especially in the heavily contaminated record D is still better.

During the development of the differentiator, there has been some concern about the use of the appropriate *window* and its effects on the behaviour of the differentiator. In practice, it did not appear to matter which window was used and both the Rectangular and the Hanning windows performed well. This indicated that *precise differentiation is perhaps not the main issue in spike detectors based on derivatives*, and that crude approximations could still yield useful results.

One of the important questions raised while the system was tested was whether it was possible to find a single optimal *threshold* that could apply either to differentiators of the

same order or the same sampling rate. It is evident from Table 6.1 that this had not been the case. Thresholds needed adjustment. It is noticeable that for f_s=1000 Hz there was no need to adjust the threshold for differentiators of orders 9 and over, for f_c=50 Hz. For the full bandwidth there was necessity to raise the threshold by several orders of magnitude to reduce the number of false detections. This was hardly necessary for the lower sampling rate of 160 samples per second, where the optimal threshold was only doubled. Evidently, its value is more related to the rejection of false detections than to the detection of genuine spikes.

Several points regarding this method of spike detection are worth noting.

- The use of a low-order differentiator may be preferred in practice, since it requires fewer operations to implement and hence is faster.

- The results for the low sampling rate were (in general) better than those for the high one. This means that more time is available between samples and hence the number of operations available for a real-time implementation on a given microcomputer system is significantly increased.

- Precise differentiation is not a crucial issue. Even low-order differentiators performed well. It seems that suppression of frequencies liable to artifacts is a more important matter.

- The need for manual setting of L^*_{th} is a serious limitation of this implementation, since in practice it depends not only on the presence or absence of artifacts but on signal power as well, which is not fixed, as the impedance between the electrodes and the scalp suffers fluctuations, especially over the period of long-term ambulatory monitoring.

- The analysis of the detector itself has limitations, as it is cumbersome to operate with non-linear functions and combinations of products in general. Therefore the behaviour of the detector cannot be fully assessed even for relatively simple inputs; for the rather complicated EEG signals it is impossible.

It should be noted that the detector presented has a vague resemblance to the arithmetic detector of Quian et al.[157] (see also section 4.6.1), although the concepts, rigour, reasoning analysis and the structure of the detector of these authors and are quite different.

6.5 State-space techniques based on time derivatives

Some of the elements of the spike detector just described were not based on elements of the signal itself, but rather on secondary characteristics of the product of its derivatives. One fundamental limitation is the loss of the individual properties of the derivatives after multiplication. The large amplitude oscillations observed in their product are quite characteristic of spike phenomena as well as some artifacts, but the extraction of structural information from them is more cumbersome than from the spike itself. The envelope of the oscillations is the only signal parameter easily extracted. Although this has been used for spike detection, it is a crude summary measure of the information contained in both derivatives and its correlation with them is only remote. It may be therefore argued that other useful information has been lost. Even more important than loss of information is the loss of continuity between the spike specification, in terms of amplitude and frequency, and the limit (threshold) of the detector output. Since the formal spike definition cannot be related to the threshold any more, the latter had to be estimated by experimental means, a practice which should be avoided, if possible, as it makes the method data-dependent and inflexible.

In an attempt to address these issues, it was decided to change the function that combined the two derivatives into one that could be related more easily to properties of individual derivatives and signal parameters, in general. For this purpose, the model used in section 6.4.1 had to be modified to be more generalized and accurate and a *state-space* approach to be adopted, at least for visualization of the signals and their properties. Based on this, a new function was defined and was made insensitive to signal power variations. The new system was based on a distance and orientation measures, which relate to formal limits for the detection of spikes. The complete method is described in the rest of this Chapter.

6.5.1 Modelling the EEG waveform

Logically, there is no reason why the spike model of section 6.4.1, expressed by equation (6.13), may not be used here also, since the derivatives are still in use. This model, however had been a crude one and only reflected the most basic properties of a spike, amplitude and duration. If more features of a spike are to be taken into account it is necessary to modify it to include them. Equation (6.13) assumes that spikes consist of a single sinusoidal oscillation, smoothed at its two ends with constant frequency as well as amplitude. This is not the case in practice, because even if the shape of the spike is roughly 'sinusoidal', the amplitude of the positive peak is sometimes larger than that of the negative one and there is a distinct reduction in the frequency of the oscillation after the negative peak is reached.

These did not affect spike detection then, as the variations would have disappeared in the multiplication of the derivatives. Here, the model had to be generalized to include them:

$$y(t) = A[\theta(t)] \sin \theta(t) \tag{6.27}$$

The constant frequency of the waveform has been replaced with a function of time, thus allowing for the increase in the period of the oscillation that is observed after the negative peak of the spike, as well as for the variations in the durations from the onset to the positive peak and that from the positive to the negative peak (although these are less apparent). It should be noted, though that $\theta(t)$ is a monotonically increasing non-negative function of time in the domain of interest $t \in [0, t_0]$ and that if $\theta'(t) = \omega_0$, a constant, then $\theta(t) = \omega_0 t$ as in the earlier model. Summarizing the properties of this function,

$$\begin{array}{lll} \theta(t), \quad \theta'(t) & \text{are continuous and differentiable in } [0, 2\,\pi] \\ \theta(0) = 0 & \\ \theta(t_0) = 2\pi & \\ \theta'(t) > 0 & \forall\, t \in [0, t_0] \end{array} \tag{6.28}$$

Similarly, the constant amplitude factor, has been replaced with a function of time, $A[\theta(t)]$. This also incorporates the window function, which was used to smooth the end-points of the oscillation. It is no longer necessary to include it as a separate entity. This time-varying function is capable of representing amplitude changes in the spike and indeed any single wave of the EEG signal. To simplify some aspects of the analysis that follows it is more convenient to express $A(\cdot)$ as a function of $\theta(t)$ rather than t. This corresponds to a distortion of the time axis. As with $\theta(t)$, certain restrictions apply, mainly on the values of the function and its derivatives at the end points, 0 and 2π, and their values outside the range of interest:

$$\begin{array}{lll} A(\tau), \quad A'(\tau) & \text{are continuous and differentiable in } [0, 2\,\pi] \\ A(\tau) \geq 0 & \forall\, \tau \in [0, 2\pi] \\ \quad = 0 & \forall\, \tau \notin [0, 2\pi] \\ A(0) = A(2\pi) = 0 & \\ A'(0) = A'(2\pi) = 0 & \\ A''(0) = A''(2\pi) = 0 & \end{array} \tag{6.29}$$

The end-point condition ensures a smooth transition from of the model value to zero. This is also a property of its derivatives.

The first two derivatives of the modified model of equation (6.27) may be readily written down:

$$y'(t) = \theta'(t)\{A[\theta(t)] \cos \theta(t) + A'[\theta(t)] \sin \theta(t)\} \quad (6.30)$$

$$y''(t) = [\theta'(t)]^2\{\{A''[\theta(t)] - A[\theta(t)]\} \sin \theta(t) + 2\, A'[\theta(t)] \cos \theta(t)\} + \theta''(t)\{A[\theta(t)] \cos \theta(t) + A'[\theta(t)] \sin \theta(t)\} \quad (6.31)$$

One might wonder about the pattern of the collection of points of the form $p(t) \doteq (y'(t), y''(t))$ in $\mathbf{R}^2$. Their locus will be mentioned quite often in this section and has been called simply *d-plot* (the name is possibly not very successful, but was used in a number of internal reports and a publication[126], so it will be adopted here also, to avoid confusion). This, of course, depends on the nature of the functions $A[\theta(t)]$ and $\theta(t)$.

Since $A[\theta(t)]$ is continuous, twice differentiable and bounded in the region of interest, its first two derivatives are also bounded. Continuity and differentiability of $\theta(t)$ also guarantees that the derivatives of this function exist. By the non-decreasing constraint they are also guaranteed to be finite in the region of interest. Moreover, since both $\sin \theta(t)$ and $\cos \theta(t)$ are bounded by the range $[-1, 1]$ then from equations (6.30) and (6.31) the locus of $p(t)$ is also bounded. The derivatives are not guaranteed to have the same sign, but it is always possible to find a closed curve, C, such that all points $p(t)$ lie in its interior. Such a curve might be the rectangle constructed by the lines $y' = a$, $y' = b$, $y'' = c$ and $y'' = d$, where a, b are the minimum and the maximum values attained by the first derivative and c and d those attained by the second derivative. It is, of course, preferable if C is only large enough to enclose the locus in question. Therefore more restrictive parametric curves have been considered, being more specific and accurate.

One such curve may be obtained using some algebraic manipulation on the locus of the points in the space of the two derivatives (d-plot). Dividing $y'(t)$ by $A(t)\theta'(t)$, produces

$$\frac{y'}{A\,\theta'} = \cos\,\theta + \frac{A'}{A} \sin\,\theta \quad (6.32)$$

where the independent variable t has been dropped for simplification of the notation. Letting $\frac{A'}{A} = \tan \phi_1$ leads to the following modified relation:

$$\frac{\cos\,\phi_1\, y'}{A\,\theta'} = \cos\,\theta\, \cos\,\phi_1 + \sin\,\phi_1\, \sin\,\theta = \cos(\theta - \phi_1) \quad (6.33)$$

Similarly, subtracting the term $\theta''(A\ \cos\,\theta + A'\ \sin\,\theta) = \frac{\theta''}{\theta'} y'$, then dividing y'' by $(\theta')^2\ (A''-A)$ and letting $\frac{2\,A'}{A-A''} = \tan\ \phi_2$ produces a similar relation for the second derivative:

$$\frac{\cos\ \phi_2\ (y'' - \frac{\theta''}{\theta} y')}{(A'' - A)(\theta')^2} = \sin(\theta - \phi_2) \quad (6.34)$$

Both expressions are useful, since they are bounded in the range of their right-hand-sides, $[-1,1]$. Using the properties of the functions involved, it is possible to determine the lower and upper limits of these expressions. Since $A(\tau) \geq 0$ and $\theta(t) \geq 0$ for $\tau \in [0, 2\pi]$, (see properties, expressions (6.29) and (6.28)), one might use their maximum values, A_{max} and θ'_{max} and if the maximum is a turning, rather than a boundary point (which is true since A is zero at the boundaries), the maximum occurs when $A' = 0$ which causes ϕ_1 to be zero. Hence the maximum of the first derivative is $a = A_{max}\,\theta'_{max}$ and its minimum is (symmetrically) $b = -A_{max}\,\theta'_{max}$. By similar arguments, expression (6.34) is maximized when $\phi_2 = 0$, $\theta' = \theta_{max}$ and $A = A_{max}$. The rôle of A'' has not been ignored in this argument, but, as it is much smaller than A, its influence is marginal. If any value for it is to be taken into account, it should be its minimum, A''_{min}. Hence the maximum and minimum for the expression $y'' - \frac{\theta''}{\theta'}y$ are $c = (A_{max} - A''_{min})\,(\theta'_{max})^2$ and $d = (A''_{min} - A_{max})\,(\theta'_{max})^2$. Strictly speaking, the method of obtaining these values is not the formal algebraic one (maxima and minima are the points of zero derivative). However, differentiation produced expressions that were so complicated they could not be used, especially since the functions have not been defined explicitly. In practice, however the error is small, since A' and A'' are negligibly small, compared with A, and may even be taken as zero for most of the spike duration, as explained later. Besides, since in real signals only the derivatives are estimated, it is possible to compute these limits directly, rather than by the parametrization used here for the development and justification of a method. The symmetry of the limits, and the expressions themselves are suggestive of the shape of the boundary curve. By defining the variables $u = y'$, $v = y'' - \frac{\theta''}{\theta'}y'$ and the constants $p = a = -b$ and $q = c = -d$, and by further assuming that $\phi_1 = \phi_2 = \phi$, one may write:

$$\begin{aligned} u &= p\cos(\theta - \phi) \\ v &= q\sin(\theta - \phi) \end{aligned} \tag{6.35}$$

These are the parametric equations of an *ellipse*, centred at the origin of the $u - v$ plane, cutting the axes at right angles and having axes $2p$ and $2q$ in the u and v directions, respectively. By eliminating the parameter $\theta - \phi$, the more common form of the equation of the ellipse is revealed:

$$\frac{u^2}{p^2} + \frac{v^2}{q^2} = 1 \tag{6.36}$$

The assumption $\phi_1 = \phi_2$ is not a critical issue in the argument, since one could define the ellipse without using the d-plot of the model, just the boundary values computed before. The development of the boundary ellipse, $\mathcal{C}_o$, from small modifications of the d-plot of the spike, which has only been described in general terms is suggestive of the proximity of the actual locus to that of an ellipse.

Figure 6.6 shows the d-plots of four models and a genuine spike. The first three plots have constant frequency, $\theta'(t) = \omega_0$ but different amplitude functions, $A(t)$, which are zero everywhere outside the domain $[0, \tau_0]$. A range of functions were considered inside this domain:

$$A_r(t) = A_0 \tag{6.37}$$

$$A_h(t) = \frac{A_0}{2}(1 - \cos\omega_0 t) \tag{6.38}$$

$$A_s(t) = A_0\,(\frac{3}{8} - \frac{1}{2}\cos\omega_0 t + \frac{1}{8}\cos 2\omega_0 t) \tag{6.39}$$

$$A_t(t) = A_0\,(\frac{1}{2} - \frac{3}{4}\cos\omega_0 t + \frac{1}{4}\cos^3\omega_0 t) \tag{6.40}$$

$A_r(t)$ is a rectangular pulse of amplitude A_0 and duration $\tau_0 = \frac{2\pi}{\omega_0}$. This, violated the requirement of continuity and differentiability at $t = 0$ and $t = \tau_0$, which is obvious in the first derivative, but produced an elliptic d-plot, apart from the end-points at which it is not differentiable.

$A_h(t)$ is a Hanning window scaled by a factor A_0: The factor 2 was introduced to counteract the diminution in the maximum amplitude of the resulting waveform in relation to that for the Rectangular window when the oscillation reaches its peak. Here, the ellipse is distorted with one extra loop introduced, corresponding to the beginning and the ending of the waveform where the slope and curvature of A are in anti-phase. The discontinuous second derivative at the end-points of the Hanning window does not cause problems here, since it is cancelled by the sinusoidal oscillation of the same frequency, which is zero at those points.

In an attempt to produce a function that fully conforms to the restrictions imposed on A (relations (6.29)), starting from the Hanning window, the functions $A_s(t)$ and $A_t(t)$ were developed (see Appendix B). The first has continuous both derivatives, but is not symmetric with respect to $\frac{\tau_0}{2}$. The second satisfies both conditions. The d-plots of the corresponding models are similar to that of the Hanning window. Examples of spike models employing $A_r(t)$, $A_h(t)$, $A_s(t)$ and $A_t(t)$, all for $\theta(t) = \omega_0 t$ are shown in Figure 6.6(d) and their corresponding d-plots in (a) and (b).

$A_t(t)$ led to a variation of the Tukey window[151, p. 549] which has a flat section of amplitude A_0, in the range $[\alpha\tau_0, (1-\alpha)\tau_0]$, where $0 \geq \alpha \geq \frac{1}{2}$, gradually decreasing to zero at the two end-points. The original Tukey window could not be used, because it has discontinuous derivatives at the points $t = \alpha\tau_0$ and $t = (1-\alpha)\tau_0$, where the 'Hanning' sections are attached to the 'rectangular' one. The remedy to this problem lied in the half-wave symmetry of $A_t(t)$ (equation (6.40)). In its final form, the duration of the 'tapering" sections

at the beginning and ending of the model was made independent using two proportional parameters at $t = \alpha\tau_0$ and $t = (1-\beta)\tau_0$, where $\alpha, \beta \in [0, \frac{1}{2}]$, permitting asymmetry in the envelope:

$$\begin{aligned} A_m(t) &= 0 && t < 0 \\ &= A_0\left(\tfrac{1}{2} - \tfrac{3}{4}\cos\tfrac{\omega_0}{2\alpha}t + \tfrac{1}{4}\cos^3\tfrac{\omega_0}{2\alpha}t\right) && 0 \geq t \geq \alpha\tau_0 \\ &= A_0 && \alpha\tau_0 \geq t \geq (1-\beta)\tau_0 \\ &= A_0\left(\tfrac{1}{2} - \tfrac{3}{4}\cos\tfrac{\omega_0}{2\beta}(t-\tau_0) + \tfrac{1}{4}\cos^3\tfrac{\omega_0}{2\beta}(t-\tau_0)\right) && (1-\beta)\tau_0 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \qquad (6.41)$$

This function has been used to describe the amplitude of the last model, which also had variable 'frequency', $\theta'(t)$, following the same function as $A_m(t)$. The angle $\theta(t)$ was obtained by integration (see Appendix B). to model more precisely the behaviour of real spikes. Notice that this model and its d-plot bear more resemblance to that of a real spike (Figure 6.6(c),(d)). Its behaviour, although it still follows a loop in all quadrants, has deviated from the ellipse. Sections of it can still be identified as elliptic, especially in the fourth and third quadrants, where the amplitude and frequency are slowly varying or constant. The significance of the elliptic shape of the d-plots will become apparent in the discussion that follows.

6.5.2 Towards a detection procedure

Having shown d-plots of models and real spikes to demonstrate graphically the suggested data representation, it will now be demonstrated how it may be used for their detection. After the onset of a spike, its amplitude may be regarded as almost constant ($A = A_0$) and assuming its frequency is constant also ($\theta' = \omega_0$, which appears to be true from just before the peak to just after the trough of the spike), then $A' = A'' = 0$, $\theta'' = 0$ and hence $\phi_1 = \phi_2 = 0$ and the d-plot resembles an actual ellipse, $\frac{(y')^2}{A_0^2\omega_0^2} + \frac{(y'')^2}{A_0^2\omega_0^4} = 1$! Note that the ratio of the vertical to the horizontal axis of the ellipse gives in this case the frequency of the oscillation, which can then be used to compute the amplitude of the signal. Hence the d-plot provides a way to readily estimate these parameters. Although their computation is possible without the d-plot, the latter provides an easy way to visualize them.

So far a boundary, C_o has been established so that the d-plots of spikes lie in the region enclosed by C_o. This is not particularly useful for their detection, since, as stated earlier, 'normal' EEG activity is *lower* in both amplitude and frequency than the spike. It is obvious, that the d-plot of such activity will be well enclosed by C_o. What is actually required is an *inner boundary*, C_i such that spikes have their d-plots exterior to this, whereas those

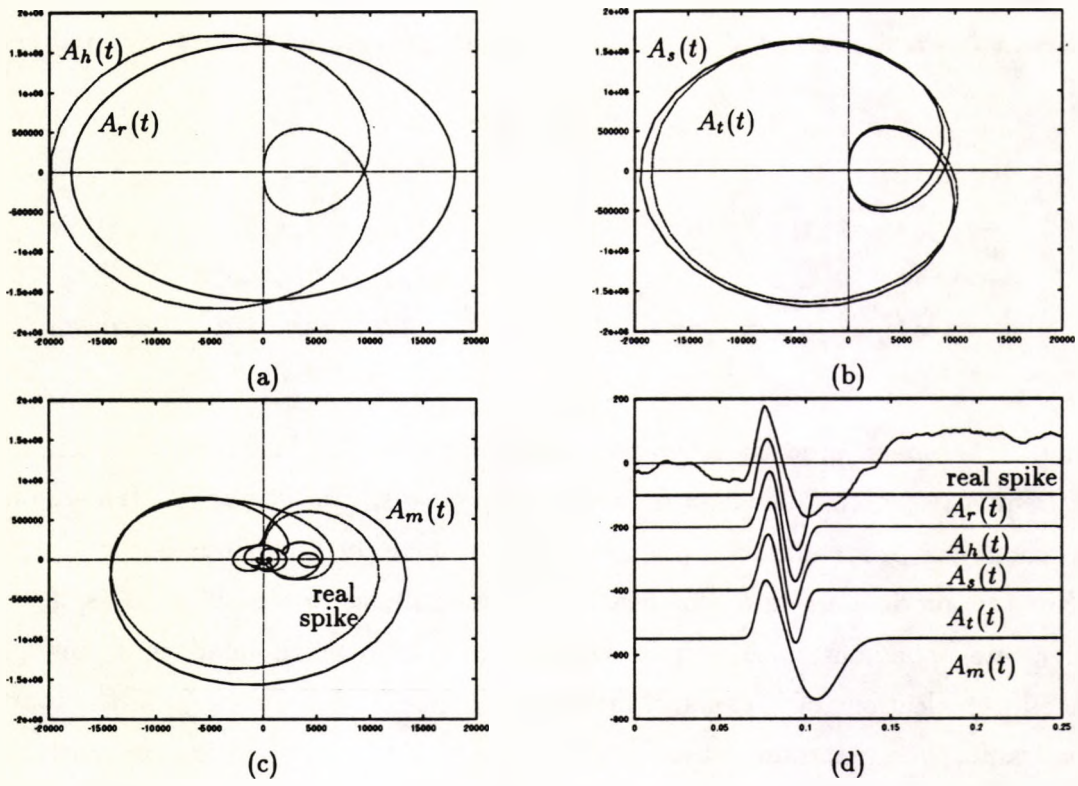


Figure 6.6: Examples of d-plots, (a)-(c) and corresponding signals (d)

corresponding to non-spike EEG lie on its interior. Due to the definition of the spike parameters, this is not possible since the d-plots of a spike start from the origin (remember that $A(0) = A'(0) = A''(0) = 0$). For real spikes, however, the end-points are superimposed with the waveforms before and after the spike and are not reliable indicators, hence they may be ignored. The model of equation (6.27) is suitable for describing any wave of the EEG and is not specific to spikes. Therefore, using the intuitive rule that the amplitude of a spike $A_s(\tau)$ is larger than that of non-spike waves, $A_{ns}(\tau)$, and that the 'frequency' of these waves follows a similar relation, $\theta'_s(t) > \theta'_{ns}(t)$, it is evident that this this is also true for their products (since their magnitudes are well above unity):

$$\begin{aligned} A_s \cdot \theta'_s &> A_{ns} \cdot \theta'_{ns} \\ A_s \cdot (\theta'_s)^2 &> A_{ns} \cdot (\theta'_{ns})^2 \end{aligned} \tag{6.42}$$

These are essentially the major and minor axes of the d-plot, assuming that it has an elliptic shape (and ignoring the effects of A'', which is rather small). Assuming that the relations (6.42) hold even for extreme cases, when the background (non-spike) waves attain frequencies and amplitudes that are atypically large when spikes have abnormally low frequencies and amplitudes, it is possible to define a boundary ellipse, $\mathcal{C}_i$, which encloses all non-spike EEG

activity, but none of the spikes. Again, this ignores the regions of transition between a typical EEG wave and a spike, when this boundary is crossed.

Summarizing the important elements of this analysis procedure,

- A spike model based on a single sinusoidal oscillation with varying amplitude and 'frequency' was proposed. This is a development from a simpler model given in the previous section.

- The 2-dimensional space defined by the first two derivatives of the EEG signal, termed the d-plot, was suggested as a way to highlight certain attributes of the signal, implied in the model, but not clearly visible in the conventional time-domain representation of the signal. These are multiplicative combinations of the amplitude and frequency functions, which are the elements of the model suggested.

- The shape of the locus of a spike represented in a d-plot, as well as those of non-spike parts of the signal (background activity), were found to form 'loops' around the origin, provided that their amplitudes are slowly varying, compared to the oscillation. In the case of sinusoidal signals, the d-plots are ellipses centred at the origin. This is approximately true for distorted sinusoidals, exhibiting small changes in their amplitude and frequency.

- Based on the previous observation, it was demonstrated that an upper and a lower boundary could be suggested for the region in the d-plot occupied by spikes, excluding their beginning and ending. It was suggested that elliptic boundaries could be used, in which case spikes have d-plots in the ring between the inner, $\mathcal{C}_i$, and the outer, $\mathcal{C}_o$, boundary ellipses. $\mathcal{C}_i$ encloses the background EEG activity, whereas $\mathcal{C}_o$ is a safeguard against artifacts and other signals of large amplitudes and frequencies.

- Therefore spikes may be identified as sections of the EEG signal that lie in the region between the bounds described.

The material presented so far was kept on a general, level. The concern here was the presentation of the ideas and the rationale behind the suggested methodology. Thus, no attempt has been made, as yet, to address the realization of the suggested procedures. It is well understood that there are many practical issues that might complicate the transition from a conceptual to a working spike detector. These issues will be addressed in the next sections, as the spike detector will be put together.

6.5.3 The Spike Detector

The relation of the derivatives of the signal to the detection of spikes, as explained in the previous section provided the framework for the development of a practical spike detector. The general method assumes that spikes have frequencies and amplitudes that are uncharacteristic of the 'normal' EEG signal and are highlighted in the space defined by its first and second derivatives. In that space, during the occurrence of a spike, the trajectory of the signal (the d-plot) passes through the region between two non-intersecting closed curves about the origin, the outer and the inner boundaries, C_o and C_i, respectively. It was also assumed that these boundaries have the form of concentric ellipses centred the origin. In practice, these boundaries are not known and must be estimated. Moreover, differentiation of the raw EEG signal may be implemented in a number of ways. Other elements and parameters of the system may also need adjustment to optimize the performance of the detector. This section deals with the practical solution of these problems.

In general, the operations needed for spike detection according to the ideas of the previous section are shown in Figure 6.7. The signal must be differentiated twice, since the procedure requires the first two derivatives. Then, the test for 'spikiness' is carried out to determine whether the signal lies between the spike boundaries C_i and C_o. The output of this test persists for the duration of the spike, whereas a spike detection is a discrete-time event. The conversion is done in conjunction with phase information, indicating the orientation of the wave. The output of this operation is an event detector, marking and counting the occurring spikes. Parameters, like the spike boundaries, were derived from the signal, as explained later in this section, whereas the low-pass filter and the slow wave detectors are additional elements whose significance will be elaborated in section 6.5.8.

6.5.4 Differentiation-the first step

It was noted earlier (section 6.3.1) that a discrete-time differentiator my be realized in many forms. A selection of those was implemented for the earlier spike detector of section 6.4.2. The tests carried out then revealed some of the intrinsics of each type of differentiator and permitted the isolation of the less successful ones as well as valuable experience for the prospective use of each one of them. The noise immunity of the parametric differentiator, based on polynomial approximation of the signal, was inferior to most non-parametric implementations. Moreover, it required more computations than non-parametric techniques with superior performance and lacked 'tuning' parameters, which could tailor its performance to the requirements of this application. For these reasons this differentiator structure was not

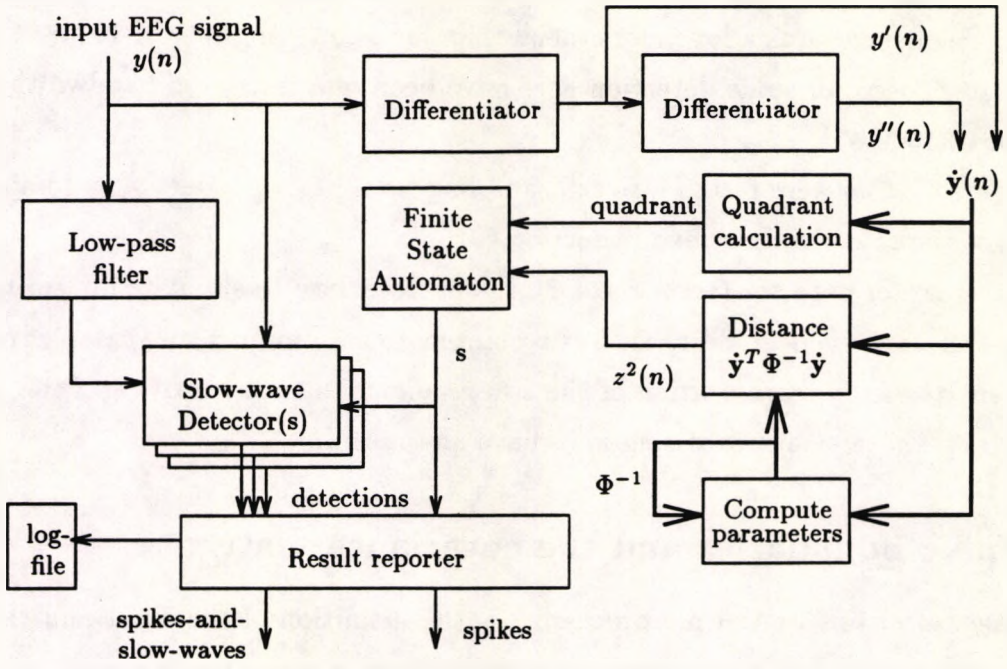


Figure 6.7: Block diagram of the spike and wave detector based on d-plots

considered in this section.

The other differentiator structures, the forward-backward difference and the frequency-specified realizations belong to the non-parametric class. The main quality of the first type is the simplicity of the computations it requires. The suppression of high frequencies, which begins after its maximum value at half the Nyquist frequency and becomes zero at the Nyquist frequency, is another desirable feature. This is especially important if the sampling rate is higher than the theoretical minimum, in which case noise at high frequencies is not enhanced. The forward-backward difference structure was the first one tried out with encouraging results[126]. It was the apparent success of this detector that motivated further investigation of other differentiator structures.

The main drawbacks of this differentiator is the lack control of the bandwidth and the inanbility improve its performance to approximate the ideal operator. These are inherent in the frequency specified structures, which have a bandwidth-related parameter, α and a user-defined length, M, to control the accuracy of the approximation to the ideal differentiator. The details of this structure were discussed before in section 6.3.1, where it was defined. In section 6.4.2, where it was applied for the first time, there appeared to be a small effect on the system performance with changes in the order, M, of the differentiator. The use of either a Rectangular or a Hanning window seem to have no effects on the performance of the overall detector and, in any case, little difference in its output. The bandwidth parameter,

α, was more critical, especially for the high sampling rate of 1000 s^{-1}, where the necessity to suppress high frequencies was a lot more crucial than for a sampling rate of 160 s^{-1}. Hence, the differentiators used for spike detection here have been of a restricted bandwidth for the sampling rate of 1000 s^{-1}.

Three differentiators were tested for each sampling rate, having orders of 9, 15 and 31 (a total of six), constructed as described in section 6.3.1.

As with the earlier detector (section 6.4.2), it was necessary to simulate first-in-first-out (FIFO) data buffers as cyclic buffers, as th two pointers to anticipate future samples required by the differentiators. Synchronization of the two gradients in time is critical if the concept of the d-plot and subsequent processing is to have any meaning.

6.5.5 Spike boundaries and the covariance matrix

The next stage, after differential pre-processing is the definition of the inner and the outer spike boundaries, C_i and C_o. During analysis these have been approximated by non-intersecting ellipses, centred at the origin. The definition of these boundaries and a proposed test to determine whether the d-plot of the signal lies between them or not will be described here.

These boundaries may be determined either by considering the formal limits of the spike amplitudes and frequencies or by on-line estimation of the ellipse parameters. Both options were considered.

According to the formal spike definition, its amplitude should be higher than that of 'normal' activity and its duration, T, between 20 and 70 ms, which defined the period of the 'sinusoidal' model of the earlier detector. Assuming that its amplitude is A_0, it follows from the models defined in section 6.5.1 that the derivative will have an amplitude of $A_0\,\omega_0$ and the second derivative an amplitude of $A_0\,\omega_0^2$, where $\omega_0 = \frac{2\pi}{T}$. Using the upper limit of T, defined above, it is possible to compute the minimum amplitudes of the derivatives as

$$\begin{aligned} p_i &= |y'|_i = 89.76 A_0 \\ q_i &= |y''|_i = 8046.09 A_0 \end{aligned} \tag{6.43}$$

These will be the axes of the ellipse defining the inner spike bound, C_i,

$$\frac{(y')^2}{p_i^2} + \frac{(y'')^2}{q_i^2} = 1 \tag{6.44}$$

The upper limits were defined in a similar manner yielding $p_o = 314.16 A_0$ and $q_o = 98696.51 A_0$.

A_0, was estimated on-line, from the signal itself in an attempt to eliminate its dependence on the signal power, which may vary from subject to subject and over time, as the

resistivity of the electrodes is not always the same and changes as the contact with the scalp is disturbed by sweating and drying of the electrolyte solution. Because interictal spikes are fairly infrequent, their amplitudes cannot be used for a reliable estimation of A_0. By using the *power* in the ongoing signal, it is possible to derive such a measure, using assumptions about the nature of the signal, thus allowing a reasonable estimate of the peak amplitude of this activity, which may be used as an estimate of the smallest amplitude which a genuine spike may possess. These assumptions have been:

- the signal does not usually contain many significant frequencies, during any period of time, say up to 4
- The amplitude of the EEG is roughly inversely proportional to frequency
- The existing frequencies are harmonically related.

The first is not unreasonable, as EEG activity often has a strong sinusoidal appearance, the second is a general observation[47] and the third is not really critical, but makes analysis easier. The power of the signal containing the first four harmonics of some frequency ω_0, at amplitudes a, $\frac{a}{2}$, $\frac{a}{3}$ and $\frac{a}{4}$ will be equal to half the sum of the squares of their amplitudes, whereas the peak may be at most equal to the sum of the amplitudes, if they are in phase. The square root of the power, the RMS value may be converted to a peak value using the ratio of peak-to-RMS value, $\frac{a(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4})}{\sqrt{\frac{a^2}{2}(1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16})}} \approx 2.5$ This gives the value of the minimum possible amplitude for a spike, being equal to the peak of the background activity, as

$$A_{min} = 2.5\sqrt{\sigma_y^2} \tag{6.45}$$

where σ_y^2 is the power in the EEG signal. The maximum value of the spike amplitude is not so critical, as it is a safeguard against high-amplitude artifacts. It was set to be five times the value of A_{min}, since none of the spikes in the records available exceeded this in amplitude. It should be noted that the power was computed using a window of K past EEG samples:

$$\sigma_y^2 = \frac{1}{K-1}\sum_{k=0}^{K-1}[y(n-k)]^2 \tag{6.46}$$

With these specifications the region of occurrence of spikes is the intersection of the exterior of the first with the interior of the second ellipse. To test the location of an arbitrary point (y'_p, y''_p) with respect to an ellipse, say C_i, one simply has to compute the left-hand side of the equation with the coordinates of that point to obtain $z_{i_p} = \frac{(y'_p)^2}{p_i^2} + \frac{(y''_p)^2}{q_i^2}$. If $z_{i_p} < 1$ the point lies on the interior of the ellipse, if $z_{i_p} > 1$ it lies on the exterior and, of course, if

$z_{i_p} = 1$ the point lies on the ellipse. Hence to determine if a point lies in the spike region, one has to compute z_i for C_i and z_o for C_o. The point is in the spike region if $z_i > 1$ *and* $z_o < 1$.

Testing revealed that, although the definition of the limits just suggested was reasonable, the results were not very satisfactory (see section 6.5.9). This is believed to be a consequence of the simplicity and inaccuracy of the definition of the spike, which was particularly stressed by Ktonas[107, 108] and Gotman[74]. The attachment of formal duration, frequency and amplitude limits to EEG phenomena is a well-known problem, since these have been defined by their appearance in the traditional EEG recordings without taking into account their, often unknown, generating mechanisms. It is not easy to mark the beginning and the ending of spikes and their parametrization is even more ambiguous, since important elements may be lost. Besides, many elements of spikes may be specific to the patient from which these were recorded.

A practical solution to this problem is to estimate the parameters of both the inner and the outer boundaries of the ellipses on-line, using the derivatives of the signal, which define the d-plot. The concept may be regarded as a generalization of the estimation of the unknown amplitude of the spikes, needed for the realization of the previous method. Hence, instead of estimating the amplitude and using the formal frequency limits for the spikes, the problem has been converted into one that involved the on-going activity. The inner boundary, C_i, as mentioned earlier, is more important than the outer, since it is separating the 'ambient' EEG signal. Improperly low specification of C_i would inevitably lead to over-detection. The incorrect specification of C_o, however is a safeguard against artifacts, which may not be present and are, in any case, suppressed by the restricted bandwidth of the differentiators (see previous section). Hence, instead of defining C_i in terms of the unknown spike parameters, it was defined as the upper bound of the more tangible on-going EEG activity. A method needed to be devised for the estimation of such a boundary and it was readily found in the form of the *covariance matrix* of the first and second derivatives. Combining the derivatives in vector form, $\dot{\mathbf{y}} = (y' \; y'')^T$, the covariance matrix, Φ, is defined:

$$\Phi \doteq E[\dot{\mathbf{y}}\,\dot{\mathbf{y}}^T] = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \tag{6.47}$$

where E signifies the expectation operator and the elements of the matrix are defined by the relations

$$\sigma_{11}^2 = \lim_{N\to\infty} \frac{1}{N-1} \sum_{n=0}^{N-1} y'(n)^2$$

$$\sigma_{12}^2 = \sigma_{21}^2 = \lim_{N\to\infty} \frac{1}{N-1} \sum_{n=0}^{N-1} y'(n)\, y''(n) \tag{6.48}$$

$$\sigma_{22}^2 = \lim_{N\to\infty} \frac{1}{N-1} \sum_{n=0}^{N-1} y''(n)^2$$

a distance measure, $z(\dot{\mathbf{y}}, \Phi)$ was specified with the aid of the covariance matrix. This is sometimes called the *Mahalanobis distance*[12] and is similar to the conventional Euclidean distance, only the space is distorted so that vectors in different directions may contribute differently to z. The distance is defined by post- and pre-multiplication of the *inverse* of Φ by $\dot{\mathbf{y}}$ and its transpose, $\dot{\mathbf{y}}^T$:

$$z^2(\dot{\mathbf{y}}, \Phi) = \dot{\mathbf{y}}^T\, \Phi^{-1}\, \dot{\mathbf{y}} \tag{6.49}$$

When Φ is equal to the identity matrix ($z(\dot{\mathbf{y}}, I)$), the weighting is the same in all directions and the result is the usual Euclidean distance of the point $\dot{\mathbf{y}}$ from the origin. In this case points that have the same value $z = z_0$ lie on a circle of radius z_0, centred at the origin.

If Φ is a diagonal matrix ($\sigma_{12}^2 = \sigma_{21}^2 = 0$), so is its inverse and the operation is equivalent to squashing or stretching the space in one direction. Hence points that have the same value, z_0 lie on the ellipse

$$\frac{y'}{\sigma_{11}^2} + \frac{y''}{\sigma_{22}^2} = z_0^2 \tag{6.50}$$

In the general case when Φ is non-diagonal, the locus of points with $z = z_0$ is still an ellipse, but rotated about the origin. Here Φ^{-1} has the form:

$$\Phi^{-1} = \begin{bmatrix} u_{11}^2 & u_{12}^2 \\ u_{12}^2 & u_{22}^2 \end{bmatrix} \tag{6.51}$$

where

$$\begin{aligned} u_{11}^2 &= \tfrac{1}{\Delta}\sigma_{22}^2 \\ u_{12}^2 &= -\tfrac{1}{\Delta}\sigma_{12}^2 \\ u_{22}^2 &= \tfrac{1}{\Delta}\sigma_{11}^2 \end{aligned} \tag{6.52}$$

and $\Delta = \sigma_{11}^2\sigma_{22}^2 - [\sigma_{12}^2]^2$ is the determinant of Φ. The equation of this ellipse can then be derived from the definition of $z(\dot{\mathbf{y}}, \Phi)$ (equation (6.49)):

$$z^2 = u_{11}^2\,(y)^2 + 2u_{12}^2\, y'y'' + u_{22}^2\,(y'')^2 \tag{6.53}$$

by setting $z = z_0$. This equation describes the most general form of the distance and can be implelented easily in software.

The concepts just introduced provide a way to estimate the inner spike boundary. As seen, the covariance matrix of the background activity with respect to the space represented in the d-plot is indicative of the likelihood that a given sample belongs to that activity, rather

than to a spike or other phenomenon. In fact, it is possible to define C_i as the ellipse enclosing the background activity with a certain level of confidence, assuming that the data follows a known probability distribution. The structure of the distance measure strongly suggested a bivariate Normal distribution with zero mean and a covariance matrix Φ. Equation (6.53) may be regarded as the conversion from a generalized to the standard Normal distribution with $\Phi = I$. Then, z_0 may be found from tables as the value corresponding to a specific confidence limit, such that the resulting C_i encloses, say, 99% of the possible samples.

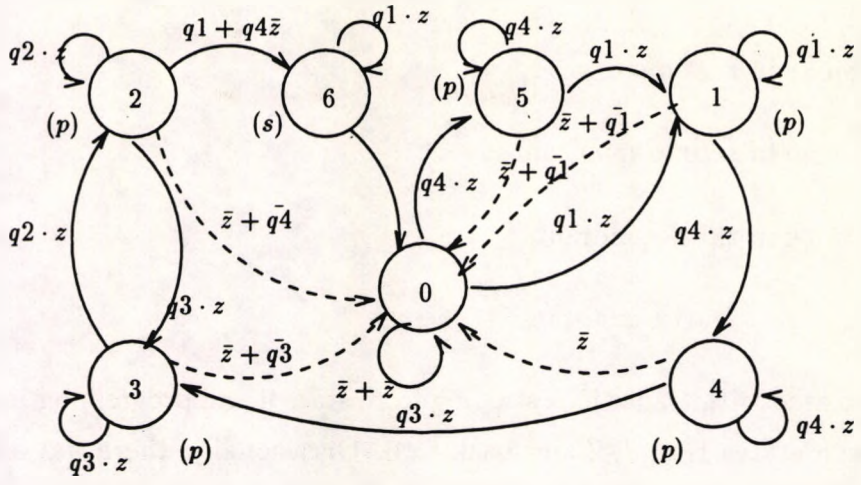
A better way is to estimate z_0 to minimize the number of false and missed spike detections (see also section 6.4.2 and Figure 6.5). A wide range of values for z_0 produced the same results, since the available spikes had reasonably distinct d-plots from those of the background activity and finally, a value of $z_0 = 1.96$ was selected lying approximately in the middle of the range for a specific record (A). This corresponds to a 95% confidence interval for the Normal distribution. There was no need to modify this for other records and it proved fairly robust for both sampling rates as well as for all differentiator structures.

The outer spike limit, C_o, cannot be defined in this manner, as the nature and effects of the d-plots of artifacts is unknown and spikes are not frequent enough to extract unbiased covariance information form them. It was, therefore decided to use the same function as that for the background activity and a higher confidence interval, z_1, so that spikes give a distance value between z_0 and z_1. It was also decided to make this limit stricter than a simple ellipse, by defining it as four values, one per quadrant of the d-plot. This allowed for deviations of spikes from a single ellipse (which are evident for Figure 6.6) to be taken into account for the rejection of high-derivative non-spike signals. Hence these values have been higher in the fourth and third quadrants, where derivatives tend to be higher, and lower in the first and second. It is difficult to define these limits optimally, but the values of 6 in the first, 8 in the second and 10 in both the third and the fourth quadrant, derived form a small number of spikes, were used with a good degree of success.

The use of piecewise elliptical boundaries required the additional computation of quadrant information. This does not involve computationally intensive operations, as it may be evaluated by considering the sign of the derivatives to determine the quadrant where a point lies.

6.5.6 Following possible spikes by a finite state automaton

The distance measure proposed for the definition of the region where spikes have their d-plot contours is the major element of the proposed spike detector, but its output is specific to a single sample and contains no information about the behaviour of the adjacent samples,



NOTES: $q1$, $q2$, $q3$ and $q4$ are the inputs corresponding to the four quadrants
Transitions from state 0 are associated with z^2 being in the spike region (corresponding to input z_b being true)
Outputs are: p =a spike is in progress, s =spike detected

Figure 6.8: The state-transition diagram used in the d-plot spike detector

which may not have similar properties. After all, a spike:

- constitutes a *succession* of samples with spike attributes.
- consists of a single wave, with the first half wave positive (electro-negative EEG potential inverted by the amplifier in the recorder) and the second negative with respect to the baseline.

Both conditions must be satisfied if a d-plot segment is to qualify as a spike. In other words once a sample is detected as belonging to a spike, subsequent samples should be tested as well and if the whole progression constitutes a *full oscillation* returning to the region of non-spikes, the detection of a spike is signalled. This solves the problem of the conversion of the continuous-time binary indication, defined by the distance measure, into a spike detection event, which is registered instantaneously.

Testing the aforementioned conditions suggested the use of a finite state machine. Its inputs are readily available as the attributes defined in the previous section, the indication of whether a sample is in the spike region (defined by the distance function) and the quadrant where the sample belonged. The former is binary, whereas the second may take one of four values, 1 to 4. The important conditions, as far as spike detection is concerned, are whether a sample is in the spike region and if it is, in which quadrant it lies. Using these, five states were defined:

- state 0: Non-spike (in background or artifact)
- state 1: Spike in first quadrant
- state 2: Spike in second quadrant
- state 3: Spike in third quadrant
- state 4: Spike in fourth quadrant

Spikes are associated with transitions starting from state 0 and progressing in the (clockwise) direction through states 1, 4, 3, 2 and back to 0. Occasionally, there was a transition from state 2 back to 3 and back to 2, which occurs on the maximum slope of some fast spikes, due to the differentiator reaching the limit of its bandwidth and being unable to follow faster transitions.

Two additional states, 5 and 6, were introduced to cover the eventualities that at the beginning and the end of a spike there may be a transition from the background region into the second and then the first quadrant. These were necessary to resolve the ambiguity whether the transition took place at the beginning or at the end of the spike.

The output of this state machine could take one of three values, indicating whether there was no spike, a spike in progress, or a spike detected, at the end of a successful cycle.

The state machine is, however, described better by the state-transition diagram of Figure 6.8 than by the preceding verbal description.

6.5.7 On-line estimation of system parameters

Evidently, parameters like Φ and Φ^{-1} are not known precisely. Their computation usually relies on estimates taken by ignoring the limit in the defining relations (see equation (6.48)). The resulting sum was taken over the N past samples including the present one.

$$\sigma_{11}^2(n) = \frac{1}{N-1} \sum_{k=0}^{N-1} y'(n-k)^2 \tag{6.54}$$

The index n on the LHS of this equation illustrates that the variance is updated on every sample. Alternatively, a block-update may be used to save computational time, but this is not necessary since a faster implementation of the given equation is possible. This is a simple recursive 'moving average' filter, a detailed explanation of which was given by Lynn[114][115, pp. 197–206] (see also section 5.6.2). It relies on the fact that the value of the summation at a given sample in time, n, is equal to the sum during the previous sample $(n-1)$ with the

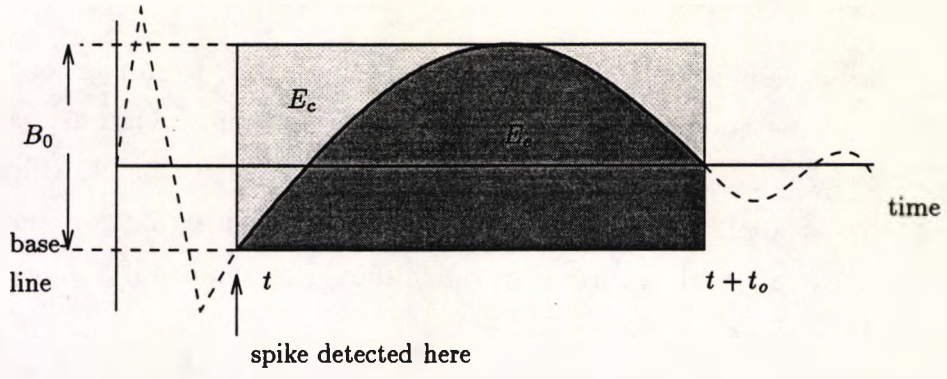


Figure 6.9: Model for slow wave

oldest sample $(n - N)$ subtracted and the new sample (n) added. This converts the above expression into the following, which requires a subtraction, a division and an addition:

$$\sigma_{11}^2(n) = \sigma_{11}^2(n-1) + \frac{1}{N-1}(y'(n)^2 - y'(n-N)^2) \tag{6.55}$$

The other elements of the covariance matrix, σ_{12}^2 and σ_{22}^2 may be evaluated by analogous relations. The price to pay in this implementation is that of extra memory to store the N samples of the input sequence, but the memory requirement for this application, where N was (arbitrarily) set to 1024, is not a serious limitation.

Other recursive methods for computing estimates of averages with limited memory requirements exist. Although the resulting estimators are often biased, they may be employed if memory availability in a small portable system is limited. Such methods are extensively used when crude estimates are needed, like in the case of adaptive filters, covered in a later Chapter.

The elements described so far proved adequate for detecting spikes. In the case of minor epilepsy, spikes are usually followed by 'slow waves'. This additional feature was exploited to improve the detection capabilities of the existing detector.

6.5.8 The Slow wave detector

Slow waves often succeed spikes, even in the interictal EEG of petit mal patients. They have a frequency of about 3 Hz (or rather a period of 333 ms, since they are not periodic phenomena). Attempts to model slow waves include the half-sinusoid by Comley and Brignell[45] and various exponentially-based functions by Stelle[185].

In this work, they have been considered as sections of a sinusoidal waveform of duration $t_0 < \frac{T}{2}$, where T is the period of the oscillation (see Figure 6.9).

Why use an area measure?

Traditionally, slow wave detectors are based on either low-pass, band-pass or matched filtering. The first two methods do produce an output higher than the normal values for the EEG, but are influenced greatly by periodic phenomena in their pass-band, which produce a higher output than isolated, non-periodic slow waves. Matched filtering does not have these problems, but requires a practical threshold, which should be avoided if a robust detector is to be constructed.

The area under the EEG signal, evaluated with respect to a given base-line for a duration t_0, is a non-periodic measure. It resembles more the matched filtering approach, although the use of the model solves some of the problems associated with thresholding and requires less computation. Such an area measure was derived from the definition of the model of Figure 6.9.

Description of the detection procedure

The slow wave was assumed to begin at the end of the spike, when the derivative of the signal returns to some small value. At this point, the EEG is not at the baseline, but below it, towards the 'trough' of the spike. This is the point where a spike detection is indicated starting the detection of the slow wave.

The idea of the detector is based on the ratio of the area under the model (heavily shaded area in Figure 6.9) to that of a constant signal of equal duration and an amplitude equal to the peak, B_0 (lightly shaded area).

The area under the model is given by the following integral:

$$E_{sw} = \int_0^{t_0} B_0 \sin \frac{2\pi}{T} t \, dt = \frac{B_0 T}{2\pi} \left[1 - \cos \frac{2\pi}{T} t_0\right] \tag{6.56}$$

Since the area of the constant signal is $E_c = B_0 \, t_0$ their ratio is given by

$$r_0 = \frac{E_{sw}}{E_c} = \frac{T}{2\pi \, t_0} \left[1 - \cos \frac{2\pi}{T} t_0\right] \tag{6.57}$$

To determine whether there is a slow wave following the spike, the area in the EEG for a duration t_0 after the detection of the spike, E_e, must be measured using the value of the first sample as a baseline and the maximum value of the signal in the same period, B_0, must be estimated. Of the ratio $r = \frac{E_e}{B_0 \, t_0}$ is close to the the theoretical, r_0, the occurrence of a slow wave is assumed, which indicates a genuine epileptic transient. The comparison of the two ratios was achieved by setting a maximum and a minimum limit above and below r_0.

The parameters of the model were selected to reflect the properties of a typical slow wave. The period T was chosen twice as long as the typical period of repetition of rhythmic spike-and-wave activity (667 ms). Sinusoidal oscillations at this frequency, when rectified, produce rhythmic slow wave activity at an approximate rate of 3 per second. The intervention of spikes of duration from 20 to 70 ms meant that t_0 should be restricted in the range between half of T with these extra durations subtracted, that is between (333-70)=263 ms and (333-20)=313 ms. In practice, somehow wider limits (230 and 370 ms) were assumed, allowing for the usual 'tolerance' in the formal definition of the slow wave.

The energy in the slow wave was computed as the summation of the value of the sampled EEG with respect to the defined baseline, multiplied by the sampling period. The error in the approximation of the integral is not large, since there are at least fifty samples (for the sampling rate of 160 $s^{-}1$), which make the benefits of interpolation marginal. Care was necessary for the determination of the peak, B_0. In order to prevent over-estimation, which would make the method unreliable, the EEG was passed through a second order Chebyshev low-pass filter with a cut-off frequency of 4 Hz. This removed high-frequency noise, sharp transients and artifacts for the purpose of the detection of the peak. The energy measure E_e was computed directly from the unfiltered EEG (see Figure 6.10).

The capability of running *more than one* slow detector at any time was included in the program, as a precautionary measure in the event of one or more (false) detections of a spike, followed by a genuine spike-and-wave complex. If a single slow wave detector was only able to run, there would have been a danger of it not completing its course by the time the real spike was detected, hence the genuine slow wave could be missed. The ability to run up to sixteen slow wave detectors simultaneously, each with its own parameters eliminated the risk of missed detections, should the aforementioned eventuality arise.

6.5.9 Experimental Results

The complete spike and wave detector was tested with the same four EEG records used in section 6.4.3. The system here has fewer variables and, despite being more complex to implement, was more robust and easier to test. The signals associated with its operation are demonstrated in Figure 6.10, which shows the behaviour of the detector in the presence of a spike-and-wave complex. The differentiator had a length, $N = 31$, limited by a Hanning window. The sampling rate was 1000 s^{-1}. The first and second derivatives are displayed scaled by 0.01 and 0.0001, respectively, whereas z^2 by 50 and the quadrant signals by 20. The detection signal contains a spike detection (small impulse) and a spike-and-slow-wave detection (larger impulse) as well as a low-amplitude pulse, whenever the wave enters the

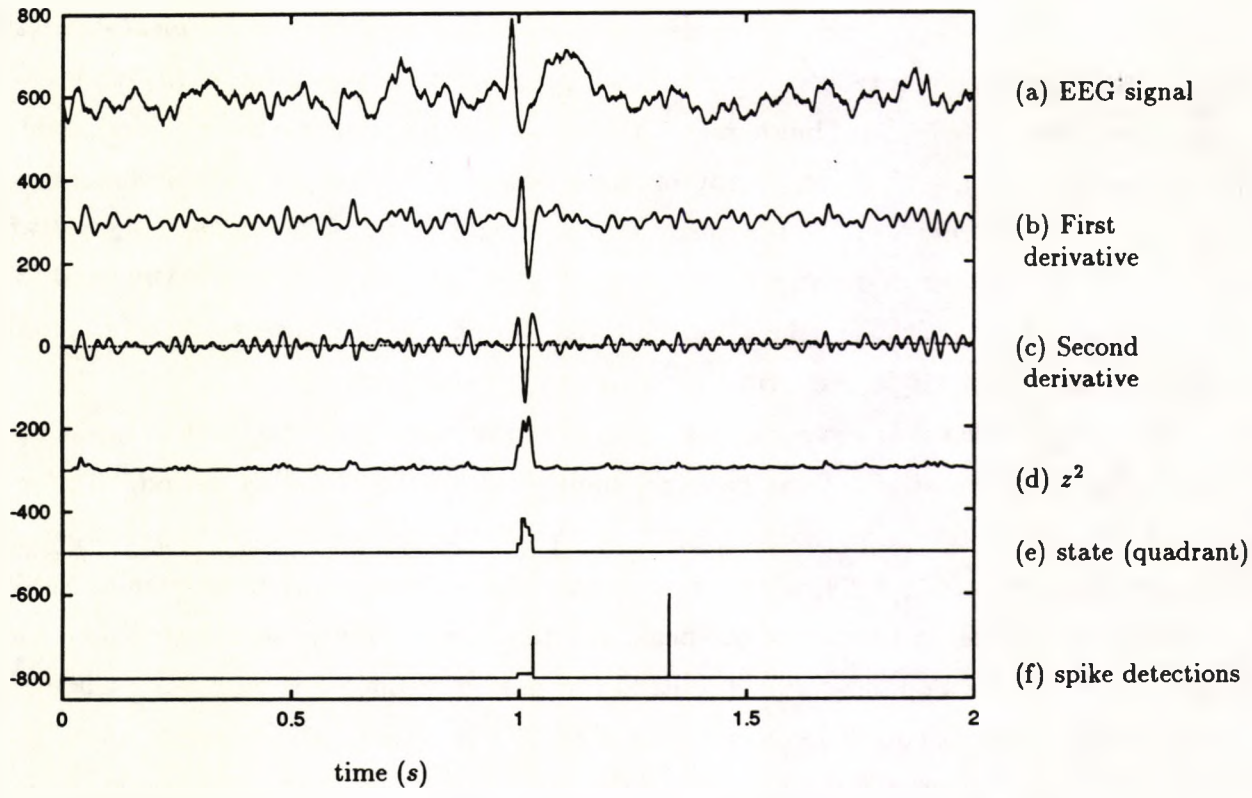


Figure 6.10: Example behaviour of the d-plot spike-and-wave detector

spike detection cycle.

The only element of this detector made variable during testing was the order of the differentiator and the sampling rate. The forward-backward difference, the simplest operator is equivalent to a frequency-specified structure with order 3. The others had orders 9, 15 and 31. The same data records were produced at both the high sampling rate (1000 s^{-1}) and the low one, (160 s^{-1}). A summary of testing results is shown in Table 6.2.

6.5.10 Discussion, comments and observations

The results presented in Table 6.2 permit the comparison of the effectiveness of the various differentiators. The first observation is that the type of differentiator has only a marginal effect on the number of detected spikes, evident especially in record A. Artifact rejection is better for higher orders, when the sampling rate was 1000 s^{-1}, whereas the improvement for the sampling rate of 160 s^{-1} was smaller. This is attributed to the improved reduction of high-frequency components, which are more susceptible to muscle artifacts, with increasing order. The problem is almost absent for the lower sampling rate after the effect of the anti-aliasing filters prior to discretization, as already explained earlier, in the discussion of the results of the multiplicative differential detector.

Record						A		B		C		D	
Duration in seconds						210		130		123		242	
Number of spikes present						52		8		10		0	
Muscle artifacts present?						some		yes		yes		many	
System parameters						Detections							
#	Diff/tor Name	f_s (Hz)	f_c (Hz)	N	Win	√	×	√	×	√	×	√	×
1	D03_H100	1000	500	3	Hann.	47	2	7	8	8	11	0	57
2	D09AR100	1000	500	9	Rect.	46	3	6	8	7	42	0	61
3	D15AH100	1000	500	15	Hann.	46	1	8	3	9	6	0	46
4	D31AH100	1000	50	31	Hann.	46	1	8	3	10	7	0	30
5	D03_H016	160	80	3	Hann.	47	1	8	3	9	7	0	28
6	D09_H016	160	80	9	Hann.	46	1	7	2	9	6	0	24
7	D31AH016	160	50	31	Hann.	46	1	8	2	10	6	0	24

Table 6.2: Results of the d-plot spike-and-wave detector

The results presented are for the complete system, including the slow wave detector. The use of the latter caused a number of genuine spikes to be missed, because the trailing slow waves were not large enough to cause a detection. Increasing the slow wave detection margins by 50% improved the detection of spikes (with only 1 or 2 spikes missed in each of the records A, B and C), but worsened the rejection of artifacts (almost doubling the number of false detections in record D). Even with this modification, the d-plot detector was better than the initial detector, whose results are listed in Table 6.1.

The advantages of the d-plot over the earlier multiplicative spike detector are immediately apparent from a simple comparison of the two. Firstly, the a priori estimation of a detection threshold which was an essential requirement for the operation of the earlier detector, is no longer necessary. The threshold is now fixed, justified by certain probability criteria and assumptions. The results confirmed that these assumptions (normal distribution of background activity and localization in the d-plot) were not far from reality. Using the covariance matrix to normalize the raw spike detection measure increased the robustness of the system in the presence of variations in the signal strength, thus reducing the effect of these on the successful detection of spikes. This was not possible in the earlier detector, which was susceptible to power variations in the signal.

Another improvement, evident from the results in the presence of artifacts is the dramatic

reduction of false detections. This is particularly highlighted in the results for record D and was attributed to the following reasons:

- More information is utilized in the new detector, as the d-plot treats the derivatives separately, whereas the earlier detector combined them into a single product, with the inevitable loss of useful details.

- The finite state automaton monitored the behaviour of the signal and thus limited false detections by rejecting signals that do not follow a general 'spike cycle'. This was not a feature of the earlier detector, where multiplication and filtering destroyed the fine structure of the derivatives.

- The detection of a slow wave, possibly following detected spikes, acted as a confirmation to the correctness of the detection of an epileptic event. 'Spikes' not followed by slow waves, such as muscle artifacts were rejected at this stage.

It should be noted that some rudimentary artifact rejection was already included in the earlier detector, when the interval between detected spikes gave an indication to their nature (genuine spikes are either isolated or separated by at least a third of a second). In the new spike detector, artifact rejection is more sophisticated and is performed at three levels. At differentiator level, with the suppression of high-frequency components, likely to be contaminated by artifacts. The restriction of bandwidth was incorporates in the structure of the differentiator and/or the selection of the sampling rate. The other two levels have already been highlighted. At the spike detection level, the use of the finite-state automaton and at system level with the incorporation of the slow wave detector.

6.6 Concluding remarks

A number of general observations on the use of derivatives for the detection of spikes will conclude this section.

Results indicated that *accurate differentiation was not critical* in the success of either of the spike detectors. The crude differentiator with an order of 3 compared favourably with the more accurate higher order ones, especially at high sampling rates. Bandwidth appeared to be more critical. If left unconstrained and the sampling rate is high (1000 s^{-1}) the operation enhanced spiky artifacts at the top end of the spectrum more than it did spikes. To prevent this, workers that have applied techniques based on derivatives found necessary to perform low-pass filtering either before[37, 107, 181] or after[45, 108] differentiation to remove 'noise'

and high-frequency spikes due to artifacts. The application of such a filter is likely to distort the output of the differentiator and the end result is *not the derivative, but the output of the differentiator in cascade with a low-pass filter.* This operation raised the question whether differentiation as a process, was an essential element of the detector.

The use of band-pass filtering[141, 155] and matched filtering[196, 197] for the detection of spikes and other transients provided an alternative explanation. Spikes are wide-band signals having most of their spectral energy in the vicinity of their fundamental frequency. Band-pass and matched filtering attempt either the reduction of frequency components outside the region where spikes are more prominent, using the remaining energy as a spike indication, or compute this indication by enhancing the frequencies corresponding to spikes (matched filer). Indirectly, methods based on derivatives may be explained in terms of band-pass filtering, since differentiation is a special form of high-pass filtering. When this is coupled with a noise-reducing low-pass filter, the end result is a band-pass operation. Similarly, the behaviour of a band-pass filter in the frequencies below its pass-band may be regarded as a crude differentiator, since it enhances high frequencies and suppresses those close to zero (although not necessarily in a linear manner). In other words, spike detectors based on entirely different principles may be quite similar in the end. This is possible, since some spike properties are used explicitly during the development of a method, whereas design often relies on the additional implicit use of other. Detectors based on derivatives, band-pass filtering and matched filtering indicate exactly this similarity. The former use the large derivatives on the edges and the sharpness of the peak of the spike as the criteria for spike detection. The second uses frequency properties, such as the spectral energy, to define a 'pass band' for spike, which is incorporated as a filter for their detection. A similar filter is constructed in the latter case, but this reflects the spectral properties of an 'average' spike model. All methods work to some extent, but other properties are often used but not stated explicitly. In the first class of detectors, to prevent 'noise' from differentiation of high frequency components, low-pass filtering is performed on the signals. Thus the combined operation of differentiation and filtering is effectively a band-pass operation. In the second case, the band-pass operation resembles differentiation of the low-frequencies, whereas in the latter, the contribution of the more detailed spectral shape of the matched filter compared to the band-pass is often small and the results similar.

This duality of explanation is characteristic of the argument. It is not the precision of differentiation as an operation that is critical, but the isolation of most of the energy of spikes and the suppression of other signals spectrally different from them!

This observation put spike detection in a different perspective. In the Chapters that follow

spike detection is viewed in general terms of signal modelling, rather than as a particular way of extracting one subset of its characteristics or another.

Chapter 7

The use of a detailed model

7.1 Introduction

The spike detectors of the previous Chapter have not been unsuccessful. Their conception was, however, based on neurological descriptions, which are biased by the appearance of spikes on a conventional paper EEG recording. Spike properties, such as 'sharpness' 'pointed peaks' and 'abruptness' are relative, often subjective, and could be distorted by changes in the time-scale or the sampling rate. The problems of using such explicit properties to describe the pattern were made apparent in section 4.6.1 and were highlighted in section 4.7. Because these properties may be expressed in many ways and working spike detectors often use other implicit ones, it is not unlikely to find that the same underlying property was employed to develop different spike detectors. An example of this duality in interpretation was given in section 6.6. A spike may be regarded as a signal with high derivatives, compared to normal EEG, when derivatives are the basis for spike detection. On the other hand, it viewed as a signal occupying a certain frequency band, when band-pass filtering is used. These apparently different interpretations are consequences of the spectral properties of spikes, which occupy the high-end of the EEG spectrum.

In an attempt to prevent the use of such secondary properties during detection, the question "what constitutes a spike?" was answered in a general way, using somehow abstract tools, like signal modelling and system analysis. This approach has numerous benefits:

- The detection of spikes could be hierarchical, starting with some general method and refining it by developing its elements later. This approach is flexible in the sense that one element may be improved later without affecting the others, since the basis of them all is free from design details
- Because of the formal methodology, the EEG signal could be described effectively and

generally, without the need for specific tools that may determine the analysis method.

- There is the possibility to apply the methodology developed for the detection of spikes to the identification of other transients, since spike-dependent elements should not have been encountered in the general development of the detector, but only in the realization later.
- Useful properties of other techniques may be unified under a common abstract methodology.

The main argument against these techniques might be their computational complexity, often resulting in their inability to operate in real-time, deeming them unsuitable for on-line operation. This problem, however, is marginal. Once a satisfactory method is developed, the computer code may be optimized and, if necessary, a faster microcomputer or a digital signal processor may be utilized. Failure to achieve real-time operation despite these suggestions does not discredit the developed method, since there is a tendency for hardware to get faster and better numerical methods are always under development.

In this Chapter the EEG will be broken down into a number of signal components. Each one will then be modelled in a rather general way and thus an overall model of the EEG signal will be constructed. The elements of this model will then be refined and a linear model will be presented, using linear signal and system theory, which will form the basis for a method for the identification of the spike component. Such an analysis method will be proposed in the second half of this Chapter. The material presented here was intentionally left on a rather general level, although refinement always results in some specialization. An implementation is proposed in later Chapters.

7.2 Components of the EEG signal

7.2.1 Where neurophysiology and signal theory meet

Since the early days of electroencephalography, neurophysiologists identified a number of patterns and conditions that may be encountered in an EEG. These were labelled with descriptive names based on their appearance on a conventional EEG recording. Since these were primarily for communicating ideas between experts, especially in the form of clinical reports, emphasis was paid to the standardization of the nomenclature (section 2.5), rather than to the attachment of precise definitions to the terms used. After all, the descriptive nature of the names and the experience of the people that use them was often adequate for diagnostic purposes.

These practices are, however, incompatible with the explicit detailed definitions necessary for the analysis of the signal by automatic means. It does not take long to realize the extent of the gap in what is offered by neurology and what is required for signal analysis (see Chapter 4). The definitions of the former are general, often subjective, based on descriptive qualitative properties, rather than accurate, objective measurements and lack detail and rigour. Conversely, the second requires definitions that are specific, objective, in the sense that they ought to be observer-independent, quantitative, based on accurate measurements and detailed, having the form of either functional or statistical origin and in any case governed by the strictness of mathematics (Chapters 4 and 5).

One might wonder how could a system be designed using the formalism of signal processing with a sound theoretical basis whilst maintaining a close relation to existing neurological definitions. Achieving this, could act as a bridge for neurophysiologists to gain some understanding on how the automatic EEG analyzer functions and possibly assist in the evaluation and the improvement of its constituent parts, expected to be more beneficial compared to it being treated as a 'black box' with the neurophysiologists simply assessing its performance.

Putting medical definitions in a signal analysis context proved easier than originally anticipated. Despite their imprecision, terms like *background activity*, *transients*, *spikes*, etc. resembled signal processing concepts, permitting their modelling by statistical and system analysis elements, as will be demonstrated in the next section. Although it is possible that the formal definitions do not fully comply with their neurophysiological counterparts, improvements in the former are possible with the use of more elaborate, or different tools, whereas the consistency and precision of the latter, which was disputed by some[16, 76], may be improved through this interaction.

Being able to produce such definitions which solved, to some extent, the communication problem between the world of signal processing and that of neurophysiology, permitted the representation of the system as a block diagram, hiding the functional details from the neurologists, but annotated with terms familiar to them. Similarly, its function could be described, generally, without using signal analysis terms. This could serve as a platform of interaction of neurologists with signal analysts and for the evaluation of the performance not only of the system as a whole, but also of its main elements.

The list of terms related to EEG description is quite extensive (see [41] and section 2.5); since epileptic spike detection is the main target of the work presented here, only the relevant ones were considered, namely the background activity, spikes, other transients, artifacts and noise. An attempt is made in the rest of this section to highlight those elements of the definition of each term that characterize and distinguish it from others, as these were used

for their modelling in the next section.

7.2.2 The ongoing component—background activity

According to its definition by IFSECN[41] background activity is

> *any EEG activity (sequence of waves) representing the setting in which a given normal or abnormal pattern appears and from which such pattern is distinguished.*

The term is used to describe the component of the signal that is always present as a consequence of the underlying neuron activity, in the absence of any specific stimulus. It is an indication of the complex mental processes that persist even during unconsciousness. This ongoing signal component is by no means fixed in its bahaviour neither over the surface of the scalp nor over time. Although it may be regarded as quasi-periodic, having the form of repeated oscillations which are sometimes almost sinusoidal, it exhibits changes in both its frequency and amplitude. Frequencies tend to increase during maturation with consequent decrease in amplitude of the background activity (section 2.5.1). The opposite effect is observed with the level of consciousness, with frequencies decreasing as sleep gets progressively deeper (section 2.5.1). Changes in its appearance are also noticeable in certain brain disorders, like schizophrenia as well as after the administration of drugs.

Although this activity was subdivided into four frequency bands delta, theta, alpha and beta (section 2.5.1), there is no clear separation between them in the EEG, which has a rather smooth frequency spectrum. Hence a continuous frequency method for its description was considered more beneficial than the discrete classification into frequency bands.

It should be stressed here that the background EEG activity suffers frequency and amplitude variations that may persist for relatively short periods as well. Such changes are, for example, observed with eye closure and opening (which gives rise to increased amplitude activity in the α frequency range) and with intense mental activity (with the appearance of higher frequencies and subsequent reduction in amplitudes). Despite their transient nature, these variations are generally distinguished from most transients, since they persist for relatively longer periods of time and their onset and offset are more gradual.

Transients, described next, have characteristic shapes and often appear spontaneously and cease quickly, although context is sometimes necessary for the distinction of background and transient activity.

7.2.3 Spikes: the transients of interest

A spike, a term used to describe a class of patterns that are characteristically associated with epilepsy, is defined formally[41] as:

- a spontaneous, transient signal whose amplitude is (usually) clearly distinguished from background activity
- with a pointed peak at conventional EEG paper speeds
- having a duration from 20 to under 70 ms
- with its main component negative relative to other areas
- and variable amplitude from one occurrence to another.

Commentary on this definition, however, stresses the fact that although spikes ought to be distinguished from sharp waves, which have similar properties but longer durations (up to 200 ms), the discrimination between the two is rather arbitrary and serves descriptive purposes only.

Generally spikes are not observed in normal subjects, whereas they are characteristic of many types of epilepsy (see Chapter 3). Sometimes they are associated with other components, typically a slow wave, which has sinusoidal appearance and duration longer than $\frac{1}{8}$ s. the resulting patterns are known as spike-and-wave complexes. During a seizure these patterns may occur in a rhythmic succession. The focus of this work is primarily on isolated instances of these patterns, which occur between seizures. These are fairly rare in occurrence and the reduction of their numbers in a given period of time is used as an indication of improvement of the condition of a sufferer, during a course of treatment. These interictal patterns appear spontaneously with no warning and may be induced by various stimulation techniques (section 2.4.3).

One important attribute of spikes, as well as other transient patterns, is their superposition on the background activity. It is quite possible, of course, that the subcortical mechanisms and origins of both transients and the ongoing signal elements of the EEG are related but this is not prominent in the EEG and the two components may be regarded as generally independent for signal analysis and neurophysiological purposes alike.

7.2.4 Other transients

Spikes are not the only transients present in the EEG. In fact any isolated wave or complex that is clearly distinguished from the background activity may be viewed as a transient

(section 2.5.2). Although isolated slow waves are often indicative of some pathological condition, other transients may result from normal brain operations. Among these are the lambda waves, related to visual exploration and the kappa rhythm, which characterizes intense mental activity. One may regard responses to stimuli (*evoked potentials*) which have specific shapes, but are often quite small to detect directly in the EEG as transients.

Transients, like spikes, were considered to be waveforms that appear spontaneously, being distinguished from the background activity and disappearing at the completion of their usual 'cycle' leaving no distinct trace of their occurrence.

Unlike the background activity, which may exhibit drastic changes depending on the age of the subject, level of consciousness etc., transients have *definite forms*, though their morphology is not always easy to quantify with mathematical tools, due to their imprecise definition and variability. As a consequence, the methods used for their description ought to be flexible enough to allow for their variability, but it must have constraints imposed on their parameters to reflect the similarity from one instance of a specific transient to another. In the case of the background activity unrestricted models were considered preferable, to reflect the more loose definition of this component of the signal.

7.2.5 Noise and other artifacts

Apart from the genuine EEG components, which result from the underlying neuron activity, there are other components picked up by the electrodes on the scalp which have their origins outside the brain. These are called artifacts and their sources and description were covered in section 2.5.4. These signals are often orders of magnitude larger than the brain rhythms and are the result of either external electrical sources, equipment and electrode faults or biological electrical sources.

External electrical interference is caused primarily by the mains frequency being picked-up by the electrodes. The amplitude of this artifact is variable and depends on the proximity of its source to the subject and equipment. In a portable monitoring device this is expected to be problematic, since the subject is allowed to walk freely in a possibly electrically unscreened environment, unlike the EEG laboratory. The narrow-band nature of this signal (sinusoidal at either 50 or 60 Hz) enables its removal when present by the use of notch filtering or other means. Its characteristic frequency and regularity make this artifact easy to identify. A more variable external artifact is caused by the build-up of electrostatic charges on the subject and is more difficult to eliminate (section 2.5.4).

Equipment fault artifacts are characteristic of the EEG recorder. The hardware of the portable monitor are simpler than those of a conventional EEG device, with most mechanical

parts removed. Hence defective pen and switch artifacts are simply absent although the deterioration of the amplifier characteristics with ageing cannot be avoided. Some self-testing and calibration could be incorporated as a hardware-software sub-system, but this is beyond the scope of this work.

Electrode artifacts are expected to cause more problems with a long-term portable monitor than they would in the EEG laboratory, since it may be unsupervised, possibly operating for long periods of time, factors that increase the probability of faults developing. When the electrolytic jelly at the electrode contact gets dry, its electrical resistance increases and the signal diminishes in amplitude. Similar effects are caused by sweat in the area of an electrode. Electrode movements with respect to the scalp cause changes in the DC. potential and AC coupling prior to amplification results in exponential decays of large amplitude.

Finally artifacts arise from biological sources of electrical potentials within the patient's body. Ocular artifacts, caused by eye movements and blinking are transients including a range of characteristic shapes some of which occasionally mimic cerebral activity (section 2.5.4), like delta waves, sharp waves, etc. They are more prominent if the recording electrodes are close to the eyes. A similar artifact is caused by the pick-up of the cardiac potentials (ECG artifacts). These can generally be prevented by avoiding the placement of electrodes near arteries.

Perhaps the more difficult artifacts to prevent are those caused by muscle potentials. They are particularly frequent in anxious patients and in long-term monitoring plenty are expected when the patient carries on with his/her daily functions, such as speaking and eating. Being a constant headache to spike detectors because of their resemblance to spikes, these potentials contaminate the higher-end of the EEG spectrum and impair even its visual interpretation. There are no simple ways for their elimination, as they consist of fairly irregular stretches of 'spiky' potentials of large amplitude.

Finally, noise from various sources affects the EEG signal. Some is directly induced by the equipment, other by contacts of the electrodes with the scalp some even during the propagation of the waves through the tissue of the head. After sampling, quantization noise is also introduced into the signal samples, its level dependent on the accuracy (number of bits) of the analogue-to-digital converter used. Noise may also be introduced by the various numerical operations during the process of the signal (e.g. digital filtering), due to numerical rounding-off, especially after multiplications and divisions. Its effect on the overall levels of noise was not critical in this study, since no apparent change occurred whether simple or double precision operations were carried out. In a practical realization on a small portable microcomputer, however, where fixed-point arithmetic could be employed this source of noise

may be considerably higher and need to be taken into account.

Having identified the main components of the signal it is possible to construct a general model for it, which could assist in the development of a method for its interpretation.

7.3 Linear modelling of EEG signal components

Models of the EEG signal have been used before in this report. In Chapter 5 and especially in Chapter 6 a number of models were proposed. Despite being used for the development of spike detectors, their purpose there was primarily the gathering of relevant intuitive properties, such as 'sharp transition' and 'pointed peak' with mathematical operations, such as the derivatives of the signal with respect to time. Those models were not , however, explicitly utilized as part of the analysis. They served as a simplified and concise description of the signal and for the formulation of arguments for and against a proposed procedure for analysis of the EEG. The association of the signal models described in this section with the analysis procedure that follows is much closer. In fact the proposed models are an integral part of the analysis.

The decomposition of the EEG signal suggested in the previous section, highlights the absence of homogeneity in its components. The on-going activity, spikes, other transients and noise certainly differ from one another, both in their structure and composition as well as in their duration of occurrence. Modelling each component separately before putting them together to formulate an overall EEG model was intuitive for constructing the composite signal, but for its decomposition, which was the principle of its analysis as well. Despite the differences between these components of the signal, it was possible to use similar tools (at least in principle) for their analysis. The question was how to parametrize signals that have changing spectral properties with time, the background activity, and others that have relatively fixed sub-patterns, occurring at unknown instances, like spikes and other transients. The further decomposition of each one of these components, $y(nT)$ into a signal of simpler structure, $x(nT)$, and a system, H, was the basis of their description. According to this method, an estimate of the required signal, $\hat{y}(nT)$ may be constructed by passing $x(nT)$ through the system H, as shown in Figure 7.1.

For this idea to have any practical use, there must exist a method for determining $x(nT)$ as well as H unambiguously. The purpose of this decomposition is to 'trap' a number of properties of $y(nT)$ into the system H and thus leave only the ones, like time of occurrence in the case of transients, which are important for their detection. The composition and structure of $x(nT)$ is characteristic of the nature of $y(nT)$. In the case of transients it

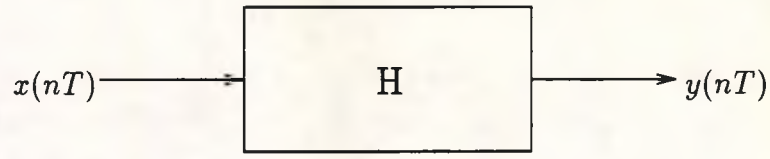


Figure 7.1: Principle of signal modelling

must be indicative of the instances of occurrence of the pattern in question. In the case of background activity, it could characterize changes in the behaviour of the signal, like increase in amplitudes and in frequency content. In this case it should have a general supportive role, to maintain the system H active, if the structure of the latter is such that its output diminishes in the absence of input. This would ensure that some on-going element is always present, like in the real signal, and to allow amplitude variations to be decoupled from H, thus simplifying its structure. It will become apparent later on in this section that the power spectral density of $x(nT)$ for both on-going and transient components have similarities.

At this stage, there is no need to impose any restrictions on the structure of H, which may be non-linear as well as time-variant. Assumptions of linearity and time-invariance, however, permit the use of a rather broad set of tools for analysis and treatment, since superposition, convolution and Fourier analysis may then be employed. These are quite important because there is a well-founded notion of an *inverse model* whose existence is central in the development of the signal identification procedure that follows in section 7.4. A simple example of this modelling approach follows. Let $y(nT)$ be a sampled sinusoid, $A_0 \sin \omega_0(n - n_0)T$, for $n \geq n_0$. This may be produced by impulse excitation of a discrete-time resonator. That is, $H(z) = \frac{\sin \omega_0 T \, z^{-1}}{1 - 2 \cos \omega_0 T \, z^{-1} + z^{-2}}$ and $x(nT) = A_0 \, u_0[(n - n_0)T]$. In this example, H expresses the (sinusoidal) behaviour of the signal, $y(nT)$, (in the form of the frequency, ω_0), whereas $x(nT)$ is indicative of its amplitude, A_0, and the moment the signal acquires its properties, $n = n_0$.

After this introduction, the principle outlined here will be used to derive models in the form of pairs (H, x) for the background activity, spikes and other transients and noise components. The models for H were presumed to be linear and time-invariant (LTI), although the second assumption will be relaxed later on. It should be stressed that the restriction to LTI systems is a practical one and could be eliminated when suitable non-linear techniques become available.

7.3.1 Model of the background activity

From the brief description of the appearance of the background activity (section 7.2.2), it is evident that this signal component changes with time, sometimes being almost periodic, even sinusoidal and sometimes more irregular, almost random. These changes are usually gradual, sometimes abrupt, but their effects persist for some time and are generally long-lasting, becoming apparent during visual inspection of the EEG record. This, however is an oversimplification of the processes, since small changes and variations in the properties of the signal could be missed by visual examination. Moreover, attempting to identify and label (classify) sections of the signal as belonging to different states assumes that the signal consists of distinct segments, whereas the transition from one to the next may be gradual. All these cause serious complications in the modelling procedure, which has to demonstrate, according to the aforementioned technique, how this signal may be produced when a system is excited by another signal with simpler properties.

This is not possible from a general standpoint, since the main characteristic of the background activity is variability. It is easy to prove that the statistical properties of the signal vary with time. The mean of the signal is zero (a.c. coupling of the EEG has been assumed to remove large low-frequency potentials often caused by the electrodes as described in section 2.4) and has no statistical significance. A useful statistical measure is the autocorrelation function (sequence) of the signal, $b(nT)$:

$$r_{bb}(n,k) = \sum_{i=-\infty}^{+\infty} b(n+i)b(k+i)p[b(n),b(k)] \tag{7.1}$$

where the (constant) sampling period T has been dropped for simplification of notation. Because the mean is zero, this is also equal to the auto-covariance of $b(n)$[152, p. 10]. The above expression does not put any restrictions on the nature of the signal, as it changes with time to reflect the changes in the properties of $b(n)$. The probability distribution function $p[b(n),b(k)]$ of the discrete (random) variable $b(n)$, could be considered uniform, in which case it becomes a universal scaling factor (like in ergodic processes)[152, p. 12].

The auto-correlation sequence of the background activity is not independent of time. It is possible, however, to investigate its behaviour by making certain assumptions, for founding and studying a model, to provide an initial interpretation of the analysis method. As will be shown in the realization of the method, some of these may be relaxed, increasing its versatility and generality. Finding a *tool* to model the signal is hindered by the non-stationarity and continuous variability of $b(n)$, which does not have even constant statistical properties. Despite their variability, these properties tend to suffer large changes relatively infrequently, compared to the signal, say every few seconds, whereas in the period

of time between these changes, variations tend to be small and gradual. Assuming they are negligible, one may consider the background activity as a cascade of segments of signals, $B = \{\ldots, b_0(n), b_1(n), b_2(n), \ldots\}$ each one of which has different, but fixed, statistical properties. These segments do not necessarily have the same 'length' (duration). A convenient way to view this process is by considering that the elementary signals in B extend beyond their scope, but they are restricted to a finite duration by the use of a window,

$$w_N(n) = \begin{cases} 1 & 0 \leq k < N \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

where N, the duration of the window is equal to the length of the segment in question. Evidently, the background activity signal may then be constructed from the set $B = \{b_i(n); i \in \mathbb{Z}\}$ using a corresponding set of windows, $\{w_{N_i}; i \in \mathbb{Z}\}$.

$$\begin{aligned} b(n) &= \ldots + b_{-1}(n)\, w_{N_{-1}}[n + N_{-1}] + b_0(n)\, w_{N_0}(n) \\ &= b_1(n)\, w_{N_1}(n - N_0) + b_2(n)\, w_{N_2}[n - (N_1 + N_2)] + \ldots \end{aligned} \tag{7.3}$$

where N_i signifies the length of the ith segment. In order to be precise, the windows may be made to extend for a small number of samples, M, in both directions, so that there is some overlap between the segments, thus resembling the real signal more closely. These 'tails' of the window should have unity value where they join the central part of the window and diminish to zero in a smooth way to guarantee continuity in the transition from one segment to the next. There is no need to discuss this in detail as it does not comprise a significant element of the methodology outlined here, but windows were discussed extensively in section 6.5.1 (see also Appendix B).

Having represented $b(n)$ in this manner, it is possible to refer to the properties of each segment for the construction of the required model. Each of the $b_i(n)$ has zero mean, since by definition $b(n)$ has no constant component. The autocorrelation sequence for the period of the segment is

$$\begin{aligned} r_{b_i b_i}(n,k) &= \sum_{i=-\infty}^{+\infty} b_i(n+i) w_{N_i}(n) b_i(k+i) w_{N_i}(n) p[b_i(n), b_i(k)] \\ &= \sum_{i=0}^{N_i} b_i(n+i) b_i(k+i) \end{aligned} \tag{7.4}$$

since $p[b_i(n), b_i(k)]$ is independent of n and k. This implies that the autocorrelation function is only dependent on the time difference between the samples and hence simplifies to the following expression:

$$r_{b_i b_i}(n,k) = \frac{1}{N_i} \sum_{n=0}^{N_i} b_i(n) b_i(n-k) \tag{7.5}$$

$$= r_{b_i b_i}(n-k) = r_{b_i b_i}(k-n) \tag{7.6}$$

which holds for real, stationary signals[152, p. 7]. As a consequence of the Wiener-Khintchine Theorem[151, p. 260], a power spectral density, $S_{b_i b_i}(\omega)$ exists for the segment which describes its second-order statistics (auto-correlation). Moreover $S_{b_i b_i}(\omega)$ may be factored in the form:

$$S_{b_i b_i}(\omega) = H_i(\omega)H_i^*(\omega) = H_i(\omega)H_i(-\omega) \tag{7.7}$$

where H_i^* denotes the conjugate of H_i. Remembering that the Fourier Transform of a linear invariant (LTI) system, is found by evaluating its transfer function, $H_i(z)$, around the unit circle $z = e^{j\omega T}$, one may write:

$$S_{b_i b_i}(\omega) = H_i(z)H_i(z^{-1})]_{z=e^{j\omega T}} \tag{7.8}$$

In this manner, the statistical properties of the signal $b_i(n)$ have been summarized in a transfer function, $H_i(z)$, that is a linear system. It remains to determine the structure of this transfer function and the input necessary to produce a signal with the same properties as those of $b_i(n)$.

The transfer functions of a great number of linear systems are expressed as rational fractions of two polynomials,

$$\begin{aligned} H_i(z) &= \frac{B(z)}{A(z)} \\ &= \frac{\sum_{m=0}^{M} b_m\, z^{-m}}{1 - \sum_{n=1}^{N} a_n\, z^{-n}} \qquad (7.9) \\ &= b_0\, z^{N-M} \frac{\prod_{m=1}^{M}(z - z_m)}{\prod_{n=1}^{N}(z - p_n)} \qquad (7.10) \end{aligned}$$

The power spectral density of this function depends on the values of the coefficients of the numerator and the denominator. It may be factored, as shown in equation 7.10 and, depending on the location of its finite number poles, $\{p_n; n = 1, \ldots, N\}$, its stability may be controlled ($|p_n| \leq 1 \forall n$). In fact, the location of its poles and zeros, $\{z_m; m = 1, \ldots, M\}$, determine the shape of the power spectral density of a stable system. Poles are associated with peaks in the PSD. The sharpness and amplitude of these peaks increases when the poles approach the unit circle in z (the limit for stability). Zeros are associated with valleys in the PSD. These become deeper, when the zeros are near the unit circle in z. Using the converse argument, given a predefined PSD, $S_{b_i b_i}(\omega)$, having a finite number of peaks and valleys (a condition that is true for many signals, including the EEG), it is possible to construct a rational transfer function $H_i(z)$, by placing poles and zeros at appropriate locations in the complex plane, such that the power spectral density associated with it (equation (7.8)) is arbitrarily close to $S_{b_i b_i}(\omega)$.

In fact, for the work presented later the transfer function was simplified by assuming there are no zeros ($B(z) = b_0$), that is an *autoregressive* (AR) structure. With this model, spectral peaks may still be modelled accurately. Valleys are constructed naturally, between the peaks, but if deep valleys are present, extra poles are needed. For an accurate approximation a model whose denominator has higher order, N, than the pole-zero structure is generally required. A simple explanation of this effect utilizes the factored representation of the model (Equation 7.10)), rewritten in the form:

$$\begin{aligned} H_i(z) &= \frac{1}{A(z)/B(z)} \\ &= b_0\, z^{N-M} \frac{1}{\prod_{n=1}^{N}(z-p_n)\prod_{m=1}^{M}\frac{1}{z-z_m}} \end{aligned} \tag{7.11}$$

Rewriting $\frac{1}{z-z_m}$ in the form $\frac{z^{-1}}{1-z_m\,z^{-1}}$ and expanding using the MacLaurin series $\frac{1}{1-w} = \sum_{p=0}^{\infty} w^p$, the zero term results in the convergent infinite series

$$\frac{z^{-1}}{1-z_m\,z^{-1}} = z^{-1}\sum_{i=0}^{\infty} z_m^p\; z^{-p} \tag{7.12}$$

provided that $|z_m| < 1$. The higher-order terms of this series diminish rapidly and it may be truncated without great loss in the approximation, resulting in a polynomial of order P_m. As such, it may be written as a product of P_m factors. Expanding all terms in a similar way would generate additional factors, which have the position of poles in the resulting system. The number of additional poles required to obtain a satisfactory representation is considerably smaller than might be expected from this method of explanation, since many coefficients of the resulting polynomial of the denominator of $H_i(z)$ that correspond to high powers of z^{-1} are expected to be small and may be neglected, thus maintaining the order of the model still low. It should be noted here that it is the frequencies *present* in the EEG that are usually important (spectral peaks) and not the ones that are missing (spectral valleys). Hence the correct approximation of spectral valleys is not so critical anyway. The use of the all-pole model offers significant analytic and computational properties which is the reason for which its was preferred. These aspects will be emphasized later during the system implementation.

In conclusion, it is possible to construct a transfer function,

$$H_i(z) = \frac{b_0}{1-\sum_{n=0}^{N} h_n\, z^{-n}} \tag{7.13}$$

that has the same PSD as that of the ith segment of the background activity, $b_i(n)$. Since the PSD and the auto-correlation function contain the same information, the second-order statistics for the segment are indirectly related to $H_i(z)$. Therefore a signal, $\hat{b}_i(n)$ having the

same statistical properties as $b_i(n)$ may be produced by exciting $H_i(z)$ with an appropriate input, $u_i(n)$. This should have an auto-correlation,

$$r_{u_i u_i}(k) = \begin{cases} 1 & k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.14}$$

Hence its PSD is equal to 1 at all frequencies and does not 'interfere' with the PSD of $H(z)$, to which it is multiplied. Sequences that have this characteristic include the unit sample, $u_0(n)$ and the white 'noise' sequence of power 1. The first generates the impulse response of $H_i(z)$ in its output, whereas the second produces a random, correlated sequence. In general, the second type of input is more applicable to the EEG, since it produces a signal that is somehow irregular, whereas strong periodic components may be produced by a transfer function with a PSD that has sharp peaks at the corresponding frequencies.

Since $u_i(n)$ is a random sequence of unit power and this is true for all segments, i, it is unnecessary to consider them as inputs to individual segments. It is more practical to refer to a single input, $u(n)$, which simplifies the modelling procedure, as only the $H_i(z)$'s need to be changed at the end of every segment (in fact, there are transition periods during which the parameters of $H_i(z)$ gradually change to those of $H_{i+1}(z)$). Having an input that is segment-independent meant that segmentation is not visible in the overall model. It is hidden as the operation of changing the transfer function inside H at the end of each segment and maintaining it for the appropriate duration. The signal model of $b(n)$ can then be regarded as the output of the time-variant system H, when excited by a random white sequence $\{u(n)\}$ of unit power (variance).

This scheme generates a signal which has the general properties of $b(n)$. Given H, it is possible to construct signals with the same auto-correlation as $b(n)$. These may have a similar 'feel', but would not, in general, have a close sample-by-sample correspondence.

7.3.2 Spike process modelling

Interictal epileptic spikes are different from the background activity, as they occur at irregular intervals in the interictal EEG and have a characteristic general shape. Despite these morphological differences, they, too, can be described using an input-system pair of functions.

One property of spikes is their discrete occurrence. Being transients they appear at a moment in time, persisting only for a short period. Another is their rarity. Interictal spikes are infrequent, so that there is typically a long duration between one spike and the next. Spikes, have a specific form and although there are variations from one to another, these

are less dramatic than those during the background activity. For the moment these will be ignored, apart from an amplitude factor. In this sense, a spike, s_0, occurring at an instant $n = 0$ consists of a (finite) sequence of samples, $\{s_0(n) = A\,g(n); n = 0, 1, \ldots, N\}$. This has been converted with the aid of the normalized sequence, $g(n)$, which has the same shape but unity maximum amplitude and a scaling factor, A.

The definition of the spike may be extended to an infinite sequence, by assuming that $g(n)$ has zero value for $n < 0$ and for $n > N$. In the same way, spikes that occur at sample values $K_0, K_1, K_2, \ldots$ with respective amplitudes $A_0, A_1, A_2, \ldots$ may be described in a similar way, by shifting $g(n)$ in time. It is then possible to refer to the 'spike signal' as the collection of all spikes:

$$\begin{aligned} s(n) &= s_{K_0}(n) + s_{K_1}(n) + s_{K_2}(n) + \ldots \\ &= \sum_{i=0}^{\infty} s_{K_i}(n) \qquad (7.15) \\ &= \sum_{i=0}^{\infty} A_i\, g(n - K_i) \qquad (7.16) \end{aligned}$$

This is a continuous signal, although it is zero for the extensive periods between spikes.

Having formulated the signal, the problem is how to establish a pair of a system and another signal to model its behaviour. Like in the case of the background activity, the system was considered first, as the properties of the input signal are controlled by its behaviour.

The assumption that spikes have the same shape, with no variations, suggested a system that could produce this constant signal shape. Assuming that all shape information should be included in the system, one could consider a linear system, again, with a transfer function, $G(z)$, whose impulse response is the normalized spike sequence, $\{g(n)\}$. It would then be possible to generate $s(n)$ by exciting this system with impulses, scaled according to the amplitude of every individual spike, $A_0, A_1, A_2, \ldots$ and shifted in 'time' by an appropriate sample number, $K_0, K_1, K_2, \ldots$. The input signal to this system would then have the form:

$$d(n) = \sum_{i=0}^{\infty} A_i\, \delta(n - K_i) \qquad (7.17)$$

In this manner, all spikes generated have the same shape, apart from an arbitrary scaling factor. In practice, the shape is also variable from one spike to the next. It is, of course possible to introduce a set of transfer functions, $\{G_0(z), G_1(z), \ldots\}$, one per spike, but the changes are not as great as those present in the background activity and hence they have been disregarded at this stage. This issue will be addressed later in section 8.4.4, as in the implementation of the analysis procedure, small variations and deviations from the LTI

model will be allowed. At this stage it is more critical to determine the structure and parameters of both $G(z)$ and $\{d(n)\}$.

The input signal is a series of impulses, with unknown amplitudes as well as times of occurrence. Both these parameters were treated as independent random variables, A and K, respectively following different probability distributions. The amplitude, A, was assumed to be distributed normally around a specific mean amplitude, μ_A, with a variance σ_A^2 ($A \sim N(\mu_A, \sigma_A^2)$). There is no particular reason why A should not follow a different distribution, but this assumption has not been contradicted by any of the findings during the testing of the method. The estimated amplitudes for the available spikes appeared to vary almost symmetrically around a mean value.

The probability distribution function for the times of occurrence of spikes, K, was based on the fact that they occur *infrequently*. It is well-known to statisticians that the number of rare events, k, occurring per unit time, follows a Poisson probability distribution, $P(k) = \mathrm{Po}(\mu_{\Delta K}) = \frac{\mu_{\Delta K}^k}{k!} e^{\mu_{\Delta k}}$. Moreover, the time between occurrences of such rare random phenomena, ΔK, follows an Exponential distribution with a parameter, $\mu_{\Delta K}$, which represents the expected (mean) time between the previously and the currently occurring spike, ΔK. Hence $\Delta K \sim \mathrm{Exp}(\mu_{\Delta K}) = \int_0^{+\infty} \mu_{\Delta k} e^{\mu_{\Delta K} \Delta_K} \, d\Delta K$.

The purpose of these statistical distributions is not the accurate representation of the spike generating process, but the parametrization of their main properties. Inaccuracies in their specification can be tolerated, as these distributions do not form part of the analysis procedure, and have not been used in its development. Their specification is related to the generation of artificial data, which were used for testing the developed system prior to its application to real EEG signals. Specific values for μ_A, σ_A^2 and $\mu_{\Delta K}$ have been used for EEG simulation but these belong to the testing procedures (section 8.6.1).

Having established that the impulse response of the system $G(z)$ which is linear and time invariant, is a spike, its structure needed to be defined. As with the model of the previous section, a rational fraction of two polynomials was considered:

$$\begin{aligned} G(z) &= \frac{B(z)}{A(z)} \\ &= \frac{\sum_{m=0}^{M} b_m \, z^{-m}}{1 - \sum_{n=1}^{N} a_n \, z^{-n}} \end{aligned} \tag{7.18}$$

In many practical situations special cases of this structure suffice. There are two such cases, the 'autoregressive' (AR), structure, which has poles, but no zeros ($m = 0$, i.e. $B(z) = b_0$) and the 'moving average' (MA) structure, which has no poles ($n = 0$), and hence the denominator polynomial is effectively $A(z) = 1$. Both structures, as well as the general case

with poles and zeros (ARMA), were considered before deciding which one would be more suitable to use.

The MA model lacking any poles would require a large order to approximate sharp spectral peaks. Spikes, however, have reasonably smooth spectra, due to their transient nature and hence the concentration of their energy in a short time interval. They can approximate spectral valleys quite well, as mentioned in the previous section and their stability is guaranteed, since their transfer function has no singularities. The parameters of $G(z)$ in this case are also easily estimated. If the impulse response of this system is the (normalized) spike sequence, $\{g(n); n = 0, 1, 2, \ldots, P\}$, and zero for $n < 0$ and for $n > N$, then $G(z) = B(z)$ is simply the z-transform of $g(n)$. Hence the order of the polynomial $B(z)$ is $M = P$ and its coefficients are:

$$b_m = g(m) \forall m = 0, 1, 2, \ldots, P \tag{7.19}$$

The all-zero model, despite its simplicity both in structure and determination of its coefficients, has a number of drawbacks when used as part of the proposed analysis method, which uses both the direct and the inverse of the spike models as will be explained later. This inverse, $G^{-1}(z) = \frac{1}{B(z)}$ is an all-pole structure. It is apparent that all the zeros of $G(z)$, would now appear as poles and the stability of the direct model, which is guaranteed by the absence of singularities, may not be a property of its inverse. Of course, it is possible to determine the zeros that are responsible and attempt to correct them, by replacing them with their reciprocals, to maintain the amplitude characteristics of $G(\omega)$[151, pp. 434–436]. Such an attempt would distort the time representation of the signal, since there will be phase changes and hence changes in the relative delays of the time-domain components. As a result, the shape of the signal would be distorted and would not represent a spike any more. Another drawback arises from the significance of the controlling spectral parameters, in this case the zeros of $G(z)$. It has been pointed out earlier that peaks of the PSD of EEG components are more significant in the analysis. Essentially, the signal is described by what is more profound in the PSD, rather than what is missing, which is indicated by the zeros of the z-transform of the signal. This is not a disadvantage if spike modelling is the only objective, as the accuracy of the model is the key issue. It does become one, though, if the model is to represent the spike and possibly other transients in parametric form for further treatment. A disadvantage of a more practical nature is related to small variations in the spike model that may occur in real spikes and will be compensated by allowing small perturbations on the model parameters. If a spectral peak of the signal is displaced, the location of many zeros may need to change accordingly, causing significant changes in the system coefficients. In this respect, there are alternative structures with better properties.

With these prospective uses of the model in mind, the alternative (AR) model was considered. This has only poles and hence an infinite impulse response and does not conform with the initial definition of the spike as a sequence of finite duration. Nonetheless, this is not a fact because small spike components following its presumed ending may be masked by the much larger amplitudes of the background activity in which they are embedded. Assuming that a spike is a signal of high-amplitude immediately after its onset, which decays rapidly afterwards quickly becoming insignificant, the autoregressive (AR) approximation does not seem unreasonable. This structure has properties that are useful for the theoretical analysis as well as the practical implementation.

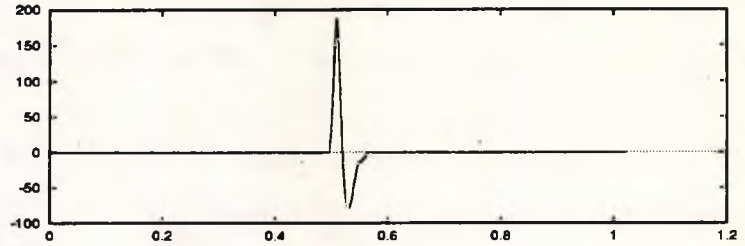
The relation of the poles to the power spectral density has already been highlighted in the previous section. Spikes exhibit one positive and one negative peak of significant amplitude. The duration of each peak may differ from that of the other and there may be some asymmetries in the signal, but the existence of this oscillation is characterized by a power spectral density that has one large peak and possibly others of lesser amplitude. The existence of this peak was indicated in the presentation of the frequency spectra of several spike-and-wave complexes by means of the discrete Fourier Transform (DFT) by Stelle[185, pp. 92–95]. In that study, most of the diagrams had two visible peaks, one at a very low frequency possibly around 3 Hz, corresponding to the slow wave, and one at a higher frequency, around 20 Hz, which corresponded to a period of 50 ms and is characteristic of the spike. One example of the amplitude spectrum of a spike (with the slow wave suppressed) is shown here in Figure 7.2. This was taken with the FFT algorithm of a window of 1024 samples with the spike in the centre, padded with zeros. The frequency resolution for the sampling rate used (1000 s^{-1}) was therefore just under 1 Hz.

This peak is not sharp, indicating the 'transience' of the spike. It may be compared to a model with one particular 'resonant frequency', but a large 'damping factor' causing the observed smoothness of the PSD. This represents the behaviour of the following discrete-time, second-order system:

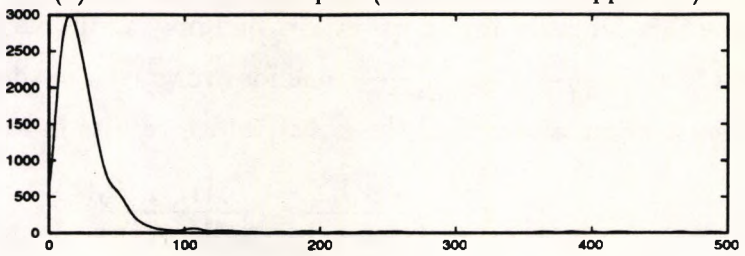
$$H(z) = \frac{r_0 \sin \omega_0 T \; z^{-1}}{1 - 2r_0 \cos \omega_0 T \; z^{-1} + r_0 \; z^{-2}} \tag{7.20}$$

where ω_0 is the resonant frequency, $r_0 = \mathrm{e}^{-aT}$ and a is the damping factor. The existence of a zero at the origin is not a significant diversion from the all-pole model, as it merely corresponds to an input delay. This model, despite its simplicity proved useful in the autoregressive approximation of spikes. The illustration of this relatively simple AR model here highlights some of its advantages over the previously mentioned MA structure.

The AR structure, however, has only poles and as such it is potentially unstable. But its impulse response will be made similar to the shape of a spike. All spikes have amplitudes that



(a) An isolated EEG spike (with slow wave suppressed)



(b) Amplitude spectrum of above waveform

Figure 7.2: The amplitude spectrum of a spike

become insignificant very quickly (transient nature of the signal). This is the behaviour of a stable system by definition. Hence a working model is guaranteed to be stable, provided that any perturbations on its parameters, should these be modified later, do not force any poles on or outside the unit circle in the z-domain. The stability of the inverse model is guaranteed, since it only has zeros (which lie inside the unit circle in z, resulting in a minimum-phase system).

The use of fewer parameters than the MA model is characteristic of the AR structure, when signals with relatively smooth PSDs, like spikes, are considered. Because it models spectral peaks directly, it requires approximately as many complex conjugate pole pairs as there are peaks. The validity of this statement is by no means absolute, but it gives an indication of the order of the filter. As a consequence, the number of parameters (essentially the order of the system) does not increase with the sampling rate, since the number of spectral peaks remains the same, whereas in the case of the MA structure, it depends on the length of the time sequence representing the spike, which will evidently increase. The values of the coefficients will, of course, change but if the poles of the system are known, their locations under the new sampling rate may be easily computed. In the example given by equation (7.20), the poles are at $e^{(-a+j\omega)T}$ and $e^{(-a-j\omega)T}$ and if the new sampling period is $T' = \lambda T$, $0 < \lambda < 1$, the only requirement is to raise the poles to the power λ. This simple model is also characteristic of the versatility of the model. A response of several samples in duration was effected with only three non-trivial coefficients (the order of the

model was $N = 2$). Because of the last two properties, significant changes in the amplitude and frequency response of the system may be achieved with small modifications to the values of the coefficients.

A last advantage of the AR model is the existence of stable optimization techniques for estimating the coefficients of the filter. Perhaps it should be noted that the initial structure, the ARMA model is far more general as it permits spectral peaks to be estimated independently. To put this property into perspective, one may assume a set of second-order (2-pole) AR models, $G_i(z) = \frac{1}{A_i(z)} = \frac{1}{z^2+a_{1i}z^1+a_{2i}}$, one for every peak in the PSD of the signal. The overall PSD is the accumulation of all the $G_i(z)$, which results in an ARMA model:

$$G(z) = \sum_i G_i(z) = \sum_i \frac{1}{A_i(z)} = \frac{\sum_i \prod_{j,j\neq i} A_j(z)}{\prod_i A_i(z)} \tag{7.21}$$

This should be born in mind if the objective is to produce a model whose parameters may be directly related to physical properties of the signal, such as resonant frequencies. In general ARMA models require fewer parameters to achieve an adequate approximation of a given PSD than their AR or MA counterparts, but their estimation requires the solution of non-linear equations for which several sub-optimal methods have been devised[152, pp. 262–269]. In any case, ARMA models may be regarded as an improvement on AR structures. The latter (AR) is a good compromise, as it combines most properties of the ARMA model and relatively simple optimization methods, hence it was preferred in this study, with ARMA modelling left as a possible future prospect.

7.3.3 Modelling other transients

The detection of transients other than spikes was not an objective of this study. Because the modelling approach had to consider as many elements of the EEG as possible, their inclusion could not be ignored. Besides, as explained later (Chapters 10 and 11) the detection of a number of other transients is possible with simple modifications and additions to the basic spike detection system.

Detailed modelling is not necessary here. It is sufficient to state that the arguments that apply to spike modelling are also valid for any other transient. Hence, every transient, type, i, was described by a linear time-invariant system, $F_i(z)$, excited by a (Poisson-distributed) sequence of impulses with variable (Normally distributed) amplitude, $c_i(n)$. This implied that the occurrence of each transient is not related to the occurrence of any other (statistical independence), which is not an unreasonable assumption to make in general.

Needless to say that the structure of the $F_i(z)$ may be regarded as minimum-phase, stable ARMA, so that their inverses are also stable. This is a general rule which is critical if the

expansion of the system to detect any of these transients is to function properly.

If there are N_T transient types, an equal number of transfer functions will be required, each having its own input. Although the existence of transients other than spikes had to be taken into consideration, there is no need to treat those individually. Hence they have been combined in a single signal, by superposition of the outputs of each transfer function, $F_i(z)$. In the z-domain this operation is illustrated simply as a sum of products:

$$T(z) = \sum_{i=1}^{N_T} T_i(z) = \sum_{i=1}^{N_T} F_i(z)C_i(z) = \mathbf{F}^T(z)\mathbf{C}(z) \tag{7.22}$$

where $C_i(z) = \mathcal{Z}\{c_i(n)\}$ and $T_i(z)$ is the z-transform of the model of the ith transient, $t_i(n)$. The composite signal $t(n)$ is the sum of all transients (in the above expression it appears as $T(z)$). In order to refer to the transients collectively, the transfer functions have been grouped in a column vector, $\mathbf{F}(z)$ and so were the input impulse sequences, $\mathbf{C}(z) = \mathcal{Z}\{\mathbf{c}(n)\}$. $t(n)$ will be referred to as the *transient signal component.*

This combination of elements using vector notation hides the individual transients, but it does not restrict future generalization to detect some of them. Individual transients may be 'pulled out' from the batch and treated individually, by processes similar to the one applied to spikes.

Finally, it is worth noting that not all transients considered in this way, are of cerebral origin. Certain artifacts are transient in nature and fit better in this category but these will be covered in the section that follows.

7.3.4 Artifacts and noise

Artifacts, the interfering signals of non-cerebral origin represent a category of signals whose properties and origins are quite varied. Because some of these are indistinguishable from other EEG component groups that have already been described, it was considered more beneficial to include these artifacts in the existing signal models rather than generate new ones. In this manner, unnecessary complications of the analysis that follows have been avoided without ignoring the presence of these signal components in the recorded EEG.

Interfering signals may be grouped into those whose onset is fast and their shape is reasonably constant and into those with gradual onset and longer durations. The first resemble transients, whereas the second bear more the characteristics of the background EEG activity, both of which have been described in the preceding sections.

Typical examples of the first kind of artifact are electrode movement artifacts, which appear as abrupt baseline changes, registered as exponential decay curves after AC coupling which takes place as part of the recording procedure. Ocular artifacts, like eye movements

and blinks, also belong to this category, as their shape is characteristic. They sometimes resemble genuine EEG transients, but their location and often their regularity can be used for their identification. Some ECG artifacts also have regular shape but their rate of repetition may be fairly constant.

Finally, muscle artifacts may be categorized as transients also, if they are isolated. These sometimes appear in the form of spikes, but their shape is not very regular and they often come in groups. These bursts are usually quite long in duration and their irregularity makes them resemble noise of excessively large amplitude. Their PSD is not flat, but covers most of the EEG frequency range and has large components at frequencies at the high-end of the spectrum, where spikes have most of their energy. For this reason their rejection is rather difficult. For signal modelling purposes, EMG artifacts may be considered similar to the background activity, in the sense that their properties are similar to those of a signal produced by white noise excitation of a transfer function (see section 7.3.1). Whether this transfer function is included along with the transients, with a different excitation function, or whether integrated along with the background activity to form a transfer function of higher order is not a critical issue. In practice both models are necessary to represent short and long time stretches of EMG interference.

Mains interference is a much easier signal to identify and classify. It has sinusoidal shape at either 50 or 60 Hz frequency. Although it is often present for long intervals, and may thus be modelled as part of the background activity, it may be more conveniently detected separately before the signals are processed and removed by either a notch filter, or better by optimum filtering, as described by Widrow et al.[201].

Having considered these sources of artifact, one should not forget the existence of random 'noise' in the recorded signal. Sources of noise vary, but the instrumentation and electrodes are responsible for Gaussian noise, with zero mean and an unknown variance (power), σ_n^2. In this study, this component was considered much smaller than the background activity. Uniform White noise is introduced because of the quantization of the signal by the analogue-to-digital converter. This follows a uniform distribution in the range $[-2^{-b-1}, 2^{-b-1}]$ with variance $\sigma_q^2 = 2^{-2b}\frac{q^2}{12}$[151, pp. 749–750], where b is the number of bits in the ADC (assuming that it performs rounding of values) and q is the quantization step in Volts. For a 12-bit ADC, like the one used here, this is of the order of $5 \times 10^{-9}q^2$. Even when the gain is such that a signal of 1 mV corresponds to the full dynamic range of the ADC ($q \approx 5 \times 10^{-7}V$), this represents a signal of the order of 34 pV, which is insignificant compared to the background activity of several μV. Reducing the precision to 8 bits would yield an error that is 100 times larger, still insignificant compared to the background activity. Finally, rounding-off noise is

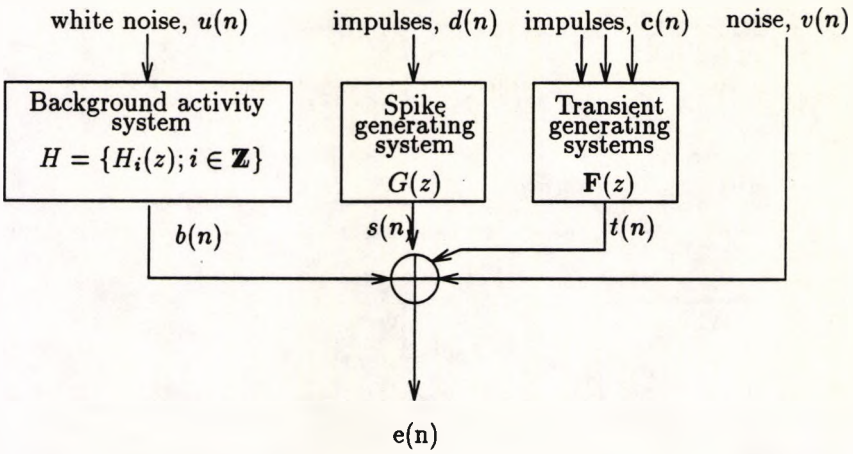


Figure 7.3: The EEG synthesis model

a consequence of the numerical operations performed on the signal during processing. This is small here, because double precision arithmetic was used overall, making these errors quite negligible, compared to other sources of interference.

Therefore, noise is mainly caused by electrodes and instrumentation, and was modelled as an independent source of uncorrelated, white, Gaussian sequence $v(n) \sim N(0, \sigma_n^2)$.

7.3.5 The overall EEG synthesis model

Having considered the components of the recorded EEG signal, one should be reminded that these exist in isolation only in theory and *cannot be measured separately* in practice. Only the final composite EEG signal is recorded and the process may be regarded as a system with both its inputs and parameters inaccessible. The single quantity that can be measured is the output, $e(n)$, and for all analysis purposes, the system parameters may be inferred by using this signal only!

The process of composition is shown in Figure 7.3. Notice that H represents the background activity generator process and that it is a system with a piecewise linear behaviour in time. $G(z)$ represents the impulse response of a spike, whereas $\mathbf{F}(z)$ is the collection of transfer functions of transients.

In general, the composite EEG signal, $e(n)$, consists of four signal components:

$$e(n) = b(n) + s(n) + t(n) + v(n) \tag{7.23}$$

where $b(n)$ is the background activity, $s(n)$ and $t(n)$ the sequences of spikes and the transient signal component respectively and $v(n)$ is the noise model. Artifacts have not been modelled separately, as they share properties with $b(n)$ and $t(n)$.

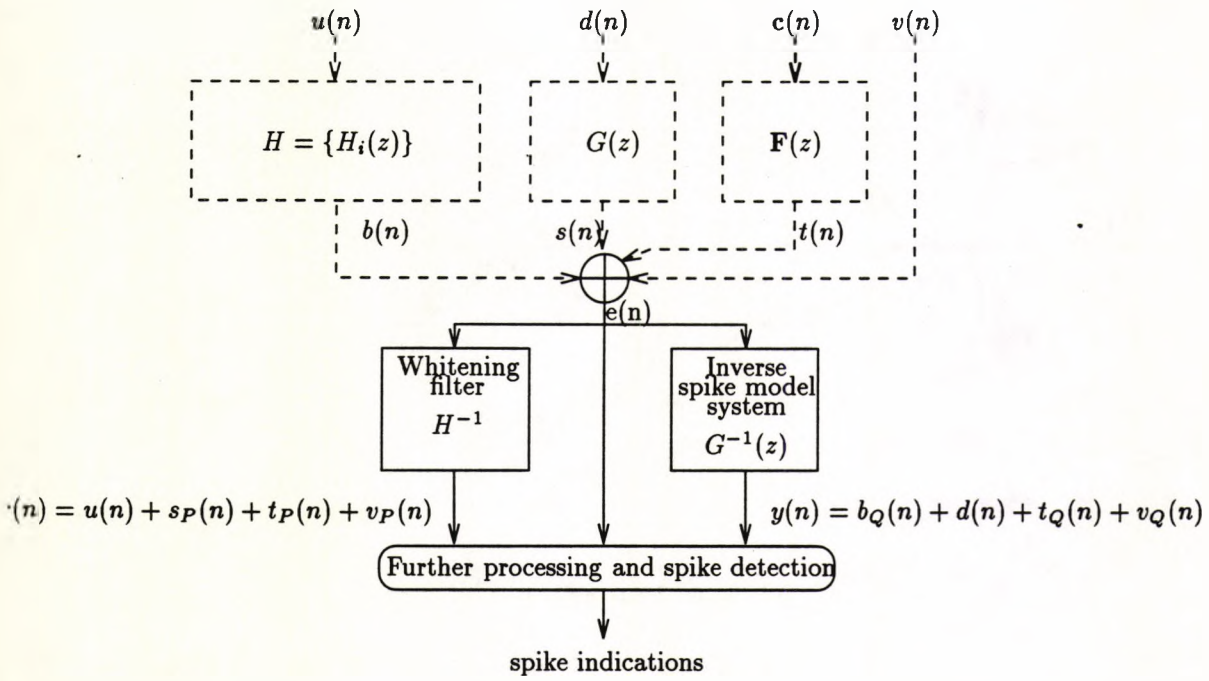


Figure 7.4: Synthesis-analysis model

7.4 Signal analysis and inverse modelling

The modelling procedure was described in great detail, because it forms the basis of EEG analysis, which is an ambiguous process. Many of the problems related to analysis have been eliminated by carefully describing the various signal components. In this section the inversion of the synthesis procedure, will be attempted, thus breaking down the EEG signal into the elementary signals suggested by the model. This will be achieved by applying the inverse of the systems that generate each component as illustrated in Figure 7.4.

The procedure and explanation given here are valid *if* these systems are known, an assumption that is not true, in practice, where only the composite signal, $e(n)$ is available. The sceptical reader should be assured that some methods for their estimation will be suggested, after their purpose is justified in this section.

Many of the signal processing operations in this section are considerably easier to interpret in the z-domain, rather than the 'time' domain. Hence the z-transform of the composite signal, $e(n)$ will be used quite extensively:

$$\begin{aligned} E(z) = \mathcal{Z}\{e(n)\} &= B(z) + S(z) + T(z) + V(z) \\ &= H[U(z)] + G(z)D(z) + \mathbf{F}^T(z)\mathbf{C}(z) + V(z) \end{aligned} \tag{7.24}$$

The significance of the various elements of this expression has already been elaborated in the preceding sections and is illustrated in Figure 7.3.

7.4.1 Separation of the background activity

The process of generating the background activity (section 7.3.1) assumed the excitation of a system H by a white uncorrelated sequence, $u(n)$. Furthermore, the system H was constructed from a set of linear time invariant systems, which were described by transfer functions, $\{H_i(z); i \in \mathbb{Z}\}$. Each one of these is valid for a time-window, but transition from one to the next is smooth, as described earlier and there may be a small overlap between adjacent windows, where the transition from one to the next segment takes place. The behaviour of the system when these transition segments occur will be examined in a practical way later, as it depends on the implementation. With the exclusion of these relatively short periods, there is only one transfer function, say $H_i(z)$, that contributes to the output sequence, $b(n)$, at a specific instant in time, hence $H[U(z)] = H_i(z)U(z)$.

It is therefore possible to study the signals associated with the application of the inverse system, $H_i^{-1}(z)$ (see Figure 7.4). Evidently, since $H_i(z)$ is an all-pole (AR) system, the inverse system has zeros at the same locations. The absence of poles ensures that $H_i^{-1}(z)$ is stable. It is also minimum-phase, because all its zeros lie inside the unit circle in z. The ability to invert a stable all-pole transfer function without worrying about the stability of its inverse was one of the factors that made the AR preferable to the more general ARMA structure during the modelling procedure (section 7.3.1).

The components of the signal $r(n)$, the output of $H_i^{-1}(z)$ when applied on the recorded (composite) EEG signal, $e(n)$, are easily identified in the z-domain:

$$\begin{aligned} R(z) &= H_i^{-1}(z)E(z) \\ &= H_i^{-1}(z)H_i(z)U(z) + H_i^{-1}(z)G(z)D(z) + H_i^{-1}(z)\mathbf{F}^T(z)\mathbf{C}(z) + H_i^{-1}(z)V(z) \\ &= U(z) + A_i(z)D(z) + \mathbf{M}_i^T(z)\mathbf{C}(z) + H_i^{-1}(z)V(z) \qquad (7.25) \\ \text{where } A_i(z) &= H_i^{-1}(z)G(z) \\ \text{and } \mathbf{M}_i(z) &= H_i^{-1}(z)\mathbf{F}(z) \end{aligned}$$

$E(z)$ and $R(z)$ are the z-transforms of $e(n)$ (equation (7.24)) and $r(n)$, respectively.

As expected, the output of this inverse system contains the generating sequence $u(n)$ of $b(n)$, the ongoing activity. The other elements of this signal are:

- A sequence due to the spike signal, consisting of distorted spikes. Because their occurrence is rather infrequent, it is very likely that the distortion will be different from one spike to the next, since $H_i(z)$ would change to another transfer function by the time the next spike is encountered. Alternatively, distortion may be viewed with respect to the input to the spike generating filter, $d(n)$, as a new system with transfer function $A_i(z)$. This is the cascade of $G(z)$ and the $H_i^{-1}(z)$.

- A sequence which is related to the transient signal component. As with the spikes, this signal is distorted and looking at the process from the input sequence(s), $\mathbf{c}(n)$, the operation corresponds to a quasi-linear system, which has the form of a vector $\mathbf{M}_i(z)$, the cascade of $H_i^{-1}(z)$ with $\mathrm{t}(z)$.

- A distorted version of the noise sequence, $v(n)$, due to the application of the inverse system, $H_i^{-1}(z)$.

These components have certain properties that are worth noting, as they play a role in the detection procedure.

$u(n)$ has lower power than $b(n)$. This is a direct consequence of the generating procedure, where $u(n)$ was initially assumed to have unit variance whereas the output, $b(n)$ of the system is typically much larger. Later on (section 8.1), it will be shown that because of the method used for the estimation of $H_i^{-1}(z)$, this property is guaranteed. Because $u(n)$ is a white sequence, $H_i^{-1}(z)$ is sometimes called a *whitening filter* and its purpose is to decorrelate $b(n)$.

Regarding the distortion of the spikes, it is possible to quantify its effects in a general way by considering some properties of the overall transfer function, $A_i(z)$. This has poles from $G(z)$ and zeros from $H_i^{-1}(z)$. The latter tends to attenuate its input, as noted in the previous paragraph. Moreover, since $H_i(z)$ reflects the PSD properties of the background activity, which occupies primarily the lower end of the spectrum, the whitening filter is expected to have most of its zeros close to the positive half of the real axis in z and would attenuate these frequencies much more than the higher ones. However the poles govern the time extent of the signal which is expected to be brief. Thus, the distorted spikes would bear some resemblance to actual spikes but with their lower frequencies considerably reduced in amplitude. The exact details of the distortion cannot be anticipated, because $H_i(z)$ (and its inverse) may change from one spike occurrence to the next.

A similar argument holds true for the transient signal components, which are also corrupted by the application of the inverse filter. Here, the overall transfer function(s) are described by the vector $\mathbf{M}_i(z)$. The poles of these are the same as those of $\mathbf{F}(z)$, but extra zeros are introduced from $H_i^{-1}(z)$. As with spikes, these zeros tend to reduce the low-frequencies of the transients, but their duration is not affected very much.

One other property shared by spikes and other transients is their large amplitudes compared to the background activity. This is a consequence of the amplitude of the excitation functions, but is more related to their nature. For the generation of spikes and transients impulse sequences were used, whereas for the background activity a white sequence was necessary. An impulse and white noise have flat amplitude spectra, but they differ on their

phase which is zero in the first case, but random in the second. Hence after the application of a transfer function, the result in the first case has the phase characteristics of the transfer function, which are generally smooth, whereas in the second case they contain a strong random element. In conclusion, with impulse excitation, many adjacent frequencies have similar phases, giving rise to large instantaneous amplitudes, whereas with a white sequence, amplitudes are spread over time, due to the random phases of adjacent frequency components.

Finally, noise is not uncorrelated any more, because of the effects of $H_i^{-1}(z)$. Its power is generally decreased with respect to what it was prior to the application of this system for the reasons given in the previous paragraphs. Low-frequency components may also exhibit attenuation due to the zeros of the applied transfer function.

In brief, the application of $H_i^{-1}(z)$ on the recorded EEG signal, $e(n)$ tends to reduce the power of the background activity and destroy its structure to a white noise sequence. Noise components are also reduced in power and spikes and transients are registered with relatively large amplitudes, but distorted. For this reason, a number of workers, like Da Silva et al.[179], Birkemeier et al.[17] and Bodenstein, Prætorius et al.[25, 153] found inverse filtering useful in the detection of spikes. Their reasoning and models have some relation to the material presented here, but their methods and systems focused on the background activity. Other components of the signal were excluded from the modelling procedure.

7.4.2 The inverse spike model

Relying on the output of the inverse of the background activity generating system, $r(n)$, for the detection of spikes proved inadequate. As explained in the previous section, spikes give rise to large-amplitude signals with brief durations, but so do non-spike transients. In some cases, their amplitudes are often one or more orders of magnitude larger than those for spikes. Hence the presence of a visible transient in $r(n)$ is an indication of the presence of either a transient or an epileptic spike, although the discrimination between them is not possible by inspecting $r(n)$. If H had a fixed transfer function, it would have been possible to identify spikes in $r(n)$, as the distortion would have been similar from one spike to the next (which would then correspond to the impulse response of a system with transfer function $H^{-1}(z)G(z)$). With a changing transfer function, only the transient nature and large amplitude of the signal are invariant.

Because spikes were also modelled as an input-system structure (section 7.3.2), it is possible to observe the behaviour of the inverse of this system on the recorded signal $e(n)$. The spike generating system, $G(z)$ is linear and has stable poles and no zeros. Its inverse,

$G^{-1}(z)$ has only zeros and hence it is also stable. Since these coincide with the poles of $G(z)$, they lie inside the unit circle and hence $G^{-1}(z)$ is a minimum-phase system as well. By its definition, $G^{-1}(z)$ is such that when applied on the spike sequence, $s(n)$ the output would be the spike-generating sequence of randomly occurring scaled impulses. The consequences of its application on the recorded composite EEG signal, $e(n)$, are easily observed if, as before, the z-transform of signals and systems is employed using the synthesis equation (7.24):

$$\begin{aligned} Y(z) &= G^{-1}(z)E(z) \\ &= G^{-1}(z)H_i(z)U(z) + G^{-1}(z)G(z)D(z) + G^{-1}(z)\mathbf{F}^T(z)\mathbf{C}(z) + G^{-1}(z)V(z) \\ &= B_i(z)U(z) + D(z) \qquad (7.26) \\ &+ \mathbf{N}^T(z)\mathbf{C}(z) + G^{-1}(z)V(z) \qquad (7.27) \\ \text{where } B_i(z) &= G^{-1}(z)H_i(z) \\ \text{and } \mathbf{N}(z) &= G^{-1}(z)\mathbf{F}(z) \end{aligned}$$

Here, $Y(z)$ represents the z-transform of the output $y(n)$ of the inverse spike filter. The output of the system contains the spike generating sequence, $d(n)$ and components due to the on-going activity, other transients and noise.

Background activity, $b(n)$ is now distorted. The effect of the inverse filter is equivalent to the application of the input white noise sequence, $u(n)$, to a system with a transfer function $B_i(z) = G^{-1}(z)H_i(z)$. This is a correlated signal whose auto-correlation function, and hence its PSD are described by the poles and zeros of $B_i(z)$ for every segment i. Because the background activity of the EEG, $b(n)$ is concentrated at the low-end of the spectrum, the critical poles of $H_i(z)$ (which are the poles of $B_i(z)$) will be at small angles to the positive real axis in the z-domain. The zeros of $B(z)$ are due to $G^{-1}(z)$, located at a larger angle and hence tend to suppress frequencies at the higher end of the EEG 'bandwidth'(section 7.4.1). As a consequence, this signal will have, in general low frequencies, because of the action of the zeros at high frequencies and the absence of components with significant amplitudes at frequencies beyond that.

The effects of the spike filter on other transients, $t(n)$, cannot be assessed specifically. Transients (section 7.3.3) are generated by sequences of impulses $\mathbf{c}(n)$, occurring at random intervals with amplitude and sometimes polarity changes from one to the next. The transfer functions that generate them, collectively written in vector form as $\mathbf{F}(z)$, are all modified by cascading with the inverse filter and the new transfer functions, $\mathbf{N}(z)$, are effectively applied to $\mathbf{c}(n)$. These will still appear as transients in $y(n)$, but their shapes will exhibit a distortion that may be predicted, if both $G(z)$ and $\mathbf{F}(z)$ are known. Although their amplitudes may (in

theory) increase by the application of $G(z)$, they cannot resemble an impulse, otherwise their time sequence would have been similar to that of a spike. In general, for most transients, which have most of their energy in frequencies below the peak of the spike model, the amplitudes are expected to be reduced and, in any case, phase distortion introduced by $G^{-1}(z)$ would change the relative time-delays of the various frequency components.

The application of $G^{-1}(z)$ on the uncorrelated noise, $v(n)$, would cause its correlation. The frequencies around its zero(s) will be suppressed, but higher frequencies may be enhanced. In general, this is not as critical as it may appear, because of the more significant amplitude of the spike impulse sequence.

Effectively, the output of the inverse system of the spike model contains the spike generating impulse sequence, but the background activity signal may still have significant power. Other transients are distorted, but not necessarily suppressed. In other words, although the spike generating sequence is present in the output of $G^{-1}(z)$, it may be difficult to detect reliably, because of the interference of other signals, mainly the background activity.

7.4.3 Inverse modelling as a means of spike detection

In section 7.4.1 it was shown that the output of the inverse background activity modelling system, $r(n)$, contained noise $u(n)$ and $v_P(n)$ and some components of brief duration, which appear spontaneously. These may be caused by spikes, $s_P(n)$ or other transients $t_P(n)$. Hence,

$$r(n) = u(n) + s_P(n) + t_P(n) + v_P(n) \tag{7.28}$$

where, from equation (7.25), $s_P(n) = \mathcal{Z}^{-1}\{A_i(z)D(z)\}$, $t_P(n) = \mathcal{Z}^{-1}\{\mathbf{M}_i^T(z)\mathbf{C}(z)\}$ and $v_P(n) = \mathcal{Z}^{-1}\{H_i^{-1}(z)V(z)\}$. From these, the ones of greatest interest are $s_P(n)$. Unfortunately, it may not be possible to isolate them from the other transients, $t_P(n)$, because the applied system, $H^{-1}(z)$ may change from time to time in an unpredictable manner, as already explained. It is, however, possible, with some reasonable assumptions to detect *both* $s_P(z)$ and $t_P(n)$ together. This is an indication that a transient phenomenon was present in the signal. The assumption that was made here is that the amplitudes of these components are larger than the RMS value (the standard deviation for signals with zero mean) of the random components, $u(n)$ and $v_P(n)$. This does not come into conflict with neurophysiological definitions, which stress that spikes (for example) should be *clearly distinguished from the background activity*[41]. From this statement and the method by which $H^{-1}(z)$ will be estimated (see section 8.4.3) one may arrive at the following proposition.

Proposition 7.1 *If spikes (or other transients) can be distinguished from the background activity, they can be distinguished more clearly after processing the signal with the inverse filter, $H^{-1}(z)$.*

This is a direct consequence of the noise-like appearance of $u(n)$ and $v_P(n)$, which have no structure and hence may be more easily suppressed by filtering than the variable-frequency background activity with large low-frequency components. As a consequence of Proposition 7.1, transients may be detected with relative ease in $r(n)$, although their shape may not help in ascertaining their origin.

The inverse spike system, on the other hand, is related to a different set of components. Its output, $y(n)$, given by equation (7.27), contained an impulse sequence due to the spikes, some other spontaneous signals, $t_Q(n)$, when other transients are present, and some sequences of correlated noise $v_Q(n)$ and $b_Q(n)$, which are always present, even in the absence of spikes:

$$y(n) = b_Q(n) + d(n) + t_Q(n) + v_Q(n) \tag{7.29}$$

where $b_Q(n) = \mathcal{Z}^{-1}\{B_i(z)U(z)\}$, $t_Q(n) = \mathcal{Z}^{-1}\{\mathbf{N}^T(z)\mathbf{C}(z)\}$ and $v_Q(n) = \mathcal{Z}^{-1}\{G^{-1}(z)V(z)\}$. Attempts to detect spikes using $y(n)$ did not produce reliable results, because, unlike $r(n)$, the on-going components here, $b_Q(n)$ and $v_Q(n)$, do not necessarily have small RMS value compared to $d(n)$, neither are they uncorrelated. However, it may be possible to utilize this output as an indication of the occurrence of a spike, despite its unreliability.

Proposition 7.2 *In the presence of a spike conforming to the defined model, $G(z)$, the amplitude of the generating impulse may be estimated from the output of the inverse model with an expected error equal to the RMS value of the signal.*

This assumes that the background activity and spikes are uncorrelated. Hence, a transient at any instant is subjected to an uncertainty caused by the variance (power), σ_0^2, of the interfering signal, consisting of $b_Q(n) + v_Q(n)$, as other transients are rare, having a small bearing on the signal variance. Their scarcity is not an important requirement, as their presence may be detected and their contribution on the computation of variances eliminated. Variance is reflected as an uncertainty in the *amplitude* of the spikes contributing its square root in either direction of the value of $d(n)$. Now, assuming that a spike occurred at an instant N and was detected by some independent technique, it may be described in terms of the excitation function $d(n) = A_N\delta(n - N)$ (see equation (7.17)), contained in $y(n)$. A_N cannot be determined precisely, but assuming that $b_Q(n) + v_Q(n)$ is Normally distributed, a range $[a, b]$ may be determined so that with a certain level of confidence, p, the actual value of A_N will be included in $[a, b]$. If σ_0^2 was known, p would have been a small probability value

on the tails of the Normal distribution, $N(A_N,\sigma_0^2)$. Because it is unknown and its estimate, $\hat{\sigma_0}^2$ has to be used, the t probability distribution for the mean is the one to use. The two corresponding values from the distributed variable, $t = \pm t_0$ may be used to determine $[a, b]$, since $\frac{y(n)-A_N}{\hat{\sigma}_0}$ follows a t distribution. Hence

$$y(N) - t_0\hat{\sigma}_0 \leq A_N \leq y(N) + t_0\hat{\sigma}_0 \tag{7.30}$$

The assumption of normal distribution is not critical. In fact any other probability density function may be employed, provided that its characteristics can be measured.

In conclusion, as a consequence of Propositions 7.1 and 7.2, $r(n)$ contains a *reliable* indication of transients, but is unable to establish which of them are spikes, whereas $y(n)$ contains an indication *specific to spikes* and capable of estimating their amplitude, but not reliable enough to be used for their detection. The combination of the two was used to form a hypothesis that an observed transient was a spike. This was tested by estimating the spike and subtracting it from the signal and repeating the test. The absence of a transient increased the certainty that the initially observed phenomenon was indeed a spike.

7.4.4 Structure and behaviour of the analysis model

It is evident from the comments at the end of the previous section that the proposed working system requires some additional elements for its implementation, which were not initially included in the theoretical model suggested by Figure 7.4. A block diagram which includes these amendments is shown in Figure 7.5.

The inverse of the system model for the generation of the background activity, $H^{-1}(z)$ appears twice in the complete system. Also, the spike generating filter, $G(z)$ appears in direct realization as well as in inverse, $G^{-1}(z)$. These elements are essentially self-contained having one input and one output. Their purpose, structure and operation has already been described in the preceding sections and is entirely related to traditional signal processing concepts.

The other four units are the constituents of the detection and decision entity (Figure 7.4), which relates more to pattern recognition than to signal analysis. Its function was based on the observations and comments of the preceding section. These units are the two transient detectors, the spike initiator and the the spike detector. Unlike the other elements of the system, their function could not be illustrated in isolation, because they combine inputs from other units. In fact they form the links between the other elements of the system and are as critical in its overall behaviour as any other unit, possibly even more.

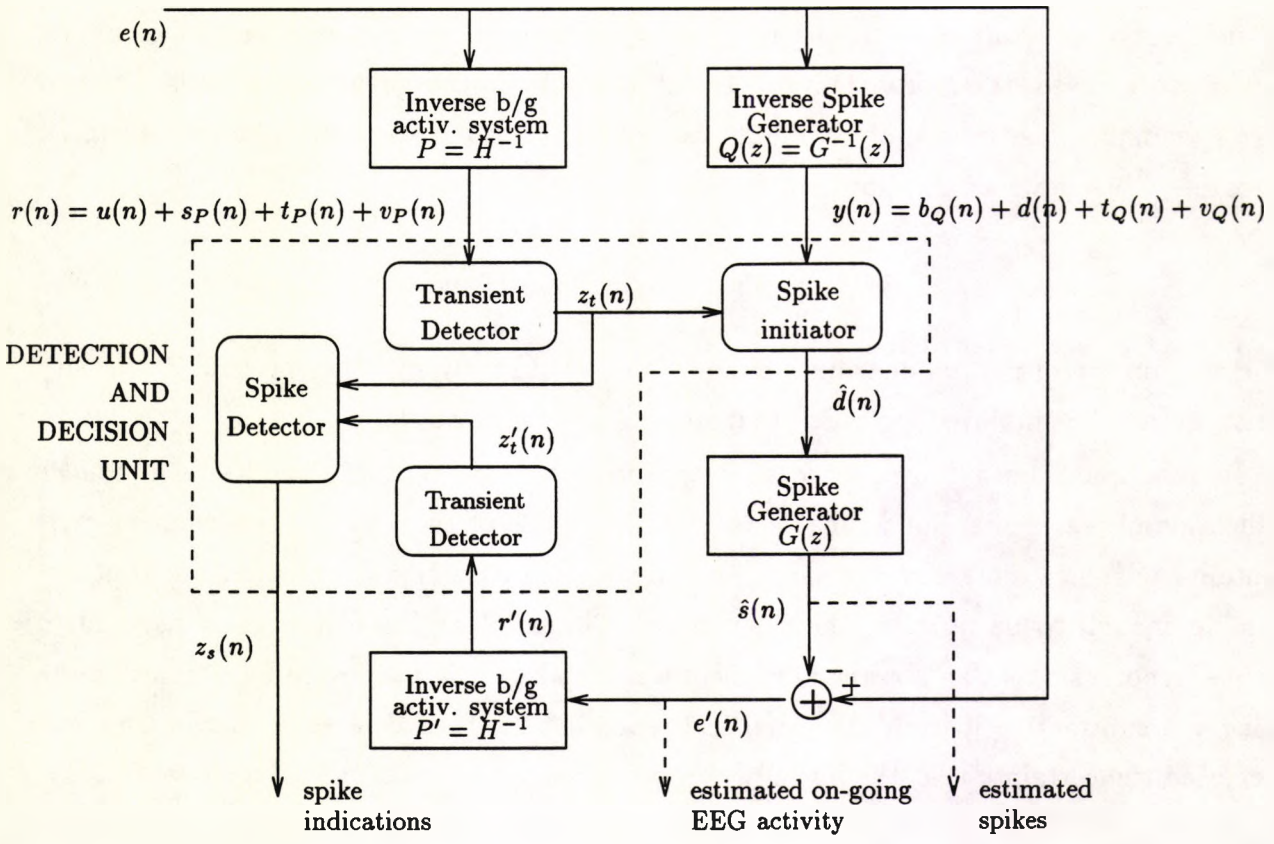


Figure 7.5: Block diagram of EEG analysis system

There are many formal and informal methods for pattern identification, ranging from simple thresholding of signals to elaborate pattern recognition schemes. In the absence of a single universal description of their structure a subjective realization was attempted. This relied on the implementation of the other elements and will be described in the Chapters that follow, but it does not inhibit the description of the behaviour of the overall system at this stage. It is essential that when describing the implementation details in later sections, the reasoning behind the use of each structure is known in advance.

Functional description of the overall system

Before looking into the implementation details of each element of the system illustrated in Figure 7.5, the operation of the system as a whole under different input conditions will be considered. Three input scenarios are of interest, depending on the components present in $e(n)$:

- A spike has occurred
- Spikes and other transients are absent
- A non-spike transient is present.

In the first case, $e(n) = b(n) + s(n) + v(n)$, since the transient component, $t(n)$, is zero. This signal is applied to $H^{-1}(z)$ and $G^{-1}(z)$. The outputs of these systems, are $r(n)$ and $y(n)$, respectively. From the explanation given in section 7.4.1, $r(n)$, will contain the random sequences $u(n)$ and $v_P(n)$ and a transient component, due to the spike, $s_P(n)$. Similarly, from the arguments stated in section 7.4.2, $y(n)$ has components $b_Q(n)$, $d(n)$ and $v_Q(n)$, corresponding to the input components $b(n)$, $s(n)$ and $v(n)$. The function of the transient detector is to produce a discrete output, $z_t(n)$, indicating the occurrence ($z_t(n)$ =true) or the absence ($z_t(n)$ =false) of a transient. This is maintained *true* for the duration of the transient $s_P(n)$. Spikes cannot be reliably detected based on the presence of an impulse-like signal, $d(n)$, in $y(n)$ and hence the spike initiator uses $z_t(n)$ as a cue. Evidently, spikes detected in $y(n)$ but not associated with a transient detection in $r(n)$ cannot be genuine and are hence discarded. In the presence of a spike, though, the spike initiator computes an estimate of the spike generating impulse $\hat{d}(n)$ from $y(n)$, as explained in section 7.4.3.

This by itself is confirmation that a spike occurred, but the process was taken a stage further. By exciting the system of the spike model, $G(z)$, by the estimated impulse, $\hat{d}(n)$, an estimate for the spike sequence, $\hat{s}(n)$, that supposedly occurred, is generated. This is subtracted from the original EEG signal, $e(n)$ and the residue signal, $e'(n)$ is then re-processed through $H^{-1}(z)$ and a duplicate of the transient detector. If a spike was correctly identified, generated and subtracted, then $e'(n)$ should be transient-free during the period of occurrence of the spike in $e(n)$ and the second transient detector should not indicate the presence of the transient. By combining the outputs of the first and second transient detectors, the spike detector is able to identify this event and correctly indicate spike occurrences as a binary event $z_s(n)$.

The effect of other transients

If a non-spike transient is present, $e(n) = b(n) + t(n) + v(n)$, then $r(n)$ will contain a transient indication, $t_P(n)$ instead of $s_P(n)$. The output of $G^{-1}(z)$, $y(n)$, will also have a transient present, $t_Q(n)$. For most transients, the spike initiator would have a zero output, because $y(n)$ will not resemble an impulse. It is possible, though, that in the multitude of possible transients some may produce sequences in $y(n)$ that may be impulse-like. The spike initiator would treat those as spikes and hence produce a non-zero output. This demonstrates the requirement for the additional elements $G(z)$ and $H^{-1}(z)$. Even if $G(z)$ is excited by a non-genuine spike sequence, its output will still be a spike estimate. Hence after its subtraction from $e(n)$, the transient is not eliminated, because $e'(n)$ will now contain the spike estimate, $-\hat{s}(n)$ as well as the initial transient, $t(n)$, which are different sequences and

do not eliminate each other. As a result, $r'(n)$ would still contain a transient, which will be picked-up by the second transient detector. The spike detector will then be presented with two signals, $z_t(n)$ and $z_t'(n)$ suggesting that even after a spike was subtracted, the transient is still present. Therefore, either the transient was not a spike, or a transient *and* a spike occurred simultaneously, a very unlikely event. Hence no spike is indicated in $z_s(n)$.

In the complete absence of transients, $e(n)$ is a combination of the background activity, $b(n)$, and noise, $v(n)$. Hence, $r(n)$ will have no transients to be detected by the first transient detector. As a result, there will be no spike initiation and hence no spike generated at the output of $G(z)$. In this case, $e'(n)$ is identical to $e(n)$ and the second detector will indicate no transients either. Consequently, the spike detector will not show the presence of a spike, as expected.

Double inverse modelling as an integrity test

The importance of the duplicate transfer function $H^{-1}(z)$ and the spike generating filter $G(z)$ should not be undermined. These are crucial in the behaviour of the detector, especially in the uncontrolled environment of ambulatory monitoring, where many unprecedented phenomena may be encountered. In terms of system analysis, these elements are a form of *feedback*, used so that the system may test its own output and adjust its behaviour accordingly.

This may be clarified by describing the operation of the system in terms of probabilities. Let the first (or the second) transient detector have a probability p_1 of making a false transient detection and the spike initiator have a probability p_2 of falsely detecting a spike, when there was not one present. The probability of a spike being falsely detected, based on the output of the spike initiator only, is p_1p_2. But if the generation and subtraction of a spike and the repetition of the transient detection process takes place, the probability of a false detection is $p_1^2p_2$, which is considerably smaller than p_1p_2, since both p_1 and p_2 are positive and much smaller than unity. The probability of a spike being missed depends on the probability of the transient detector missing a transient, p_3, which is not affected by the feedback of the 'spike-free' signal. This is expected to be fairly low, anyway. Assuming independence, the probability of making an error (either a false or missed detection) is the sum of the probability of missing a spike and the probability of falsely detecting one which is $p_1p_2 + p_3$, without feedback and $p_1^2p_2 + p_3$ with feedback of $e'(n)$. Evidently, the improvement in the performance of the spike detector after the addition of 'feedback' is significant, especially if p_1 is small, a fact that is generally true for a good transient detector.

A procedure based on processing $e(n)$ by $H^{-1}(z)$ and then detecting spikes in its output

using a bank of matched filters for various types of spikes was suggested by Pfurtscheller and Fischer[148]. This is an example of a system without feedback of the detected events for testing the integrity of the detected phenomena.

7.5 Concluding remarks

A procedure for the detection of spikes based on signal modelling concepts was presented. This was based on the separation of the signal of interest (the EEG) into a small set of components (background activity, spikes, etc.) having different properties. Each one of those was further modelled as the output of a system, driven by an input signal with simpler properties. This decomposition was utilized for the separation of the constituents of the signal, based on the systems described and their inverses. Specific elements (spikes) may be detected from the signals produced during this decomposition. Therefore, signal modelling here is an integral part of the analysis method.

The advantage of this approach over the earlier differential spike detector (see section 6.5) lies in its generality. In the analysis presented, the detailed properties of the signal to be analyzed were not taken into account. It was simply considered as a mixture of elements with long duration and others with brief duration but characteristic shapes (transients). As such it is free from heuristics based assumptions about properties of a specific transient, but rather describes a procedure to follow for their modelling and detection. The only requirement is that these transients have consistent shapes known a priori. Their detection, when embedded in a continuous signal which may have possibly unknown characteristics, is then possible.

The proposed analysis technique has other properties that cannot be undermined. First of all, it is still independent from the internal structure of the systems used for modelling. Although quasi-linearity was hinted, this restriction may easily be lifted once suitable more general (non-linear, for instance) modelling techniques become available.

The separation of the input signal into components that may be studied independently is a useful by-product of this approach. In terms of EEG analysis, both the estimated spikes, $\hat{s}(n)$ and the (hypothetically) spike-free signal, $e'(n)$ are available separately (see Figure 7.5). The ability to study their properties independently may yield useful neurological findings. Moreover, these signals are available in parametric representation, which may be easier to study, or may be useful for the compression of the signal during on-line analysis. These issues will be emphasized in the Chapters that follow, as they are linked to the implementation.

One of the key issues in the modelling approach was the possibility of generalization. This

may be viewed in the context of the EEG signal or in the more general theme of composite signals.

As far as EEG analysis is concerned, with the introduction of new systems and the modification of the detection elements, the proposed system may be be generalized to detect other transients, apart from spikes. The extension of the proposed system is the subject of Chapters 10 and 11.

In its general form, the method may be applied to other signals, of biological origin or other, consisting of a structured (correlated) sequence with superimposed transient components.

It should be noted that in this Chapter the systems describing the signal components of the input signal were presumed known in advance. In real life they are largely unknown and must be estimated, if the method proposed may be applied. Several methods for system estimation both on-line and off-line exist, (see for example [151, 152]). A class of methods, particularly suitable for on-line processing adapted for the purposes of this work will be discussed in the next two Chapters.

Chapter 8

Model parameter estimation

The signal analysis method proposed in the previous Chapter can only function if the systems associated with the signal models for spikes, $G(z)$ as well as the background activity, H, are known a priori. This condition is not satisfied, since it was stressed in section 7.4 that the modelling procedure is not based on physical processes and that the only element available in reality is the composite recorded EEG signal, $e(n)$. This Chapter serves two purposes, to bridge the gap between analysis and implementation and to propose such an implementation. The first comprises of the demonstration that the estimation of the required systems is feasible. The second is essentially a realization of such an estimation procedure, subject to additional constraints mainly for on-line implementation.

The methods derive from the general optimum least-squares method of system design, introduced to engineering in 1949 by Norbert Wiener and is named after him. It will be shown that *Wiener filtering* may be employed to estimate H from $e(n)$. With further considerations, $G(z)$ may also be regarded as a *Wiener filter* and thus computed using similar techniques. The *principle* of system estimation is described first. Algorithms for its realization are presented later.

8.1 Optimal Wiener filtering

Wiener filtering addresses the problem of linear signal estimation. This deals with the approximation (estimation) of one discrete-time sequence using a weighted linear combination of elements of another. Four elements are required for the formulation of the solution:

- A sequence of samples, $x(n)$, termed the *input sequence*
- A second sequence, $y(n)$, called the *desired response*

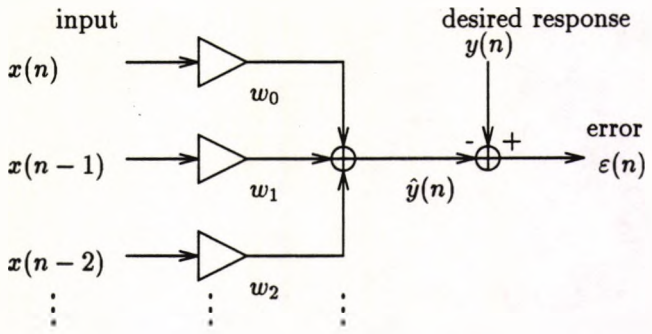


Figure 8.1: Principle of Wiener filtering

- A set of coefficients, w_i, known as *weights*
- A goodness of fit criterion, ξ, the *cost function*

The input sequence $x(n)$ is often, but not always, a temporal sequence, created by discretization of a continuous time function, $x(t)$ and hence may have infinite length. EEG signals belong to this category, but it is possible to have a sequence that is spatially related, or a finite-length sequence of discrete variables of (possibly) heterogeneous sources (e.g. $x(0)$ may represent pressure, $x(1)$ distance etc.). Mixtures of different input types is also possible, but these are beyond the scope of this work.

The weights, w_i are applied on $x(n)$ to form a linear combination of elements of this sequence (see Figure 8.1),

$$\hat{y}(n) = \sum_{i=0}^{M} w_i \, x(n-i) \tag{8.1}$$

This description implies that the set of $\{w_i\}$ is finite and consists of $M+1$ elements, since for every n, $\hat{y}(n)$ is a function of only $M+1$ elements of $x(n)$. Such an operation may be regarded as an FIR, causal filter. According to Wiener theory i may extend indefinitely ($M = \infty$), resulting in IIR behaviour[152, pp. 240–248], or take negative values also, a non-causal realization[152, pp. 249]. With these generalizations, the estimation may take into account all values of $\{x(n)\}$, extending infinitely into its 'past' and 'future'. Wiener filtering is better expressed using vector notation as this is simpler and more illustrative. Hence the input,

$$\mathbf{x}(n) = (\ldots, x(n+1), x(n), x(n-1), \ldots)^T \tag{8.2}$$

is a vector of possibly infinite length, consisting of the elements of the input sequence, $\{x(n)\}$, 'centred' at the nth element. The weight vector,

$$\mathbf{w} = (\ldots, w_{-1}, w_0, w_1, \ldots)^T \tag{8.3}$$

is formed by all the weights in the unconstrained (IIR, non-causal) filter. The operation of the filter (equation 8.1) may then be expressed as the *inner product* of $\mathbf{x}(n)$ and $\mathbf{w}$:

$$\hat{y}(n) = \mathbf{w} \cdot \mathbf{x}(n) = \mathbf{w}^T\mathbf{x}(n) = \mathbf{x}(n)^T\mathbf{w} \tag{8.4}$$

Having described the operation of the filter, the rôle of the desired response and the cost function will now be outlined. These are the core of Wiener filtering, as the objective is to find a suitable weight vector, $\mathbf{w}^*$, so that the output sequence of the filter, $\{\hat{y}(n)\}$ (equation 8.4) is 'as close as possible' to the desired response sequence $\{y(n)\}$ (Figure 8.1). That is for a given $\mathbf{x}(n)$, the output $\hat{y}(n)$ of the filter with weights $\mathbf{w}^*$ must be the 'best estimate' of the desired response $y(n)$ and this must be true for all n without changing the weights. The purpose of the cost function, ξ, is to quantify the goodness of the estimation. It is typically a function of the *error of the estimation*, $\varepsilon(n) = y(n) - \hat{y}(n)$, as depicted in Figure 8.1. The cost function usually employed is the *mean squared error function* (MSE), defined as the expected value of the squared error for all n:

$$\xi = \bar{\varepsilon^2} = E[\varepsilon(n)^2] = E\left\{[y(n) - \mathbf{w}^T\mathbf{x}(n))^2]\right\} \tag{8.5}$$

The choice of $\bar{\varepsilon^2}$, is associated with many desirable properties of estimation, some of which are listed at the end of this section, and is well-supported by linear algebra, leading to great simplification and increased versatility of the analysis. Using the linearity of the Expectation operator the above expression may be transformed to the following[202, pp. 19–21] (see Appendix C):

$$\xi = r_{yy}(0) - 2\mathbf{w}^T\mathbf{r}_{xy} + \mathbf{w}^T R_{xx}\mathbf{w} \tag{8.6}$$

where, by definition, $r_{yy}(0) = E[y(n)^2]$, $\mathbf{r}_{xy} = E[y(n)\mathbf{x}(n)]$ is the cross-correlation vector between $\mathbf{x}(n)$ and $y(n)$ and $R_{xx} = E[\mathbf{x}(n)\mathbf{x}(n)^T]$ is the auto-correlation matrix of $\mathbf{x}(n)$.

The optimum weight vector in this sense is the one corresponding to the minimum value of ξ, which occurs when the gradient of ξ with respect to $\mathbf{w}$,

$$\nabla_w\xi = (\ldots, \frac{\partial\xi}{\partial w_{-1}}, \frac{\partial\xi}{\partial w_0}, \frac{\partial\xi}{\partial w_1}, \ldots)^T \tag{8.7}$$

vanishes. Applying the gradient operator on expression 8.6,

$$\begin{aligned} \nabla_w\xi &= \nabla_w[r_{yy}(0)] - 2\nabla_w[\mathbf{w}^T\mathbf{r}_{xy}] + +\nabla_w[\mathbf{w}^T R_{xx}\mathbf{w}] \\ &= -2\mathbf{r}_{xy} + 2R_{xx}\mathbf{w} \end{aligned} \tag{8.8}$$

and setting $\nabla_w\xi = 0$ yields the optimal value of the weight vector, since, from the last expression, $R_{xx}\mathbf{w}^* = \mathbf{r}_{xy}$ or

$$\mathbf{w}^* = R_{xx}^{-1}\mathbf{r}_{xy} \tag{8.9}$$

which is the well-known *Wiener-Hopf equation*[202][pp. 22][152, p. 241],[14, p. 297][140, p. 166] for evaluating the optimal weight vector from the auto-correlation matrix of the input vector and the cross-correlation vector between the input vector and the desired response.

Wiener filtering, as a method for estimating one signal from another, possesses several desirable properties, analysed in texts on optimal signal processing[202, pp. 26–43][152, p. 239–248]. The most important of these are listed below (their derivation may be found in Appendix C):

- The mean squared error, being a function of $\mathbf{w}$ has the shape of a hyperparaboloid, which is concave upwards and, hence, has only one minimum at $\mathbf{w} = \mathbf{w}^*$.
- The value of the MSE at this point is given by

$$\xi_{min} = r_{yy}(0) - \mathbf{r}_{xy}^T \mathbf{w}^* \tag{8.10}$$

- Weight vectors corresponding to the same value of ξ lie on concentric (hyper)ellipses centred at $\mathbf{w}^*$.
- By defining $\mathbf{v} = \mathbf{w} - \mathbf{w}^*$, it is easy to show that the MSE and its gradient are given by the following expressions[202, pp. 23–25]:

$$\begin{aligned} \xi &= \xi_m in + \mathbf{v}^T R_{xx} \mathbf{v} \\ \nabla_w \xi &= 2 R_{xx} \mathbf{v} \end{aligned} \tag{8.11}$$

- Perhaps the most important property of the filter, arising from the use of the MSE criterion for optimization is the *decorrelation of the error and the input* when the optimal weight vector, $\mathbf{w}^*$ is used:

$$E[\varepsilon(n)\mathbf{x}(n)] = 0 \tag{8.12}$$

  In other words, the error vector is orthogonal to the input vector. This has a geometric interpretation. Since the MSE function is proportional to the modulus of the error vector, $||\mathbf{e}(n)||$, it is minimized when $\mathbf{e}(n) = [\ldots, \varepsilon(n-1), \varepsilon(n), \ldots]^T$, is perpendicular to the (hyper)plane spanned by $\mathbf{x}(n)$.

- Consequently, the output of the filter $\hat{\mathbf{y}}(n)$ and $\mathbf{e}(n)$ are also orthogonal.

Wiener filtering is a powerful modelling tool, made more appealing because of the existence of fast and computationally effective implementations, like adaptive filtering, introduced in section 8.4.

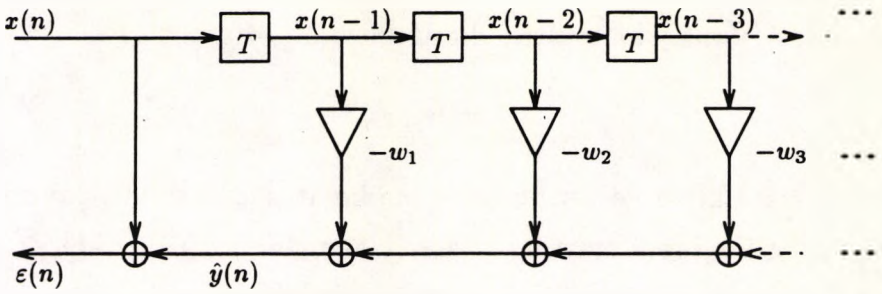


Figure 8.2: A one-step linear predictor

8.1.1 Optimal linear prediction

The relevance of Wiener filtering to EEG signal modelling and analysis may not be immediately apparent. One special case, linear prediction, is particularly useful for signal modelling and deconvolution[151, pp. 452–457][152, pp. 231–251][140, pp. 265–301].

In *linear prediction* a value, $x(n+N-1)$, in a sequence $\{x(n)\}$ is to be estimated (predicted) from the (known) past samples of the same sequence, $\mathbf{x}(n-1) = (\ldots, x(n-2), x(n-1))^T$.

If this is to be performed by linear weighting, an optimum weight vector, $\mathbf{w}^*$, that minimizes a cost function, ξ, of the prediction error, $\varepsilon(n)$, is sought. Evidently, if ξ is the MSE function, linear prediction becomes a special case of Wiener filtering, where the desired response is $y(n) = x(n+N-1)$ and the output of the filter, $\hat{y}(n) = \hat{x}(n+N-1)$, is an estimate of this. This is a *constrained Wiener filter*, and is causal since the weights for 'future' samples are set to zero ($\mathbf{w} = (\ldots, 0, 0, w_1, w_2, \ldots)^T$). Special attention is drawn to the case when the next sample in the sequence is to be predicted for reasons that will be explained.

One-step forward prediction

In *one step forward prediction*, $N = 1$, and hence $y(n) = x(n)$. A diagram of this system is shown in Figure 8.2. The output and error signal for this structure are given by the following relations:

$$\hat{x}(n) = \sum_{i=1}^{\infty} w_i x(n-i) \tag{8.13}$$

$$\begin{aligned} \varepsilon(n) &= x(n) - \hat{x}(n) \\ &= x(n) - \sum_{i=1}^{\infty} w_i x(n-i) \end{aligned} \tag{8.14}$$

By comparison with the Wiener-Hopf solution (equation (8.9)), the optimal prediction weight vector, $\mathbf{w} = (w_1, w_2, \ldots)^T$ may be evaluated as follows:

$$\mathbf{w}^* = \hat{R}_{xx}^{-1}\mathbf{r}_{xx} \tag{8.15}$$

where the cross-correlation vector between the input and desired signals is now the autocorrelation vector of the input starting at zero lag, $\mathbf{r}_{xx} = [r_{xx}(0), r_{xx}(1), r_{xx}(2), \ldots]^T$. The auto-correlation matrix is slightly different from the definition given in the previous section, since the first row and column have been eliminated as the input vector at time n is $\mathbf{x}(n-1)$ instead of $\mathbf{x}(n)$. Hence the matrix used here is $\hat{R}_{xx} = E[\mathbf{x}(n-1)\mathbf{x}(n-1)]^T$. Because of the way the elements of the autocorrelation sequence of the input are combined in the above equation, substantial simplifications in the computation of $\mathbf{w}^*$ are possible, but these are considered later in the implementation.

The expressions for $\hat{x}(n)$ and $e(n)$, given above, represent causal systems with infinite impulse response in the general case. Equation 8.14 gives the error of the prediction and is of particular interest in signal modelling. It should be noted that $P_f(z)$ may be computed by using information extracted solely from the input sequence, namely its autocorrelation function, $r_{xx}(k)$. By re-labelling the weights, equation 8.14 may be rewritten as:

$$\varepsilon(n) = \sum_{i=0}^{\infty} h_i x(n-i) \tag{8.16}$$

where $h_0 = 1$ and $h_i = -w_i, i > 0$. Using the z-transform $P_f(z)$ may be readily written:

$$P_f(z) = \frac{E(z)}{X(z)} = \sum_{i=0}^{\infty} h_i\, z^{-i} \tag{8.17}$$

This system has the property that when the sequence $x(n)$ is applied at its input, a white uncorrelated sequence is produced at its output. Conversely, if the white noise sequence $\varepsilon(n)$ is passed through the inverse of this system, $P_f^{-1}(z)$ the output is the sequence $x(n)$, since

$$X(z) = \frac{E(z)}{P_f(z)} = \frac{E(z)}{\sum_{i=0}^{\infty} h_i\, z^{-i}} \tag{8.18}$$

This signal representation is similar to the modelling process described earlier in section 7.3.1. Now, however, it is related to Wiener filtering, which readily provides a method for the computation of $P_f(z)$ and $\varepsilon(n)$.

Because of its effects in the frequency and the time domain, the transfer function $P_f(z)$ is descriptively called a *prediction-error*, *whitening*, or *analysis filter*[140, p. 64]. The white output sequence, $\varepsilon(n)$, is called the *innovations process*[152, p. 210][140, p. 174] for the sequence $x(n)$.

The prediction-error filter and the innovations process share the properties of Wiener filtering, which acquire a new meaning because the desired response is the input signal itself. Detailed discussions on those may be found in texts on linear prediction[152, pp. 208–251][140, pp. 195–301].

The orthogonality principle of the input and the error sequences described by equation 8.12 in the previous section demonstrates that the innovations $\varepsilon(n)$ and the input $x(n)$ sequence are uncorrelated. Hence the system $P_f(z)$ may be regarded as a parametric form of summarizing the 'predictable' attributes in $x(n)$, whereas $\varepsilon(n)$ represents the "unpredictable" element, or new information that is present in every new sample of this sequence[140, p. 174]. The representation of a stationary stochastic process $x(n)$ in terms of a system $P_f(z)$ and the innovations process $\varepsilon(n)$ is known as the *Wold representation*[152, p. 210].

A consequence of the criterion used for the optimization is the reduction in power between the input and the output of the system $P_f(z)$. By substituting $y(n)$ by $x(n)$ and delaying the input sequence by one sample in time, the minimum MSE given by equation 8.10, may be rewritten as

$$\xi_{min} = r_{xx}(0) - \bar{\mathbf{r}}_{xx}^T \mathbf{w}^* \tag{8.19}$$

where $\bar{\mathbf{r}}_{xx}$ is the autocorrelation sequence of $x(n)$ for $n < 0$. The corresponding autocorrelation matrix $\check{R}_{xx}$ is such that by the Wiener-Hopf equation $\check{R}_{xx}\mathbf{w}^* = \bar{\mathbf{r}}_{xx}$. Substituting $\bar{\mathbf{r}}_{xx}$ in the last equation gives

$$\xi_{min} = r_{xx}(0) - (\check{R}_{xx}\mathbf{w}^*)^T\mathbf{w}^* = r_{xx}(0) - \mathbf{w}^{*T}\check{R}_{xx}\mathbf{w}^* \tag{8.20}$$

where by symmetry $\check{R}_{xx}^T = \check{R}_{xx}$. Since this matrix is positive semi-definite, then for any vector $\mathbf{w}$, $\mathbf{w}\check{R}_{xx}\mathbf{w} \geq 0$. Hence $r_{xx}(0) - \xi_{min} \geq 0$. By definition, $\xi_{min} = \sigma_\varepsilon^2$ represents the power (variance) of the innovations sequence and $r_{xx}(0) = \sigma_x^2$ the power in the input sequence. It is therefore evident that

$$\sigma_\varepsilon^2 \leq \sigma_x^2 \tag{8.21}$$

and therefore the innovations sequence has less power than the input sequence. Equality is possible if the input is a white sequence (or an impulse sequence), when $\check{R}_{xx} = I$. When the signal sequence is deterministic (fully predictable), the power of the innovations sequence becomes zero.

Linear prediction based on a finite number of past samples

Finally, it should not be forgotten that the only constraint imposed on the linear predictor is that of causality. The weight vector, $\mathbf{w}$ extends infinitely taking into account the whole of

the past samples of the input sequence, $x(n)$ for the prediction. This has obvious difficulties in its realization, as $x(n)$ is only known for a finite number of past samples and an infinite-length weight sequence represents a non-computable problem. In practice, the length of the weight vector (and hence the dimension of $\check{R}_{xx}$ and $\check{\mathbf{r}}_{xx}$) is restricted to some finite value, M. This may result in a predictor with inferior performance than the one described above. In many practical applications, including the EEG (see section 8.4.3), the magnitude of the weights decreases rapidly for large lags and eventually becomes insignificant. Hence it is generally possible to truncate the weight vector to a length M that is adequately large so that the contribution of the weights not included in $\mathbf{w}$, $\{w_i; i > M\}$ have a negligible effect on the performance of the predictor. Such a system has a finite impulse response and its inverse is an all-pole (AR) structure:

$$P_f^{-1}(z) = \frac{E(z)}{X(z)} = \frac{1}{1 - \sum_{i=1}^{M} w_i z^{-i}} \tag{8.22}$$

It is clear that if $x(n)$ is produced when an autoregressive system of order L, $Q(z) = \frac{1}{1+\sum_{i=1}^{L} a_i z^{-i}}$ is excited by a sequence with white PSD, $r(n)$, then an optimal linear predictor of order $M \geq L$, $P_f(z) = 1 - \sum_{i=1}^{M} w_i z^{-i}$ will be identical to the inverse of the generating system:

$$w_i = \begin{cases} -a_i & 1 \leq i \leq L \\ 0 & L < i \leq M \end{cases} \tag{8.23}$$

If the system $Q(z)$ has zeros, as well as poles, the FIR linear predictor described will still yield an approximation to the inverse system, but this will never be exact, although it will get better with increasing M. There have been attempts to formulate an ARMA linear predictor which could approximate transfer functions with infinite impulse response more economically (smaller number of parameters). Unfortunately, the presence of feedback results in a set of non-linear equations[152, pp. 212–213]. There is no guarantee for a unique solution in this case, because the performance surface $\xi(\mathbf{w})$ is neither quadratic nor unimodal for some weights. Besides, there are problems of stability of the resulting predictor[202, pp. 135–137]. These are general problems arising from the use of rational fractions as Wiener filters and are not specific to the prediction problem.

The error sequence out of the linear prediction error filter $P_f(z)$, $\varepsilon(n)$. will be the same as the input sequence, of $P_f^{-1}(z)$ and will have an autocorrelation function that consists of an impulse at the origin with amplitude equal to the power of the sequence, σ_ε^2, e.g. white noise or an impulse. The PSD of the input sequence, $P_{xx}(\omega)$, may be readily evaluated from the inverse of the predictor $P_f(z)$ and σ_ε^2:

$$P_{xx}(\omega) = \frac{\sigma_\varepsilon^2}{P_f(z)P_f(z^{-1})}\Big]_{z=e^{j\omega T}} \tag{8.24}$$

which is a secondary, but possibly useful, by-product of this method. Linear prediction, in general, is covered by a comprehensive review paper by Makhoul[116], who mentions its use for the computation of the spectrum of the EEG. This paper, however concentrated on the concept, properties and possible applications of linear prediction systems, rather than its on- or off-line, recursive or non-recursive implementation.

8.2 The On-going activity as a Wiener predictor

From the concepts unfolded in the previous section, the relation of linear prediction to the EEG signal, especially the synthesis and analysis model established in previous Chapter, is obvious.

The synthesis model described in section 7.3.1 is essentially an autoregressive system, H, excited by a white sequence, $u(n)$. The system H is time variant, as its coefficients change at random time instants, but for the period between two changes they are constant, resulting in a form of short-term transfer function, $H_i(z)$. This was used as part of the analysis model in section 7.4.1, to develop an inverse, $H_i^{-1}(z)$, which is central in the present system (sections 7.4.3 and 7.4.4). The structure of this inverse is precisely that of a linear one-step prediction error system, a fact that is easily established by comparing the structure of $H_i(z)$ (equation 7.13) with the inverse of the prediction-error structure, $P_f^{-1}(z)$ (equation 8.22). If both systems have the same order, $M = N$, then there is a direct correspondence between their coefficients, $a_i = w_i, \forall i = 1, \ldots, N$. Hence it is possible to evaluate $H_i^{-1}(z)$ using linear prediction, provided that $b(n)$ is known. In this case, the innovations sequence out of the predictor would be the generating white sequence, $u(n)$, which agrees with the analysis procedure of section 7.4.1. In practice, $b(n)$ cannot be separated from the other components of the EEG signal, but the whole signal $e(n)$ may be used for the estimation of this transfer function. Of course $e(n)$ contains spikes, other transients and noise, but spikes and other transients are fairly rare events and have significant amplitudes only for brief durations of time. being generally uncorrelated with the background activity, these components would contribute little to the auto-correlation function of $e(n)$, which is largely similar to that of $b(n)$ (see section 8.4.3). In the implemented system the detected spikes and transients were excluded from the computation of the parameters of $H_i^{-1}(z)$ thus eliminating their effects altogether.

The presence of the noise component, $v(n)$, would affect the values of the predictor coefficients, because it is an integral part of the estimated signal. Being a random sequence it contributes to the autocorrelation function of $b(n)$ only for zero lag. If $v(n)$ is small, it may

be argued that the autocorrelation sequence of the recorded EEG signal is almost identical to that of the background activity, $r_{ee}(n) \approx r_{bb}(n)$. This permits the computation of the inverse transfer function for the generation of the background activity, $H_i^{-1}(z)$, using the recorded EEG signal, $e(n)$, and the theory of optimal prediction-error filtering.

Two questions on this issue still remain: What should be the order, N, of the linear predictor (the order of $H_i(z)$ is not known) and how to make the linear predictor estimate the transfer functions $H_i^{-1}(z)$ for consecutive values of i, since transitions between one and the next occur at unpredictable instants (section 7.3.1). Solutions to these problems will be provided during the implementation in section 8.4.3.

8.3 Finding the Parameters of the Spike Model

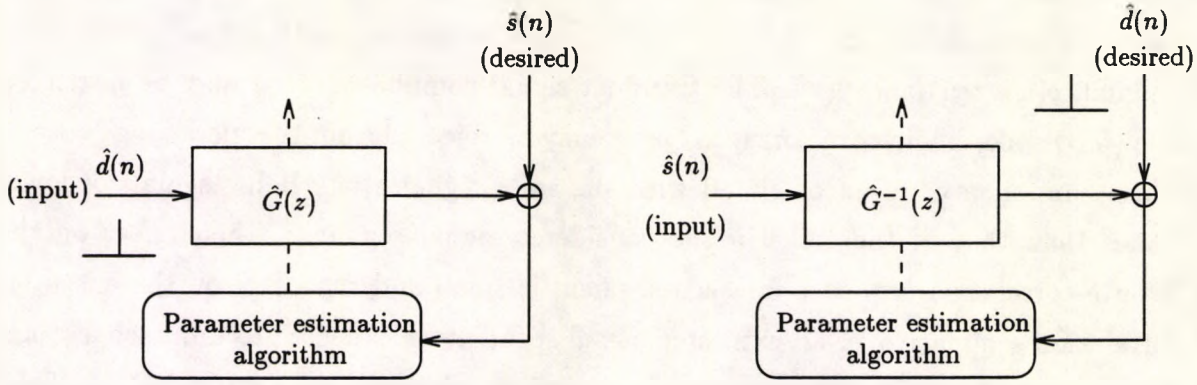
The spike signal, $s(n)$, has completely different properties from the background activity. Nevertheless, it is possible to apply Wiener theory to find a decomposition into an input sequence of scaled impulses, $d(n)$, and an AR process with transfer function, $G(z)$ (section 7.3.2).

Because spikes are rare and their energy is concentrated in short time spans, their overall contribution to the power of the EEG signal is quite small. Thus, they must be identified and isolated from the rest of the signal, to ensure that mean squared error optimization will concentrate on the spikes, not on the background activity or other transients. For this purpose, the initial differential spike detector described in a previous Chapter (section 6.5) was employed. This identified most of the spikes in a number of data records. The spike detection signal (a binary function of 'time') was displayed along with the EEG signal and amendments were made to remove the few false detections and introduce a detection when a spike was missed. A window of length K was then formed around the points of detection to extract the segments of the signal containing spikes.

This semi-automatic procedure, supplemented with an appropriate user interface could also be used by an expert clinician to mark spikes in new records, acting as a supervisor correcting the response of the system and providing a means for its adjustment without the need to understand the underlying signal processing concepts.

The EEG signal with the spikes marked does not by itself suggest any particular method for the evaluation of the required transfer function. Two approaches are possible, based on the definition of $G(z)$ as an AR system whose impulse response is a spike.

According to the first scheme, an input sequence of spikes may be generated, one per spike. This is an estimate of the generating sequence, $\hat{d}(n)$, which is subsequently applied



(a) Direct modelling (b) Inverse modelling

Figure 8.3: Alternative schemes for the estimation of the spike model

to the system to be estimated, $\hat{G}(z)$ and its output is subtracted from the genuine spike, presumed generated from the unknown $G(z)$ with excitation by the impulse sequence $d(n)$ with possibly added noise. The parameters of $\hat{G}(z)$ are then optimized according to the Wiener-Hopf equation 8.9, assuming that the desired response is a genuine spike, $s(n)$ and the input signal is the sequence of impulses, $d(n)$ (or rather its estimate $\hat{d}(n)$):

$$\mathbf{g}^* = R_{dd}^{-1}\mathbf{r}_{ds} \tag{8.25}$$

This method, illustrated in Figure 8.3(a) but is subject to two limitations. Firstly, the estimated sequence $\hat{d}(n)$ cannot be exact, since neither the times of occurrence nor the amplitudes of the impulses are known and their estimation is subject to errors that affect both the consistency of the estimation and the behaviour of the derived filter. Then, the system estimate, $\mathbf{g}^*$, being an all-zero model, would represent nothing but an average 'matched filter' in reverse, an implementation lending itself to little flexibility and poor signal parametrization. The methodology was developed with the prospect of future expansion to the description of other transients, some of which may have simple shapes but long durations, where this filter might represent an optimal solution in terms of Wiener theory, but a rather poor one in terms of number of coefficients to store and number of mathematical computations required for their application on the signal. Besides, the system is not guaranteed to be minimum-phase and its inverse, which is also part of the system (section 7.4.4) could be unstable.

The alternative, depicted in Figure 8.3(b), represents an attempt to estimate the inverse of the spike generating process, $G^{-1}(z)$. According to this scheme, the segments containing the extracted spikes are the input to a linear prediction error filter. In the absence of noise, the output of the filter should be an impulse, provided that the order of the filter is

adequately high, at least the same as the order of the spike generating autoregressive system described in section 7.3.2. In practice there is 'noise' consisting of the other components of the signal, $b(n)$, $t(n)$ and $v(n)$. The transient signal component, $t(n)$ may be neglected since the (few) spikes which are corrupted by it may be rejected from the modelling process. The other components are uncorrelated with the spike signal $s(n)$. Being signals of lower amplitudes than that of the spike in the considered signal segments, their effect on the sample auto-correlation sequence for each segment is small and the effect on the ensemble auto-correlation sequence over all extracted signal segments is expected to diminish as their number increases. Therefore, the sample signal auto-correlation is an approximation of the spike auto-correlation, $r_{ss}(n) \approx r_{ee}(n)$. The coefficients $\mathbf{g}^* = (g_1\ g_2\ \ldots\ g_L)^T$ of the inverse system, $G^{-1}(z) = 1 - \sum_{i=1}^{L} g_i\, z^{-i}$ (sections 7.3.2 and 7.4.2) are given by the (Wiener-Hopf) equation:

$$\mathbf{g}^* = \check{R}_{ee}^{-1}\mathbf{r}_{ee} \tag{8.26}$$

which is basically the solution to the optimum linear prediction problem (equation (8.15)). This expression is similar to the one already given for the background activity in the previous section. It can therefore describe precisely an autoregressive spike model of order $M \leq L$, in the absence of noise. When noise is present or the model is ARMA or MA, a reasonable approximation may still be achieved, but the output signal may contain excess error components due to poor modelling.

Provided that the order of the predictor is adequate and that the noise level in the input is low, a clearly distinguished impulse should be detected in the error sequence, $\varepsilon(n)$ out of $G^{-1}(z)$ when a signal segment containing a spike is input. This may be used to verify the correctness of the model and serves as an integrity test on the selection of the order of the predictor.

A systematic way for finding the inverse filter, $G^{-1}(z)$ is based on a form of constrained Wiener filtering that conforms to the optimal solution presented as equation (8.9), but uses a finite-length weight vector $\mathbf{g} = (g_0\, g_1\ \ldots\ g_L)^T$ as in section 8.1.1. An additional constraint is the limited length of the input and output sequences. Under these conditions, the required optimal vector $\mathbf{g}^*$ should be such that the output, $d(n)$ of the system is an impulse. Because of the shape of its output, $\mathbf{g}^*$ is often called a *spiking filter*, a misleading term in the context of this report, where the term 'spike' has a different significance. A good analysis of spiking filters which are primarily used in deconvolution is given by Orfanidis[140, pp. 286–301]. The equation that gives the optimal spiking filter for an epileptic spike pattern is:

$$\mathbf{g}_k^* = R_{ee}^{-1}\mathbf{r}_{ed_k} \tag{8.27}$$

where $\mathbf{r}_{ed_k}$ is the cross-correlation between the signal containing the epileptic spike and the

desired response, $d_k(n) = u_0(n-k)$ (an impulse delayed by k samples). The introduction of the index k permits the consideration of several spiking filters, for various delay factors. If the input sequence of N samples, is written as a vector $\mathbf{e}(n) = [\mathbf{0}^T\, e(n)\, e(n+1)\, e(n+2) \ldots e(n+N)\, \mathbf{0}^T]^T$, where $\mathbf{0}$ is an (infinitely long) zero vector, one may introduce a delay k and write $\mathbf{e}(n-k)$. By defining the input matrix $E = [\mathbf{e}(n)\, \mathbf{e}(n-1)\, \mathbf{e}(n-2) \ldots \mathbf{e}(n-L)]^T$, the output when $e(n)$ is applied to $\mathbf{g}^*$ is simply the vector $\hat{\mathbf{e}}_k = E\mathbf{g}_k^*$. Similarly, the desired response matrix is effectively the identity matrix since $D = (\mathbf{d}_0(n)\, \mathbf{d}_1(n)\, \mathbf{d}_2(n) \ldots \mathbf{d}_L(n))^T = I$. Using this notation, one may write the optimal spiking filters for delays $k = 0, 1, \ldots, L$ in matrix form

$$G^* = R_{ee}^{-1} R_{ed} = R_{ee}^{-1}\, E^T \tag{8.28}$$

where

$$\begin{aligned} R_{ed} &= (\mathbf{r}_{ed_0}\, \mathbf{r}_{ed_1} \ldots \mathbf{r}_{ed_L})^T \\ &= E^T D = E^T \end{aligned}$$

Noting that the outputs of these filters, G^*, are given by the columns of matrix $\hat{D} = E\, G^* = E\, R_{ee}^{-1}\, E^T$, the best spiking filter $\mathbf{g}^* = \mathbf{g}_k^*$ is the one for which the kth element on the diagonal of D has the largest amplitude over all other elements[140, p. 289].

8.3.1 Towards a practical solution

In this section it has been illustrated that it is possible (at least in theory) to estimate the transfer function(s) for the generating process of the background activity, $H_i(z)$ using optimum linear prediction. Similarly, two schemes have been suggested for the estimation of the generating system for epileptic spikes, $G(z)$, which are also based on Wiener theory and linear prediction. One limitation of the Wiener-Hopf solution (equation (8.9)) its the use auto- and cross-correlation functions which are sums of products of signals that theoretically extend infinitely into the past and future. In any implementation, these must be replaced by sample correlations computed over a finite time interval. If the signals are stationary, any short-time correlation will be an estimate of the theoretical correlation, but in the presence of non-statonarities there will be inconsistencies.

Short-time correlations may be suitable for the computation of the spike transfer functions, since only short segments of the signal are employed. In the case of the background activity, the signal changes with time and its non-stationarity apparently invalidates the use of sample correlations. Recalling, however, that the model of this signal has been described by a *set* of transfer functions being in effect at different times, it makes sense to use correlations that are valid locally to estimate each one of these. Sample correlation functions may be

estimated in a number of ways, depending on how the expectation operator is implemented. These include from summation over a window of samples and recursive estimations.

The explicit computation of the auto- an cross-correlation functions is not a necessity, however Wiener filters may be computed by numerical optimization, using gradient search of the performance surface, $\xi(\mathbf{w})$ or otherwise. Some of these methods estimate the optimal vector recursively. These are sometimes suitable for on-line processing and are often better, quicker and more elegant than their block-by-block sample correlation counterparts. A class of recursive methods of system estimation, descriptively called *adaptive filters* are the subject of the next section.

8.4 Practical computation of model parameters by adaptive techniques

Wiener theory depends on correlation functions and although it is possible to estimate these from a (time) window of the signals considered, better approaches exist for finding the optimal weight vector. Because samples of the EEG signal are continuously generated, it is more sensible to use new samples to improve the estimates of the correlation functions which appear in the Wiener-Hopf equation (8.9). This suggests an *iterative method* for updating R_{xx} and $\mathbf{r}_{dx}$, as it would be pointless to recalculate them anew with the arrival of every sample. Indeed, recursive methods for their computation do exist (see Bellanger[14, pp. 55–94], for example). It is also possible to estimate R_{xx}^{-1} directly, enabling the computation of $\mathbf{w}^*$ without matrix inversion (see section 9.1).

More radical methods do not use the analytical solution of the Wiener-Hopf equation (8.9), but are based on its significance. Noticing that it gives the weight vector $\mathbf{w}^*$ for which $\xi(\mathbf{w})$ reaches its minimum value, the optimal solution may be reached through a method that estimates this point. Gradient search techniques are popular in numerical optimization, because they are simple and avoid matrix arithmetic. Their underlying theory is summarized next.

8.4.1 Recursive solution of the Wiener-Hopf equations by gradient descent

The recursive estimation of the optimal weight vector utilizes properties of the MSE surface, $\xi(\mathbf{w})$, a scalar function of the weight vector (see Appendix C). The algorithms described here exploit the quadratic nature of ξ, which possesses a single minimum at $(\mathbf{w}^*, \xi_{min})$.

Points of equal ξ have the shape of hyperellipses centred around the above point, whereas cross-sections of the error surface along the radii of these ellipses have parabolic shape. By translating the minimum point to the origin and rotating the coordinate system so that the axes of the ellipses are aligned with the coordinate axes (see Appendix C), ξ may be expressed in the new vector variable $\mathbf{u} = Q^T(\mathbf{w} - \mathbf{w}^*)$ to the following:

$$\xi(\mathbf{u}) - \xi_{min} = \mathbf{u}^T \Lambda \mathbf{u} \tag{8.29}$$

where Q is the modal matrix and Λ is the (diagonal) eigenvalue matrix.

Steepest descent is a numerical method for the estimation of the minimum of a function like $\xi(\mathbf{w})$, by progressively moving the estmated minimum point guided by the direction of the gradient (vector), $\nabla\xi(\mathbf{w})$. Intuitively, starting from an initial guess, $\mathbf{w}(0)$, of the point of minimum ξ, a better guess would result from the introduction of a correction $\Delta\mathbf{w}(0) = -\mu\nabla\xi(\mathbf{w}(0))$, where μ is the size of the step or *learning rate* and $\nabla\xi(\mathbf{w}(0))$ is the gradient at that point. The negative sign indicates that the step should be taken in a direction *opposite* to that of the gradient vector. The process is applied iteratively, according to the following updating recursion[200, 201, 202]:

$$\mathbf{w}(k) = \mathbf{w}(k-1) + \Delta\mathbf{w}(k-1) = \mathbf{w}(k-1) - \mu\nabla\,\xi(\mathbf{w}(k-1)) \tag{8.30}$$

and as the number of iterations, k, increases, $\mathbf{w}(k)$ gradually approaches the true minimum $\mathbf{w}^*$.

For a quadratic function like ξ, where a single minimum exists, the stability of the method (convergence to this point) depends almost entirely on the size of μ. To investigate convergence, the gradient vector is replaced by by equation (8.8), $\nabla\,\xi[\mathbf{w}(k)] = 2R_{xx}\mathbf{w}(k) - 2\mathbf{r}_{xy}$, after the substitution of $\mathbf{r}_{xy} = R_{xx}\mathbf{w}^*$ from equation (8.9) followed by the subtraction of $\mathbf{w}^*$ from both sides:

$$\begin{aligned} \mathbf{w}(k) &= \mathbf{w}(k-1) - 2\mu\, R_{xx}[\mathbf{w}(k-1) - \mathbf{w}^*] \\ \mathbf{w}(k) - \mathbf{w}^* &= \mathbf{w}(k-1) - \mathbf{w}^* - 2\mu\, R_{xx}[\mathbf{w}(k-1) - \mathbf{w}^*] \end{aligned} \tag{8.31}$$

Working in reverse, compared to equation (8.29) and replacing $(\mathbf{w} - \mathbf{w}^*)$ by the equivalent $Q\mathbf{u}$[202, pp. 57–60] converts the above relation to the following:

$$Q\mathbf{u}(k) = Q\mathbf{u}(k-1) - 2\mu\, R_{xx}Q\mathbf{u}(k-1) \tag{8.32}$$

which may be rotated to the principal (eigen) coordinate system $\mathbf{u}$ by multiplying the whole relation on the left by $Q^{-1} = Q^T$ and substituting $R_{xx} = Q\Lambda Q^T$ (see Appendix C), as with equation (8.29):

$$\mathbf{u}(k) = (I - 2\mu\,\Lambda)\mathbf{u}(k-1) \tag{8.33}$$

since the matrix, $I - 2\mu\ \Lambda$ is diagonal, the vectors on the two sides of these relation are decoupled and the recursion may be converted into the (non-recursive) geometric sequence,

$$\mathbf{u}(k) = (I - 2\mu\ \Lambda)^{k-1}\mathbf{u}(0) \tag{8.34}$$

which converges when $\mathbf{u}(k)$ approaches the required point, the origin, as k gets large:

$$\begin{aligned} & \lim_{k\to\infty}(I - 2\mu\ \Lambda)^{k-1} = \mathbf{0} \\ \Rightarrow\quad & \lim_{k\to\infty}(1 - 2\mu\ \lambda_n)^{k-1} = 0 \quad n = 0, 1, \ldots \end{aligned} \tag{8.35}$$

where λ_n is the nth eigenvalue of R_{xx}. This is true when $0 < \mu < \frac{1}{\lambda_n}$, $\forall n$ which is globally satisfied for the worst case, occurring with the largest eigenvalue, λ_{max}:

$$0 < \mu < \frac{1}{\lambda_{max}} \tag{8.36}$$

This result is very important, because it may be employed during implementation, thus ensuring convergence by avoiding its estimation by trial and error, which would limit the usefulness of the method for on-line realizations in the absence of a human monitor, like in long-term EEG analysis.

The speed of convergence of $\mathbf{w}(k)$ towards $\mathbf{w}^*$, which increases as μ approaches its upper bound, may be expressed in the form of time constants, each one corresponding to an eigenvalue λ_n given approximately by $\tau_{n_w} \approx \frac{1}{2\mu\lambda_n}$. The *learning curve*, $\xi(k)$ (MSE as a function of the number of iterations, k), indicates how ξ is reduced at every iteration. It has the form of a sum of decaying exponentials[202, pp. 60–63], whose time constants are half of those mentioned, and are given by $\tau_n \approx \frac{1}{4\mu\lambda_n}$[202, pp. 81–83]. The speed of convergence depends, primarily, on the the minimum eigenvalue, since, for a given μ, it corresponds to the slowest mode (largest time constant) of the learning curve, because the geometric ratio $(1 - 2\mu\lambda_{min})$ is closer to unity than any other ratio[140, pp. 422–423]. Substituting the upper limit for μ gives $\left(1 - \frac{\lambda_{min}}{\lambda_{max}}\right)$, which suggests that the speed of convergence is inversely proportional to *eigenvalue spread*, $\frac{\lambda_{max}}{\lambda_{min}}$ of R_{xx}.

Gradient search methods, like the steepest descent, rely on the implicit assumption that the gradient vector of the MSE function, can be evaluated. Since $\xi(\mathbf{w})$, an infinite sum, can only be computed approximately, so does its gradient. Moreover, in practice, $\xi(\mathbf{w})$ is known only at discrete points and hence $\nabla\xi(\mathbf{w})$ can only be computed numerically. In its simplest form, gradient computation is based on central differences (see section 6.3.1, for instance). These approximations, cause the final MSE (after convergence) to be higher than the theoretical minimum ξ_{min}. Widrow[202, pp. 67–93] defined two measures for the deviation of the weight vector from the optimal, the *Perturbation*, P, and the *misadjustment*,

M. The first is caused by the approximation of the gradient by taking finite differences, δ, on the weight vector while adjusting the weight vector itself. It is encountered when adaptation is performed on-line and corresponds to an excess MSE of size $\gamma = \lambda_{av}\delta^2$. Perturbation is the ratio of this to the theoretical minimum MSE, ξ_{min}:

$$P = \frac{\lambda_{av}\delta^2}{\xi_{min}} \tag{8.37}$$

where λ_{av} is the average of the eigenvalues of the input auto-correlation matrix. The misadjustment is caused by the presence of 'noise' in the gradient estimates, due to the approximation of ξ by a finite sum, rather than the 'true' expectation. For the steepest descent method, this is responsible for an excess MSE, which is approximately $\Delta\xi \approx \frac{(L+1)^2\xi_{min}}{8P}(\frac{1}{\tau_n})_{av}$, where $L+1$ is the number of weights, P the perturbation (defined in the previous paragraph) and $(\frac{1}{\tau_n})_{av}$ is the average of the reciprocals of the time constants, τ_n, corresponding to the $L+1$ modes of the performance surface. The misadjustment is simply $\Delta\xi$ normalized by ξ_{min}:

$$M = \frac{(L+1)}{16NP}(\frac{1}{\tau_n})_{av} = \frac{K}{P} \tag{8.38}$$

The overall penalty in excess MSE, given by the sum of the perturbation and the misadjustment, $M_{tot} = M + P = \frac{K}{P} + P$, cannot be decreased without bound for the steepest descent method, but is optimized when $P = \frac{1}{2}M_{tot}$. In the algorithm presented next, this restriction will be removed.

8.4.2 The Widrow-Hoff LMS algorithm

One special case of the steepest descent method is of particular interest in signal processing, as it leads to significant computational simplifications, since it eliminates the need for on-line computation of gradients.

The MSE function estimate, given by the finite sum:

$$\xi(\mathbf{w}) = E[\varepsilon(n)^2] \approx \frac{1}{N}\sum_{i=0}^{N}\varepsilon(n)^2 \tag{8.39}$$

becomes more accurate with increasing N, but reasonable estimates may be obtained for small values. Widrow and Hoff[200, 201][202, pp. 99–144] used the special case, when $N = 1$, giving

$$\hat{\xi}(\mathbf{w}) = \varepsilon(n)^2 = [y(n) - \mathbf{w}^T\mathbf{x}(n)]^2 \tag{8.40}$$

which is equivalent to *ignoring the expectation operator*. This simple estimate has a larger variance than the ones for $N = 2, 3, \ldots$, but the gradient takes a simple form, making its

numerical estimation unnecessary, since

$$\nabla\hat{\xi}(\mathbf{w}) = -2\varepsilon(n)\mathbf{x}(n) \tag{8.41}$$

This is an explicit gradient estimate and defines the *Least-Mean-Squares* (LMS) algorithm for updating the weight vector:

$$\begin{aligned} \mathbf{w}(k) &= \mathbf{w}(k-1) - \mu\nabla\,\hat{\xi}(\mathbf{w}(k-1)) && (8.42)\\ &= \mathbf{w}(k-1) + 2\mu\varepsilon(k)\mathbf{x}(k-1) && (8.43)\end{aligned}$$

Although this recursive relation is a special form of steepest-descent, some of its properties are special and worth noting. One of the attributes that has not changed is the range of convergence for μ. Taking the expected value of the above expression,

$$\begin{aligned} E[\mathbf{w}(k)] &= E[\mathbf{w}(k-1)] + 2\mu E[\varepsilon(n)\mathbf{x}(k-1)]\\ &= E[\mathbf{w}(k-1)] + 2\mu E[y(n)\mathbf{x}(k-1)] - 2\mu E[\mathbf{x}(k-1)\mathbf{x}(k-1)^T\mathbf{w}(k-1)]\\ &= E[\mathbf{w}(k-1)] + 2\mu\mathbf{r}_{xy} - 2\mu R_{xx}E[\mathbf{w}(k-1)]\end{aligned}$$

Substituting $R_{xx}\mathbf{w}^* = r_{xy}$ and subtracting $\mathbf{w}^*$ from both sides of this relation gives

$$\begin{aligned} E[\mathbf{w}(k)] - \mathbf{w}^* &= E[\mathbf{w}(k-1)] - \mathbf{w}^* + 2\mu R_{xx}\mathbf{w}^* - 2\mu R_{xx}E[\mathbf{w}(k-1)]\\ \Rightarrow\ E[\mathbf{w}(k)] - \mathbf{w}^* &= (I - 2\mu R_{xx})\{E[\mathbf{w}(k-1)] - \mathbf{w}^*\} && (8.44)\\ \Rightarrow\ E[\mathbf{w}(k) - \mathbf{w}^*] &= (I - 2\mu R_{xx})E[\mathbf{w}(k-1) - \mathbf{w}^*]\end{aligned}$$

Applying the rotation implied by the eigenvector matrix and converting the recursive relation into a geometric series, as in the previous section, yields equation (8.34), as for the steepest-descent[202, pp. 101–103]. The LMS algorithm is stable when $0 < \mu < \frac{1}{\lambda_{max}}$, (see equation (8.36)). Computing the largest eigenvalue of the correlation matrix is a costly operation[14, pp. 81–83] and it is avoided in practice, by noting that, when all eigenvalues are non-negative (R_{xx} is positive (semi)definite), the following holds true:

$$\begin{aligned} \lambda_{max} &\leq \sum_{i=0}^{L}\lambda_i\\ &= \mathrm{tr}[\Lambda] = \mathrm{tr}[R_{xx}]\\ &= \sum_{i=0}^{L} r_{xx_{ii}}\end{aligned} \tag{8.45}$$

where $\mathrm{tr}[R_{xx}]$ is the *trace* of the matrix (the sum of its diagonal elements). In the case of the (temporal) EEG signal, each element on the leading diagonal of R_{xx} is equal to the variance (power) of $x(n)$, $r_{xx_{ii}} = \sigma_{xx}^2$, hence the range for the step parameter, μ, is practically set to

$$0 < \mu < \frac{1}{(L+1)\sigma_{xx}^2} \tag{8.46}$$

which is more strict than the theoretical relationship, but is much easier to compute.

The LMS algorithm converges in the direction of the gradient vector, although the estimation of the gradient is now 'noisy'. As with the steepest descent, the learning curve has $L+1$ time constants, one for each eigenvector, given approximately by $\tau_n \approx \frac{1}{4\mu\lambda_n}$, for $n = 1, 2, \ldots, (L+1)$. Evidently, the smallest eigenvalue determines the speed of convergence, as before.

Unlike the steepest descent, though, the Widrow-Hoff algorithm does not suffer from the effects of perturbation, because the gradient is not approximated numerically[202, p. 109]. Hence misadjustment is caused entirely by 'noise' in the estimate of the weight vector, caused by the poor approximation of equation (8.40). It can be shown[202, pp. 109–110] that the noise in the the weight vector, after convergence, is given by $\mathbf{q}(n-1) = \nabla\,\hat{\xi}(\mathbf{w}(k-1)) = -2\varepsilon(n)\mathbf{x}(n-1)$. The covariance matrix of $\mathbf{q}(n)$ is then $\Phi_{qq} = 4\xi_{min}R_{xx}$ which translates into $4\xi_{min}\Lambda$ in the principal-axis coordinate system. It may be demonstrated[202, p. 78] that the covariance of the weight vector in this coordinate system is given by

$$\begin{aligned}
\Phi_{uu} &= E[\mathbf{u}(n)\mathbf{u}(n)^T] = \frac{\mu}{4}(\Lambda - \mu\Lambda^2)^{-1}Q^T\Phi_{qq}Q \\
&= \mu\xi_{min}(\Lambda - \mu\Lambda^2)^{-1}\Lambda \\
&\approx \mu\xi_{min}(\Lambda)^{-1}\Lambda \\
&= \mu\xi_{min}I
\end{aligned}$$

where $\mu\Lambda^2$ was neglected since its elements are considerably less than unity. The excess MSE may then be expressed as $\Delta\xi = E[\mathbf{u}(n)^T\Lambda\mathbf{u}(n)] = \sum_{i=1}^{L+1}\lambda_i E[u_i(n)^2]$, where $E[u_i(n)^2] = \mu\xi_{min}$ are the diagonal elements of Φ_{qq}. Hence $\Delta\xi = \mu\xi_{min}\sum_{i=1}^{L+1}\lambda_i = \mu\xi_{min}\mathrm{tr}[R_{xx}]$. Dividing by ξ_{min} gives the misadjustment:

$$M_{tot} = M = \mu\mathrm{tr}[R_{xx}] = \frac{L+1}{4}\left(\frac{1}{\tau_n}\right)_{av} \tag{8.47}$$

In general, the LMS algorithm has a lower misadjustment and shorter time constants than the general steepest descent, due to weight updating on every sample as it arrives and the avoidance of gradient estimation. The misadjustment may be reduced as desired by decreasing μ, but at the expense of speed of adaptation (i.e. longer time constants).

8.4.3 Estimation of the background activity model

The development of the LMS algorithm was based on the implicit assumption that the input and desired signals are stationary, ergodic processes, so that their theoretical (statistical) auto- and cross-correlation functions are independent of time and hence equal to the average

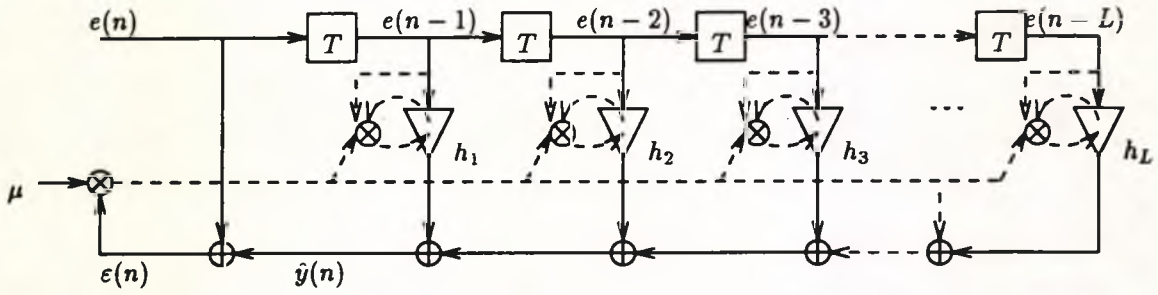


Figure 8.4: The inverse of the background activity as an LMS filter

of sample estimates. The background activity signal, $b(n)$, the main component of the EEG does not have the same uniform properties over time, as already explained in section 7.2.2. A synthesis model was based on the (reasonable) assumption that changes in the properties of the signal which are responsible for its non-stationarity occur infrequently. These were modelled as transfer functions which change at random instants, thus creating a piecewise linear system (section 7.3.1). The analysis procedure of section 7.4.1 assumed the existence of a method for the estimation of these transfer functions. Here, it will be shown that the LMS algorithm, despite the initial assumption of stationarity is, indeed, such a procedure.

The synthesis model consisted of a set of transfer functions, $\{\mathbf{H_i(z)}\}$, excited by a common random sequence, $u(n)$. One of the outputs from each of these systems is then selected by means of consecutive windows, each having a different duration $\{N_i\}$ for every segment i. The background activity $b(n)$ is the concatenation of these segments. The only signal available is the recorded EEG, $e(n)$, which may contain transients and noise, but according to the analysis procedure, transients may be detected and in most cases have a brief duration, so their impact on the auto-correlation of $b(n)$ is small. The noise, on the other hand is uncorrelated Gaussian and has only an effect on the leading diagonal of R_{ee}. Hence, as explained in the analysis procedure, $r_{bb}(n) \approx r_{ee}(n)$, for low noise levels.

The analysis method (sections 7.4.1 and 7.4.4) requires the use of the inverse of the transfer functions, $\{H_i^{-1}(z)\}$, implied by the synthesis model. Assuming that the $H_i(z)$ are AR, their inverses are MA linear prediction error filters (section 8.2). A transversal LMS adaptive prediction error system implementing $H_i^{-1}(z)$ is shown in Figure 8.4.

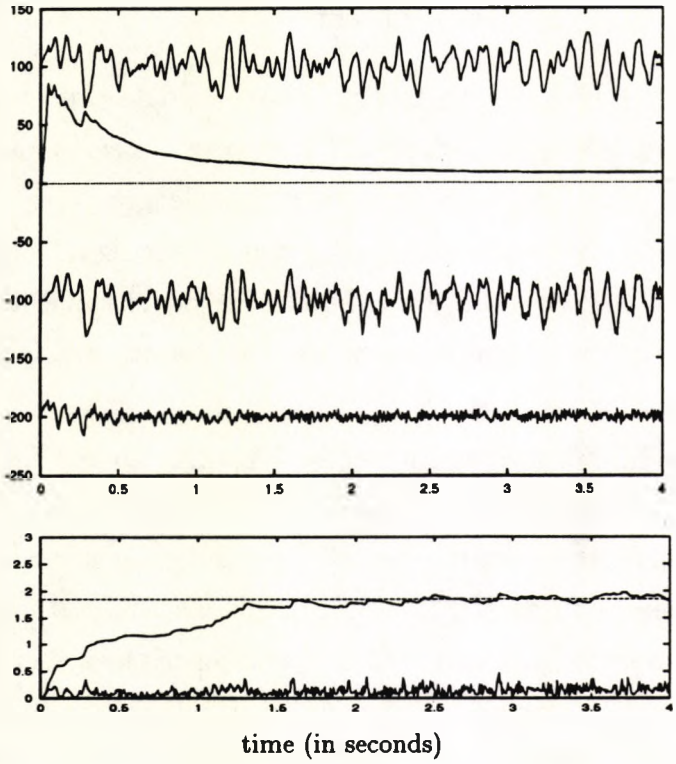
Of course, there are a number of obstacles in applying adaptive filtering to this signal. These are related to the behaviour of the algorithm during the first segment (initial convergence), during transitions between transfer functions as well as sources of weight misadjustment specific to the non-stationary nature of the signal. These issues are examined next.

Behaviour of the LMS algorithm during the first segment

Although the model suggested in section 7.3.1 extends infinitely into the past, data collection will have to begin at an instant, before which the signal history will be ignored. It is assumed that the first segment of $b(n)$, described by $H_1(z)$, with input $u(n)$ and length N_1 is in progress. The behaviour of the adaptive algorithm was studied using a synthetic signal, that conforms to the EEG model (section 7.3.5). With no a priori knowledge of its value, the initial weight vector was set to the origin of the weight space $\mathbf{w}(0) = \mathbf{0}$, corresponding to the system $H_1(z) = 1$, which is trivially minimum-phase.

The sample-by-sample updating procedure of the weight vector according to the LMS algorithm, gradually changes the weight vector bringing it closer to the optimal, the inverse of the AR system $H_1(z)$, accomplishing the decorrelation of the input sequence $b(n)$, a fundamental element of the analysis procedure (section 7.4). The implementation is computationally efficient, requiring approximately $2L$ operations (L is the order of the system), but takes a number of samples to converge, during which the system is sub-optimal and neither complete decorrelation is achieved nor minimal MSE. For the first segment the time taken for the weight vector to converge is considered. This is related to the dominant time constant of the adaptive process, $\tau_{max} \approx \frac{1}{4\mu\lambda_{min}}$, as explained in section 8.4.2. Assuming that the steady-state of adaptation is reached after the lapse of four time constants[202, p. 92], There is an overhead of approximately $M_\tau = \frac{1}{\mu\lambda_{min}}$ samples before analysis may function in the manner described. When μ is set to its theoretical maximum (equation (8.36)), the ratio of the largest to the smallest eigenvalue gives the number of samples required for the convergence. Using the more stringent practical limit of equation (8.46), M_τ is likely to increase, but always bound by $(L+1)\frac{\lambda_{max}}{\lambda_{min}}$ which is proportional to the order of the prediction error filter. For an eigenvalue spread of 100 and a filter order of 15, this would correspond to an overhead of 1500 samples, which may translate into a period of 10 seconds, if a low sampling rate of 150 s^{-1} is assumed. This is a long period, but its practical implications are not as serious as they may initially seem. Even in the extreme case just suggested, the initial setting time is insignificant compared to the time of operation of the filter, which would be several hours for on-line long-term monitoring. It should be noted, however, that if the length of the segment is shorter than M_τ, the algorithm will not have time to converge to the optimal weight vector and the error level may be significantly higher than the theoretical optimal ξ_{min}.

A more realistic situation is depicted in the waveforms of Figure 8.5, where the convergence is clearly shown in both the estimate of the mean squared error and the squared modulus of the weight vector. These results were produced using the synthetic data described



(a) Input simulated EEG signal

(b) $\hat{\xi}(n)$

(c) Predictor output

(d) Error of prediction

(e) Actual $\|\mathbf{w}\|$
(d) Estimated $\|\mathbf{w}\|$

(e) $\|\Delta\mathbf{w}\|$

Figure 8.5: Convergence of the LMS Prediction Error Filter for $b(n)$

later in section 8.6.1.

Transition between segments

The time constant(s) of the adaptive process may not be critical in the initial convergence for the reasons explained in the previous section, but they are expected to affect the performance of the algorithm during the transition between consecutive segments. This is related to the ability of the algorithm to *track* changes in the weight vector, which cannot be generally quantified, because they depend on the signal behaviour, especially on the difference between the optimal weights for consecutive signal segments. Let segments $i-1$ and i have lengths M_{i-1} and M_i and transfer functions $H_{i-1}(z)$ and $H_i(z)$, respectively. Assuming that the LMS algorithm had converged to the optimal weight vector $\mathbf{w}^*_{i-1}$, the behaviour of the algorithm will be examined during and after the transition from $H_{i-1}(z)$ to $H_i(z)$, which will be initially assumed instantaneous. Weight adaptation in this case is similar to that already described for the first segment, but with different initial conditions. $\mathbf{w}(0)$ was assumed to be zero then, whereas now it is the vector $\mathbf{w}^*_{i-1}$ corresponding to the estimated inverse transfer function for the previous segment, $\hat{H}^{-1}_{i-1}(z) = 1 - \sum_{k=1}^{L} w_{(i-1)_k} z^{-k}$. The convergence of the algorithm will depend on the properties of the auto-correlation matrix, of the current segment R_{xx_i}, especially its eigenvalue spread, $\frac{\lambda_{max_i}}{\lambda_{min_i}}$. Perhaps a more critical factor is the difference between

R_{xx_i} and $R_{xx_{i-1}}$, of the previous segment, which is associated with a change in the optimal weight vector, $\Delta\mathbf{w}^* = \mathbf{w}_i^* - \mathbf{w}_{i-1}^*$. A small change in R_{xx} is associated with a small $\Delta\mathbf{w}^*$, which represents the deviation from optimal at the beginning of the ith segment and hence the number of iterations until convergence.

The investigation of the behaviour of the algorithm, without making further assumptions about $\Delta\mathbf{w}^*$, is not generally possible. Widrow et al.[201] investigated the the artificial but tractable situation where the weights are produced by stationary, ergodic random processes. For the suggested EEG model, weight generation is different, but Widrow's analysis may be used as a guideline, especially concerning small changes in the weights occurring *during* a segment.

In the case of a fixed weight vector (section 8.4.2), misadjustment is caused effectively by gradient noise and is approximately equal to $M_N = \mu \mathrm{tr}[R_{xx_i}]$, which applies here also. When the optimal weight vector changes with time, the algorithm is expected to follow these changes and maintain the current optimal weight vector. LMS weight estimation being a first-order difference equation (equation (8.43)), has the dynamic response of a low-pass filter. This introduces smoothing and lag (due to phase shifts) to quickly varying weights causing tracking errors, resulting in *additional misadjustment*[201]. This is related to the learning rate μ, which controls the dynamics of the LMS recursive equation. A small value for μ means slow convergence, thus more smoothing and hence worse tracking. This source of weight misadjustment for the specific case studied by Widrow was found to be approximately equal to

$$M_L = \frac{\sigma_{ww}^2}{2\xi_{min}} \sum_{i=0}^{L} \tau_i \lambda_i = \frac{1}{\mu} \left\lceil \frac{(L+1)\sigma_{ww}^2}{4\xi_{min}} \right\rceil \tag{8.48}$$

where σ_{ww}^2 is the variance of the weight vector (assumed to consist of weights that are independent identically distributed passed through random ergodic, first-order processes[201]). The total misadjustment in this case is given by:

$$M_{tot} = M_N + M_L = \mu \mathrm{tr}[R] + \frac{1}{\mu} \left\lceil \frac{(L+1)\sigma_{ww}^2}{4\xi_{min}} \right\rceil \tag{8.49}$$

The convergence rate appears in a linear and a reciprocal term, suggesting the existence of a value for μ for which the misadjustment is minimal. This is given by:

$$\mu = \left\lceil \frac{(L+1)\sigma_{ww}^2}{4\xi_{min}\mathrm{tr}[R_{xx}]} \right\rceil^{\frac{1}{2}} = \frac{\sigma_w}{2\sigma_{xx}\sigma_{ee}} \tag{8.50}$$

where $\sigma_{xx}^2 = \frac{\mathrm{tr}[R_{xx}]}{L+1}$ is the variance of the input signal and $\sigma_{ee}^2 = \xi_{min}$, the variance of the noise after convergence. This relation is only useful in theory, since σ_w^2 is not known and it depends on the weight-generation process. Using the practical limit given in equation (8.46),

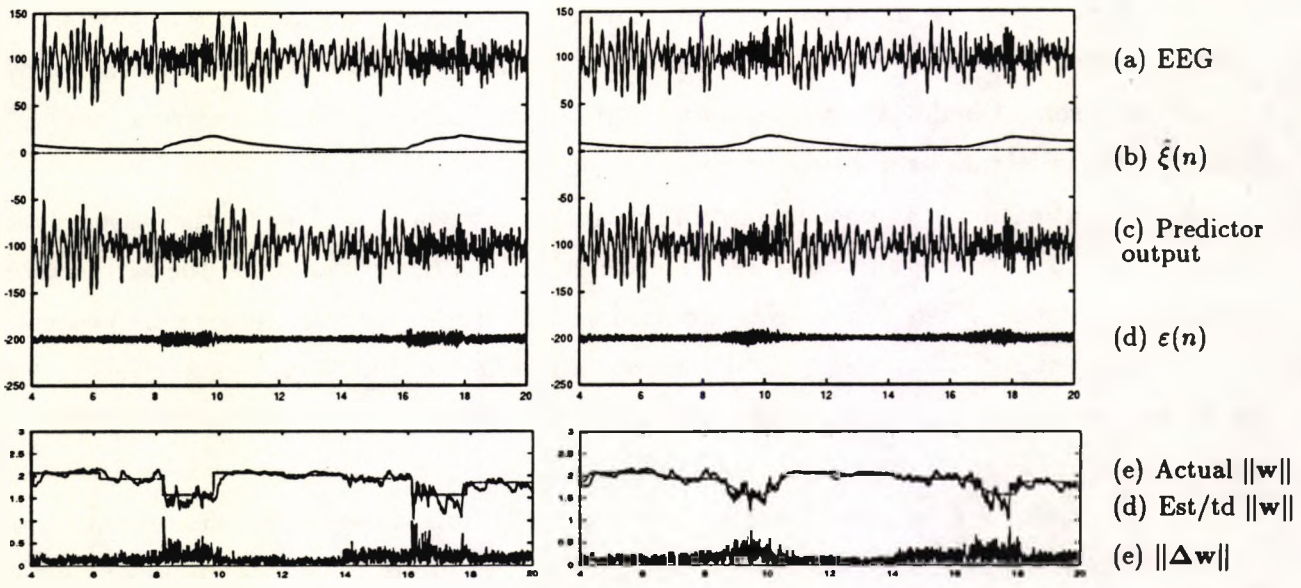


Figure 8.6: Tracking Capabilities of the LMS Prediction Error Filter for $b(n)$

it is evident that the optimal value for μ, as expressed by the above relation, can be reached if $\sigma^2_{ww} < \frac{4}{(L+1)^2}\frac{\xi_{min}}{\sigma^2_{xx}}$, which may be quite small indeed. In practice, however, this did not have a significant bearing in the proposed system, possibly because changes in the weight vectors are infrequent.

The time constant of the adaptation is critical in the case of the piecewise approximation, since $\tau_{max} = \frac{1}{4\lambda_{min}}$ dominates the speed of transition from one weight vector to the next. The behaviour of the algorithm was tested with artificial data (section 8.6.1) and the moduli of $\mathbf{w}^*$ and $\Delta\mathbf{w}^*$ are displayed along with an estimate of the MSE and the modulus of the actual weight vector (Figure 8.6(a)). The lag between the actual and estimated weight vectors is evident near the points of transition between segments.

In practice, transitions are not as abrupt as indicated by the above model, but come gradually, hence the transition between adjacent segments is smoother and the effects of lag less profound (Figure 8.6(b)). The gradual transitions were introduced by explicitly smoothing the weight vector using an arbitrary time constant or by making the windows of application of adjacent transfer functions tapered and overlapping, as suggested in section 7.3.1. The first method was adopted for the generation of the data presented here, as it provides a one to one correspondence between the coefficients of the generating and the estimating procedures during these transitions.

Determining the order of the predictor and the learning rate

The theoretical analysis and the results with simulated EEG signals supported use of the LMS for the estimation of the background activity inverse system but nothing has been said about the selection of its critical parameters, the order of the adaptive system, L, and the learning rate, μ. These are the *only* constants that need to be supplied to the algorithm externally and the implications of their incorrect choice, especially μ, could have devastating effects on the behaviour of the adaptive system, which should be prevented, especially if it is intended to operate unsupervised for long periods as in monitoring of ambulatory patients.

Ideally, L should be set equal to the order, L_i of the generating AR process, $H_i(z)$, whose inverse is sought. In practice, however, L_i may vary from segment to segment. Although it could be set to the maximum, $L = \max_i L_i$, this maximum is not easy to estimate in practice, since the assumption that the process is all-pole, or even a linear is somehow arbitrary (section 7.3). If L is smaller than the order of the system to be predicted, a poor approximation may result. The MSE will be higher than the theoretically achievable minimum and the decorrelation of $b(n)$ in $\hat{u}(n)$ will be incomplete. If it is higher, the weights for $L > L_i$ will be small, changing randomly and their corresponding spectral parameters will be insignificant. In general, over-estimation of L does not have serious effects on the performance of the algorithm, although convergence may be slowed down and the tracking capability of the filter may be reduced. A computational penalty is also possible, as the number of operations required is proportional to the filter order. Several criteria for the selection of the optimal L have been suggested, the more frequently used ones being the Final Prediction Error (FPE) and the Akaike Information Criterion (AIC)[99, 194], but these information theoretic functions require representative samples of EEG signals if they are to yield reasonable estimates.

Often a more 'practical' approach is more effective. Signals are processed by adaptive systems of progressively larger order and the MSE, ξ is recorded for the whole length of the signal for each value of L[161, 201]. This function, $\xi(L)$ is expected to have large values for small L, but rapidly decreasing and when L reaches the value of the order of the generating AR process, it stops decreasing and levels off. In the non-stationary case, for large L , ξ may increase for a fixed learning rate, as the misadjustment due to gradient noise becomes significant[201]. For real EEG signals the all-pole assumption is only approximate as the presence of even a single zero would make the order of the equivalent AR filter infinite. Hence $\xi(L)$ will be decreasing for all L. Such a graph was constructed for the existing EEG signals and for orders greater than 1 (Figure 8.7). For small L, the decrease in ξ is rather dramatic, but there is a marked reduction in the rate of decrease, once some critical value

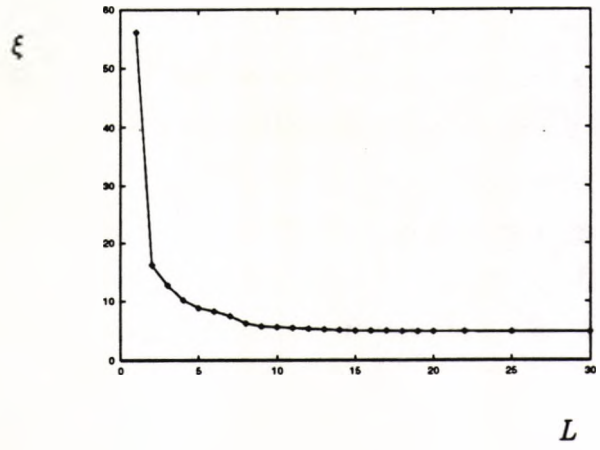


Figure 8.7: Selecting the optimal predictor order

L_0 is exceeded. After that point, the MSE decreases very slowly, hence the selection of a larger number of weights would only produce marginal improvements in the level of the error. Besides, a large order, for a given learning rate is likely to slow down the convergence (and tracking properties) of the algorithm, as it introduces additional time constants and possibly eigenvalues of small magnitude (see last two subsections).

This critical value, with the addition of a few more weights as a safety margin, was found to be around 14. This is not much different from 10, the value estimated by Roberts and Tarassenko[161, 162] for a system for EEG quantification during sleep. The discrepancy is not significant, taking into account the use of different signal sets and their nature. In their work these authors used sleep EEGs, whereas here records were taken during wakefulness and contained abnormal (epileptic) patterns.

The experimentally evaluated optimal order of the predictor was verified from observations on the available EEG records and information provided by EEG analysis texts (and covered in the earlier Chapter 2.5). Sections of the EEG signal consist of successions of quasi-periodic waves, which have in many cases sinusoidal appearance, although harmonics are often present. For 'sinusoidal' signals, like alpha, beta, delta and theta waves, a single pair of complex conjugate poles may suffice for their description. The inverse system will have these as zeros and a second-order transfer function, $P_{(i)}(z) = H^{-1}_{(i)}(z) = 1 - 2r_i \cos\omega_i T\, z^{-1} + r_i^2\, z^{-2}$ where ω_i is the frequency of the oscillations and r_i is related to the rate of decay (see also equation (7.20)). If harmonics, or other frequency components, are present, like in the case of mu rhythm and lambda waves, these would correspond to similar systems and the composite signal is the output of a generating system whose transfer function is the superposition of

these two-pole sections, $H_{(i)}(z)$:

$$\begin{aligned} P(z) &= \frac{1}{\sum_{i=1}^{K} \frac{1}{P_{(i)}(z)}} \\ &= \frac{\prod_{j=1}^{K} P_{(i)}(z)}{\sum_{i=1}^{K} \prod_{j=1, j\neq i}^{K} P_{(i)}(z)} \\ &= \frac{\sum_{i=1}^{2K} a_i z^{-i}}{\sum_{i=1}^{2K-2} b_i z^{-i}} \end{aligned}$$

This system has poles as well as zeros. The denominator polynomial may be written in factored form $\prod_{i=1}^{2K-2}(1 - p_i z^{-1})$ and each of the single pole terms $\frac{1}{1-p_i z^{-1}}$ expanded as an infinite geometric series, $1 + p_i z^{-1} + p_i^2 z^{-2} + \ldots + p_i^k z^{-k} + \ldots$, since the poles p_i are all inside the unit circle. There will be $2K-2$ such series and their product is a polynomial of infinite order $\sum_{i=0}^{2K-2} b_i z^{-i}$. Because each of the coefficients b_i consists of a sum-of-products of the p_i, their magnitudes get progressively smaller and soon become insignificant, especially if there are poles close to the origin. If all poles had the same magnitude, r, b_i would contain terms of magnitude r^i. For $r = 0.9$, only the coefficients for $i < 22$ have magnitudes greater than 0.1 and for $r = 0.8, 0.7, 0.5, 0.3$ the upper limit for i progressively decreases to $11, 6, 4, 2$. For the EEG, a maximum of four resonant frequencies at any time ($K = 4$) were assumed, with the poles of $P(z)$ having magnitudes in the above range. Using the intermediate value $i \approx 8$, the effective order of the system is $2K + i = 16$, slightly higher than the one estimated by the experimental technique. Tests were carried out with orders between 10 and 20 and the results were satisfactory for orders above 12 (section 8.6.2).

The other element worthy of consideration is the estimation of a suitable learning rate μ. Equation (8.46) gave an easily computable upper limit as $\frac{1}{(L+1)\sigma_{xx}^2}$, where σ_{xx}^2 is the power of the input signal (the composite EEG, $e(n)$). In practice, an estimate of the power is computed, usually based on a window of the input data:

$$\hat{\sigma}_{xx}^2(n) = \frac{1}{M}\sum_{i=0}^{M} x(n-i)^2 = \hat{\sigma}_{xx}(n-1) + \frac{1}{M}[x(n)^2 - x(n-M)^2] \tag{8.51}$$

This expression gives a statistically unbiased estimate, but requires considerable memory if M is large. An alternative method uses a simple recursive first order low-pass filter:

$$\hat{\sigma}_{xx}^2(n) = (1-\alpha)x(n)^2 + \alpha\sigma_{xx}^2(n-1) \tag{8.52}$$

The above expression is equivalent to the transfer function $A(z) = \frac{1-\alpha}{1-\alpha z^{-1}} = (1-\alpha)(1 + \alpha z^{-1} + \alpha^2 z^{-2} + \ldots)$ which is essentially a window of infinite length, but with diminishing emphasis for older samples. The parameter α is a positive constant less than unity. When it is close to unity, past samples are emphasized more, whereas when close to zero the past

of $x(n)$ quickly loses its effect on the estimate, in which case it is largely dependent on the instantaneous value of the power, $x(n)^2$. For this reason α is sometimes called the 'forgetting factor'. The gain $1-\alpha$ has been introduced to ensure that the estimate gives the correct value for the power of a constant signal (hence $\hat{\sigma}^2_{xx}(n) = \sigma_{xx} = 1$ for $x(n) = 1 \ \forall \ n$). The estimate is biased, since recent samples are emphasized more than old ones, but is adequate for the computation of the upper bound for μ, which need not be unbiased nor very accurate. The selection of α is not very critical, as long as it is large enough to smooth-out large fluctuations in $x(n)^2$, while being sensitive to permanent changes in power, due to the non-stationarity of the signal. Widrow[202, p. 148] suggested that a good compromise would be to set its value such that the 'half-life' of the exponentially decaying window equal to the length of the stationarity, n_s, of the input signal ($\alpha^{n_s} = \frac{1}{2}$). From this the following formula for computing α may be derived:

$$\alpha = 2^{-\frac{1}{n_s}} \tag{8.53}$$

Assuming that the sampling rate is 160 s^{-1} and the EEG is stationary for about 1 s, then $n_s = 160$ and $\alpha = 0.9956$. Halving and doubling the stationarity yields forgetting factors with values 0.9914 and 0.9978, respectively, differing only marginally from the above value, which was employed in this implementation (although in the actual program it was user-defined).

The value for the learning rate was chosen to be a fraction λ of the upper limit of the learning rate:

$$\mu(n) = \frac{\lambda}{(L+1)\hat{\sigma}^2_{xx}(n)} \qquad 0 < \lambda < 1 \tag{8.54}$$

It should be noted that μ is now updated on every sample. This time-dependence has been reported to work quite well for non-stationary signals and had been applied extensively for the linear predictive coding of speech signals[14, pp. 117–121].

Two additional elements were introduced related to the power estimation and the learning rate formulæ(8.52 and 8.54). The first is a safeguard in case the power of the signal reaches zero (making μ infinite). After $\hat{\sigma}^2_{xx}$ is computed, if found to be zero (or below a small set threshold $\check{\sigma}^2$) it is set to the minimal value $\check{\sigma}^2$. Alternatively, this value may be added as a constant term in equation (8.52), thus avoiding the comparison. Its biasing effect is small since the power is typically several orders of magnitude above $\check{\sigma}^2$. The other consideration is the initial value of the variance $\hat{\sigma}^2_{xx}(0)$. Normally this is set to a typical value, $\bar{\sigma}^2$, based on a priori signal information, but its value becomes insignificant after the lapse of a short time. By deliberately setting this at a *lower* value than estimated, a higher convergence rate is achieved initially, thus speeding up convergence. This increases exponentially with α reaching its 'correct' value after the filter takes its course. In order to guarantee the stability

of the filter, $\bar{\sigma^2}$ was set to generate the upper limit for μ, $\hat{\sigma}^2_{xx}(0) = \lambda\bar{\sigma^2}$.

8.4.4 Estimation of the spike model

The main consideration in the estimation of the background activity was the behaviour of the adaptive method on-line, to model the unpredictable variations in the background activity signal. In the case of epileptic spikes (and other transients) the signal $s(n)$ contains patterns with a similar morphology, which has been captured off-line in the form of the impulse response of a linear system, $G(z)$ (section 7.3.2). The elements of this system should be known before analysis may begin (sections 7.4.2 and 7.4.4). These considerations, however, are based on the assumption that all spikes have exactly the same shape, apart from an arbitrary scaling factor (section 7.3.2). Real spikes, although they have similar shapes in general, differ in their details. Hence, in order to approximate spikes more precisely, small variations were permitted on $G(z)$ at run-time, making the model time-varying. Different methods were employed for the initial off-line and the subsequent on-line estimation.

Initial off-line model estimation

A method for the estimation of $G^{-1}(z)$ has already been suggested in section 8.3, using the concept of spiking filters, based on finite-length input and output sequences. Spikes were marked in the input EEG signal by the interactive (semi-automatic) procedure described in section 8.3. Using spike markings as a cue for its position, a segment of the signal extending K_1 samples before and K_2 samples after that instant was extracted. Each one of these segments, $\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_p$, contained a noisy estimate of a spike in the form of $K = K_1 + K_2 + 1$ consecutive samples of the input EEG signal, $\hat{\mathbf{s}}_i = \{e(n_i), e(n_i+1), \ldots, e(n_i+K-1)\}$. These are noisy estimates of spikes, since they are contaminated by background activity. The value of K was chosen for the duration of the window to be approximately a third of a second.

To estimate the best spiking filter, the average spike, formed by the superposition of all p available noisy spikes, $\hat{s} = E[\hat{\mathbf{s}}_i]$. This can then be applied to the spiking filter procedure suggested by equation (8.27). It was anticipated, though, that this approach would not have been versatile, because spikes have small differences in their duration and their exact alignment is not possible, deeming the consistency of the average questionable.

A better alternative is to find the filter coefficients corresponding to individual spikes and to select the best set obtained for every spike. The average of these p coefficient sets are then a representative set for any spike.

Although this procedure is carried out off-line, therefore using equation (8.27) directly is not restrictive, the recursive solution of the Wiener-Hopf LMS algorithm (section 8.4.2),

which was already applied to the background activity (section 8.4.3) was adopted instead. This special application of the LMS algorithm is considered next. There are p spikes, $\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_p$, each consisting if a sequence of K samples, $\hat{\mathbf{s}}_i = \{\hat{s}_i(0), \hat{s}_i(1), \ldots, \hat{s}_i(K)\}$. The procedure that follows was repeated for every spike sequence, $\hat{\mathbf{s}}_i$.

The desired response sequence, $\{(d_k(0), d_k(1), \ldots, d_k(K-1)\}$ for the kth spiking filter is such that $d_k(n) = 1$ if $n = k$ and zero otherwise. The LMS algorithm (equation (8.43)) was applied to the spike sequence to estimate the spiking filter of order $L < K$, $\mathbf{g}_k = (g_{0,k}, g_{1,k}, \ldots, g_{L,k})^T$ as follows:

- Initially $\mathbf{g}_k(0) = \mathbf{0}$.
- For all $n = 1, 2, \ldots, K$ compute the following:

$$\varepsilon(n) = d_k(n) - \mathbf{g}_k(n)^T \mathbf{s}_i(n-1) \tag{8.55}$$

$$\mathbf{g}_k(n) = \mathbf{g}_k(n-1) + 2\mu\varepsilon(n)\mathbf{s}_i(n-1) \tag{8.56}$$

It was necessary to repeat this process several times, reusing a spike again and again until $\mathbf{g}_k(n)$ reached a steady state. The procedure was terminated once the magnitude of the change in the weight vector $|2\mu\varepsilon(n)\mathbf{s}_i(n-1)|$ was lower than a preset threshold. After convergence, the filter $\mathbf{g}_k^*$ was applied to the input and the output sequence, $\hat{d}_k(n)$, and the mean squared error, ξ_k, were evaluated. The best three spiking filters were then recorded, (in ascending values of ξ_k). It was observed that in most cases these corresponded to adjacent values of k, which is intuitively appealing, since the degradation of the performance is gradual, near the optimal delay value, k. In fact only the best filter was necessary, but recording the best three ensured that nearly-optimal cases were not ignored.

Once the best spiking filters were estimated for all p available spikes, their coefficients were visually inspected. In general these were clustered in a small region of the weight space, and the mean of the coefficient vector

$$\mathbf{g}^* = \frac{1}{4p} \sum_{i=1}^{p} \sum_{k=1}^{4} \mathbf{g}_{k,i} \tag{8.57}$$

was used in the detection procedure (section 7.4.4). The first six coefficients for a spiking filter of order $L = 10$ are shown in pairs in Figure 8.8. All filters were normalized so that $g_k(0) = 1$, to correspond to the AR filter $G_k(z) = \frac{1}{1+\sum_{i=1}^{L} g_{k,i} z^{-i}}$, which is independent of the amplitude of the spike.

Evidently, a higher order system is preferred in terms of accuracy, but with on-line coefficient adaptation, suggested in the next section, a system of order $L = 2$ proved adequate for describing the spike feature.

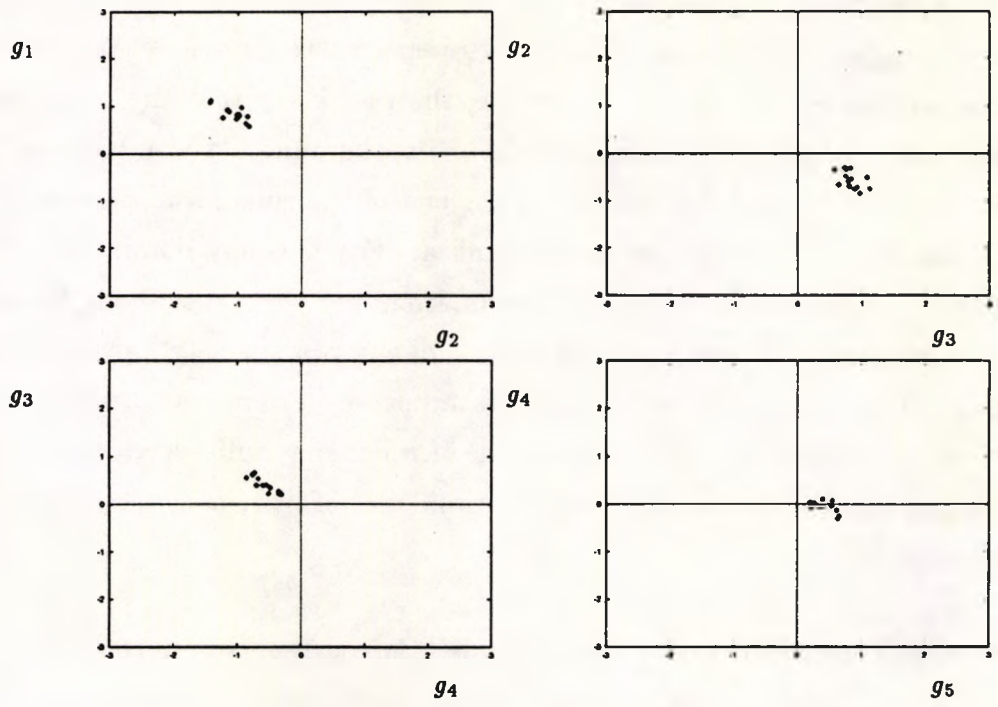


Figure 8.8: Distribution of optimal spiking filter coefficients

On-line adaptation

The spike generating system, $G(z)$, as well as its inverse (spiking filter) $G^{-1}(z)$, estimated by the off-line method, is perhaps the best linear time-invariant (LTI) approximation and is capable of modelling amplitude variations (by controlling the strength of the input excitation impulse as described in section 7.3.2). Other variations often observed in real spikes, like duration and fine detail, cannot be expressed by a time-invariant model, irrespective of its order. As seen from the coefficients between the optimal systems for individual spikes (Figure 8.8), small variations are present, which translate to deviations of individual spikes from the output of the typical spike model.

Better spike modelling could be achieved if the system coefficients were allowed to deviate from their typical values and take those appropriate for the spike in consideration. This process must be performed on-line, as the values of the system for individual spikes cannot be predicted before they are encountered.

It is possible, however, to adapt the coefficients until they converge to the optimal during the progress of a given spike. According to this approach the coefficients of the system will be sub-optimal at the onset of a spike, but will progressively get better.

The adaptation of the coefficients was performed according to an LMS-type algorithm, not on the inverse (spiking filter) but on the spike generator filter ($G(z)$ in Figure 7.5). There

is a good reason for that. As explained in the previous section it was necessary to restrict the input to the spiking adaptive filter so that only segments where the spike signal is strong could influence the adaptation procedure. Otherwise, the possibility that the coefficients would deviate from those for the spike in favour of some background activity component cannot be ignored. This will, in fact, be the case at the end of the spike, where the amplitudes of the spike gradually diminish and the background activity becomes dominant. Of course it is possible to set a window, as with off-line estimation, but this causes complications, since spikes vary in duration and justification of the use of any window length over another would be debatable. Also, bearing in mind that the proposed system was structured with the possibility of expansion to other transients, some of which may suffer from greater variations in their duration than spikes do, suggested the avoidance of windows, which is possible with the method employed.

LMS-type algorithms for recursive systems Adaptation is performed on the autoregressive inverse, $G(z)$, of the spiking filter, both of which share the same coefficient values (which appear in the denominator polynomial for $G(z)$). Least Squares estimation for IIR filters is possible, by introducing additional terms in the auto-correlation matrix corresponding to the feedback coefficients. The output of such a system is given by the recursive equation

$$\begin{aligned}\hat{y}(k) &= \sum_{i=0}^{L} b_i x(k-i) - \sum_{i=1}^{L} a_i \hat{y}(k-i) \\ &= \mathbf{b}^T\mathbf{x}(k) - \mathbf{a}^T\hat{\mathbf{y}}(k-1) \\ &= [\mathbf{b}^T, -\mathbf{a}^T]\begin{bmatrix}\mathbf{x}(k)\\ \hat{\mathbf{y}}(k-1)\end{bmatrix} \\ &= \mathbf{w}^T\mathbf{u}(k) \end{aligned} \tag{8.58}$$

where the input vector $\mathbf{u}(k)$ to the Wiener filter now contains the previous outputs of the filter and the weight vector $\mathbf{w}$ contains all the a_i as well as the b_i. It is evident that if the development of the Wiener-Hopf equation presented in section 8.1 was repeated in this case, the input auto-correlation matrix would have the form

$$R_{uu} = E[\mathbf{u}(n)\mathbf{u}(k)^T] = \begin{bmatrix} R_{xx} & R_{x\hat{y}} \\ R_{x\hat{y}} & R_{\hat{y}\hat{y}} \end{bmatrix} \tag{8.59}$$

Which leads to a set of non-linear differential equations, because the terms that incorporate $\hat{\mathbf{y}}(k)$ are functions of the weight vector $\mathbf{w}$. These equations cannot generally be solved in closed form[152, p. 262] like the Wiener-Hopf equation (8.9). Numerical optimization in the form of the LMS algorithm is still applicable, since the arguments and procedure of

section 8.4.2 still hold. The resulting recursive relation is complicated by the appearance of cross-coupling terms in the derivative $\nabla \varepsilon = -\nabla \hat{y}(k)$. Considering the non-recursive and the recursive coefficients separately,

$$\begin{aligned}\beta_j(k) &= \frac{\partial \hat{y}(k)}{\partial b_j} = x(k-j) - \sum_{i=1}^{L} a_i \frac{\partial \hat{y}(k-i)}{\partial b_j} \\ &= x(k-j) - \sum_{i=1}^{L} a_i \beta_j(k-i) \qquad (8.60)\end{aligned}$$

$$\begin{aligned}\alpha_j(k) &= \frac{\partial \hat{y}(k)}{\partial a_j} = -\hat{y}(k-j) - \sum_{i=1}^{L} a_i \frac{\partial \hat{y}(k-i)}{\partial a_j} \\ &= -\hat{y}(k-j) - \sum_{i=1}^{L} a_i \alpha_j(k-i) \qquad (8.61)\end{aligned}$$

These can then be put in a single vector $\dot{\mathbf{u}}(k) = [\beta_0(k)\,\beta_1(k)\,\ldots\,\beta_L(k)\,\alpha_1(k)\,\alpha_2(k)\,\ldots\,\alpha_L(k)]^T$ and the LMS algorithm written as before:

$$\mathbf{w}(k) = \mathbf{w}(k-1) + 2\mu\varepsilon(k)\dot{\mathbf{u}}(k) \qquad (8.62)$$

This differs from the equation proposed by Widrow and Stearns[202, pp. 154–161] only on the learning rate. These authors have used a diagonal matrix of learning rates one for each weight. A simpler algorithm can be derived, if the derivatives are approximated by $\alpha_j(k) = -\hat{y}(k-j)$ and $\beta_j(k) = x(k-j)$. Hence $\dot{\mathbf{u}}(k) \approx \mathbf{u}(k)$, an algorithm presented by Bellanger[15, pp. 319–322]. Additionally, smoothing the error sequence prior to its application to the weight updating equation results in the Hyperstable Adaptive Recursive Filter (HARF)[202, pp. 154–161]

The first two algorithm variants (not the HARF) were tried but their results were effectively the same. This similarity is not unexpected, since the initial weight vector is close to the optimal solution anyway and the gradient has small values, thus the effect of the terms omitted in the simpler algorithm (the summations in equations (8.60) and (8.61)) is small. Hence the second is preferred, since it is simpler, requiring considerably fewer operations and less memory to implement. Because $G(z)$ is purely autoregressive, the non-recursive numerator coefficients do not apply ($b_i = 0 \forall i > 0, b_0 = 1$) and the implemented filter structure is somewhat simpler than the general ARMA filter just presented. The autoregressive parameters of $G(z)$, called g_i in the previous section, were intentionally named a_i here to confirm with the usual notation employed in texts giving details of the algorithm.

The value for μ was set in a normalized manner similar to the background activity (equation (8.54)). The input power factor $\hat{\sigma}^2_{xx}$, however, was substituted by that of the output of the filter $\hat{\sigma}^2_{\hat{y}\hat{y}}$. Although this worked well in practice, there is no theory to support

the convergence of the algorithm for ARMA systems, because of the non-linearity of the equations involved. There is also some doubt whether gradient search algorithms would find the point of minimum MSE, because the error surfaces for these systems are non-quadratic and might be multimodal, having local minima to which the weight vector may be trapped[202, pp. 154–161]. Widrow and Stearns give an example of such a performance surface[202, p. 136], which is clearly multimodal.

Range limiting for the autoregressive coefficients With recursive adaptive systems there are two stability issues. Stability of the algorithm and stability of the systems itself (all its poles must be inside the unit circle in z). The question of stability of the AR system during adaptation is an important issue and is linked to the the maintenance of the autoregressive coefficients $\{g_i; i = 1, 2, \ldots, L\}$ within the range of those for spikes. It is possible during the course of the EEG signal, for a false transient detection to trigger the spike generating filter $G(z)$ (see section 7.4.4), which is adapted to track the input signal. Hence, even if the detection is finally rejected by the spike detector, with no restrictions put on the range of the coefficients of $G(z)$, they may drift away from those for spikes in an attempt to model the behaviour of the 'false spike'. This is hazardous for the detection procedure and the detection of future spikes, since not only the signal under consideration may cause a false spike detection, but subsequent spikes may be missed, because $G(z)$ would not correspond to an epileptic spike any more. To avoid this problem, the coefficients of the system must be confined to a small region around their values for the typical spike, estimated during off-line adaptation (see previous section).

To develop a procedure for the restriction of the range for the coefficients it is convenient to write them in vector form. Thus, $\mathbf{g}_i = (g_{i_1}\, g_{i_2} \ldots g_{i_L})^T$, is the best spiking filter corresponding to the ith spike. The collection of $\mathbf{g}_i$ for all spikes $i = 1, 2, \ldots, p$ has been used to compute the average, presumed to correspond to the most typical spike, $\mathbf{g}^*$ (equation (8.57)), as they appear in a rather tight cluster (Figure 8.8). It is anticipated that all spikes correspond to coefficient vectors within the extent of the cluster and that during adaptation these should not leave this region if a detected spike is genuine. The coefficient space, corresponding to spikes is a connected region in $\mathbb{R}^L$. As such it is defined by a boundary ψ. The adaptive process starts with the centroid of the region, $\mathbf{g}(0) = \mathbf{g}^*$, which is inside ψ and subsequent values of $\mathbf{g}(k)$ should be maintained inside the boundary. Since the algorithm, operates on the unconstrained range, $\mathbb{R}^L$, restrictions must be performed on the coefficient vector after each adaptive procedure.

Restricting a priori the adaptation rate, μ depending on the location of the current $\mathbf{g}(k)$

is not sufficient, since it will apply on the *next* recursion, while $\mathbf{g}(k)$ may already be outside ψ. Again, substituting $\mathbf{g}(k)$ if outside ψ by the one found during the previous iteration, $\mathbf{g}(k-1)$, appeared to affect the progress of the adaptive algorithm, as well as its output. No theoretical explanation was found for this phenomenon, which was attributed to the arbitrary and discontinuous deviation from the expected course of the filter coefficients. This was a general observation for restriction functions that had abrupt discontinuities.

A method for the restriction of the range based only on the location of a weight vector was developed, according to which unconstrained adaptation was performed initially so that the weight vector, $\mathbf{g}(k)$, may be outside ψ. An onto mapping $\mathbf{g}'(k) = \phi(\mathbf{g}(k))$ from $\mathbb{R}^L$ to the interior of ψ was then applied, ensuring that all coefficient vectors were mapped back into the spike region. Finding such a function for an arbitrary boundary ψ is not easy. Approximating it with a (hyper)sphere, centered at $\mathbf{g}^*$ is one of the simplest shapes. The radius of the (hyper)sphere, ρ was taken either as

1. the distance from the centre to the furthermost spike vector in the cluster
2. a multiple of the average distance, which is related to the covariance matrix of the set of vectors.

Both criteria assume equal spread of vectors in all directions around the centroid and worked well in practice, producing similar results, but the second is more easily adapted if the spread is not equal in all directions. In this case the covariance matrix may be computed and define a hyperellipse boundary. For the cluster considered the spherical assumption was not unreasonable. Whichever the criterion, the boundary was defined by $\psi = \{\mathbf{g}; (\mathbf{g}-\mathbf{g}^*)^T(\mathbf{g}-\mathbf{g}^*) = \rho^2\}$. A simple range mapping is one that maintains all points in the interior of ψ to their current location, while compressing the whole of the exterior of ψ on its boundary. This maintains the *direction* of the difference vector $\Delta\mathbf{g} = \mathbf{g}-\mathbf{g}^*$, while modifying its magnitude, $R = \|\Delta\mathbf{g}\|$. The direction of the vector is given by the normalized vector

$$\bar{\mathbf{n}} = \frac{\Delta\mathbf{g}}{\|\Delta\mathbf{g}\|} = \frac{\Delta\mathbf{g}}{R} \tag{8.63}$$

Hence $\Delta\mathbf{g} = R\bar{\mathbf{n}}$ and since ϕ affects only R, it may be expressed in terms of a mapping of a single variable, $r = f(R)$, which is easier to handle. The vector mapping, ϕ is then given by:

$$\mathbf{g}' = \Delta\mathbf{g}' + \mathbf{g}^* = f(R)\bar{\mathbf{n}} + \mathbf{g}^* \tag{8.64}$$

where $f(R)$ is an onto function from $[0,\infty)$ to $[0,\rho)$, which should map the centre of the sphere to itself ($f(0) = 0$). Moreover, it should map vectors close to the centroid to or near

themselves, so that the distortion of typical spike vectors is kept small. This is achieved if the derivative $f'(0) = 1$. A simple mapping having these properties is the following:

$$f_1(R) = \begin{cases} R & R < \rho \\ \rho & R \geq \rho \end{cases} \tag{8.65}$$

but the lack of smoothness for $R = \rho$ seemed to affect the behaviour of the adaptation once a vector outside ψ was encountered. Several smoother functions, conforming to the above criteria, were used successfully:

$$f_2(R) = \frac{2\rho}{\pi}\text{Arctan}(\frac{\pi}{2\rho}R) \tag{8.66}$$

$$f_3(R) = \rho(1 - e^{-\frac{R}{\rho}}) \tag{8.67}$$

$$f_4(R) = 2\rho\frac{(\frac{R}{2\rho})}{1 + (\frac{R}{2\rho})^2} \tag{8.68}$$

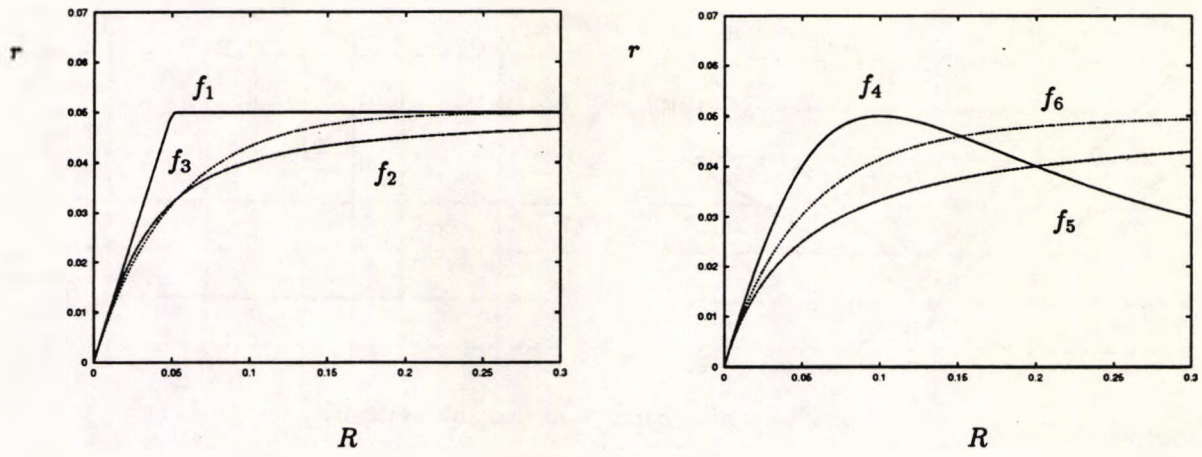
$$f_5(R) = \rho - \frac{\rho^2}{R + \rho} \tag{8.69}$$

$$f_6(R) = \rho - \frac{\rho^{n+1}}{(\frac{R}{n} + \rho)^n} \tag{8.70}$$

the mappings $f_1(R)$ to $f_6(R)$ are illustrated in Figure 8.9. Although they all perform the same task, each function has different properties. f_2 and f_3, for instance, behave almost linearly near the origin but compress the space for increasing R towards ρ. The simpler polynomial fractions, f_4, f_5 and f_6, are cheaper to compute than f_2 and f_3. f_4 folds vectors at large distances back on the centroid, hence for large deviations, when the vector is unlikely to correspond to a spike, it is returned to those for typical spikes. f_5 behaves like f_2 and f_3, whereas its generalization, f_6, has low second derivative at the origin and hence its linear approximation extends to a larger domain than f_5.

System stability issues The question whether range limiting affects the stability of the filter was raised. Obviously, if the whole of the range defined by ψ corresponds to stable systems, this issue need not be considered further. Spikes and other transients are characterized by a deterioration in their amplitude in the time following their onset. They are therefore described by the impulse response of stable systems. If these are second order the poles can be readily computed. In this case, it was found that the poles of the system lie well inside the unit circle in z, for the whole of the coefficient range ψ. Therefore, range limiting would guarantee stability as well, but the spherical parametrization of ψ, being approximate, might include unstable coefficient vectors, as will be demonstrated by a simple example.

Let a transient be approximated by a second-order system with two complex conjugate poles, p and p^*, located away from the real axis. Stable poles have modulus less than unity



In all cases $\rho = 0.05$

Figure 8.9: An assortment of range-limiting mappings

and for complex conjugate poles the square of the modulus is equal to their product, pp^*, that is the coefficient g_2 in $G(z)$. For complex poles, stability is therefore achieved if $g_2 < 1$ If the actual cluster of individual coefficient vectors is elliptic with a major axis r_1 in the direction of g_1 and minor axis r_2 in the direction of g_2, the radius of the approximating spherical (circular) cluster would have a value $r_2 < \rho < r_1$. If the centre of the cluster is such that $g_2 > 1 - \rho$, although all vectors in the actual cluster are stable, the spherical approximation will contain unstable vectors.

Although the number of unstable vectors is expected to be small, testing the stability of the coefficient vector and eliminating those that are unstable is very important and can be achieved by a range-limiting mapping, similar to the one already described. The simplest method to guarantee that all coefficient vectors correspond to stable systems is a modification of the parametric form of ψ, so that the entire sphere around the typical central vector $\mathbf{g}^*$ is contained in the stable region. Reducing the radius ρ of ψ so that after range-limiting all the resulting vectors are stable might cause vectors belonging to the cluster to be missed. A better alternative is to check stability on-line and to map unstable coefficient vectors to stable ones independently from range limiting. The range of stability for a second-order system is the well-known triangle[151, p. 207][15, p. 171]. Higher order systems cannot be factored in general and hence a similar range cannot be determined in coefficient space. Numerical methods may be employed for the estimation of the poles of the system[85, pp. 497–551], but these are generally iterative procedures (e.g. Newton's method) and are suitable for off-line operation.

Fortunately, a simpler (and faster) test for stability of a system without computing its

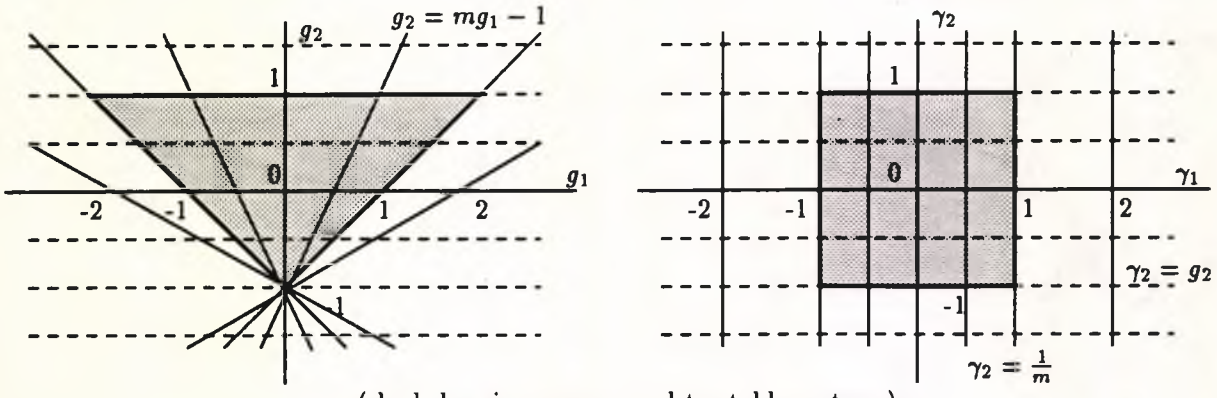


Figure 8.10: Correspondence between $\mathbf{g}$ and γ for $L = 2$

poles is available. This is the *Schür-Cohn stability test*[151, pp. 203–206]. According to the test, the denominator polynomial $G^{-1}(z) = F_L(z) = 1 + \sum_{i=1}^{L} g_i\, z^{-1}$ is decomposed into a set of values $\{\gamma_p; p = 1, 2, \ldots, L\}$, called *reflection coefficients* with the aid of the reverse polynomial, $B_L(z) = z^{-L} + \sum_{i=1}^{L} g_i\, z^{i-L}$. The procedure recursively computes one reflection coefficient at every recursion as well as a polynomial of reduced order. The basic steps of the algorithm are:

- Start with $F_L(z) = 1 + \sum_{i=1}^{L} g_{i,L}\, z^{-i}$
- For orders $p = L, L-1, \ldots, 1$ do the following:
  - compute the reverse polynomial, $B_p(z)$
  - set $\gamma_p = g_{p,p}$
  - Compute $F_{p-1}(z) = \frac{1}{1-\gamma_p^2}[F_p(z) - \gamma_p\, B_p(z)]$

The process fails if any of the $|\gamma_p|$ becomes equal to unity (in which case the system is unstable anyway). This unlikely problem has been overcome by changing the appropriate coefficient by a small amount $\delta\gamma$ and proceeding with the algorithm. The Schür-Cohn test consists of inspecting the magnitude of the reflection coefficients, since unstable systems have some of them greater than or equal to unity ($|\gamma_p| \geq 1$). There is a correspondence between the stability regions in $\{g_i\}$ and $\{\gamma_i\}$, at least for $L = 2$, where they are both known. This is demonstrated in Figure 8.10.

Because the procedure may be reversed equally easily, it is possible to compute the coefficient vector $\mathbf{g}$ from the reflection coefficients $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_L)^T$. This is significant, as it provides a simple method to *stabilize* an unstable system, by computing the reflection

coefficients, applying a range-limiting mapping to bring those outside the range $(-1, 1)$ back to the stable range and then computing the modified (stable) coefficient vector by the inverse mapping.

Range-limiting here could be one of the functions f_1 to f_6, (see previous section) with a one-dimensional version of equation 8.64 and a radius $\rho = 1$.

It should be noted that it is possible to implement both the adaptive procedure as well as the application of the filter directly using the reflection coefficients, without the use of $\mathbf{g}$ at all. This, however, will be discussed separately, as it is a case of an *adaptive lattice filter* (see section 9.3).

Finally, an advantage of the on-line adaptive system to implement $G(z)$ will be emphasized. By choosing a low-order filer (e.g. $L = 2$) it was possible to approximate epileptic spikes quite accurately, despite the simplicity of the model. This may only be attributed to one factor. Spikes may be approximated by a linear, but *time varying* generating process. Indeed, a spike resembles the output of a second-order system, whose poles are initially at a specific location, but gradually drift towards a lower frequency (angle in the z-domain). These variations could only be followed by a time-varying model, which was implemented as an adaptive system. In conclusion, the introduction of the constrained adaptation of the coefficients on-line should not be regarded as an additional feature, but as an essential element in the successful implementation of spike modelling.

8.5 Other elements of the analysis method

Finally, the decision making elements of the system will be considered. These are shown in Figure 7.5 and their function described in section 7.4.4.

8.5.1 The Transient detectors

The role of the two transient detectors is to detect short-lived changes in the (random, uncorrelated) background activity generating sequence, $u(n)$ in the presence of the filtered noise sequence, $v_P(n)$ (section 7.4.1). These are caused by spikes and other transients, filtered by $H^{-1}(z)$, the inverse background activity system and give rise to the components $s_P(n)$ and $t_P(n)$, which have certain properties that make their detection possible:

- they have amplitudes that are generally larger than the signal component $u(n) + v_p(n)$
- they are the output of all-pole stable systems, $G(z)$ or $F_i(z)$, when excited by impulses
- $H^{-1}(z)$ is the inverse of a stable all-pole system

The consequence of the second is that the amplitude of these transient signal gradually comes to a steady-state of zero, often fairly quickly after the time of their onset. The third suggests that $H^{-1}(z)$ is minimum phase, as its zeros lie inside the unit circle in z. The application of this on a transient is equivalent to the impulse excitation of a system $H^{-1}(z)G(z)$, for example. This is a stable, minimum phase system. The minimum phase property means that most of the energy of its output will be concentrated in the time interval immediately after the excitation of the system[151, pp. 434–437].

Of course both these properties are secondary compared to the amplitude difference between transients and $u(n) + v_P(n)$ (sections 7.4.3 and 7.4.4). The transient detector is a system that takes the output of this inverse filter $r(n) = u(n) + v_P(n) + s_P(n) + t_P(n)$, and generates a binary indication, $z_t(n)$ which is *true* for the duration of the transient and *false* otherwise. The term 'duration' here refers to the few consecutive samples near the time of onset of the transient. Assuming that the on-going activity is mainly due to $u(n)$, which has been assumed to be a white sequence with zero mean and variance σ_{uu}^2, and that the transients occur infrequently, they can be detected as samples with atypically large amplitudes, not belonging to the mentioned distribution. The parameters necessary for this detector are σ_{uu}^2 and a threshold value for the comparison, t_0. σ_{uu}^2 is not known, but may be computed easily using either a recursive method (equation (8.52)) or the more direct approach of equation (8.51). For this detector, the second was preferred, because it was unbiased and the maintenance of a window of the $r(n)$ sequence was necessary for other parts of the detector, anyway. A small number of samples (about 3, but assume $2Q + 1$ in general) in the middle of the window were excluded from the computation of the variance. This is where transients are detected. The modified equation for computing the variance is

$$\hat{\sigma}_{uu}^2(n) = \frac{1}{2(M-Q)-1}\left(\sum_{i=-M}^{-Q} r(n-i)^2 + \sum_{i=Q}^{M} r(n-i)^2\right) \tag{8.71}$$

$$= \hat{\sigma}_{uu}^2(n-1) + \frac{1}{2(M-Q)-1}[r(n+M)^2 - r(n+Q)^2 + r(n-Q)^2 - r(n-M)^2] \tag{8.72}$$

Notice that the 'current sample', $r(n)$ is situated in the middle of the window and the above operation is non-causal. This convention has been adopted because spike detection requires the generation and subtraction of spikes from the signal, which can only occur after the transient has been detected, hence on the nth sample. The use of 'future samples' in the computation of the variance has another significance. Permanent changes in the power of $r(n)$ would be associated with an increase in the amplitudes after a certain sample. These can only be distinguished from transients by looking ahead in time. If the change is

permanent this will be reflected in most future samples, whereas for transients this is local, characterizing only a few consecutive samples (because of the minimum phase property).

On-line estimation to incorporate power fluctuations in $u(n)$ is expected to increase the reliability of the detector. Moreover, $v_P(n)$ may now be taken into account, because the estimation is based on $r(n)$, assuming that the infrequent transients have little effect on the variance (power) of the signal. Provided that the window taken is large enough, these will have a small effect anyway which was verified by monitoring $\sigma_{uu}(n)$.

Once the variance became available, the middle of the window, $r(n)$ was tested for atypically large values. Because the distribution of $u(n)$ is uniform in the range $(-a, a)$, and its variance $\sigma_{uu}^2 = \frac{a^2}{3}$[60, pp. 194–195]. This was estimated by $\hat{\sigma}_{uu}^2(n)$ and the bounding value, a, was evaluated on-line from the variance as being equal to $\sqrt{3\hat{\sigma}_{uu}^2(n)}$. Then the normalized variable $t = \frac{r(n)}{\hat{\sigma}_{uu}(n)}$ would be less than a threshold $t_0 = \sqrt{3}$ for samples belonging to the distribution of $u(n)$. On-line, to avoid the square root operation on σ_{uu}^2, t^2 was computed and tested:

$$\begin{aligned} t^2 &= \frac{r(n)^2}{\sigma_{uu}^2} \\ z_t(n) &= \begin{cases} \text{true} & t^2 > t_0^2 \\ \text{false} & t^2 \leq t_0^2 \end{cases} \end{aligned} \tag{8.73}$$

where $z_t(n)$ is the (binary) output of the transient detector. Alternatively (see Chapters 10 and 11), $r(n)^2$ may be compared to $t_0^2\sigma_{uu}^2(n)$. Also, to allow for small timing discrepancies between the spike and the transient detection, instead of carrying the test on a single sample $r(n)^2$, this was performed on the sum of the central $2Q+1$ values excluded from the variance calculation. Therefore in the presence of a transient, its detection persists for a few samples in time to allow the spike initiator (section 8.5.2) to form the spike generating impulse at the correct time.

During the development of the system, a safety margin of factor 2 was introduced on the threshold ($t_0^2 \approx 6$) and z_t was modified to a three-level signal to show whether the transient was positive or negative (subdividing *true* into *true+* and *true-*. This distinction has a bearing on the correct detection of some transients (like spikes), which always have the same polarity.

In the constructed system there are two transient detectors. One acting on $r(n)$ and the other on the output of the (duplicate) system $H^{-1}(z)$, $r'(n)$, whose input is a (presumably) spike-free signal $e'(n)$ (see Figure 7.5). The function of both is identical apart from their inputs and variance estimates.

8.5.2 The spike initiator

This has two inputs, the binary output of the first transient detector, $z_t(n)$, and the output of the inverse spike filter, $y(n) = b_Q(n) + d(n) + t_Q(n) + v_q(n)$ (section 7.4.2 and Figure 7.5). Its purpose is to extract the spike-generating (impulse) sequence, $d(n)$, from $y(n)$ for the excitation of the spike filter. As explained in section 7.4.4, this cannot be done reliably on the basis of $y(n)$ only. Thus, when $z_t = true$ (actually $true+$), the output $\hat{d}(n)$ of the spike initiator should be an estimate of $d(n)$ (zero otherwise).

Because $y(n)$ cannot be decomposed, the best estimate is the instantaneous value of its input, $\hat{d}(n) = y(n)$, if $z_t(n) = true+$. Although satisfactory in most cases, occasionally the generated spikes had a lower amplitude than they should. This is believed to be due to the fixed-gain AR process that approximates the spike, which cannot increase its energy level to compensate an underestimated input sequence. Hence the peaks of these spike were sometimes poorly approximated, especially when $b_Q(n) + v_Q(n)$ were large. On the other hand over-estimation of $d(n)$ did not have a noticeable effect on the generated spikes, perhaps because the adaptive spike generator can adapt this by increasing its damping factor.

Over-estimation required the addition of a positive quantity Δy to make $\hat{d}(n) = y(n)+\Delta y$. A proportion of the RMS value, $\sigma_{yy}(n)$ of the signal components that are responsible for the error $b_Q(n) + v_Q(n)$ was used as the correction factor, as explained in section 7.4.3:

$$\hat{d}(n) = y(n) + t_0\hat{\sigma}_{yy}(n) \tag{8.74}$$

The estimation of the variance $\hat{\sigma}^2_{yy}(n)$ followed the same formula as the transient detectors (see equation (8.72) and explanation in the previous section). Here, the square root on the variance could not be avoided, since an amplitude measure is required, but this is only necessary on few occasions.

Finally it would be an omission not to stress that $\hat{d}(n)$ may not resemble an impulse, since it may have a duration of more than one samples (as long as $z_t(n) = true+$). This, however did not affect the overall performance of the system. It is appears that some form of compensation is provided by the adaptive nature of $G(z)$.

8.5.3 The spike detector

By comparison to the transient detector and the spike initiator, the spike detector has a very simple structure. It has two inputs, one from each of the transient detectors $z_t(n)$ and $z'_t(n)$. The first, when positive indicates the presence of a transient in the input EEG signal, $e(n)$. The second indicates the presence of a transient in the signal with a spike subtracted, $e'(n)$. Genuine spikes cause a positive indication in the first detector, but no indication in

the second (section 7.4.4). Other conditions are characteristic of the absence of a a transient (both detectors are *false*), or the absence of a non-spike transient (both are *true*). It may also be possible in practice for the second detector to be *true* when the first is *false*. The task of the spike detector is to isolate the first of these cases in the central sample (as both its inputs are windowed) and produce an output $z_s(n)$ that shows whether a spike was present or absent:

$$z_s(n) = \begin{cases} \text{spike} & z_t = true+ \text{ and } z_t' = false \\ \text{non} - \text{spike} & \text{otherwise} \end{cases} \tag{8.75}$$

This basic operation may be amended by the addition of secondary known properties of epileptic spikes, such as their frequency of occurrence (e.g. at most 3 per second, during a seizure of petit mal, if this type of epilepsy is only considered), which may be useful for the identification of false spike detections at the beginning of a muscle artifact, which sometimes occurs.

8.6 Results

The system described above was tested with simulated as well as real EEG data. Simulated records were initially used for testing the functionality of individual elements of the system and to observe the behaviour of the complete spike detector when the input is 'ideal'. Real data were used to verify that it functioned well in practice.

8.6.1 Synthetic data

Despite the detailed analysis of many aspects of the proposed analysis system, it is not possible to predict its precise behaviour. There was some concern, especially regarding the convergence of the adaptive algorithm and the selection of a suitable order for the background activity model, $H(z)$. Then the behaviour of the constrained adaptive IIR spike modelling system had to be assessed with spikes conforming to the model, prior to its application on the imperfectly modelled real epileptic transients. It was actually at this stage when it was observed that the simple range-limiting method, initially suggested in section 8.4.4, had undesirable side-effects and was subsequently substituted with a smooth non-linear mappings.

A synthetic EEG record conforming to the model suggested in section 7.3 was generated. The EEG sequence, $e(n)$, consisted of the superposition of four elementary signals, as suggested by the synthesis model: the background activity signal, $b(n)$, the spike signal, $s(n)$, transients, $t(n)$, and noise, $v(n)$. These signals were generated independently, according to

their assumed individual processes. The details of this procedure may be found in Appendix D.

The choice of the sampling rate may appear low (160 s^{-1}) but there are two reasons for avoiding a high sampling rate with the proposed method. The first is related to the precision of the coefficients of the filter when data are oversampled several times. Most of the poles of $H(z)$ would then be clustered together near the real axis and the coefficients, especially those for high powers of z^{-1}, which involve products of the poles would require a large precision to represent accurately. On an on-line portable system, precision is often limited by the capabilities of the processor. To prevent problems, the poles were dispersed over a large sector of the unit disk in the z-domain, which is maximized when sampling at twice the Nyquist frequency of the signal (70-80 Hz normally). Higher sampling rates would limit the time available between samples, which should be made sufficient for carrying out one cycle of the whole detection procedure described. Even if this was achieved, the possibility to process more than one channel at one time should be considered as a possible future prospect. Of course a higher sampling rate would mean that changes in the signal behaviour occur over a longer span of samples, hence the adaptive procedure could follow changes more quickly in terms of real-time. In this case, the system transfer function may require a higher order and there is a risk of the filter representing non-EEG signals in the upper frequency range (see section 6.5.9 and 6.5.10). These components may be filtered out, but then there would be no benefit in oversampling apart from increasing the speed of adaptation. This is not a strong argument, however, because the data at the nth sampling instant may be reused, carrying out K adaptation steps for every sampling step and thus reducing the time constant of the adaptive process and speeding up the LMS algorithm, as Bellanger[14, pp. 119–121] suggested.

Some waveforms to demonstrate the operation of the spike detection system are shown in Figure 8.11. The first waveform depicts the input simulated EEG signal. This is then processed through the adaptive background activity analysis (inverse) filter (sections 7.4.1, 7.4.4 and 8.4.3) to produce the deconvolved signal $r(n)$. The noise-like appearance of this signal and the superposition of the transients are clearly indicated in Figure 8.11. In the same Figure one may observe $y(n)$, the output of the spike analysis system, $G^{-1}(z)$. Here, spikes resemble impulses, but they are embedded in other high-amplitude signals (sections 7.4.2, 7.4.4 and 8.4.4). The (three-level) output (section 8.5.1) of the first transient detector, $z_t(n)$, clearly marks the transients present, whereas the spike initiator impulse output ($\dot{d}(n)$) and the corresponding spike, $\dot{s}(n)$ signal out of the generating system $G(z)$ (section 7.4.4) is an accurate representation of the spike, when one had occurred. This is clear after the

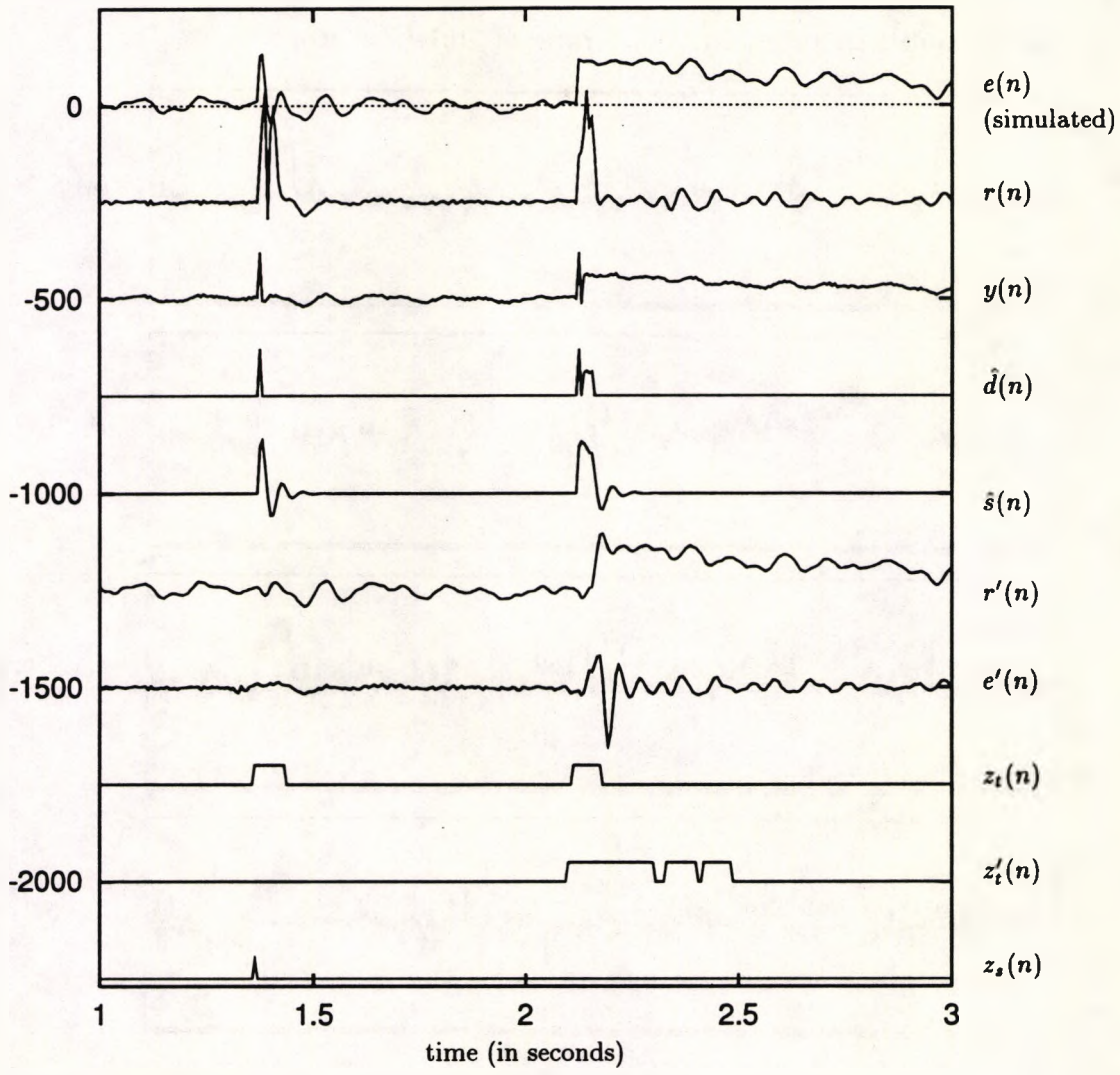


Figure 8.11: Waveforms showing various stages of the detection procedure for an artificial EEG record

subtraction of $\hat{s}(n)$ from $e(n)$, producing $e'(n)$, which is spike-free. Both components of $e(n)$, $\hat{s}(n)$ and $e'(n)$ are also shown. The absence of the spike is evident in the error, $r'(n)$, after the application of the background activity analysis filter, $H^{-1}(z)$ on $e'(n)$ is also shown. The final transient and spike detections, $z_t(n)$, $z'_t(n)$ and z_s, (section 8.5) correctly mark the spikes present.

This waveform also shows that transients, initially taken as spikes are rejected because they are still present after the removal of the (presumed) spike in the error signal, after passing $e'(n)$ through $H^{-1}(z)$.

Evidently, the method behaved well with artificial data, as demonstrated by the waveforms presented. It is unnecessary to say that all twenty spikes present in the data record

shown in the example above were detected successfully and there were no false alarms. This record corresponded to a signal-to-noise ratio of 10:1 ($\kappa = 10$).

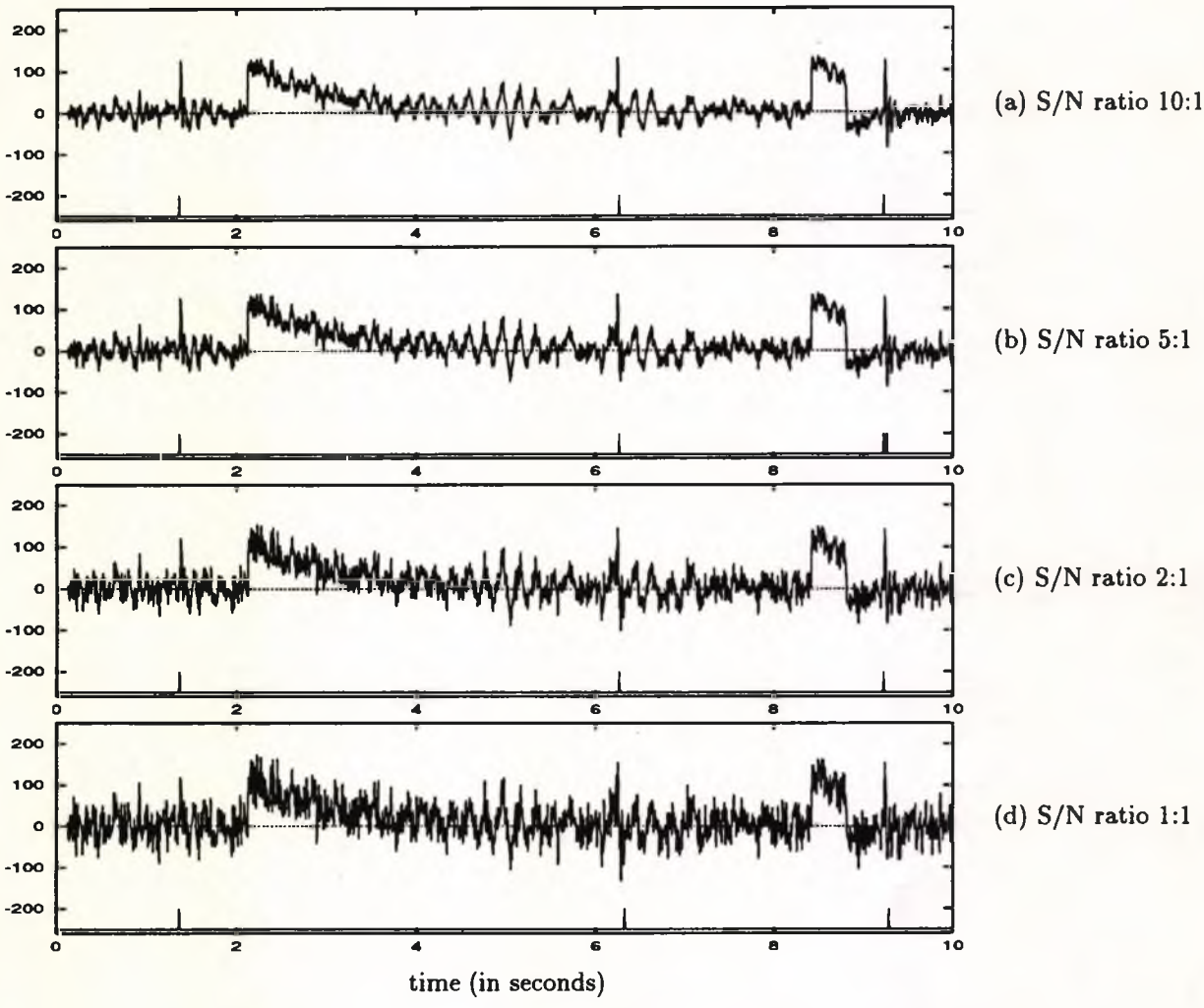


Figure 8.12: Artificial EEG data showing the deterioration of the signal with increasing noise levels

As noise levels increased further to a fifth, a half and then to the same power level as the signal, the structure of the signal input gradually became unrecognized (see Figure 8.12) and consequently, the performance of the spike detection system gradually deteriorated. The results for the same record for the four values of κ tested are shown in Table 8.1.

Clearly, the increase of the noise level in the EEG affected the performance of the spike detector considerably. This has not been unexpected. The gradual reduction in the number of detected spikes, as the noise level in the input increased approaching that of the main signal (the background activity) is the result of two factors. The first is the effect of the increasing noise level at the output, $r(n)$, of the inverse background activity model, $H^{-1}(z)$. In the absence of noise, this output contains a random, uncorrelated signal, $u(n)$, from

Simulated EEG record*		E1	E2	E3	E4
Noise power ratio ($\frac{1}{\kappa}$)		0.1	0.2	0.5	1.0
Spike Detections	Correct	20	16	7	2
	Missed	0	4	13	18
	False	0	1	3	5

*: All records had the same length (10000 samples) and number of spikes (20)

Table 8.1: Results for Artificial Data Records

the background activity, whose power is σ_{uu}^2. On this signal, there are spikes and other transients (distorted by $H^{-1}(z)$) which are clearly distinguished from $b(n)$ because their squared amplitudes for a brief duration are well above σ_{uu}^2 (see section 7.4.1). Therefore they can be distinguished by thresholding at a level t_0 above the level of σ_{uu}. When noise is present, this causes an excess error signal in $r(n)$, which cannot be isolated from $u(n)$ and since σ_{uu}^2 is measured on-line what is actually computed is the variance of both $b(n)$ and the (filtered) noise signal, $v_P(n)$. Since these signals are independent, their variances are additive and because the noise power was a fraction, κ of σ_{uu}^2, in $r(n)$ it may still be regarded as proportional, $\eta\sigma_{uu}^2$. Hence the effective variance will be $(1+\eta)\sigma_{uu}^2$. For a fixed value of t_0, the threshold level for spike amplitudes would increase by a factor $\sqrt{1+\eta}$. Hence the 'gap' that separated spike amplitudes from $b(n)$ is narrowed, as it is defined by the amplitude threshold function. Therefore some spikes with low-level amplitudes would not pass the threshold any more and are missed by the transient detector. The number of spikes missed increased as the noise level increased, which raised the threshold even further. When η reached unity, the signal and the noise had the same power and the amplitude threshold increased by a factor of $\sqrt{2}$, which is almost one and a half times the original and can inhibit some spike detections. nonetheless, the results for even for high levels of noise were much better than anticipated.

The presence of extraneous signals in the input of the adaptive predictor caused a deviation from the optimal weights for H, and affected the convergence characteristics of the adaptive algorithm. $\mathbf{h(n)}$ The first is clearly demonstrated by considering the theoretical solution, expressed in the Wiener-Hopf equation (8.9). This has the second-order statistics (correlation functions) of the input and the desired signal. In linear prediction, considered here, these are identical, with the input being delayed by one sample (section 8.1.1). The optimal linear predictor weight vector is given by equation (8.15). In the absence of noise, the input $x(n) \equiv b(n)$ and the optimal coefficient vector is $\mathbf{a}^* = \hat{R}_{bb}^{-1}\mathbf{r}_{bb}$. When the input

is $x(n) = b(n) + v(n)$, where $v(n)$ is random, uncorrelated noise, the above equation will contain the auto-correlation of $x(n)$, instead of $b(n)$. Because $b(n)$ and $v(n)$ are statistically independent, their correlations may be superimposed (added). The auto-correlation function of $v(n)$ is zero everywhere, apart from zero lag, where it is equal to its power, $r_{vv}(0) = \sigma_{vv}^2$. Hence $R_{xx} = \bar{R}_{bb} + \sigma_{vv}^2 I$ and $\mathbf{r}_{xx} = \mathbf{r}_{bb} + [\sigma_{vv}^2 \mathbf{0}^T]^T$. Evidently the optimal coefficient vector obtained through these equations is not the same as the one computed in the noise-free case. The difference becomes more profound with increasing noise power.

The extraneous noise also affects the adaptive implementation, since it is reflected in the estimated gradient vector (see equation 8.43). Noise in the input affects the estimation of the gradient, which also becomes noisy and may not point accurately in the correct direction to proceed. Hence, adaptation is slowed down. Reduction in the speed of the algorithm is also observed because of the on-line estimation of the adaptation rate, μ, as explained in section 8.4.3 (equation (8.54)). Since μ was made inversely proportional to the power of the input signal, the presence of noise directly increases the input signal power and thus slows adaptation down, since the time constant is related to μ.

The inverse spike filter output, $y(n)$ is affected to a lesser extent. The reason for these effects is the overestimation of the spike amplitude by the spike initiator (section 8.5.2), whose operation is based on the variance of $y(n)$ to compute how much a spike amplitude excitation function should be. The presence of noise makes its power higher (the sum of variances of the background activity and the noise, after being filtered by $G^{-1}(z)$) which increases the uncertainly of the true amplitude of the spike excitation function. As a result, overestimation is now more profound and for some spikes this could not be compensated by the on-line adaptive behaviour of the subsequent spike generator, $G(z)$. This made some spike estimates larger than they should be, hence causing negative transients when subtracted from the EEG, which were observed by the spike detector (section 8.5.3) and rejected as false detections.

The combined effect of all sources of error made the detection of many spikes impossible when the level of noise was large. In real life, however, noise levels are lower than those that caused the deterioration of the system behaviour. Besides, it is expected that some noise-reduction method should applied as part of the pre-processing stage before the signals are submitted for analysis to either the system or a human analyst. It is important to note that in conditions where the system failed, human interpretation could not be applied reliably, as visual inspection is seriously impaired by the presence of large noise amplitudes.

8.6.2 Testing with real EEG signals

After assessing the reliability of the analysis system with artificial signals conforming to the assumed synthesis model, tests were carried out using real EEG records. As stated before, the models used were only approximate, since the processes of real EEG signal generation may be correlated and possibly non-linear.

The response of the system was observed for some EEG data records. Four records were sampled and stored in computer files for the purpose of testing. These are the same as those used to test the earlier methods presented in Chapter 6. The sampling rate was 160 s^{-1} for all four records. The first record contained many spikes but was clear of artifacts and other extracerebral signals. The next two contained both spikes and muscle artifacts, whereas the last contained only artifacts. Unfortunately more extensive tests could not be carried out due to the unavailability of other data records at the time of completion of this work.

All fifty-two spikes, as identified by visual analysis, in the first record were detected correctly by the system. There were two false detections which were caused by transients that might have been identified as spikes if their amplitude were higher. These were rejected during the manual assessment of the record because they were not clearly distinguished from the background activity. This characterization is however subjective and hence it was decided not to increase the thresholds for the detection of transients to eliminate them, as it is better to detect these disputed spike occurrences for human experts to analyse later.

The next two records contained 8 and 10 spikes respectively as well as some artifacts. Spikes present were detected but there were some false detections (2 and 6) respectively caused by spiky muscle artifacts whose amplitudes and shape resembled spikes. These were located near the beginning of the artifact, which appears as a spike burst.

The same phenomenon was observed in the last record, which technically contained only artifacts with no epileptic spikes. In this record there were 25 false detections, observed primarily near the onset of muscle artifact bursts, as before. Artifact bursts seem to be approximated by the background activity prediction error filter, as they normally have longer durations than single spikes. This is evident by the changes of the coefficient vector for $H(z)$ observed during the artifact and the noisy appearance of the prediction error, indicating that the structure of the signal is captured by $H(z)$. The few initial detections at every artifact burst have been associated with the adaptive method. Because coefficients are updated in small steps, they cannot change instantaneously with the behaviour of the input signal. This did not seem to be a problem with transitions between segments of the background activity (section 8.4.3), because they were gradual and, to a large extent, could be followed by the adaptive process. On the contrary, muscle artifact bursts start spontaneously and

bear no relation to the EEG activity currently in progress, which they completely swamp. The sudden initial transition cannot be captured by $H(z)$ (which changes more slowly) and may give some initial transients in $r(n)$, which are found by the transient detector and lead to false spike detections. The double inverse modelling mechanism cannot detect the error, because some artifacts have spectral characteristics similar to epileptic spikes, from which they may be distinguished only by considering context information.

Testing with real EEG records will be covered more extensively in the next Chapter, where some alternative adaptive algorithms are considered.

Chapter 9

Alternative system realizations

The implementation of $H^{-1}(z)$ as an adaptive LMS system (section 8.4.3) reflected the time-varying nature of the on-going EEG signal. Although the system behaved well with simulated signals, there was some concern about its response to real EEG data, especially those containing bursts of spiky muscle artifacts. When the system misinterpreted a number of artifacts and other large amplitude signals as epileptic spikes, this was attributed to the slow dynamic response of the LMS algorithm to changes in the signal behaviour. The rate of convergence of this algorithm, is inherently linked to the differences in the spectral properties of the signal before and during the artifact burst. Hence, if $H^{-1}(z)$ was 'close' to the one required for the artifact, then the number of false detections was small, otherwise a number of spikes were detected until the LMS algorithm changed the system to model the burst sufficiently.

The maintenance of the minimum phase property of the linear predictor $H^{-1}(z)$ (hence the stability of $H(z)$) is also crucial. This could be investigated indirectly, using the Schür-Cohn stability test, but other realizations offer simpler methods as well as other desirable properties.

To address these issues and to select a configuration which converged fast, to track non-stationarities in the EEG signal whilst providing the means to control certain properties of the adaptive system, possibly including the location of its poles or zeros, other adaptive methods were considered. Some of these use alternative means to search the performance surface, not based on the LMS algorithm, and have superior properties. Others use alternative realizations of the prediction error systems. By restructuring the transfer function it is possible to change the shape of the performance surface so that the slow adaptation modes of the LMS are eliminated.

In this Chapter three such realizations will be described. The first is an alternative

adaptive algorithm, called the RLS. The other two are variants of the LMS method for the cascade and lattice realizations of the transfer function $H^{-1}(z)$. Although these were initially considered for the background activity prediction-error filter, the possible realization of $G(z)$ was also considered. On some occasions, variants of the 'standard' algorithms were developed for the purpose of EEG analysis. Various system configurations will be presented and their results discussed.

9.1 The Recursive Least Squares (RLS) algorithm

This method utilizes a more sophisticated gradient search method than the one that forms the basis of the LMS algorithm. Thus, instead of progressing in the direction of the steepest descent, the point of zero gradient is sought by following Newton's method. In the one-dimensional case this is developed by considering the following linear approximation of the gradient of the function whose zero is sought, $\frac{d\xi(w)}{dw}$ at the point where $w = w(k-1)$:

$$\frac{d^2\xi(w(k-1))}{dw(k-1)^2} = \frac{\frac{d\xi(w(k-1))}{dw(k-1)} - \frac{d\xi(w(k))}{dw(k)}}{w(k-1) - w(k)} \tag{9.1}$$

Solving this for $w(k)$ and assuming that $\frac{d\xi(w(k))}{dw(k)} \approx 0$, results in the following recursive relation for updating the weight which corresponds to the point of zero gradient[202, pp. 52–54]:

$$w(k) = w(k-1) - \frac{\frac{d\xi(w(k-1))}{dw(k-1)}}{\frac{d^2\xi(w(k-1))}{dw(k-1)^2}} \tag{9.2}$$

Since $\xi(w)$ is quadratic in w, its second derivative is a constant and Newton's method converges to the correct solution in a single iteration, rather than in small steps, like the steepest descent method (section 8.4.1). This property was maintained by the multi-dimensional extension of Newton's method. By substituting $\mathbf{r}_{xy} = R_{xx}\mathbf{w}^*$ (from the Wiener-Hopf equation (8.9)) the following expression is obtained:

$$\nabla\xi(\mathbf{w}) = 2R_{xx}\mathbf{w} - 2R_{xx}\mathbf{w}^* \tag{9.3}$$

Solving for $\mathbf{w}^*$ and converting it into a recursive equation (by letting $\mathbf{w} = \mathbf{w}(k-1)$ and $\mathbf{w}^* = \mathbf{w}(k)$) the vector form of Newton's method is revealed[202, pp. 55–56]:

$$\mathbf{w}(k) = \mathbf{w}(k-1) - \frac{1}{2}R_{xx}^{-1}\nabla\xi[\mathbf{w}(k)] \tag{9.4}$$

Widrow explained that in practice, because of computational errors in R_{xx}^{-1} it is required to stabilize the method by introducing a dimensionless convergence rate, μ, instead of the

factor $\frac{1}{2}$ in the above relation. The result is a 'generalized' Newton's algorithm, which is stable for $0 < \mu < 1$, but critically damped for $0 < \mu < \frac{1}{2}$, which is the useful range. Its behaviour may be studied by translating and rotating the weight space using the optimal weight vector as the origin and the principal (eigenvector) basis functions as the directions of the coordinate axes. It can then be shown that this method of gradient search is equivalent to a (non-recursive) geometric series with ratio $(1 - 2\mu)$. This is constant for all modes of the performance surface and, hence, it is unaffected by the eigenvalue spread, the main reason for slow convergence of the LMS algorithm (see section 8.4.2). Expressions for the time constant, τ, and the misadjustment, M, may also be derived using the same procedure and arguments as with the steepest descent method[202, pp. 147–154]:

$$\tau \approx \frac{1}{4\mu} \tag{9.5}$$

$$M_{tot} \approx \frac{(L+1)\lambda_{av}}{16NP\tau}\left(\frac{1}{\lambda}\right)_{av} + P \tag{9.6}$$

Evidently, the time constant is smaller than for the method of the steepest descent, since it is independent of the eigenvalues of R_{xx}. Newton's method also has a lower misadjustment[202, pp. 89–90]. The better convergence properties of Newton's method are associated with the introduction of the inverse of the autocorrelation matrix of the input signal, which normalizes the convergence along every mode and makes the direction of convergence a straight line from the initial guess to the optimal $\mathbf{w}^*$, rather than the direction of the gradient.

The recursion here is more complicated than that of the steepest descent, because it contains two elements that are unknown and need to be estimated, the gradient vector, $\nabla\xi[\mathbf{w}(k)]$ and the inverse of the input auto-correlation matrix, R_{xx}^{-1}.

A method to estimate the gradient has been suggested by Widrow and Hoff, who substituted the stochastic function $\xi[\mathbf{w}(k)]$ with its instantaneous value, $\varepsilon(\mathbf{w})^2$, leading to the approximation $\nabla[\mathbf{w}(k)] \approx -2\varepsilon\mathbf{x}(k)$, as in the development of the LMS algorithm (section 8.4.2).

The estimation of R_{xx}^{-1} is more complicated. A recursive method starts by defining a time-updating relation for R_{xx}. A sample autocorrelation matrix may be computed by utilizing a finite-length window of past input values, or by an exponentially weighted window, extending infinitely into the past. The latter method is a generalization of the (one-dimensional) update relation for the power (equation (8.52)):

$$\begin{aligned}\hat{R}(k) &= (1-\alpha)\sum_{i=0}^{\infty}\alpha^i\mathbf{x}(k-i)\mathbf{x}(k-i)^T \\ &= \alpha\hat{R}(k-1) + (1-\alpha)\mathbf{x}(k)\mathbf{x}(k)^T\end{aligned} \tag{9.7}$$

where the index in $R_{xx}(k)$ has been dropped to simplify the notation. The objective, however, is to estimate $R^{-1}(k)$, which may be computed by matrix inversion, but its recursive computation is both simpler and faster. This procedure, which is similar to the one suggested by Widrow and Stearns[202, pp. 146–153] is demonstrated here briefly. Pre-multiplying equation (9.7) by $\hat{R}^{-1}(k)$ and post-multiplying by $\hat{R}^{-1}(k-1)$:

$$\begin{aligned}\hat{R}^{-1}(k)\hat{R}(k)\hat{R}^{-1}(k-1) &= \alpha\hat{R}^{-1}(k)\hat{R}(k-1)\hat{R}^{-1}(k-1) \\ &+ (1-\alpha)\hat{R}^{-1}(k)\mathbf{x}(k)\mathbf{x}(k)^T\hat{R}^{-1}(k-1) \\ \hat{R}^{-1}(k-1) &= \alpha\hat{R}^{-1}(k) + (1-\alpha)\hat{R}^{-1}(k)\mathbf{x}(k)\mathbf{x}(k)^T\hat{R}^{-1}(k-1)\end{aligned}$$

post-multiplying through by $\mathbf{x}(k)$ and solving for $R^{-1}(k)\mathbf{x}(k)$ gives the following important step in the development of the algorithm:

$$\hat{R}^{-1}(k)\mathbf{x}(k) = \frac{\hat{R}^{-1}(k-1)\mathbf{x}(k)}{\alpha + (1-\alpha)\mathbf{x}(k)^T\hat{R}^{-1}(k-1)\mathbf{x}(k)} \tag{9.8}$$

The denominator in the above relation is a scalar. Post-multiplying once more by $\mathbf{x}(k)^T\hat{R}^{-1}(k-1)$ and simplifying the expression by substituting $\hat{R}^{-1}(k-1)\mathbf{x}(k) = \mathbf{s}(k)$,

$$\begin{aligned}\hat{R}^{-1}(k)\mathbf{x}(k)\mathbf{x}(k)^T\hat{R}^{-1}(k-1) &= \frac{\hat{R}^{-1}(k-1)\mathbf{x}(k)\mathbf{x}(k)^T\hat{R}^{-1}(k-1)}{\alpha + (1-\alpha)\mathbf{x}(k)^T\hat{R}^{-1}(k-1)\mathbf{x}(k)} \\ &= \frac{\mathbf{s}(k)\mathbf{s}(k)^T}{\alpha + (1-\alpha)\mathbf{x}(k)^T\mathbf{s}(k)}\end{aligned}$$

Finally, noting that from equation (9.7), $\mathbf{x}(k)\mathbf{x}(k)^T = \frac{\hat{R}(k)-\alpha\hat{R}(k-1)}{1-\alpha}$ the desired expression for $R^{-1}(k)$ is found to be as follows:

$$\hat{R}^{-1}(k) = \frac{1}{\alpha}\left[\hat{R}^{-1}(k-1) - \frac{(1-\alpha)\mathbf{s}(k)\mathbf{s}(k)^T}{\alpha + (1-\alpha)\mathbf{x}(k)^T\mathbf{s}(k)}\right] \tag{9.9}$$

The *Recursive Least Squares* (RLS) algorithm is then defined as a set of initial conditions, and a recursive procedure to be followed for every sample. Initially, the 'forgetting factor' is set to a value $\alpha = 2^{-\frac{1}{n_s}}$ to reflect the length of stationarity, n_s, of the input signal, as suggested by equation (8.53) and the associated explanation. As an initial value for $\hat{R}^{-1}(k)$ Widrow[202, p. 150] suggested the use of the identity matrix multiplied by a large positive constant. A more appropriate value was used in the implementation, since some information about the signal is known. It should be remembered that, during the construction of the LMS adaptive filter for the background activity (section 8.4.3), an initial power estimate $\bar{\sigma^2}$ was necessary for the computation of μ. The same estimate was used here. If the correlation matrix was diagonal, its diagonal entries would have been equal to the power of the signal. In

the recursive case the best estimate of the initial auto-correlation matrix, which is considered diagonal is $\hat{R}(0) = \bar{\sigma^2} I$. The inverse of this matrix is simply:

$$\hat{R}^{-1}(0) = \frac{1}{\bar{\sigma^2}} I \tag{9.10}$$

The weight vector was initialized at the origin. The algorithm is then described by the following procedure:

1. The process is initialized by setting $\mathbf{w}(0) = \mathbf{0}$, $\hat{R}^{-1}(0) = \frac{1}{\sigma^2} I$
2. Compute the first weight vector, $\mathbf{w}(1) = \mathbf{w}(0) + 2\mu \hat{R}^{-1} \varepsilon(0)\mathbf{x}(0)$, where $\varepsilon(0) = y(0)$
3. for $k = 1, 2, \ldots$ the following steps are repeated
   (a) Evaluate the output of the filter, $\hat{y}(k) = \mathbf{w}(k)^T \mathbf{x}(k)$
   (b) Find the error, $\varepsilon(k) = y(k) - \hat{y}(k)$
   (c) Evaluate $\mathbf{s}(k) = \hat{R}^{-1}(k-1)\mathbf{x}(k)$
   (d) Compute $\gamma = \alpha + (1-\alpha)\mathbf{x}(k)^T \mathbf{s}(k)$
   (e) Estimate $\hat{R}^{-1}(k+1) = \frac{1}{\alpha}\left[\hat{R}^{-1}(k) - \frac{(1-\alpha)}{\gamma}\mathbf{s}(k)\mathbf{s}(k)^T\right]$
   (f) Change the weight vector, $\mathbf{w}(k+1) = \mathbf{w}(k) + 2\mu \hat{R}^{-1}(k)\varepsilon(0)\mathbf{x}(k)$

Strictly speaking, this is a hybrid algorithm demonstrated by Widrow and Stearns[202, pp. 146–153] and called the *Sequential Regression* algorithm (SER). The RLS algorithm consists of the special case when $\mu = \frac{1}{2}$, that is an implementation of Newton's method in the multi-dimensional case.

This algorithm converged to the initial optimal solution more quickly than the LMS algorithm described earlier. The price for the speed of convergence is the complication of the algorithm which contains many more storage elements than the LMS and requires matrix operations. The latter is perhaps more critical, because the number of multiplications per iteration necessary for the implementation of an Lth order filter is $3L^2 + 2L + 3$. This is proportional to L^2 and as L increases the method may be slowed down and its on-line implementation on a small portable microcomputer may become impossible. There is also the issue of numerical stability of this algorithm, which is particularly susceptible to round-off noise and may not converge in some cases, especially when fixed-point arithmetic with a short wordlength is used, a fact highlighted by Proakis and Manolakis[151, pp. 888–917] and Proakis et al.[152, pp. 351–389]. Solutions to the stability problem have been suggested. A brief, but thorough review of some of the most popular may be found in the above references, as well as in a text by Orfanidis[140, pp. 473–492]. These include the well-known *square*

root algorithm, which decomposes the auto-correlation matrix into the product of a lower triangular matrix, E_L, a diagonal matrix, Δ and an upper triangular matrix, $E_U = E_L^T$. Thus, $\hat{R}(k) = E_L(k)\Delta(k)E_L(k)^T$, where the factors of the matrix, $E_L(k)$ and $\Delta(k)$ are updated on every iteration, instead of $\hat{R}(k)$. This method is more numerically robust than the RLS described earlier, but requires even more computations for its implementation (of the order of L^3). Fast algorithms exist, which require a number of operations proportional to L. There are many variants of the basic algorithm and are based on the lattice filters, which recursively decompose the autocorrelation matrix into a succession of lower-order matrices (see section 9.3). The resulting update relations are performed on scalar quantities and thus avoid matrix operations, while being numerically stable. One such method, the *Fast a Posteriori Error Sequential Technique* (FAEST) developed by Carayannis et al.[33], requires only $5L$ operations per recursion. Ljung and Ljung[111], however, demonstrated theoretically that these algorithms are potentially unstable when an arbitrary numerical error is introduced into the algorithm.

9.2 The Cascade LMS method

The LMS and the RLS algorithms used the implicit assumption that the system is implemented in direct form, following a (polynomial) transfer function, like the linear predictors $H^{-1}(z)$ and $G^{-1}(z)$. In the time-domain this corresponds to a weighted sum of delayed input samples, a straightforward realization, but not free from problems. One major issue arises from the minimum-phase constraint imposed on the predictor to ensure the stability of its inverse autoregressive system. A similar problem was encountered during the attempt to limit the range of adaptation for $G(z)$ (section 8.4.4). If the zeros of the transfer function were known, both these problems would become trivial, since the stability and the range of adaptation can be tested. Unfortunately, there is no closed form for the factorization (root finding) of a polynomial transfer function of order $L > 3$. Numerical methods could be employed, but these would be too costly to implement on-line, compared to other available solutions to this problem. These suggested the use of alternative realizations, where properties related to the zeros of the transfer function could be tested more easily.

The *cascade realization* has these properties as well as other advantages over the direct realization. The filter structure, as the name suggests, consists of breaking down the polynomial transfer function of order L into a product of lower order polynomials:

$$H^{-1}(z) = P_1(z)P_2(z)\cdots P_N(z) = \prod_{i=1}^{N} P_i(z) \tag{9.11}$$

If L first-order sections are used, each will have the form $P_i(z) = 1 - p^i z^{-1}$ and the zeros of the polynomial are equal to the coefficients, p_i. These, however, are generally complex values and in order to avoid complex arithmetic, second order sections, $P_i(z) = 1 + a_1^i z^{-1} + a_2^i z^{-2}$, have been implemented instead. The zeros p_1^i and p_2^i of these sections may be easily computed, using the well-known formula $p_{1,2}^i = -a_1^i \pm \frac{1}{2}\sqrt{a_1^{i\,2} - 4a_2^i}$. The stability of $H(z)$ is trivial to test, since stable systems have $|p_j^i| < 1$ for $i = 1, 2, \ldots, N$ and $j = 1, 2$.

In this implementation, the order of the filter must be even. This does not affect the generality of the method, since possible overestimation of the order by 1 to make it so has only marginal effects on the computational requirements for its implementation. The factorization of the polynomial transfer function which is implied by the cascade realization has other useful properties, apart from the easy test for stability. First of all, the structure is less sensitive to the effects of quantization of the coefficients, than the direct one. This may be critical for its implementation on a small, portable microcomputer with finite wordlength using fixed-point arithmetic. A brief comparison of the two filter structures may be found in [151, pp. 750–752]. The effects in question are related to the sensitivity of the zeros associated with small changes in the coefficients like those imposed by quantization. It can be shown[15, pp. 193–210] that the error on the location of the zeros increases with their clustering and with the order of the system, L. Because the first is indeterminate for an adaptive system, reduction of the order is the only means of controlling the sensitivity of the zeros. The implementation of a high-order system as a cascade of second order systems is a practical interpretation of this observation. Moreover, because each section combines two poles, which are generally a complex conjugate pair, their clustering is avoided, unless these are close to the real axis in z. Another characteristic of the cascade realization is the reduction in the dynamic range of its coefficients compared to the direct realization. For an Lth order system, there are L zeros and the coefficient a_L corresponding to z^{-L} in the polynomial $H^{-1}(z)$ is the product of all the zeros. Assuming that all poles lie in the annulal region $\alpha < |z| < \beta$, where $0 < \alpha < \beta < 1$ the magnitude of a_L must lie in the range (α^L, β^L). If the system is a linear predictor, $a_0 = 1$ and the poles are dispersed in all quadrants of the unit circle then the ratio $\frac{a_L}{a_0} = a_L$ characterizes the dynamic range of the coefficients. For the range of zeros considered, the lower limit is more significant. Hence for a 10th order filter and $\alpha = 0.9$ the range of coefficients may be from an order around 1 to as low as $(0.9)^{10} \approx 0.349$. For $\alpha = 0.6, 0.5, 0.3$ the corresponding limits for a_L are $0.006, 0.00098$ and 0.000006, which indicate the requirement for long wordlengths for their storage. On the contrary, the cascade implementation has a fixed range α^2, independent of L, which gives a lower limit of 0.09 for $\alpha = 0.3$, a dramatic reduction in the dynamic range of its coefficients,

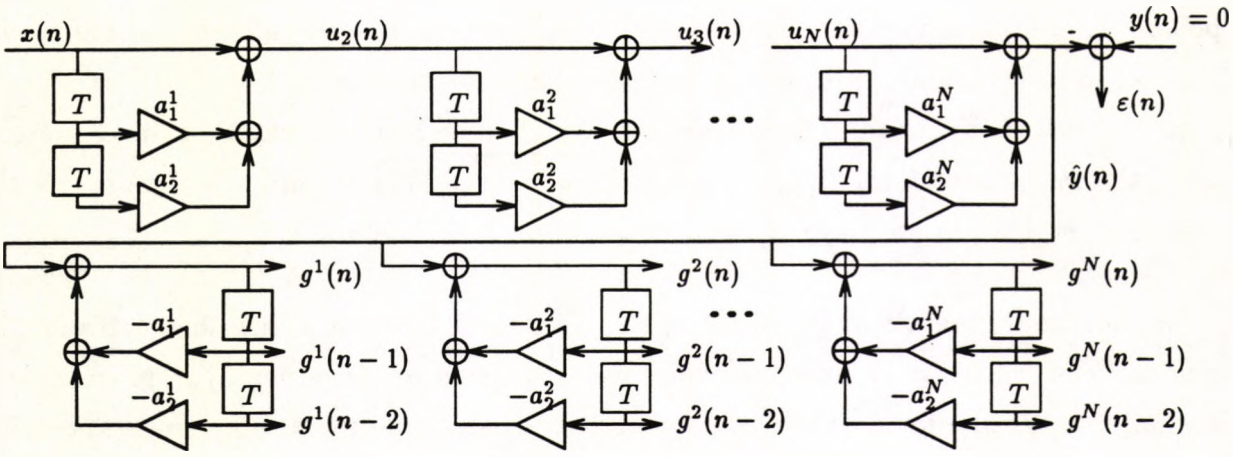


Figure 9.1: Adaptive System in Cascade Form

enabling all computations to be carried out in fixed-point arithmetic.

The cascade realization is also intuitively appealing, because the second-order factors have a direct spectral interpretation, under some circumstances. If the input consists of a superposition of sinusoids, some of the two-zero sections will model a sinusoid each, thus providing a spectral parametrization of the signal. Expressing the zeros of each section in polar form $p^i_{1,2} = r_0 e^{\pm j\omega_0 T}$, then ω_0 is the frequency of the modelled sinusoid, whereas r_0, the distance from the origin in z, related to the bandwidth, is characteristic of the rate of decay. For the background activity, this may assist in its automatic classification into, α, β, δ and θ (see section 2.5.1), or even provide a continuous description of the dominant frequency of the signal and the way this changes with time.

Adaptive cascade systems have not been very popular and the literature on the subject is rather limited. A paper by Jackson and Wood[94] provides a good coverage. These authors proposed a gradient, LMS-type algorithm which was used with modifications by other workers for the implementation of both FIR and IIR cascade adaptive filters for specific applications where on-line factorization of transfer functions was required[53, 163]. A recursive-least-squares method was also described by Romano et al.[164].

The algorithm of Jackson and Wood was used for the realization of $H^{-1}(z)$. A detailed account of the gradient (LMS) algorithm for the cascade realization has also been given by Bellanger[14, pp. 123–124][15, pp. 318–319]. The adaptive system, implemented as a cascade of second-order sections, is shown in Figure 9.1. The overall transfer function is the product of the transfer functions of the $N = \frac{L}{2}$ sections, $P_i(z) = 1 + a^i_1 z^{-1} + a^i_2 z^{-2}$.

The error of adaptation, $\varepsilon(n)$ is again the difference between the desired signal, $y(n)$, and the output of the system, $\hat{y}(n)$. The cost function to minimize is the MSE, as before $\xi(\cdot) =$

$E[\varepsilon(n)^2]$. Following the arguments that led to the development of the LMS algorithm for the direct system realization (section 8.4.2), the MSE is approximated by the instantaneous value of the squared error, $\hat{\xi}(\cdot) = \varepsilon(n)^2 = [y(n) - \hat{y}(n)]^2$. This is minimized when its gradient is zero. For a gradient descent algorithm, small adjustments of size proportional to μ are made on the coefficients of every stage $P_i(z)$ and direction opposite to that of the gradient, on every iteration:

$$\begin{aligned} a_k^i(n) &= a_k^i(n-1) - \mu \frac{\partial \varepsilon^2(n)}{\partial a_k^i} \\ &= a_k^i(n-1) - 2\mu\varepsilon(n)\frac{\partial \varepsilon(n)}{\partial a_k^i} \\ &= a_k^i(n-1) + 2\mu\varepsilon(n)\frac{\partial \hat{y}(n)}{\partial a_k^i} \end{aligned} \tag{9.12}$$

It is necessary to compute the partial derivative of the output of the filter with respect to all coefficients $\{a_k^i; i = 1, 2, \ldots, N \;\; k = 1, 2\}$. Because $\hat{y}(n)$ may be related to the input $x(n)$ and the $\{a_k^i\}$ by means of convolution in the time-domain, it is easier to use the z-domain representation, where multiplications substitute convolutions:

$$\hat{Y}(z) = H^{-1}(z)X(z) = \prod_{i=1}^{N} P_i(z)X(z) \tag{9.13}$$

$$= \prod_{i=1}^{N}(1 + a_1^i z^{-1} + a_2^i z^{-2})X(z) \tag{9.14}$$

Employing the inverse z transform,

$$\hat{y}(n) = \oint_{\Gamma} z^{n-1} Y(z) dz \tag{9.15}$$

where Γ is a closed anti-clockwise contour in the region of convergence of $\hat{Y}(z)$ enclosing the origin, the required partial derivatives may be expressed as follows:

$$\begin{aligned} \frac{\partial \hat{y}(n)}{\partial a_k^i} &= \oint_{\Gamma} z^{n-1} \frac{\partial}{\partial a_k^i} \prod_{i=1}^{N}(1 + a_1^i z^{-1} + a_2^i z^{-2})X(z)dz \\ &= \oint_{\Gamma} z^{n-1} z^{-k} \prod_{i=1, i\neq k}^{N} (1 + a_1^i z^{-1} + a_2^i z^{-2})X(z)dz \end{aligned}$$

$$\text{but} \quad H^{-1}(z)X(z) = \prod_{i=1}^{N} P_i(z)X(z) \tag{9.16}$$

$$\begin{aligned} \text{hence} \quad \frac{\partial \hat{y}(n)}{\partial a_k^i} &= \oint_{\Gamma} z^{n-1} z^{-k} \frac{H^{-1}(z)X(z)}{P_k(z)} dz \\ &= \oint_{\Gamma} z^{n-1} z^{-k} \frac{\hat{Y}(z)}{P_k(z)} dz \end{aligned} \tag{9.17}$$

The last expression may be interpreted as the output, $g^i(n-k)$, of an autoregressive system with transfer function $\frac{1}{P_i(z)} = \frac{1}{1+a_1^i z^{-1}+a_2^i z^{-2}}$. The presence of the factor z^{-k} for $k = 1,2$ indicates that the partial derivatives are given by *delaying* the output of the corresponding system, as shown in Figure 9.1.

This algorithm is significantly more complex than the LMS for the direct system implementation. It requires $L = 2N$ multiplications to implement the filter and an equal number to compute the partial derivatives. Updating the coefficients requires another 2 multiplications per stage and one extra multiplication to scale $\varepsilon(n)$ by 2μ. Overall, a total of $3L+1$ operations are required per sample, compared to $2L$ for the direct LMS, hence the penalty is not enormous, as it is still proportional to L. Some consideration is necessary for the range of the learning rate, μ, for which the adaptive process is convergent. This is related to a number of issues associated with the shape of the performance surface, $\xi(\mathbf{a})$, where $\mathbf{a} = (a_1^1\ a_2^1\ a_1^2\ a_2^2\ \ldots a_1^N\ a_2^N)^T$ is the coefficient vector. Unlike the direct implementation, where the performance surface was quadratic having a unique minimum, here its shape is far from quadratic and there are many minima. This is easy to realize, because changing the order of the stages $\{P_i(z)\}$ results in the same overall transfer function. Every such permutation, however, corresponds to a different coefficient vector, $\mathbf{a}$, all of which are equivalent. For an N-stage filter there are $N!$ solutions. Besides, attempting to produce an analytical theoretical solution using correlation functions, like the Wiener-Hopf equation, is not possible because of the dependence of the coefficients on the outputs of the individual stages (which themselves depend on elements of the coefficient vector). This interdependence leads to a set of non-linear equations which cannot be solved explicitly[94]. Numerical optimization, using the gradient algorithm suggested earlier or otherwise, may yield one of the solutions, but there is a possibility that the weight vector may be trapped on flat (almost zero derivative) regions between the desired solutions, or on a local maximum (saddle point). These are, however small possibilities, because of the multitude of equally optimal minima of the MSE surface, and in general the algorithm will eventually converge to one of these $N!$ solutions. Each solution has a basin of attraction around it, which is convex. Hence a gradient algorithm, like the one suggested, may reach a different optimal solution, depending on the initial weight vector. Once inside the basin of attraction, certain assumptions may be employed to *linearize* the differential equations and investigate the convergence properties of the algorithm analytically. These may include a Taylor expansion approximation of the error surface *around* the optimal solution, but since the argument only began with the attempt to find the range of μ for which the algorithm is convergent, a less formal analysis was sufficient. In general, the gradient components $\{g^i(n-k)\}$ for different sections are highly correlated,

because of the dependencies between the cascaded stages. For coefficient vectors near an optimal solution, however, these correlations are small[94][164]. The interdependences are absorbed by the decomposition of the input signal to each stage. Since these are responsible for the non-linearity of the equations, not only the problem is linearized, but the correlation matrix of the inputs to the stages of the system becomes block-diagonal, each block consisting of a 2×2 sub-matrix:

$$R_{zz} = \begin{bmatrix} r_{11}(0) & r_{11}(1) & 0 & 0 & \cdots & 0 & 0 \\ r_{11}(1) & r_{11}(0) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & r_{22}(0) & r_{22}(1) & \cdots & 0 & 0 \\ 0 & 0 & r_{22}(1) & r_{22}(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & r_{NN}(0) & r_{NN}(1) \\ 0 & 0 & 0 & 0 & \cdots & r_{NN}(1) & r_{NN}(0) \end{bmatrix} \equiv \begin{bmatrix} R_{11} & 0 & \cdots & 0 \\ 0 & R_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{NN} \end{bmatrix} \tag{9.18}$$

where notation was simplified, so that $r_{ii}(n)$ is the auto-correlation of the input sequence to the ith stage of the cascade (for the first stage this is equivalent to $r_{xx}(n) \equiv r_{u_1 u_1}(n)$ and for the rest $r_{ii}(n) \equiv r_{u_i u_i}(n)$). In this form one may use the arguments put forward for the quadratic performance surface in section 8.4.2, to derive that the method converges for the range suggested by equation (8.36) and then use the practical limit of equation (8.46) $0 < \mu < \frac{1}{(L+1)\sigma_{xx}^2}$. A better limit may be derived, however by considering the above matrix in its sub-matrix form and hence convert the solution into a set of N simultaneous equations with two unknowns each. Then a different value μ_i may be used for every stage of the lattice defined by the inverse of the sum of the diagonal elements of the 2×2 sub-matrix R_{ii}. This is equal to twice the power in its input, $u_i(n)$, defining the limits:

$$0 < \mu < \frac{1}{(2)\sigma_{u_i u_i}^2} \tag{9.19}$$

The power and the values of the $\{\mu_i\}$ were estimated using the recursive methods discussed in section 8.4.3, and a (normalized) fraction of the above range, λ. The application of different convergence rates for every stage enhanced the speed of adaptation, as they tend to reduce the difference in the time constants, speeding up the convergence of slow modes. In this respect the adaptation algorithm is a hybrid, sharing properties of the steepest descent as well as Newton's algorithm. The block diagonal form of the matrix presented in equation (9.18) was exploited further by Romano et al.[164] who developed an RLS algorithm for adaptive filters in cascade form, since the inverse of R_{zz} may be computed easily from the inverse of the sub-matrices, R_{ii} (the inversion of these 2×2 matrices is trivial).

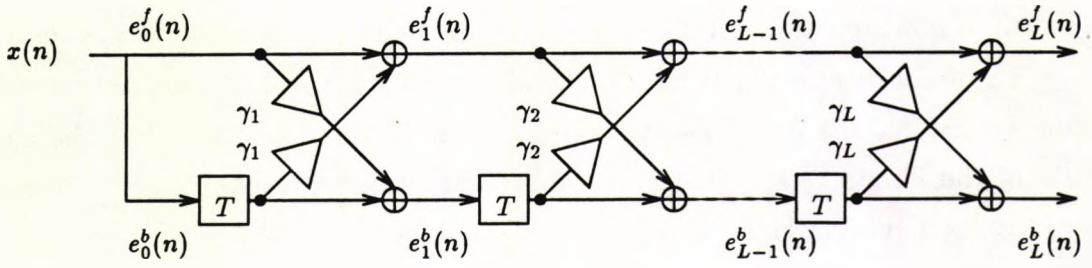


Figure 9.2: The lattice structure

The adaptive cascade structure suffers from one more potential problem. When two of the zeros of the overall transfer function become close to one another, the corresponding gradient estimates for sections i and j where this is observed, become similar in amplitude, $g_i(n-k) \approx g_j(n-k)$. This presents a problem in the non-stationary case, because if two or more zeros come close together and then separate again, the adaptive algorithm cannot separate the coefficients of the corresponding stages, once they become coincident (*root coalescence*). This problem may be prevented by updating only one section on every iteration.

The cascade structure was used for the implementation of the background activity inverse model, $H^{-1}(z)$, which is known in a factored form. Testing whether it is minimum-phase is now possible, whereas each factor may be interpreted in terms of resonant frequencies and bandwidths, which have a physical interpretation.

9.3 The Lattice LMS Linear Predictor

The concept of the lattice system realization was encountered indirectly in section 8.4.4 in the form of the Schür-Cohn stability test. The test required the computation of a set of parameters, $\{\gamma_i\}$, called the reflection coefficients, from the coefficients $\{a_i\}$ of the direct realization. The $\{\gamma_i\}$ were then used to test the stability (minimum phase) of the system, but they actually correspond to a special filter structure, called the lattice realization. *Adaptive lattice systems* have been very popular, because they possess several attractive properties.

Unlike the cascade realization, which is linked directly to the polynomial transfer function of the system, the connection between the system function and the lattice structure is not immediately apparent. Because the lattice consists of cascaded sections with two inputs and two outputs each (see Figure 9.2), it is not easy to assimilate to the single-input single-output system in direct form. The coefficients of the lattice realization, called the *partial correlation* or *reflection coefficients*, $\{\gamma_i\}$ are the same as the coefficients derived by the

Schür-Cohn recursive procedure to carry out the stability test in section 8.4.4. The lattice realization is an implementation of this recursive procedure, which takes advantage of the use of the forward as well as the reverse polynomial and hence it applies both transfer functions at once in a coupled form that produces the outputs of the two transfer functions simultaneously. Lattice system implementations are linked to linear prediction and arise naturally when prediction error systems, like the background activity inverse system, are considered. A linear one-step forward predictor has an associated prediction error system $P(z) = 1 + \sum_{i=1}^{L} a_i z^{-i}$ as seen in section 8.1.1. The output of this system is the error of the prediction of the current sample based on the L previous samples:

$$e_L^f(n) = x(n) + \sum_{i=1}^{L} a_i x(n-i) \tag{9.20}$$

The superscript f indicates that prediction is forward in time, whereas the subscript L indicates the order of the prediction system. Similarly, the reverse polynomial, $z^{-L}P(z^{-1})$ corresponds to a system which attempts to predict $x(n-L)$ using the more recent samples up to the present $x(n)$. This 'backward' prediction system has a prediction error given by the following relation:

$$e_L^b(n) = x(n-L) + \sum_{i=1}^{L} a_i x(n-L+i) \tag{9.21}$$

One of the properties of the lattice realization is the recursive computation of both the forward and the backward prediction errors for all orders up to L. With the introduction of every new stage, the prediction errors are computed using only those of the previous section (corresponding to a predictor of order one less) using the recursive relations[151, pp. 476–483][15, pp. 236–241]:

$$\begin{aligned} e_p^f(n) &= e_{p-1}^f(n) + \gamma_p e_{p-1}^b(n-1) \\ e_p^b(n) &= e_{p-1}^b(n-1) + \gamma_p e_{p-1}^f(n) \end{aligned} \tag{9.22}$$

where $p = 1, 2, \ldots, L$ and the recursion is initialized by setting $e_0^f(n) = e_0^b(n) = x(n)$. It is easy to show by mathematical induction that the relations (9.20) and (9.22) are equivalent. For $p = 1$, $e_1^f(n) = x(n) + \gamma_1 x(n-1)$, which corresponds to the direct realization with $a_1^1 = \gamma_1$. For $p = k$ it is assumed that:

$$\begin{aligned} e_k^f(n) = e_{k-1}^f(n) + \gamma_k e_{k-1}^b(n-1) &= x(n) + \sum_{i=1}^{k} a_i^k x(n-i) \\ e_k^b(n) = e_{k-1}^b(n) + \gamma_k e_{k-1}^b(n-1) &= x(n-k) + \sum_{i=1}^{k} a_i^k x(n-k+i) \end{aligned} \tag{9.23}$$

where a_i^k is the coefficient of a predictor of order k corresponding to lag i. For $p = k+1$, by substituting the above relations the correspondence between the direct and the lattice realization still holds.

$$\begin{aligned} e_{k+1}^f(n) &= e_k^f(n) + \gamma_{k+1} e_k^b(n-1) \\ &= x(n) + \sum_{i=1}^{k} a_i^k x(n-i) + \gamma_{k+1}[x(n-k-1) + \sum_{i=1}^{k} a_i^k x(n-k-1+i)] \\ &= x(n) + \sum_{i=1}^{k} (a_i^k + \gamma_{k+1} a_{k-i}^k) x(n-i) + \gamma_{k+1} x(n-k-1) \\ &\equiv x(n) + \sum_{i=1}^{k+1} a_i^{k+1} x(n-i) \end{aligned}$$

where in the last relation $a_{k+1}^{k+1} = \gamma_{k+1}$ and $a_i^{k+1} = a_i^k + \gamma_{k+1} a_{k-i}^k$ for $i = 1, \ldots, k$.

Having proved that the direct and the lattice prediction error filters are equivalent, the next step is to demonstrate an adaptive algorithm for the lattice structure. Assuming that the system has the form of a linear prediction error filter is merely for convenience, since any polynomial transfer function may be written in that form with the introduction of an output scaling factor, if the coefficient for $x(n)$ is not unity.

Adaptive lattice systems are more laborious to derive than their direct counterparts, because it is not immediately clear how the mean squared error minimization affects the partial correlation coefficients $\{\gamma_i\}$. This relation may be found by considering the Wiener equation for (forward) prediction $R_k \mathbf{a} = -\mathbf{u}_k$, where the notation has been simplified, so that R_k is the $k \times k$ autocorrelation matrix of $x(n)$ called $\hat{R}_{xx}$ earlier and $\mathbf{u}_k = \mathbf{r}_{xx}$ is the k-dimensional column vector of autocorrelations. Adding $\mathbf{u}_k$ on both sides and augmenting the auto-correlation matrix by an extra row and column $\mathbf{u}_k$ as well as the correlation element for zero lag, $r_0 \equiv r_{xx}(0)$, it becomes an $(k+1) \times (k+1)$ matrix, which has the same form as R_k, but has a higher dimension. Calling this matrix R_{k+1} and augmenting the coefficient vector accordingly by an element equal to 1 the following equation may be written:

$$R_{k+1} \begin{bmatrix} 1 \\ \mathbf{a}_k \end{bmatrix} = \begin{bmatrix} r_0 & \mathbf{u}_k^T \\ \mathbf{u}_k & R_k \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{a}_k \end{bmatrix} = \begin{bmatrix} E_k^f \\ \mathbf{0} \end{bmatrix} \tag{9.24}$$

where $E_k^f = [r_0 \mathbf{u}_k^T] \begin{bmatrix} 1 \\ \mathbf{a}_L \end{bmatrix}$ is the power of the forward prediction error. With similar arguments, using the matrix form of the optimum backward predictor equation, an equivalent relation is constructed:

$$R_{k+1} \begin{bmatrix} \mathbf{b}_k \\ 1 \end{bmatrix} = \begin{bmatrix} R_k & \mathbf{v}_k \\ \mathbf{v}_k^T & r_0 \end{bmatrix} \begin{bmatrix} \mathbf{b}_k \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ E_k^b \end{bmatrix} \tag{9.25}$$

where $\mathbf{v}_k = J_k\mathbf{u}_k$, has the same elements as $\mathbf{u}_k$ but in reverse order. It should be noted that J_k is the k-dimensional *co-identity matrix*. Moreover, the backward optimum predictor coefficient vector, $\mathbf{b}_k$ is related to the forward prediction vector $\mathbf{a}_k$ because of the symmetry properties of R_{k+1}. This matrix is symmetric (for stationary signals) and hence $R_{k+1}^T = R_{k+1}$. Because it is also symmetric about the secondary diagonal, $J_{k+1}R_{k+1} = [J_{k+1}R_{k+1}]^T = R_{k+1}^T J_{k+1}^T = R_{k+1}J_{k+1}$. Using this relation and pre-multiplying the backward prediction error equation by J_{k+1} yields:

$$R_{k+1}\begin{bmatrix} 1 \\ J_k\mathbf{b}_k \end{bmatrix} = \begin{bmatrix} E_k^b \mathbf{0} \end{bmatrix} \tag{9.26}$$

Comparing this relation to the forward prediction equation (9.24), the backward and forward prediction error coefficient vectors have the same elements, but in reverse order, $\mathbf{a}_k = J_k\mathbf{b}_k$. Moreover, under these conditions, the forward and the backward prediction errors have the same power, $E_k^f = E_k^b = E_k$. Expressions (9.24) and (9.25) are valid for all k and hence for $k-1$. With the aid of the prediction coefficients for order $k-1$ the following two expressions hold true:

$$R_{k+1}\begin{bmatrix} 1 \\ \mathbf{a}_{k-1} \\ 0 \end{bmatrix} = \begin{bmatrix} R_k & \mathbf{v}_k \\ \mathbf{v}_k^T & r_0 \end{bmatrix}\begin{bmatrix} 1 \\ \mathbf{a}_{k-1} \\ 0 \end{bmatrix} = \begin{bmatrix} E_{k-1} \\ \mathbf{0} \\ K_k \end{bmatrix} \tag{9.27}$$

$$R_{k+1}\begin{bmatrix} 0 \\ \mathbf{b}_{k-1} \\ 1 \end{bmatrix} = \begin{bmatrix} r_0 & \mathbf{u}_k^T \\ \mathbf{u}_k & R_k \end{bmatrix}\begin{bmatrix} 0 \\ \mathbf{b}_{k-1} \\ 1 \end{bmatrix} = \begin{bmatrix} K_k \\ \mathbf{0} \\ E_{k-1} \end{bmatrix} \tag{9.28}$$

where $K_k = \mathbf{v}_k^T\begin{bmatrix} 1 \\ \mathbf{a}_{k-1} \end{bmatrix} = \mathbf{u}_k^T\begin{bmatrix} \mathbf{b}_{k-1} \\ 1 \end{bmatrix}$. Multiplying equation (9.28) by $\gamma_k = K_k/E_{k-1}$ and subtracting from equation (9.27) to eliminate K_k produces the following equation:

$$R_{k+1}\left[\begin{bmatrix} 1 \\ \mathbf{a}_{k-1} \\ 0 \end{bmatrix} - \gamma_k\begin{bmatrix} \mathbf{b}_{k-1} \\ 1 \end{bmatrix}\right] = \begin{bmatrix} (1-\gamma_k^2)E_{k-1} \\ \mathbf{0} \\ 0 \end{bmatrix} \tag{9.29}$$

which by comparison with equation (9.24) produces the recursive relation for computing the forward (and backward) prediction coefficients for order k from those for order $k-1$:

$$\mathbf{a}_k = \begin{bmatrix} \mathbf{a}_{k-1} \\ 0 \end{bmatrix} - \gamma_k\begin{bmatrix} \mathbf{b}_{k-1} \\ 1 \end{bmatrix} \tag{9.30}$$

$$E_k = (1-\gamma_k^2)E_{k-1} \tag{9.31}$$

The last relation conforms with the recursive definition of the reflection coefficients for the Schür-Cohn stability test and they can implement the forward and backward prediction error systems described by the recursive equations at the beginning of this section.

The above analysis is a summary of the explanation given by Bellanger[14, pp. 135–168][15, pp. 322–326] and Proakis et al.[152, pp. 208–251]. It uses the assumption of stationarity of the input signal, which is responsible for the symmetry properties of the auto-correlation matrix. A more elaborate analysis with extensive explanation which does not rely on this assumption has been provided by Orfanidis[140, pp. 195–310].

9.3.1 Properties of lattice prediction error systems

The lattice realization of linear predictors has a number of properties that are worth noting.

Firstly, the forward predictor is a minimum phase system. This is expected, since the autoregressive process from which the input was generated is a stable system. This property may, however, be explained in terms of the function of the predictor and the minimization of the mean-squared error. Several detailed explanations of this property exist in literature[152, pp. 208–251][140, pp. 195–310], but a simple justification will be presented here. Effectively, the output of the forward predictor is the difference between the predicted and the actual value of its input. The power in this signal for a predictor of order k is the quantity E_k encountered earlier. Intuitively, a high-order predictor is expected to have a lower output power (which is equal to the minimum achievable MSE) than a low-order one. The improvement may stop when the order of the predictor becomes higher than the order of the generating AR process, but in any case it will never increase. Since the lattice implements all optimal predictors for orders $k = 1, 2, \ldots, L$, it is evident that there should be an improvement at the output of every stage, and therefore $E_k \leq E_{k-1}$. From equation (9.31), it is obvious that for this to be true $0 < (1 - \gamma_k^2)E_{k-1} < E_{k-1}$, which is true for $|\gamma_k| < 1$. This is the condition used by the Schür-Cohn stability test to prove that the roots of a polynomial are inside the unit circle. Therefore, the forward prediction error filters of all orders must have their zeros inside the unit circle and as such they are minimum phase. A proper proof that does not utilize the stability test (which is in fact a consequence of the computation of the $\{\gamma_k\}$) may be found in a text by Orfanidis[140, pp. 230–233].

As a direct consequence, the backward prediction system is maximum phase. This is easy to demonstrate, since it has the same coefficients as the forward predictor, $A_k(z) = 1 + \sum_{i=1}^{k} a_i z^{-i}$, but in reverse order, that is $B(z) = z^{-k} + \sum_{i=1}^{k} a_i z^{i-k} = z^{-k} A(z^{-1})$. This has the roots of $A(z^{-1})$, which are the reciprocal of those for $A(z)$, which is minimum phase and has all its zeros inside the unit circle. Therefore the roots of $B(z)$ must all lie outside

the unit circle making it a maximum phase system.

Of great importance in adaptive signal processing is the fact that the backward prediction errors out of all stages are mutually orthogonal. To demonstrate this it is sufficient to recall that from equation (9.25) when R_{k+1} is applied on the weight vector $\mathbf{b}_k = J\mathbf{a}_k$ the resulting vector has zero elements apart from the kth element (which is equal to E_k). By appending enough zeros at the end of each vector $(\mathbf{b}_k^T 1)^T$ to make them $L+1$ dimensional and writing them in matrix form, B, it is easy to show that $R_{L+1}B = E$, where

$$E = \begin{bmatrix} E_0 & 0 & \cdots & 0 \\ & E_1 & \cdots & 0 \\ & \vdots & \ddots & 0 \\ & * & \cdots & E_L \end{bmatrix} \tag{9.32}$$

which is a lower-triangular matrix. The $*$'s indicate elements that are produced when the coefficient vector of order k is applied to the $(L-k)$ lower rows of R_L. Moreover, B is an upper triangular matrix and hence its transpose, B^T, is lower triangular. Pre-multiplying the above relation by B^T yields $B^T R_{L+1} B = B^T E$. The left hand side of this relation is symmetric, so must be the right hand side. But $B^T E$ is the product of two lower triangular matrices and it must be lower triangular also. Being symmetric, it must be diagonal and therefore all unspecified entries in the E are zero. Noting that the backward prediction error vector $\varepsilon^b = B^T\mathbf{x}$, the autocorrelation matrix for $\varepsilon_k^b(n)$ is:

$$E[\varepsilon^b(\varepsilon^b)^T] = E[B^T\mathbf{x}\mathbf{x}^T B] = B^T R_{L+1} B = E \tag{9.33}$$

which is a diagonal matrix. Hence the backward prediction errors are uncorrelated (orthogonal)[15, pp. 322–326][152, pp. 208–251]. This procedure is equivalent to the decomposition of the input signal vector into a set of mutually orthogonal signals (the backward prediction errors), which is equivalent to the *Gram-Schmidt orthogonalization method*[140, pp. 195–310]. This is the reason why the lattice structure is often used in adaptive signal processing, as a preprocessing element. The decoupling of the signal correlations is beneficial in the improvement of the speed of adaptation of a subsequent adaptive algorithm, especially when the input signal samples are highly correlated a fact highlighted in any text on adaptive signal processing[14, 152, 140]. A number of other properties, most of which are connected with the orthogonality of the input and the prediction error signals or between the forward and backward prediction errors, are listed by Proakis et al.[152, pp. 234–235].

Some of its characteristics make the lattice structure more suitable than other realizations in some adaptive signal processing applications. The absence of correlations between the outputs of every stage of the structure permit the adaptation of the corresponding reflection

coefficients independently, which is important if a parallel implementation is required[152, pp. 227–231]. The orthogonality of the backward prediction error signals makes the lattice structure an ideal preprocessing unit for the implementation of adaptive systems, as it removes correlations between adjacent samples, as stated earlier. Finally, because minimum phase systems have partial correlation coefficients whose magnitude is less than unity, the lattice realization provides an easy way to ensure the stability of the inverse of the system during adaptation by employing this test directly. This is essentially the Schür-Cohn test, described earlier, only the computation of the $\{\gamma_k\}$ from the coefficients of the direct realization is now redundant.

The decorrelation of the prediction error signals makes the lattice algorithm suitable for the implementation of recursive-least-squares adaptive algorithms, since the inversion of the (diagonal) autocorrelation matrix E is trivial. Hence matrix manipulations for the inversion of R_L (see section 9.1) are unnecessary. In fact, fast implementations of the RLS algorithm are usually based on the implicit use of the reflection coefficients[152, pp. 351–386][140, pp. 473–487]. These algorithms are, however, considerably more complex than the LMS (section 8.4.2) and the benefits from their fast convergence have been disputed, since for real EEG signals, where stationarity is debatable and ill-conditioning of the correlation matrix possible, the stability of a method that use estimates of correlation may also be questioned. Instead, an LMS-type gradient search algorithm was realized. This is still faster than the direct realization presented in section 8.4.2, but maintains a lot of the simplicity of the original algorithm, while avoiding the explicit use of correlations. Descriptions of this algorithm and some of its variants may be found in Orfanidis[140, pp. 444–454] and Proakis et al.[152, pp. 385–389].

The optimal partial correlation coefficients, $\gamma_1, \gamma_2, \ldots, \gamma_L$ are determined by minimization of the mean squared error of the prediction, that is the power of the forward predictor, E_L^f (see equation (9.24)). But since the partial correlation coefficients define the backward predictor, they should also minimize E_L^b, the power of the error from this system (equation (9.25)). It was shown earlier that these two are equal in the stationary case, but in reality they may be different, especially around a non-stationarity. Hence, a better performance criterion to minimize is the average sum of squares of the forward and backward errors:

$$\xi_L = E[e_L^f(n)^2 + e_L^b(n)^2] \tag{9.34}$$

which is equal to $E_L^f + E_L^b$ if the two error signals are independent. Using equations (9.23) that link γ_L to the inputs of the last stage of the lattice and differentiating ξ_L with respect

to γ_L gives the following relation:

$$\begin{aligned}
\frac{\partial \xi_L}{\partial \gamma_L} &= 2E[e_L^f(n)\frac{\partial e_L^f(n)}{\partial \gamma_L} + e_L^b(n)\frac{\partial e_L^b(n)}{\partial \gamma_L}] \\
&= 2E\left\{e_L^f(n)\frac{\partial[e_{L-1}^f(n) + \gamma_L e_{L-1}^b(n-1)]}{\partial \gamma_L} + e_L^b(n)\frac{\partial[e_{L-1}^b(n-1) + \gamma_L e_{L-1}^f(n)]}{\partial \gamma_L}\right\} \\
&= 2E\left\{[e_L^f(n)e_{L-1}^b(n-1)] + [e_L^b(n)e_{L-1}^f(n)]\right\} \qquad (9.35) \\
&= 2E\left\{[e_{L-1}^f(n) + \gamma_L e_{L-1}^b(n-1)]e_{L-1}^b(n-1) + [e_{L-1}^b(n-1) + \gamma_L e_{L-1}^f(n)]e_{L-1}^f(n)]\right\}
\end{aligned}$$

expanding the terms to which the expectation operator is applied, taking γ_L as a common factor and noting that $E\{\cdot\}$ is a linear operator:

$$\frac{\partial \xi_L}{\partial \gamma_L} = 2\left\{2E[e_{L-1}^f(n)e_{L-1}^b(n-1)] + \gamma_L E[e_{L-1}^b(n-1)^2 + e_{L-1}^f(n)^2]\right\} \qquad (9.36)$$

ξ_L is minimized when the above expression is zero, from which the partial correlation coefficient may be computed:

$$\gamma_L = -\frac{2E[e_{L-1}^f(n)e_{L-1}^b(n-1)]}{E[e_{L-1}^b(n-1)^2 + e_{L-1}^f(n)^2]} \qquad (9.37)$$

This procedure has been carried out on the last stage of the lattice, but the $\gamma_{L-1}, \gamma_{L-2}, \ldots, \gamma_1$ may be computed in an equivalent manner, since they only depend on the inputs of the corresponding stages and the prediction error powers for successive stages are mutually orthogonal, as demonstrated earlier. The latter means that the optimization of the kth stage may be based on the corresponding criterion $\xi_k = E[e_k^f(n)^2 + e_k^b(n)^2]$.

9.3.2 A gradient adaptive lattice algorithm

To make the computation adaptive, a gradient descent method has been employed, similar to the LMS for the direct realization:

$$\begin{aligned}
\gamma_k(n+1) &= \gamma_k(n) - \mu_k\frac{\partial \xi_k}{\partial \gamma_k} \\
&= \gamma_k(n) - 2\mu_k\left\{2E[e_{k-1}^f(n)e_{k-1}^b(n-1)] + \gamma_k(n)E[e_{k-1}^b(n-1)^2 + e_{k-1}^f(n)^2]\right\} \\
&= \gamma_k(n) - 2\mu_k[C_k + \gamma_k(n)D_k]
\end{aligned}$$

where the 'intermediate' variables $C_k = 2E[e_{k-1}^f(n)e_{k-1}^b(n-1)]$ and $D_k = E[e_{k-1}^b(n-1)^2 + e_{k-1}^f(n)^2]$ have been introduced to simplify the notation. Unlike the LMS algorithm, a different learning rate μ_k has been introduced for each stage of the lattice. The reasons for this will become obvious in the development of the algorithm that follows. As Orfanidis explained[140,

p. 445], making γ_k time-varying introduces a complication, since C_k and D_k become dependent on the partial correlation coefficients of the previous stages, $\gamma_1, \gamma_2, \ldots, \gamma_{k-1}$, which are also time-varying. Therefore the resulting equations are non-linear and unlike the LMS algorithm, analysis cannot be carried out exactly. Assuming, however that C_k and D_k have their asymptotic values, independent of time, the above recursive relation may be shown to be equivalent to the following geometric series, around the (optimal) value of the partial correlation coefficient, $\gamma_k = -\frac{C_k}{D_k}$, as given by equation (9.37):

$$\gamma_k(n+1) - \gamma_k = (1 - 2\mu_k D_k)^{n+1}(\gamma_k(0) - \gamma_k) \tag{9.38}$$

where C_k was eliminated by substituting it with $-\gamma_k D_k$. The above series converges to the optimal value when the geometric ratio $|1 - 2\mu_k D_k| < 1$. This is true when $0 < \mu_k < \frac{1}{D_k}$. Like with the LMS, using a normalized adaptation rate $0 < \lambda < 1$, the learning rate may be written as a fraction of the above range:

$$\mu_k = \frac{\lambda}{D_k} \tag{9.39}$$

Having determined μ_k, it is possible to construct the adaptive algorithm by substituting $\frac{\partial \xi_L}{\partial \gamma_L}$ from equation (9.35) into the gradient algorithm above using the above expression for μ to obtain:

$$\gamma_k(n+1) = \gamma_k(n) - 2\frac{\lambda}{D_k} E\left\{[e_L^f(n)e_{L-1}^b(n-1)] + [e_L^b(n)e_{L-1}^f(n)]\right\} \tag{9.40}$$

The gradient adaptive lattice algorithm is derived from this equation by ignoring the expectation operator, as with the LMS:

$$\gamma_k(n+1) = \gamma_k(n) - 2\frac{\lambda}{D_k(n)}[e_L^f(n)e_{L-1}^b(n-1) + e_L^b(n)e_{L-1}^f(n)] \tag{9.41}$$

In the above practical implementation, D_k is also computed recursively in the same way as the power of the input signal described earlier for the LMS algorithm in section 8.4.3:

$$D_k(n) = \alpha D_k(n-1) + (1-\alpha)[e_{k-1}^b(n-1)^2 + e_{k-1}^f(n)^2] \tag{9.42}$$

where, as before, α is linked to the length of the signal stationarity. It should be noted that a variant of this method[140, p. 447], uses a similar approximation for the term in square brackets (in equation (9.41)), instead of simply ignoring the expectation operator (which is equivalent to $\alpha = 0$).

9.3.3 Implementing the background activity prediction system, $H^{-1}(z)$ and the spike generating system, $G(z)$

The lattice structure is suitable for the implementation of linear prediction error systems, but some further considerations were necessary to ensure the suitability of this structure to

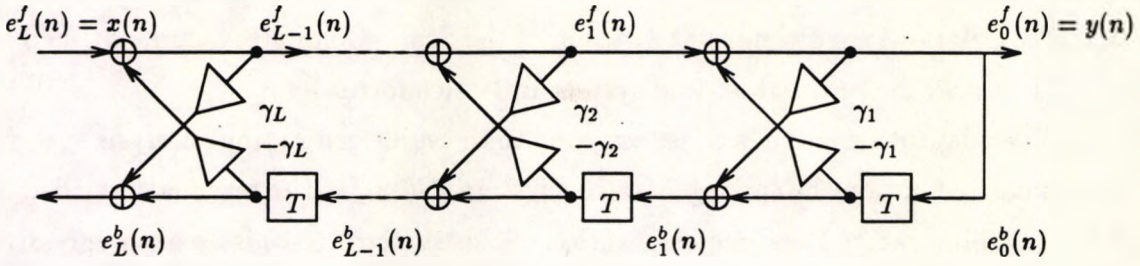


Figure 9.3: An IIR lattice system

EEG analysis according to the model suggested in section 7.3.

The first system to implement is the background activity analysis system $H^{-1}(z)$. As demonstrated in sections 7.4.1, 7.4.4 and 8.4.3 this system is naturally a linear predictor, being the inverse of the autoregressive system $H(z)$. This was implemented directly as an adaptive lattice system optimized according to the gradient algorithm suggested before.

The question arises for the spike generating system, $G(z)$. It was shown in section 8.4.4 that this is actually an IIR all-pole system. There is more than one issue connected with the lattice implementation of $G(z)$. First of all, is it not clear how this is linked to linear prediction. Then, there are two sides to its optimization, the off-line and the on-line constrained adaptive procedure.

It is not difficult to prove that the inverse of $G(z)$ has the structure of a linear predictor. Since $G(z)$ is an all-pole system, its inverse is an all-zero system with transfer function $G^{-1}(z) = a_0 + \sum_{i=1}^{L} a_i z^{-i}$. Multiplying the whole expression by $\frac{1}{a_0}$ to make the coefficient for z^0 equal to 1, the system has precisely the form of the prediction error system which is the basis of the lattice realization and may be implemented in lattice form. The scaling factor a_0, may be applied to the input of $G(z)$ and if made time-varying it may be adapted independently, using the LMS algorithm. These issues are related to $G^{-1}(z)$, but since $G(z)$ is also part of the analysis system (sections 7.4.2 and 7.4.4) it is important to ensure that it may also be implemented as a lattice structure. IIR lattice systems are discussed in most signal processing texts, like [151, pp. 495–506][152, pp. 235–239][15, pp. 236–240] and [140, pp. 225–230]. They have a structure similar to the FIR lattice, but with inputs and outputs reversed, as shown in Figure 9.3.

Having showed that $G(z)$ may be implemented as a lattice, the next issue is to investigate its optimization. Off-line optimization, described for the direct implementation in section 8.4.4, was performed with the aid of spiking filter banks. It would have been perfectly feasible to optimize a bank of lattice systems with the additional scaling factor a_0, but a simpler alternative approach was preferred. Direct form filters were constructed using a spiking filter

bank, as before. Then the coefficients of the optimal system were converted to the reflection (partial correlation) coefficients with the aid of the Schür-Cohn algorithm, outlined in section 8.4.4. This gave the optimal off-line system in lattice form.

On-line adaptation is the last issue related to the spike generating/analysis system. Adaptation is carried out on the spike generating system, $G(z)$, for the reasons described in section 8.4.4, while illustrating the direct realization of this system. $G(z)$ is an autoregressive system and has feedback, and hence any attempt for optimization of its parameters, $\{\gamma_k\}$ would result in a set on non-linear differential equations. Solution of these in closed form is generally impossible and numerical methods may be employed. In an adaptive system optimization is done numerically, but the convergence of the method is a critical issue. Gradient methods, cannot be guaranteed to converge to the optimal solution when dealing with non-linear differential equations, which may have many optimal solutions as well as multimodal performance surfaces. Provided that the algorithm is initialized with a value close to the optimal, a reasonable approximation to a linear system may be assumed, by using a Taylor expansion of the cost function around the optimal point or otherwise. Adaptive IIR lattice filters are not particularly useful, however, because they do not share any of the properties of their FIR counterpart. The independent optimization of the FIR lattice stages, is not a characteristic of the IIR lattice, because the outputs of every stage are not error signals. In fact the usefulness of an adaptive IIR lattice system is rather limited and no literature was found on the subject. A simple algorithm was developed, however, based on the difference, $\varepsilon(n)$, between the desired output $y(n)$ and the output of an Lth order lattice, $\hat{y}(n)$. The optimization criterion employed here is $\xi = E\{[y(n) - \hat{y}(n)]^2\}$, where $\hat{y}(n) = e_0^f(n) = e_1^f(n) + \gamma_1 e_0^b(n-1)$. Using the same procedures as before, for the development of the LMS algorithm, a steepest descent-type algorithm for optimizing γ_1 may be written:

$$\gamma_1(n+1) = \gamma_1(n) - \mu_1 \frac{\partial \xi}{\partial \gamma_1} \tag{9.43}$$

Using the approximation employed by Widrow for the development of the LMS, $\hat{\xi}(n) = \varepsilon(n)^2 = [y(n) - \hat{y}(n)]^2$ and differentiating with respect to γ_1 gives the approximation $\frac{\partial \xi}{\partial \gamma_1} = -2\varepsilon(n)e_0^b(n-1)$. For the rest of the γ_k, ignoring the coupling effect due to feedback equivalent updating equations may be derived. Hence the following updating equation has been developed and implemented:

$$\gamma_k(n+1) = \gamma_k(n) + 2\mu_k \varepsilon(n) e_{k-1}^b(n-1) \tag{9.44}$$

This equation may be rather inaccurate for general optimization, but for the restricted range of adaptation around the typical spike parameters the non-linearity and cross coupling have

a small effect on the otherwise small gradients and the above algorithm should suffice. Like before, μ_k, the learning rate for the kth stage, was normalized by assuming that it consists of a fraction λ of the inverse of the power in $e^b_{k-1}(n)$, computed recursively as suggested in section 8.4.3. Like in the case of the direct realization, this does not guarantee convergence, but in practice, with $\lambda = 0.1$, the method did not suffer from instability. Range limiting here was applied on all the γ_k, using the range limiting functions and procedures described before in section 8.4.4.

9.4 Comparison of algorithms and structures

9.4.1 Results for different configurations

The algorithms and structures described in the preceding section were incorporated in the system to replace either the direct form LMS prediction error systems associated with the background activity, $H^{-1}(z)$ or the spike generating system $G(z)$ and its inverse (sections 8.4.3 and 8.4.4).

The alternatives for $H^{-1}(z)$ include the RLS algorithm for the direct system implementation (section 9.1) and gradient LMS-type realizations for the cascade (section 9.2) and the lattice (section 9.3) structures. $G(z)$ was realized in the form of an IIR LMS system in direct form as in the original system described in section 8.4.4 and as a lattice IIR-type system developed in section 9.3.

Six separate configurations were tested, including the original system described before. The results for the three real EEG records on which all previous tests were carried out are shown in Table 9.1. All configurations implemented a prediction error system $H^{-1}(z)$ of order 16 and had a normalized learning rate of 0.1. Typical waveforms for a section of a real EEG signal are shown in Figure 9.4. These correspond to configuration 4 of table 9.1.

Cases 2 and 3 are worthy of some special attention, because the corresponding configurations exhibited anomalies in their behaviour. The RLS algorithm in 2 appeared to operate well at the beginning of the records, but after a while it became unstable, as the coefficients of the system drifted far from optimal with the error levels increasing continuously. Reducing the learning rate did not solve the problem. It merely delayed the occurrence of the phenomenon and made the increase in the error more gradual, but without preventing instability. The cascade LMS algorithm, used in the third tested configuration, suffered from a different problem. As time progressed, it seemed that $H^{-1}(z)$ started adapting to spikes. This was more profound in the Record A, which contained many spikes relatively close to one another. Despite preventing its adaptation when a spike was detected, which

Record			A		B		C		D	
Duration in seconds			210		130		123		242	
Number of spikes present			52		8		10		0	
Muscle artifacts present?			some		yes		yes		many	
Subsystem Implementation			Spike Detections							
#	$H^{-1}(z)$	$G(z)$	√	×	√	×	√	×	√	×
1	Direct-LMS	Direct IIR-LMS-C	52	2	8	2	10	6	0	25
2	Direct-RLS	Direct IIR-LMS-C	16†	0†	2†	1†	3†	3†	0	12†
3	Cascade-LMS	Direct IIR-LMS-C	42‡	0‡	8‡	1‡	9‡	4‡	0	30‡
4	Lattice-LMS	Direct IIR-LMS-C	52	2	8	2	9	6	0	22
5	Lattice-LMS	Lattice IIR-LMS-C	52	4	8	2	9	6	0	22
6	Direct-LMS	Lattice IIR-LMS-C	52	2	8	2	10	6	0	25

Notes:

†: These results correspond to part of the records considered (approximately a third), after which the system broke down.

‡: Missed detections increased towards the end of the data records.

Table 9.1: Results for real EEG records and various subsystem structures and adaptive algorithms detector

delayed the occurrence of the problem, this realization still absorbed some of the energy of the spikes making their detection more difficult. The coefficients of the last stages of the cascade acquired values that corresponded to poles in the region of the fundamental of the spike, where most energy of this transient was concentrated.

The more detailed interpretation of the various results is the subject of the following section.

9.4.2 Comments and interpretation of results

The results presented in Table 9.1 provide a means of comparison of the various algorithms and structures implemented. This was based on a number of criteria, which were considered important in this application. From the reported results, it became necessary to consider the reliability of the resulting system, as some adaptive algorithms developed some undesirable behaviour or even instability. Other important issues are the speed of adaptation (initial convergence and tracking properties), their response to artifacts and the possession

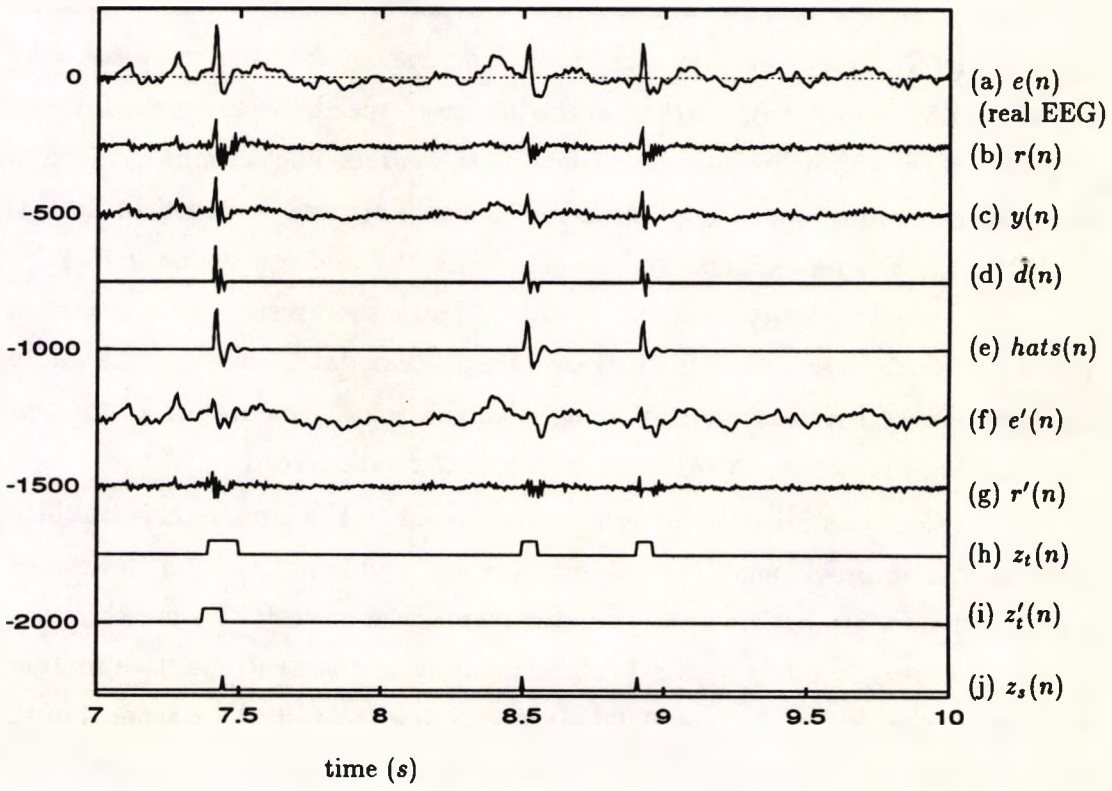


Figure 9.4: Waveforms for a segment of real EEG signal containing spikes

of additional desirable features.

Reliability and stability of adaptive procedures

The issue of *stability* was encountered when using the RLS (SER) algorithm to adapt the direct form implementation of the inverse of the background activity system $H^{-1}(z)$. Some implementations of the cascade LMS algorithm led to unreliable spike detection, but without becoming unstable. The other configurations had similar responses to one another and have generally been stable. Some minor differences will be accounted for at the end of this section, after considering the systems whose function exhibited serious irregularities.

In the RLS implementation of $H^{-1}(z)$, it appeared that the response was the same as that of the LMS algorithm, apart from the increased speed of convergence. Later on, though, instability of the algorithm became evident by the sudden fluctuations in the filter coefficients and a corresponding increase in the error levels. This was observed in both the synthetic and the real EEG records and the observed effects could not all be justified theoretically. In the first, the instability was transient and the method resumed its proper function after approximately 50 iterations. In the case of real EEG data, however, the fluctuation in the coefficients increased continuously along with the error. Restoration of the functionality of the algorithm did not occur in the duration of the data record.

The problem was initially thought to be related to the numerical instability of the technique, but no improvements were observed when double precision arithmetic was employed in the implementation. It was at this stage when the basic RLS algorithm was modified to include a convergence factor $\mu < \frac{1}{2}$, corresponding to the SER adaptive method. When the problem persisted, even for $\mu = 0.001$, it was decided to test other aspects of the algorithm. A modification was introduced to the time update formula for the correlation matrix described in most texts. This was incorporated in the development of the procedure, presented earlier on in this Chapter, but could easily pass unnoticed. Traditionally, the update relation for the correlation matrix has the form $\hat{R}(k) = \mathbf{x}(k)\mathbf{x}(k)^T + \alpha\hat{R}(k-1)$, which lacks the scaling factor $(1-\alpha)$ multiplied with the first term. The gain of this algorithm for constant signals is $\frac{1}{1-\alpha}$, which is large when α approaches unity, its limiting value. This lead to an overestimation of $\hat{R}(k)$. The introduction of the normalizing factor in the equations for the computation of the inverse of the correlation matrix ensured that the correction introduced on $\hat{R}(k-1)$ is not excessive and the new value, $\hat{R}(k)$ is more realistic. With this modification the system response improved slightly, as divergence was slowed down, but stability was not achieved.

Investigations of the results for artificial input signals revealed that the instability was

linked to a large transition in the transfer function $H(z)$, that is the boundary between two segments. Not all large transitions, however, caused instability, which appeared to be associated with some property of the transition, perhaps its direction. Arguably, the presence of a non-stationarity has not been the only factor affecting the behaviour of the algorithm. It is more likely that it is one of the main contributing factors. Others might include the sensitivity of the algorithm to a numerical condition which takes effect when the matrix $R^{-1}(k)$ has a value that becomes ill-conditioned by the introduction of the new input data. The effect may be related to the numerical instability of some fast RLS algorithms highlighted by Ljung and Ljung[111].

In the case of the cascade realization the algorithm did not become unstable, but it is apparent from the results that some spikes were missed. Closer observation of the progress of the spike detection procedure revealed that the missed detections occurred towards the end of the records and were more profound in Record A, which had longer duration and more spikes than B, C and D. The missed detections were related to the progressive reduction in the amplitude of the transient in the output of the primary prediction error filter, $r(n)$, which fell below the detection threshold and did not produce a transient indication. The source of this was traced to the initial coefficient vector. Starting with all elements of the vector set to zero results in root coalescence (section 9.2), because all stages are initialized to the same values and subsequent attempts to modify these elements would result in the same modification to all stages, which would henceforth have the same parameters. To prevent this, the initial weight vector was initialized to some small vector. Small random values could work, but it is also possible to construct each section $k = 1, 2, \ldots, \frac{L}{2}$ to have complex conjugate poles $p_{k,1,2} = re^{\pm j\frac{2\pi k}{L}}$, where $0 < r \ll 1$. This spreads the poles of the system evenly on the circumference of a circle with radius r, centred at the origin. The results reported in Table 9.1 correspond to the latter case, but similar results were observed for many random vector initializations. It seems that because a number of poles of the initial system lie near the region corresponding to epileptic spikes, some of the energy of the spikes is absorbed by one or more stages of the cascade. If the order of the system is less than or equal to that of the generating process for the background activity, then eventually these poles will move to model some of its aspects and the problem will not occur. But if the order of the system is greater than required, then the last few stages in the cascade will be fed with essentially white noise and transients, due to spikes etc. It is therefore possible for them to lock-on spikes, in the absence of any other correlated signal. This was confirmed by testing with artificial data, where the order of the system was known. The use of extra stages resulted in their attraction to the occasional occurrence of spikes, thus depleting their registration in

the output. The effects were less profound with some random initial coefficient vectors, or with the elimination of the vectors corresponding to the spike region. This, however, does not constitute a solution to the problem, because it is possible during adaptation for some of the coefficients to move to this region and for the undesirable effect to occur.

In the initial implementation of the lattice prediction error system, the lattice system missed some spikes[128]. The source of this discrepancy appeared to lie in the speed of adaptation of the gradient lattice structure which is significantly greater than that for the direct LMS algorithm. Therefore, when a spike is encountered, the algorithm attempts to adapt the system to track this signal. The direct form LMS, being slower, only causes a small change to the coefficients of the system, before the transient is detected and adaptation inhibited (see section 7.4.4). In the case of the lattice, the change in the coefficients is significant even for the few iterations it takes until the transient is detected. On a few occasions, the coefficients of the lattice structure adapted to the spike characteristics after a sample or two, thus making the detection of the transient in $r(n)$ weak, or too early. As a consequence the amplitude of the spike generating (im)pulse, $\hat{d}(n)$ was either too low or too high and its moment of occurrence premature. The resulting spike $\hat{s}(n)$ was therefore a bad estimate and its subtraction from $e'(n)$ did not make the transient disappear, thus the spike was rejected as a false alarm by the spike detector. This phenomenon, is only temporary and the coefficients of the lattice soon resume the values corresponding to the background activity. To improve the detection abilities of the system, the adaptation rate λ could be reduced at the expense of convergence speed. A better approach which does not make this compromise will be described here. The window of the transient detector was extended so that Q in equation (8.72) was increased to a value of 2. Since the detection of a transient in $r(n)$, inhibits adaptation, the system stops adapting a few samples before the spike is encountered and starts again some samples after the transient has effectively ceased. Hence adaptation during transients is prevented and spikes are properly detected.

Finally, it is worth noting that in the case of $G(z)$ both the direct and the lattice LMS algorithms for IIR systems performed in a similar way. In the general case there could have been differences, but these were possibly unnoticed because adaptation here was confined to a small set of points around the 'optimal'.

Speed of convergence

Although the slow convergence of the direct form LMS algorithm was one of the factors that initiated the use of other structures and algorithms, it is not a very critical issue[130], since for the existing records the LMS algorithm performed well. But it may be argued that the

test set of EEG signals was rather small and that in practice it is possible to encounter some EEGs containing abrupt transitions between one set of parameters and another, for which faster adaptation might be necessary for successful transfer function tracking. The speed of adaptation for the algorithms tested will now be examined.

Before it became unstable, the RLS algorithm appeared to converge much faster than the LMS. This was particularly noticed at the beginning of the record where it took less than 10 iterations for the RLS to 'level off' to the minimum MSE, whereas the direct form LMS required approximately 150 iterations.

The convergence characteristics of the cascade LMS structure were not much better than the original direct form LMS, requiring over 100 iterations before the MSE approached its limiting value.

Finally, the LMS lattice structure took about 20 iterations to reach close to the minimum MSE, although it is a steepest descent-type algorithm and should behave similarly to the LMS for the direct structure, as Widrow and Stearns reported[202, pp. 177–181]. The difference lies in the use of a normalized adaptation rate throughout the stages of the lattice, which is modified by the power at every stage to produce the effective learning rate (section 9.3). This results in different learning rates for every stage and bears a relation to the RLS algorithm, which uses the inverse of the input autocorrelation matrix to provide a means of different adaptation for every coefficient.

Response to artifacts

Although artifacts were taken into consideration during the development of the system (section 7.4.4), the adaptive implementation did not concentrate on an artifact tolerant system. As different sources of artifacts have different properties, they may be classified into either the transient category, or the background activity, depending on their properties and duration. The exponentially decaying transients that were introduced into the artificial records for testing the system belong to the first category. Muscle artifacts are a primary concern, because they often possess spectral properties similar to spikes.

The results for records C and D in Table 9.1 indicate that although artifacts caused a number of false detections, there was some degree of artifact rejection in all system configurations tested. Excluding the pathological conditions of the RLS and the cascade realizations, only some of the artifacts were detected as spikes by the system. Most of these were encountered during isolated muscle spikes or at the beginning of artifact bursts. Their numbers were influenced by the implementation of $H^{-1}(z)$, whereas that for $G(z)$ did not have any significant effects. The direct LMS method that initially appeared superior in all

respects was outperformed by the Lattice LMS with the modification described in the previous section. Both algorithms responded similarly to isolated artifact spikes (causing false detections). The faster speed of adaptation of the lattice system proved advantageous when artifact bursts were encountered. The first spike of the burst was falsely detected, but as adaptation progressed, the artifact was modelled by $H^{-1}(z)$, its energy was greatly absorbed by this system, with only an occasional false detection or two during the burst. For the direct realization, response was similar, but slow adaptation caused more false detections at the onset of the muscle spike bursts.

Although these systems might misinterpret artifacts as epileptic spikes, it was decided that artifact rejection could be performed at a later stage, or even off-line, by manual inspection of the reconstructed records or even by automatic investigation to reject repetitive detections in segments where $H(z)$ represented high power in the spectral region of the spike (see also section 9.5 and main Conclusions).

Other properties of alternative system realizations

Some of the adaptive systems tested possess secondary but useful properties that are absent from the the direct form LMS adaptive filter originally employed.

The RLS algorithm offers the advantage of fast convergence but little else, because it utilizes a direct form realization, like the initial LMS system.

The cascade system consists of second order sections and hence its zeros are easily computed, because the transfer function of each section can be readily factored. It is then easy to determine whether all the zeros lie inside the unit circle in z (minimum phase property) which ensures the stability of the inverse system, $H(z)$. Moreover, knowing the poles of $H(z)$ which coincide with the zeros of the inverse system, is a useful form of parametrization. The location of the poles is related to resonant frequencies[15, pp. 152–170] and hence to principal frequency components in the signal. In an on-line system this may provide a way to describe collectively the main features of the background activity in a way that is economical to store for off-line retrieval, as a small set of spectral parameters. It is then possible to 'reconstruct' the EEG signal so that its main features may be visually inspected. Exact reconstruction may not be possible, but manual interpretation rarely depends on the details of the EEG signal, which is usually described in very general frequency terms (α, β, δ and θ activities, for instance).

The lattice structure does not offer simple factorization methods of the transfer function it implements, although an algorithm for finding the poles of an IIR lattice structure was presented by Jones and Steinhardt[100]. Nonetheless, it provides a simple means for testing

the location of the zeros of the prediction error system, $H^{-1}(z)$ as well as the poles of $G(z)$. This property makes the lattice realization appealing for the implementation of both systems. The ability to test the location of the poles of the spike generating system $G(z)$ is of particular importance, because it provides a simple way to control their range with the algorithms described in section 8.4.4.

9.5 General comments and observations

In this Chapter, a number of alternative implementations of the 'linear' parts of the spike detector proposed in Chapters 7 and 8 have been investigated. All realizations have been in the general theme of adaptive linear prediction and filtering. It was demonstrated, however, that structures and algorithms like the lattice, outperformed in some aspects the direct-LMS realization, whereas others, the direct RLS and the cascade realizations did not perform very well.

Irrespective of their results, the possibility to change parts of a system, while maintaining the rest unchanged, even at the conceptual level, has shown how versatile a modular system can be. It may be possible to replace the linear systems with some non-linear ones, when the theory behind the latter becomes well-documented. This future prospect is not shared by the earlier detectors of Chapter 6, which were also quite reliable, but depend on specialized elements selected for this particular application.

The approach suggested here produces as part of the analysis a number of secondary signals, like a spike-free EEG estimate and an estimate of the spike signals as well as a number of parameters, which may be used in further analysis, either visual or automatic. There is also a possibility for its application to other signals, not necessarily biological. These are explained in more detail in the Conclusions of this Thesis.

As far as EEG analysis is concerned, the suggested system may be generalized to detect other transients, not only spikes. This would require the extension of the system by adding modelling systems for the transients considered in much the same way as with the spike model, $G(z)$, as described in section 7.4.2. The decision elements (sections 7.4.4 and 8.5) will also need to be modified to include transient detection as well as some means of discriminating between the types of transients. This extension to the original system will be addressed in the next two Chapters (10 and 11).

Chapter 10

Generalization of the analysis procedure

10.1 Introduction

Although the main objective of this work was the detection of epileptic spikes, the presence of other transients cannot be ignored. Their majority consists of signals of non-cerebral origin (artifacts) and these spontaneously occurring signals contaminate the EEG signal. Many are rejected by the spike detection procedure of section 7.4.4, but determining the source of each one has advantages. Verifying the origin of a transient provides an additional confirmation that it did not constitute a spike that was missed. It also permits its estimation and hence its removal from the EEG. Therefore, the scope of the analysis procedure, proposed in Chapter 7, may be extended to traditional clinical, as well as ambulatory monitoring, as it may provide a means of *artifact correction.* This is a useful feature, especially when dealing with nervous or uncooperative patients, whose EEG may contain a plethora of artifacts.

It should also be reminded that the analysis method proposed in the previous Chapters has not been developed specifically for the EEG, but as a more general procedure to analyze signals which contain a mixture of transients superimposed on a signal with infinite time extent. In those cases, there may be more than one transient of interest and the ability of the system to detect all of them is crucial. Relating this to the EEG, it would be desired to discriminate spikes from sharp waves, which have longer durations, or even between subclasses of spikes as well as other features. These possibilities have not been tested, because of the unavailability of records containing a variety of transients, but the extension of the system is a possibility deserving some attention from both the signal analysis theoretical point of view as well as from the neurophysiological perspective.

In this Chapter the extension of the analysis procedure to detect non-spike transients is studied and the necessary modifications to the existing structure and the addition of new elements are covered. Finally, results for artificial and real EEG records are presented.

10.2 Transients

Epileptic spikes are not the only transients which have their origin in the cortex. In section 2.5.2, other patterns were listed, not all of which are transients according to the signal modelling definition of section 7.3.3. Sharp waves are similar to spikes, apart from their longer durations, and are characteristic of this group of signals. Isolated slow waves may also be regarded as transients. *k*-complexes and vertex sharp waves also fall in this category, though not related to pathological conditions. An extended system would ideally attempt to detect these transients as well as combinations of them occurring as complexes (spike-and-slow-wave, sharp-wave-and-slow-wave, polyspikes-and-wave, etc.).

The second category includes transients whose source is not located in the brain and their presence in the EEG is normally undesired. Some of these artifacts (section 2.5.4) originate from potentials accumulated on the eyes and appear as discharges in the EEG associated with eye movements, opening and closure, blinks and eyelid fluttering. Another category of transients with a definite shape originates from the heart. Pulse artifacts are registered in the EEG when an electrode is placed close to an artery, but recording of the cardiac rhythms directly, in the form of QRS complexes, is also possible. One of the most profound sources of interference in the EEG is muscle potentials. Artifacts due to muscle activity, although having 'spiky' appearance do not have a well-defined shape and often persist for considerable periods of time. Hence they are generally difficult to model and eliminate.

Finally, a number of extracerebral potentials originate in the recording equipment and the electrodes. Electrode movements, dry contacts, changes in their resistance because of changes in the electrolyte solution that forms the contact with the skin (section 2.5.4) are among the commonest. These sometimes have characteristic shapes and may be detected and subsequently removed.

Not all mentioned transients fit in the signal model (section 7.3.3). Muscle artifacts, for instance, which are the most troublesome in spike detection, generally do not have specific forms and often have long durations. Those that do conform to the general definition of transients affect the analysis system (section 7.4.4) at various levels of the detection procedure, which are now considered.

Transients, like spikes, are registered as disturbances of large amplitude in the output,

$r(n)$, of the background activity prediction error system, $H^{-1}(z)$. It is not possible to discriminate whether the origin of the disturbance was due to a spike or another transient, nor is it possible to differentiate between the various types of transients based on this indication alone.

For this reason, the inverse spike model was introduced, so as to provide the signal appropriate for the generation of an estimate of a spike every time such an event is detected in $r(n)$. The estimated spike is then subtracted from the recorded EEG signal and the absence of a transient verifies that the origin of the disturbance was indeed a spike. If another transient had occurred, then the subtracted spike would not have removed it, since this has a different morphology and the transient would still be registered in the spike-free signal, $e'(n)$. This procedure is capable of detecting spikes, but is unable to produce a transient-free signal, $e'(n)$, when non-spike transients are present. Moreover, if a transient has morphology similar to spikes, it may be detected as one. In the EEGs studied this did not occur, apart from some muscle artifacts. The detection and elimination of other transients to produce a 'clean' EEG signal has been the primary source of motivation for the extended system described next. The similarity in the morphology between transients has not been neglected, because the possible extension of the range of application of the system to other signals was one of the targets during its development. Perhaps in these signals there are many transients of interest, some similar to each other, all of which have to be detected.

In the sections that follow, a method for the extension the already developed system is proposed. Although the amendments were primarily concerned with the detection of one transient, apart from spikes, the procedure is easily generalized for any number of transients.

10.3 Extending the system to detect other transients

The analysis procedure used for spike detection depends heavily on inverse modelling. A prerequisite for its generalization is a model for the transients considered, from which some inverse may be derived and included as part of the system.

10.3.1 A reminder of the model of transients

A model for non-spike transients has been constructed and discussed in section 7.3.3. The model here is concerned with N transient types, each described as a signal component, $[t_1(n), t_2(n), \ldots, t_N(n)]$. Each one of these, $t_i(n)$, was decomposed into a system, $F_i(z)$ and an excitation sequence, $c_i(n)$. The $\{F_i(z)\}$ could have any structure, but for the purposes of analysis and estimation they were considered to be all-pole transfer functions of order

L, as with the equivalent spike model, $G(z)$. All the excitation signals, $\{c_i(n)\}$ consisted of impulses occurring at random instants and having different amplitudes. Durations between consecutive impulses were considered to be an exponential random process, whereas amplitudes were assumed to follow a Normal distribution around a mean value. For some transients having sometimes positive and sometimes negative polarity, these were made to have equal probability of occurrence.

Because spikes are just one of many transients in the extended system, they may be included in the model as the zeroth transient, $t_0(n) \equiv s(n)$ where $F_0(z) \equiv G(z)$ and $c_0(n) \equiv d(n)$.

10.3.2 Elements of inverse modeling

Based on the model given above, an analysis procedure may be derived. The procedure is exactly equivalent to the one already described for spikes (section 7.4.2) and will not be discussed in detail here. In general, the application of the inverse of $F_i(z)$ on the recorded EEG signal $e(n) = b(n) + \sum_{i=0}^{N} t_i(n) + v(n)$ would produce a signal $y_i(n)$ containing the generating impulse sequence $c_i(n)$ of the ith transient, a distorted signal because of the background activity, $b(n)$, noise due to $v(n)$ and, in the presence of any of the other transients, a corrupted version of them. As explained for the spikes in section 7.4.2 the additive signals may be large, rendering the reliable detection of $c_i(n)$ impossible.

Of course, the above procedure requires the exact knowledge of the $F_i(z)$ or its inverse, $F_i^{-1}(z)$. These systems are generally unknown, but their off-line estimation is possible using existing records containing the transients and the method for estimation of optimal spiking filters, described in section 8.4.4. On-line optimization, to allow for small deviations of the individual occurrences of the transients from the off-line model, may also be introduced in the by constrained on-line adaptation of the $F_i(z)$, as with $G(z)$ in section 8.4.4. Both the direct and the inverse of the $F_i(z)$ for all transients considered are required for the implementation of the system, which is described next.

10.3.3 Extension of the analysis procedure

The generalized transient detection system is an extended form of the spike detection procedure, described in section 7.4.4. The most apparent difference is the introduction of an array of transfer functions in the place of the spike modelling system. Like the earlier system, both the inverses, $F_0^{-1}(z), F_1^{-1}(z), \ldots, F_N^{-1}(z)$ of the modelling systems and the direct transfer functions are present in the system, as demonstrated in Figure 10.1.

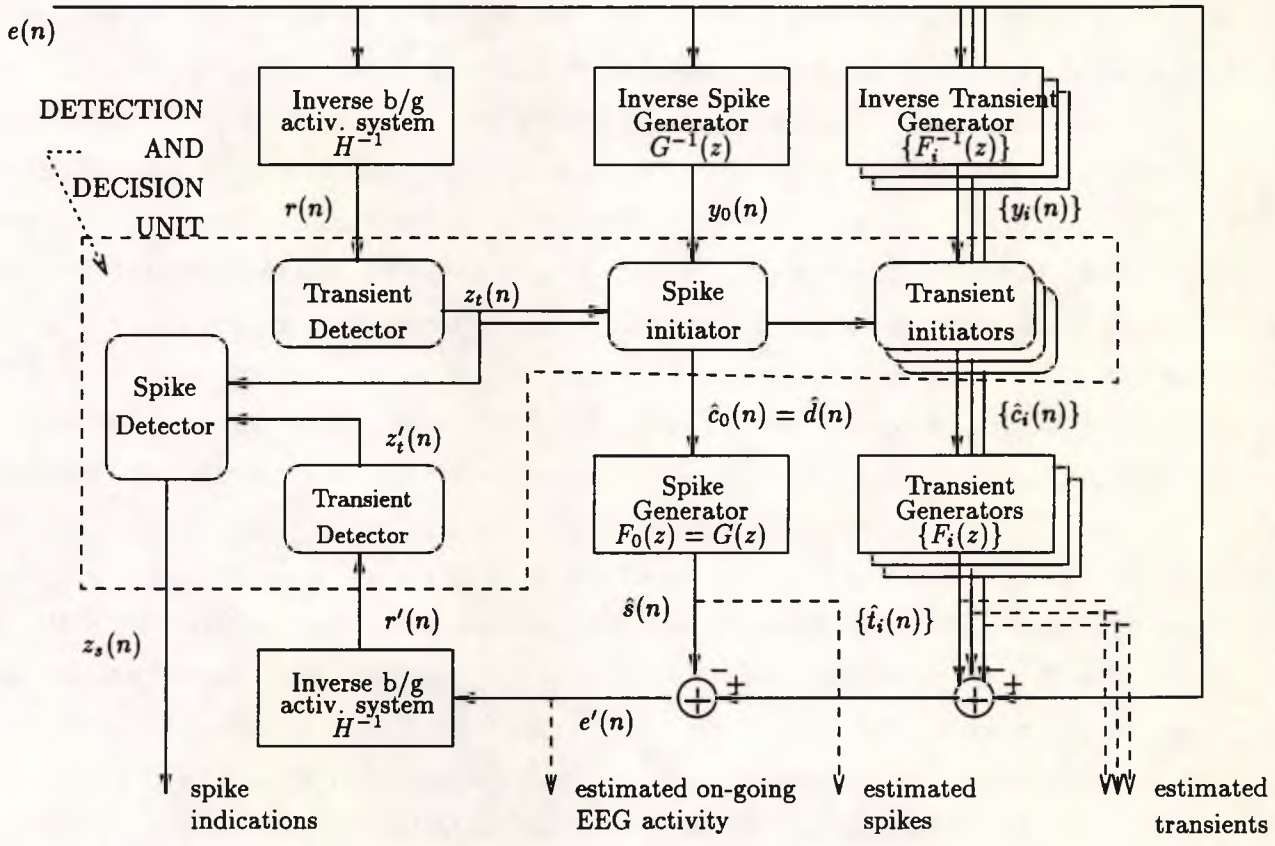


Figure 10.1: Diagram of the extended analysis system

Modifications have also been introduced in the decision and detection unit, since the single spike initiator had to be augmented by N transient initiators. The rôle of these should be evident from the previous description in section 7.4.4. A brief explanation is, however, necessary to clarify some points. This will concentrate on the signals present at the output of the prediction error systems, $H^{-1}(z)$, $r(n)$ and $r'(n)$ and the outputs of the inverse transient modelling systems, $y_0(n), y_1(n), \ldots, y_N(n)$. These define the function of the decision and detection unit under different conditions.

Firstly, when no transient is present in $e(n)$, then $r(n)$ contains only the on-going signals, due to the background activity and the noise. Therefore the transient detector operating on $r(n)$ will not detect any signal and hence the remaining systems will be inactive.

If one of the modelled transients (including epileptic spikes) occurs, this will be registered in $r(n)$ as a disturbance whose shape will depend on the type of transient and the condition of $H^{-1}(z)$, which is continuously updated. This shape, as mentioned in section 7.4.4 cannot be predicted, but it should contain components mainly in the higher end of the spectrum, due to the high-pass action of the inverse of the (generally) low-pass system $H(z)$. The transient disturbance, however, has large amplitudes that are apparent in the 'noisy' appearance of the on-going signal after decorrelation. These are found by the transient detector ($z_t(n) = true$). The type of transient cannot be determined based on this signal alone. Because, however, the shapes of the transients (which are defined by their modelling transfer functions) are different, the outputs of the $\{F_i^{-1}(z)\}$ may be used to detect which transient had occurred. The output of the system corresponding to the transient taking place, $y_j(n)$, should contain an impulse. It is intuitive that this is not possible for the remaining systems, because their corresponding transfer functions are different by definition. These outputs play a secondary rôle in the detection procedure, because they are contaminated by large signals, making the detection unreliable. Once, however a transient is hinted by $z_t(n)$, it may be possible to find which of these contained an impulse signal. The combined detection is an output from the system. There are N such outputs, one for every transient, $z_0(n), z_1(n), \ldots, z_N(n)$. Normally one of these, $z_j(n)$ will be *true*, indicating that the jth transient was the one that had occurred. This signal is used as a mask on the corresponding $y_j(n)$, to derive an excitation impulse-like sequence, $\hat{c}_j(n)$ which will be fed in the corresponding transfer function, $F_j(z)$ to generate a model of the transient that presumably occurred. In fact a correction, $\Delta_j(n)$ will be added on the corresponding $y_j(n)$, as explained in section 7.4.3:

$$\hat{c}_j(n) = z_j(n)[y_j(n) + \Delta_j(n)] \tag{10.1}$$

Strictly speaking, this is performed by a separate element of the system, the spike (or transient) initiator. One such unit is required for every modelled transient, whose simple

function has already been described in section 8.5.2.

The model of the transient, $\hat{t}_j(n)$ will then be subtracted from the recorded signal, $e(n)$ to produce $e'(n)$, as before. By repeating the transient detection procedure on $e'(n)$ and observing the absence of a disturbance in its output $r'(n)$, the detection of the transient that occurred is verified. In the present system, where only spikes are of interest, there is need for a single output, $z_s(n)$, indicating the presence or absence of spikes.

Finally, a transient which is not included in the list of those modelled will still produce a detection in $z_t(n)$. It is possible, depending on the condition of the $\{y_i(n)\}$ for a detection to appear in one of the $\{z_i(n)\}$, especially if the transient bears some resemblance to one of the modelled signals. If this occurs, then an estimate of one of the $\{\hat{t}_i(n)\}$ will be generated by mistake. In this case its subtraction from $e(n)$ will not remove the transient, which will cause a disturbance in $r'(n)$ and the incident will be correctly identified by the detection logic (no transient will be signalled in $z_s(n)$). Obviously, if none of the $\{F_i(z)\}$ is excited, there will still be a transient in $e'(n)$ (since none was subtracted) and the detection logic will arrive at the same conclusion.

Evidently, the detection and decision logic in the generalized system is considerably more involved and the corresponding system more complex than the one required for spike detection.

10.4 Modification of the decision and detection unit

The function of the system described in the previous section bears some resemblance to the simple unit required for spike detection, described in section 8.5. In the earlier system, it was not necessary to discriminate between types of transients, hence the only inputs to the detection unit were the outputs of the two prediction error filters, $r(n)$ and $r'(n)$. There were also two outputs, a transient detector, used as a spike initiation signal, $z_t(n)$, and a spike detection signal, $z_s(n)$, indicating the final decision whether a spike had occurred at the nth sampling instant.

10.4.1 Additional inputs and outputs

The generalized system must discriminate and generate initiation signals, $z_0(n), z_1(n), \ldots, z_N(n)$, for any modelled transient. To do this, the outputs of the inverse filters of all the modelled transients, $y_0(n), y_1(n), \ldots, y_N(n)$ must be inspected, as shown in Figure 10.2.

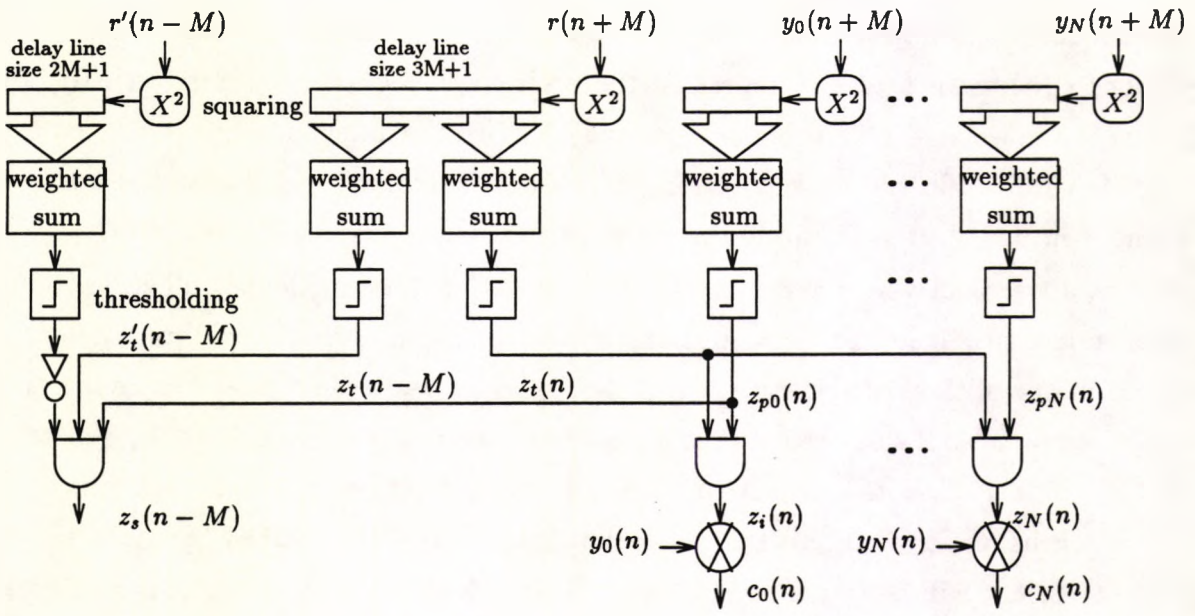


Figure 10.2: The extended detection unit

The detection-decision unit is, as before, a combination of 'continuous' preprocessing functions and logical operations to produce the (effectively binary) outputs.

10.4.2 Preprocessing of the inputs

Preprocessing of $r(n)$ and $r'(n)$ could be maintained as in the earlier system of section 8.5. This consisted of squaring the 'current sample', that is the one in the middle of a window of length $2M+1$, normalizing it by a variance estimate, based on the samples in the window excluding samples around its centre.

It was decided to express this in terms of the information contained in individual samples using a weighted sum-of-squares of samples in the window. This approach permitted the direct comparison with a later development (section 11.5).

Transients need to be detected in the nth sample value that occupies the middle of the window for the signal under consideration ($r(n)$ or $r'(n)$). It is however more effective to allow the detection to 'spread' over a small group of samples, $2Q+1$ before and after that time instant. It acts as a window, when there is a small time difference between the appearance of the transient in $r(n)$ and the excitation impulse in $y_i(n)$ (see section 8.5.2). This may be the result of the difference in the phase characteristics between the $\{F_i^{-1}(z)\}$ and the $H^{-1}(z)$. To determine suitable weights, the samples in the window were classified in two types. The central $2Q+1$ contribute to the sum increasing it, whereas the $2(M-Q)$

samples in the 'tails' of the window have a decreasing effect. Thus, if the central samples are atypically large, compared to the rest, the sum should be positive, otherwise negative. The latter should also occur if there are large amplitudes far from the centre of the window, although this cannot be guaranteed in all such cases. It is evident that the weights in the centre of the window should be positive, whereas the rest should be negative. For simplicity, all positive weights were made to have the same value, w^+, and the negative ones similarly have the value w^-. Hence the weighting procedure is defined by:

$$f(n) = -\sum_{j=-M}^{-Q-1} w^- x^2(n+j) + \sum_{i=-Q}^{Q} w^+ x^2(n+i) - \sum_{j=M}^{Q+1} w^- x^2(n+j) \qquad (10.2)$$

where $x(n)$ is the input under consideration ($r(n)$, $r'(n)$, etc.). Suitable values for w^+ and $w-$ were determined by considering that the summations of the off-centre samples are approximated by a multiple of the variance, σ^2 of the input signal with all transients removed. Their total value would then be $\Sigma^- = 2w^-(M-Q)\sigma^2$. Hence, for any one of the $2Q+1$ central samples to make $f(n)$ positive, its squared amplitude, weighted by w^+ must be greater than Σ^-. If a detection must occur when $x^2(n+i) > t_0^2\sigma^2$, as suggested in the original system (Equation (8.73)), then, for the two detection schemes to be equivalent, $f(n) = w^+x^2(n+i) - w^+t_0^2\sigma^2$. By comparison with the definition of $f(n)$, $w^+t_0^2\sigma^2 = 2w^-(M-Q)\sigma^2$, giving the ratio:

$$\frac{w^+}{w^-} = \frac{2(M-Q)}{t_0^2} \qquad (10.3)$$

By setting $w^- = 1$, and using the parameters of the system ($M = 32$, $Q = 2$ and $t_0^2 = 6$) the remaining parameter is $w^+ = 10$.

The required binary indication of the occurrence of a transient, $z(n)$ is derived from the linear function $f(n)$ by thresholding, so that $z(n) = 1$ if $f(n) > 0$ and $z(n) = 0$ otherwise. Two such weighted sum and threshold units were constructed to implement the two transient indicator procedures, one for each of the signals $r(n)$ and $r'(n)$. These are labelled in Figure 10.2 as $z_t(n)$ and $z_t'(n)$.

With similar considerations detecting impulses in the outputs of the transient modelling signals could be performed by similar units, one for each signal, $y_i(n)$. Because the power of the extraneous components due to the background activity and the noise in these signals is higher than the equivalent in $r(n)$, the detection of the impulse-like components in this signal is more difficult and can only be achieved for a lower confidence limit, $t_0 = 1.65$ assuming that the distribution of the signal is Gaussian with zero mean and a 10% probability of detecting a transient when there is none (see also section 7.4.3). This is a consequence of the reduced reliability of these signals when used alone for the detection of transients (section 7.4.4). The

corresponding binary outputs, $z_{pi}(n)$, of the preprocessing units for each transient (Figure 10.2) are only intermediate and the output detection and activation signals are formed by logical operations on these as well as $z_t(n)$ and $z'_t(n)$.

10.4.3 Binary logic for the production of outputs

In the general system there are $N+1$ transient initiation outputs, $z_0(n), z_1(n), \ldots, z_N(n)$ one for every transient type including spikes. There is also a separate indication of the final spike detection, $z_s(n)$. All these signals may be generated by logical combinations of the outputs of the preprocessing units.

Evidently, $z_i(n)$ is active ($true$) if there is a transient ($z_t(n) = true$) and there is an indication of the transient in z_{pi}, at the same time. This is a logical AND operation, since both $z_t(n)$ and $z_i(n)$ must be $true$ simultaneously. Similarly, the output of the final spike detector is active only when $z_t(n) = true$ (a transient is present in the input) and $z'_t(n) = false$ (it was removed by the subtraction of the modelled transient) and additionally if the removed transient had been a spike ($z_{p0}(n)$ was active). Because it is not possible to compute $z'_t(n)$, due to the absence of values of the signal $r'(n+i)$ for $i > 0$, spikes are detected with a delay of $n - 2M$. This has an intuitive interpretation. If a transient is initiated at the nth sampling instant, it is not possible to detect whether its main components (which *follow* the initiation) have reduced the prediction error after they have been subtracted from the input EEG signal. Consequently, it is necessary to wait for M samples before attempting to observe whether a transient had been removed or not.

As a consequence it is necessary to save at least M 'future' values of the sequences $z_t(n)$, $z'_t(n)$ and $z_{p0}(n)$ in order to perform the comparison. Alternatively, additional preprocessing units resembling the ones that produced each one of the required signals may be used to compute the corresponding values with a delay of M samples. These have essentially the same weight distribution as the former, but they are centred around the $(n-M)$th input sample. The approach is more computationally intensive, but it has been preferred because it maintains the 'combinatorial', memoryless, structure of the detection-decision unit, which fitted well with a different realization described in the next Chapter. The extra element for this implementation, has been depicted in Figure 10.2, labelled by its output $z_t(n-M)$. The same weight distribution is actually used to generate $z'_t(n-M)$ and $z_{p0}(n-M)$.

It should be noted that the logical operations $z_i(n) = z_t(n)\text{AND}z_{pi}$ as well as $z_s(n) = z_t(n-K)\text{AND}z_{p0}(n-K)\text{ANDNOT}z'_t(n-K)$ may also be implemented as weighted sums with thresholding, assuming that the $true$ or $active$ has the numerical value 1 and its complement the value 0. Equations implementing these operations are be presented in section 11.5.1.

According to this scheme, for the detection of N transients, plus spikes, it is necessary to generate N outputs (one per transient, $z_i(n)$) plus one for spike activation, $z_0(n)$ and an additional spike detection signal, $z_s(n-M)$, that is a total of $N+2$ outputs. These require an equal number of logical operations, as described in the previous paragraphs, and $N+5$ preprocessing units, out of which $N+2$ produce outputs based on the nth sample of their inputs and the remaining 3 with a delay of M samples.

10.5 Results and discussion

The proposed extended analysis method was constructed for the detection of spikes and one more transient, having the form of an exponential decay with scaled amplitude. This signal does not correspond to any particular transient, but resembled some artifacts due to eye movements, as they are registered in the EEG signal or even some electrode movement artifacts (see Appendix D). These were artificially generated, because in the four available EEG records epileptic spikes were the only transients in abundance.

The decision-detection unit utilized a window of samples from the four input signals, $r(n)$, $r'(n)$, $y_0(n)$ and $y_i(n)$. Sub-windows of 65 samples each, were used in each one of the preprocessing units. Therefore the parameters of the system, as described in section 10.4.2 were $M = 32$ and $Q = 2$. The delay for the detection of spikes was, therefore 64 samples. This combination of parameters was a good compromise between accuracy and computational requirements. There were $N+5 = 6$ preprocessing units and $N+2 = 3$ binary units, a number equal to the outputs (spike detections, $z_s(n-K)$, spike activations, $z_0(n)$ and transient activations $z_1(n)$).

10.5.1 Testing the implemented system

The system constructed was tested with both artificial and real EEG data records.

As before, the sampling rate was 160 s^{-1} and the background activity modelling system had an order of 16, a normalized adaptation rate $\lambda = 0.1$ and a power estimation parameter $\alpha = 0.9956$. This system $H^{-1}(z)$ was realized as a direct LMS linear predictor, whereas $F_0(z) \equiv G(z)$ and $F_1(z)$ were implemented as IIR lattice constrained linear systems (section 9.3). The remaining system parameters were unchanged from those of the initial system described in detail in section 8.4.

Five records were used for testing the method. The results for those are shown collectively in Table 10.1. Record A was synthetic, constructed by superimposing the outputs of the background activity, spike, noise and transient generators (as in section 8.6.1).

EEG Record	Duration in seconds	Number of Spikes	Number of Transients	Muscle Artifacts present?	Detections			
					Spikes		Transients	
					√	×	√	×
A†	62	25	20	No	21	2	18	0
B‡	210	52	–	Some	45	15	–	3
C‡	130	8	–	Yes	7	19	–	7
D‡	123	10	–	Yes	6	23	–	19
E‡	242	0	–	Many	–	55	–	42

Notes:

†: Artificial record of duration 10000 samples.

‡: Real EEG records; they correspond to records A, B, C and D of previous tables.

Table 10.1: Results for simulated and real EEG records for the extended system structure

Records B, C, D and E are real sampled EEG signals. They correspond exactly to records A, B, C and D used for testing the spike detectors described in previous Chapters and they only contained spikes (and artifacts).

Figure 10.3 shows a segment of the synthetic input signal (Record A), and Figure 10.4 shows how the system responded to a real signal. The sequences out of the inverse systems and the detection signals for the two transients present. The estimated spikes, transients and background activity signals are also indicated.

10.5.2 Comparison with the earlier system and comments

Evidently the extended system is not as effective as the original spike detector. There is an increase in the number of false detections for both spikes and the transients considered, as well as in the number of missed spikes. False detections due to muscle artifacts were more numerous here, but the primary concern is the behaviour of the system for the synthetic record A, where effectiveness was expected to approach 100%.

The source of this deterioration of performance was traced to the preprocessing units and their logical combinations, which comprise the decision elements. The focus here is on the transient preprocessing units, whose outputs $\{z_{pi}; i = 0, 1, \ldots, N\}$ have two undesirable characteristics:

- Each one depends on a single transient, as its input is a window of samples of one of the $y_i(n)$ only

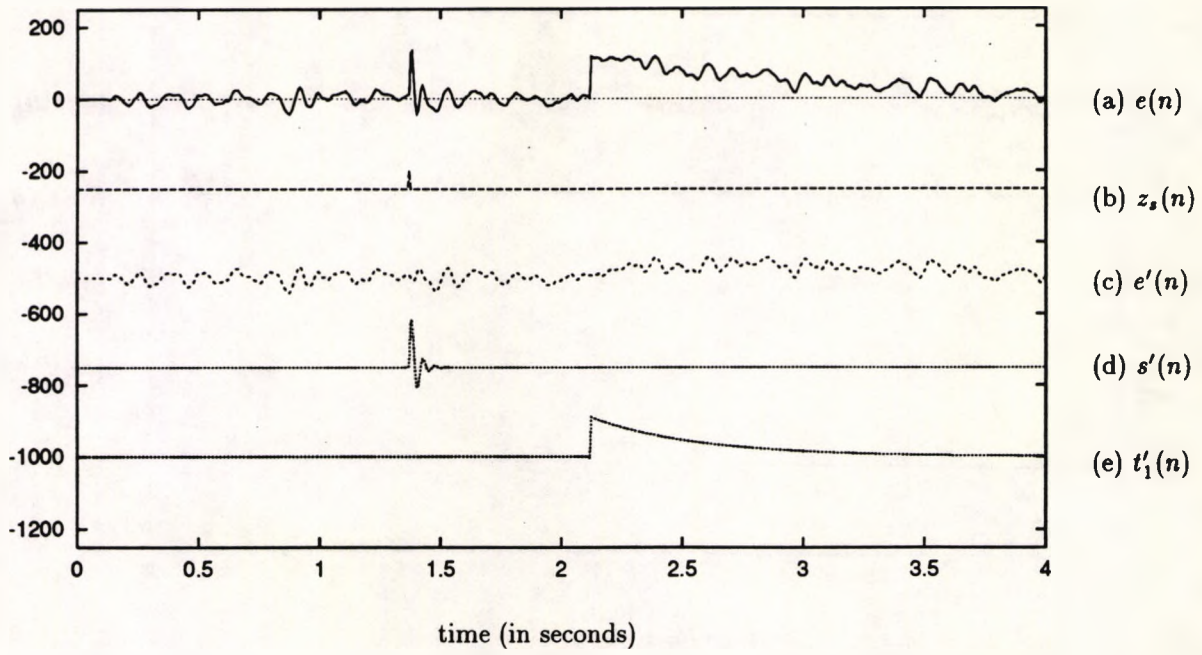


Figure 10.3: Response of the extended system to synthetic data

- The $z_{pi}(n)$ are all binary signals.

A consequence of the first property is the inability of the preprocessing unit to relate its operation to the signals for other transient types and is, therefore, 'blind' to what happens elsewhere in the system. A transient of a particular type, say i, is likely to cause some 'disturbance' of relatively large amplitude not only in the signal $y_i(n)$ that it represents, but also in all other outputs of the transient inverse systems. These are, of course smaller in amplitude than the one in $y_i(n)$, but as a consequence of the binary nature of their outputs the preprocessing units are inherently incapable of distinguishing between them. Hence, there may be multiple detections of transients by more than one preprocessing unit. This, in turn, makes every one of the transient activation units, where preprocessing indicated a transient, to become active. Therefore more than one transient model is generated and subtracted from $e(n)$ (see Figures 10.1 and 10.4). Subsequent testing of the residue signal, $e'(n)$, would reveal that a transient is still present, because of the extraneous generation and subtraction of a transient. Hence a spike indication would not occur, even when the transient was actually a spike.

This interpretation agreed with the observed results and suggested some ways to overcome the problem. It is possible to make the outputs of the preprocessing units continuous, instead of binary, so that the *amplitude* of the input transient may be reflected in its output. Then it would be the task of the output units to discriminate which transient was the one most likely

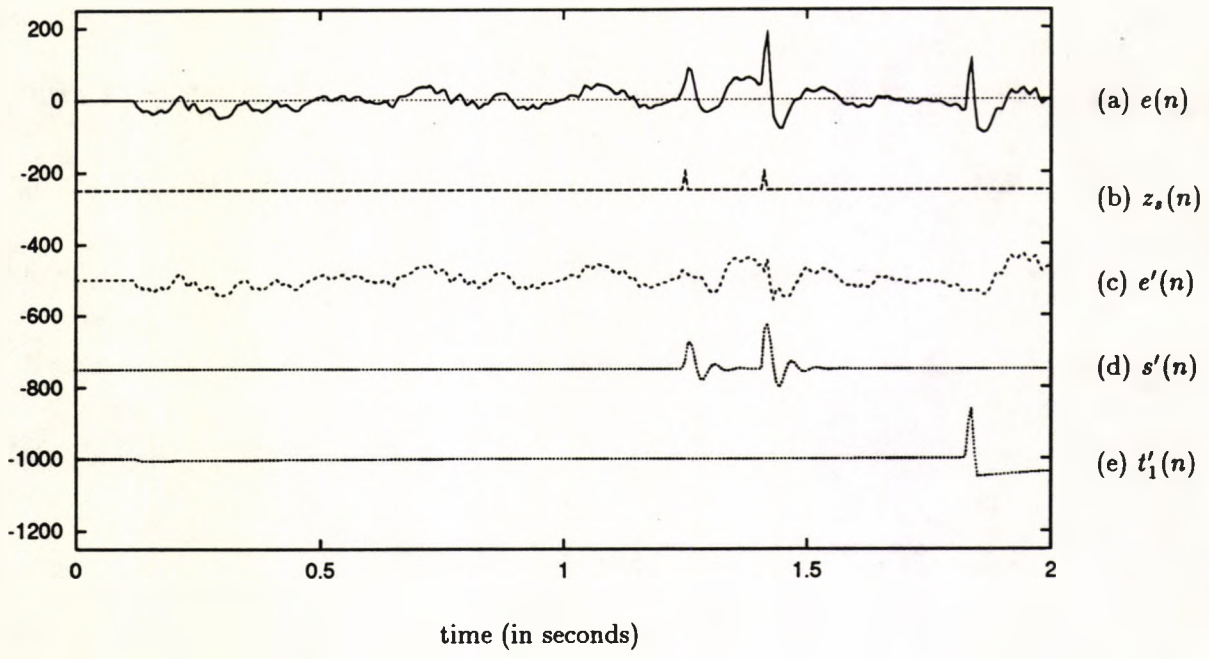


Figure 10.4: Waveforms for real EEG signals

to have occurred. Needless to say that the necessary logic would not be binary any more, but the thresholded weighted sum implementation of these operations described in section 10.4.3 may be employed. Of course, each of the elements would need to be modified to add inputs from all other transient preprocessing units. The output of the ith unit, $z_i(n)$, would only be active when $z_t(n)$ is active and $z_{pi}(n)$ has the largest amplitude of all the preprocessing units. This method would only generate one transient at a time, so if two of them occur at one specific instant only one would be activated and it is possible for one or both to pass undetected, a rather rare eventuality whose effects in practice would be negligible.

Another option is to incorporate all of the transient signals, $y_i(n)$ in all of the preprocessing units. In this way the amplitudes of occurring transients may be compared at preprocessing rather than afterwards and hence the (binary) output of each unit would be active only when the corresponding transient is the one with the largest amplitude.

Both suggestions appear reasonable, but they are not free from problems. In the first, the introduction of the continuous signal in what was a purely binary decision entity made the validity of the latter questionable. It is quite possible that the weights corresponding to the binary operations (see section 10.4.3) would need to be modified to achieve the expected system behaviour. There are, however, no analytic means of describing them, since they depend inherently on the nature of the signals concerned. It is also questionable whether a linear combination of all the inputs to the unit, according to the second suggestion would

be capable of discriminating whether a transient is at its peak in a particular sample of the multitude of inputs. This would be particularly difficult when the transient is registered with fairly large amplitudes in more than one of the $\{y_i(n)\}$.

Evidently, a more flexible approach is required to overcome the limitations of the proposed modifications of the decision-detection unit. Such an approach is described in the next Chapter.

Chapter 11

A flexible detection unit

11.1 Introduction

The problems encountered during the extension of the analysis procedure to more than one transient, indicated that the rôle of the detection and decision unit could not be undermined. Despite detailed considerations during its design in sections 7.4.4 and 8.5, it proved to be the limiting factor, as far as the performance of the whole system was concerned.

The general conclusions from the previous Chapter indicated that the insufficiency of the decision unit was its inability to locate the source of a transient (on some occasions). This was traced to the way it discriminates individual transients, which is based on very limited information and ignores their inter-correlations. A method was sought, to overcome these limitations by incorporating relative amplitudes between transients and their correlations for their detection. This was found in a well-known type of Artificial Neural Network, the *Multi-Layer Perceptron.* This Chapter illustrates that there is a close correspondence between the MLP and the existing system, and that the latter is in fact a natural generalization of the former.

Nevertheless, its implementation was not trivial, as new issues were raised. The MLP achieves its performance by 'learning by example'. It was therefore necessary to construct a 'training set'. Moreover, as shown later, its configuration, initialization and conditioning of its inputs were issues worth considering. Several configurations were tested. The results showed that the earlier, intuitive realization was a step in the right direction, but it required some 'tuning' of its parameters.

11.2 Artificial neural networks

Artificial Neural Networks (ANNs) are a broad class of structures, capable of changing their behaviour to maximize performance. This property is important in this application, since it consists of what the initial decision-detection is missing, the ability to improve its response.

They evolved in parallel with digital computer technology and have a turbulent history [136, pp. 14–20] (see Appendix E). Recent developments in the 1980's by a small number of enthusiasts highlighted the capabilities of the ANN approach to solve problems considered difficult by conventional computing methods, especially those related to pattern recognition[166].

Artificial neural networks are characterized by relatively simple functional units, called *nodes*, *processing elements* or *neurons*. Inputs to a processing element may be outputs of other elements or they may originate outside the network. Similarly, its output may be 'internal', serving as an input to other processing elements, or output to the 'outside world', or both. A node has a transfer function, which describes how the output may be generated from its inputs and local memory. The utility of an ANN comes from the collective, parallel, operation of its processing elements, which are highly interconnected.

From the outside, an ANN may be regarded as a system which processes a number of inputs and generates a number of outputs. This offers the possibility of inclusion of an ANN as an embedded unit in a conventional software or hardware system[136, pp. 9–10]. It may be implemented either in hardware or simulated in software.

Depending on their general properties, ANNs are classified as *associative*, *mapping*, *spatiotemporal*, *stochastic* or *hierarchical* (see Appendix E). These are described in detail by Hecht-Nielsen[136, pp. 79–109] and Lippman[112].

The ones of interest in this application are mapping networks, since for the purposes of the construction of the decision-detection unit for the EEG analysis system, what is required is a mapping from a set of inputs to a set of binary activation signals and detection outputs. One of these *backpropagation neural network* (a form of the MLP)[136, pp. 110–162] is the most likely candidate to perform well in the task at hand without requiring many changes in the existing structure of the system.

11.3 Architecture of the MLP neural network

The MLP neural network is a feed-forward structure, that is there is no feedback. Strictly speaking, this is not true during training, because this is based on feeding the errors back from the outputs through the network (see section 11.4). It is also a layered structure consisting

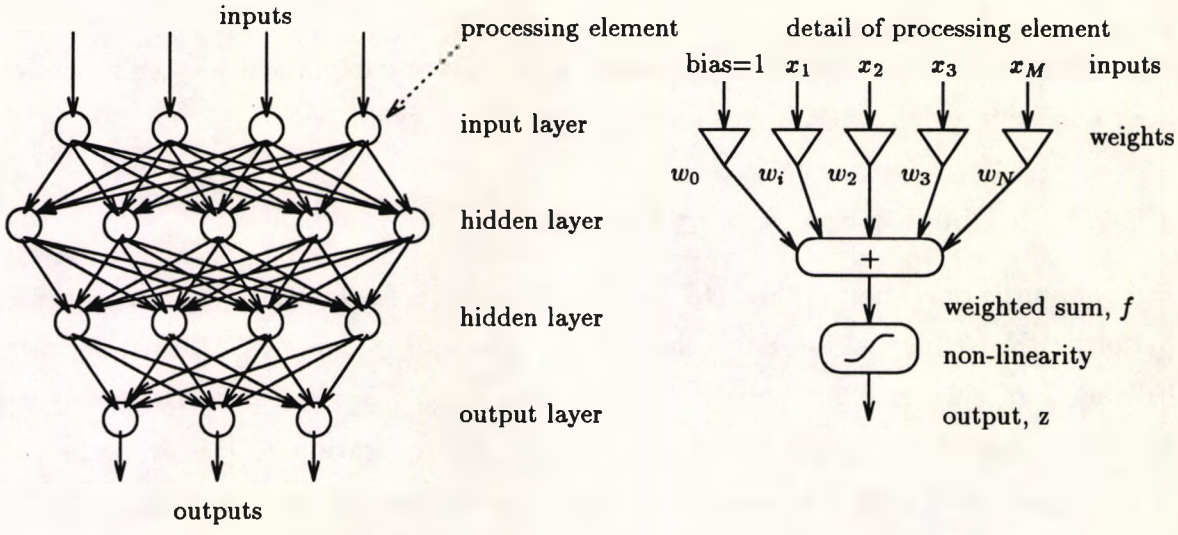


Figure 11.1: Architecture of the MLP

of groups of processing elements (see Figure 11.1). Hence, the inputs to the network are all received by one layer, called the *input layer* and all the outputs are produced by a layer of nodes, called the *output layer*. Between these there are one or more '*hidden*' layers of processing elements, so called because they are not directly connected to the outside world. The input layer consists of inactive units, whose only purpose is to provide the inputs to the rest of the network. This will be ignored in subsequent analysis, because it has no functional significance. The hidden layers are usually numbered sequentially, starting with the one closest to the input.

Traditionally, the inputs to the processing elements of one layer are the outputs of the previous layer. In some applications inputs from more than one previous layer are provided. The rule is that connections always come from preceding layers. One characteristic of the MLP is the symmetry of the connections, since there is full connectivity between one layer and the next. In other words, the output of an element in a particular layer is an input to *all* elements (nodes) of the next layer, as depicted in Figure 11.1.

As Hecht-Nielsen pointed out[136, pp. 125–138], the backpropagation MLP neural network just outlined does not comply with the definition of an ANN, because during optimization it is necessary for a processing element to have more than one output (see next section). It is, however, possible to redefine its architecture by decomposing the processing elements so as to ensure that all resulting nodes have only one output each.

11.4 Approximations of functions by optimization

Having outlined the structure of the MLP, it will now be explained why this can act as an approximation to a function and how this may be achieved.

11.4.1 The Operation of the MLP

First of all the operation of the MLP must be described. Each processing element essentially performs the same task as any other. The transfer function of an arbitrary processing element p of a layer l has a set of weights $w_{lp1}, w_{lp2}, \ldots, w_{lpN_l^i}$, each associated with one of the N_l^i inputs to the layer $x_{l1}, x_{l2}, \ldots, x_{lN_l^i}$. The operation of the processing element is a weighted summation to produce a linear combination of its inputs, $f_{lp} = \sum_{j=1}^{N_l^i} w_{lpj} x_{lj}$ followed by range-limiting, akin to thresholding. This is a non-linear operation, based around a 'threshold' value, θ_{lp} to produce the output of the element, $z_{lp} = \sigma(f_{lp} - \theta_{lp})$, where $\sigma(\cdot)$ denotes the range limiting non-linearity. To simplify the description, vector notation will be adopted. The threshold will also be implemented as a weight applied to a constant input of unity and denoted by the 'zeroth weight' in the augmented weight vector, $\mathbf{w}_{lp} = (w_{lp0}, w_{lp1}, w_{lp2}, \ldots, w_{lpN_l^i})^T$, where $w_{lp0} = -\theta_{lp}$. The input vector may then be denoted by $\mathbf{x}_l = (1, x_{l1}, x_{l2}, \ldots, x_{lN_l^i})^T$. Then the transfer function of the processing elements of the MLP may be described by the pair of functions:

$$\begin{aligned} f_{lp} &= \mathbf{w}_{lp}^T \mathbf{x}_{lp} \\ z_{lp} &= \sigma(f_lp) \end{aligned} \tag{11.1}$$

Because in the traditional form the inputs to a layer of the MLP neural network are the outputs of the previous layer, the input vector may be expressed in terms of these outputs, $\mathbf{x}_l \equiv \mathbf{z}_{l-1} = (1, z_{(l-1)1}, z_{(l-1)2}, \ldots, z_{(l-1)N_{l-1}})^T$. Consequently, the number of inputs N_l^i to a layer is equal to the number of outputs (and hence elements) in the previous layer, N_{l-1}. By this convention $\mathbf{z}_0$ is the input vector, $\mathbf{x}$. Both notations will be used interchangeably. Moreover $\mathbf{z}_L$ will be used to denote the network output vector, since there are L layers in the MLP hierarchy (excluding the inactive input layer). In some applications the non-linearity is not included in the output layer, thus permitting the functional approximation of the MLP to extend to functions whose range is the whole of $\mathbb{R}^{N_L}$ rather than a compact subset of this space.

The non-linearity, $\sigma(x)$, originated in the threshold function, whose range consists of two values, 0 for $x \leq 0$ and 1 for $x > 0$. This function, however is non-differentiable at $x = 0$ and cannot be used in conjunction with the learning algorithms that are often employed[112]. Therefore smoother functions, σ, are utilized, which have the following properties: A limited

range, $\sigma(x) \in [a,b] \forall x \in \mathbb{R}$. Typically, $[a,b] \equiv [0,1]$ or $[-1,1]$. They should also be differentiable at all points and monotonically increasing, which guarantees that $\lim_{x\to-\infty} \sigma(x) = a$ that $\lim_{x\to+\infty} \sigma(x) = b$. The most commonly used is the *logistic function*[166, 112][136, pp. 56–63, 124–137],

$$\sigma_1(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{11.2}$$

Other possibilities include $\sigma_2(x) = \tanh(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}}$ and $\sigma_3(x) = \frac{1}{\pi}\mathrm{Arctan}(x)$.

Having explained how the MLP neural network functions it may now be demonstrated how it achieves the desired approximation to a required mapping.

11.4.2 Approximation of a given mapping, optimization

Unlike some other ANNs, which attempt to construct *a* mapping by clustering similar inputs based on a set of rules, the backpropagation MLP network requires an explicit description of the mapping as a set of input-output pairs of vectors, $(\mathbf{x}, \mathbf{d})$. Of course $\mathbf{d}$ denotes the desired output vector, which may not necessarily be equal to the output of the MLP after training, but is expected to be close to it. Therefore, the desired mapping is specified by a *training set*, which is a (usually finite) collection of pairs $\{[\mathbf{x}(n), \mathbf{d}(n)]; n = 1, 2, \ldots\}$. It is desirable that the training set is comprehensive, covering all possibilities arising in the particular application, with adequate representation. These issues will be addressed towards the end of this section, as they are related to the training scheme.

'Training' the MLP implies the procedure of presenting an input vector example $\mathbf{x}(n)$ to the network from which its output vector, $\mathbf{z}_L$ is computed. The network is then presented with the desired, or target, vector, $\mathbf{d}(n)$, and is 'corrected' by modifying the weights at the connections between its processing elements in order to make its output similar to the desired response. This procedure must be repeated for the whole of the training set.

This scheme requires two elements: A performance criterion to measure the global behaviour of the network to the whole of the training set and a method to compute the weights of the network. The performance of the MLP can be measured by 'how close' its output is compared to the desired mapping implied by the training set. Since this is related to the error of the approximation, $\varepsilon(n) = \mathbf{d}(n) - \mathbf{z}_L(n)$, performance criteria are typically positive, non-decreasing (distance) functions of $\varepsilon(n)$. The mean-squared-error function, $\xi = \mathrm{E}[\varepsilon(n)^T \varepsilon(n)] = \mathrm{E}[\|\varepsilon(n)\|^2]$ is by far the commonlest, although others like the mean absolute error $\mathrm{E}[\|\varepsilon(n)\|]$ and higher order norms $\mathrm{E}[\|\varepsilon(n)\|^\nu]$ have been investigated by some authors[180].

The minimization of the MSE function was encountered before, as it is the basis of

Wiener filtering, discussed in section 8.1. The systems then were linear and it was possible to express the solution in closed form with the aid of the correlation matrix. The presence of the non-linearities in the outputs of the processing elements and their hierarchical (layered) organization make a similar formulation and solution impossible. Numerical optimization is often employed to determine the optimal weights recursively.

This is a procedure that uses the weight vector formed by including all weights of all processing elements of the network, $\mathbf{w}$. The MSE, being a function of $\mathbf{w}$ is minimized when its gradient is zero. Therefore, the point of minimum MSE may be determined by starting from an initial guess and proceeding in repeated steps $n = 0, 1, 2, \ldots$ in a direction opposite to that of the gradient $\nabla\xi(\mathbf{w})$:

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \mu\nabla\xi(\mathbf{w}(n)) \tag{11.3}$$

This procedure is equivalent to the method of the steepest descent (section 8.4.1). There are many varients of the basic equation. To smooth the noisy estimates, a 'momentum term' is usually introduced, which utilises the previous direction of the gradient, $\mathbf{w}(n-1)-\mathbf{w}(n)$[136, pp. 59–63]:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \alpha[\mathbf{w}(n-1) - \mathbf{w}(n)] - (1-\alpha)\mu\nabla\xi(\mathbf{w}(n)) \tag{11.4}$$

To implement a practical scheme it is necessary to estimate the gradient vector, $\nabla\xi(\mathbf{w}(n))$. This consists of the partial derivatives of ξ with respect to each weight. An algorithm closely related to the LMS , the *generalized delta rule*, was proposed by Rumelhart et al.[166]. Because its description is rather lengthy, but some of its aspects will be necessary for the discussion that follows, this has been included separately in Appendix F.

11.5 Implementation of the MLP decision and detection unit

The earlier detection-decision structure, described in section 10.4, arose naturally as a method for the detection of spikes as well as for the production of the signals that control the generation of suspected transients. Its structure and function were selected to perform the required task using the usual engineering approach of breaking the system into smaller entities whose function is simple enough to implement directly. When testing revealed a deterioration in its performance, efforts were concentrating on improving the existing unit, rather than on replacing it, since its structure appeared logical and so did its functionality. The only ambiguous points were the preprocessing of its inputs and the possibility of hidden

correlations, which could not be considered by the ad hoc design methodology of the earlier system. It was therefore surprising to recognize the resemblance between this system and the MLP. The latter is a generalization of the former and as such it offers capabilities that the initial system could not possess.

11.5.1 Relation to the intuitive system

Initially, the relation of the intuitive system to the MLP may not be apparent. Of course both are layered structures, but the first is a mixture of real and binary (logic) elements, its output being a binary signal vector, whereas the second performs a mapping from a real space to another real space, using affine combinations and smooth non-linearities only.

Moreover, the function of the MLP is generally defined as a mapping from its inputs and its desired output rather than by the behaviour of its individual elements, which was the basis of the earlier system. Related to this, is the way the connections between its functional elements are defined in the two structures. In the existing system these are prescribed by the purpose of every element, whereas in the MLP elements lose their individual significance and it may not always be clear what part of the whole function each one performs. Therefore one layer is fully interconnected to the next.

To demonstrate the similarity of the earlier system to the MLP, the structural equivalence between the two will first be shown, followed by a demonstration that the function of the processing elements is also the same in both cases.

Similarities of the overall structure

Observing the initial system macroscopically, it is obvious that it contains several functional elements that are linked hierarchically without loops (Figure 10.2). The processing elements are organized in layers, a preprocessing layer, consisting of elements receiving the inputs and producing some intermediate signals, which are received by the elements of the second layer from which the outputs of the system are generated. This hierarchical organization resembles the MLP structure, where processing elements are also organized in layers. The preprocessing layer takes the place of the first (and only) hidden layer and the second layer that of the output layer. The input layer of the MLP may be ignored, since it is functionally inactive, serving only as a means for distributing the inputs to the first hidden layer.

Similarities also exist between the interconnections. The inputs are only applied to the first layer and the outputs of this to the output layer. Unlike the MLP, however, these are not fully connected, neither are all the inputs applied to every node of the preprocessing (hidden) layer. This is a minor discrepancy, however, because the missing connections may

be introduced and left unused by the corresponding processing elements. It is believed that part of the reason for the deterioration in the performance of the earlier system was the absence of certain interconnections. These would give an element a more global view of what happened in other parts of the system and hence make it capitalize on the extra information to produce a more realistic output when it would otherwise fail. Because the processing elements essentially perform linear weighting of their inputs, the equivalence between the original system and the backpropagation neural network is possible by assuming that the weights in the inputs provided by the additional connections are set to zero (and thus these inputs contribute nothing to the function implemented by the processing element).

Having demonstrated their structural equivalence, it remains to demonstrate that the processing elements of the two systems can perform the same function.

Comparison of preprocessing elements

According to their description in section 10.4.2 the preprocessing elements receive a window of samples from one of the input signals and perform squaring, normalization, weighted summation and thresholding to produce a binary output signal.

Squaring of inputs is not normally part of the transfer function of the processing elements of the MLP. Normalization is even more ambiguous in terms of neural network functions, because it is a global operation requiring the computation of the sum of squares of the inputs. Of course an extra processing element could be placed before the preprocessing layer to square and average all the inputs and fan out the normalizing factor as an extra input to preprocessing. Because, however, squaring of the inputs would have been necessary in every preprocessing element anyway, it is more practical to separate these operations from the corresponding transfer function and the network. Squaring and normalization may be considered as a form of preconditioning of the inputs that may be performed by a separate operation acting as a buffer between the signals of the rest of the system and the decision-detection unit, as depicted in Figure 11.2. Squaring and normalization is only one of the possible forms of input preconditioning, an issue addressed in section 11.6.1.

Having separated the preconditioning stage from its transfer function, the equivalence between the preprocessing elements of the intuitive system and those of the MLP is apparent.

Both transfer functions use a weighted sum of their inputs to produce an intermediate signal, $f(n)$ (equations (10.2) and (11.1) respectively). The earlier system, combines only selected inputs for the generation of $f(n)$, whereas the transfer function of an MLP element combines all its inputs and has a bias input as well. The former transfer function may thus be regarded as a special case of the latter with the weights corresponding to the unused

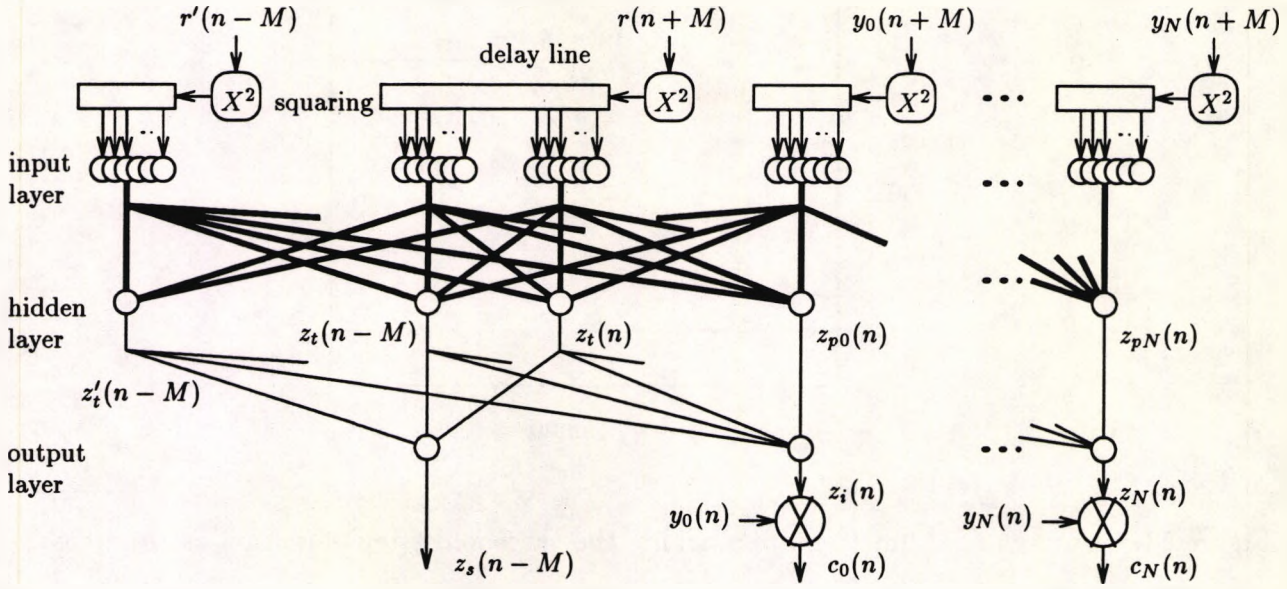


Figure 11.2: The MLP realization of the decision-detection system

inputs and the bias set to zero (see Figure 11.2).

The last element to compare is the non-linearity at the output of the processing element. In the earlier system this was a thresholding operation, taking one of two values, either 0 or 1. For the MLP, however, this is a sigmoidal, which takes values in the range between 0 and 1. It will be demonstrated, however that both may behave in a similar way.

For a given set of weights the linear output of the transfer function is $f(n)$, and the output of the element is $z(n) = \sigma[f(n)]$, which is typically linear at $f(n) = 0$ and saturates towards either 0 or 1 depending on whether $f(n)$ is negative or positive. This is actually true for the logistic function, but the argument may be extended to other non-linearities. Scaling all the weights by a factor δ_w would cause an increase in the amplitude of the linear output, $\delta_w f(n)$. If $\delta_w > 1$ this would correspond to a compression of the domain of the mapping defined by $\sigma(\cdot)$. As a result, if the turning points of the sigmoid, beyond which saturation is more profound, are at $f(n) = \pm\phi$, for the original weights they are now pushed closer together to $\pm\frac{\phi}{\delta_w}$. In the limiting case, as $\delta_w \to \infty$, the above range approaches a single point, $f(n) = 0$, and saturation occurs for all positive values to 1 and for all negative ones to 0. The similarity in the outputs between the sigmoid and the threshold function is, in fact, noticed long before δ_w approaches infinity, as portrayed in Figure 11.3 for a sigmoid with unit inputs and unit weights for various values of δ_w.

From these considerations it is clear that the preprocessing units of the earlier system may be implemented by the (first) hidden layer of an MLP neural network.

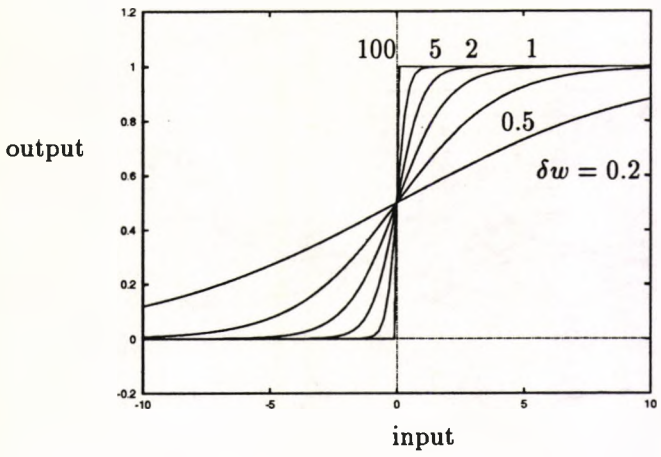


Figure 11.3: A sigmoidal function approaching the threshold (step) function as the weights approach infinity

Implementation of logic functions

It remains to show that the MLP can also implement the logic operations required by the initial system for the generation of the various output signals. It was actually hinted in section 10.4.3, that the logic functions required could be implemented as weighted sums followed by thresholding.

The required operations are rather simple and can be implemented each by single processing elements in the output layer. The first is a simple ANDing of two signals, $z_i(n) = z_t(n)\text{AND}z_{pi}$. The equivalence between this operation and the following sum and threshold pair of functions is easy to determine by considering the corresponding truth table.

$$f_i(n) = wz_t(n) + wz_{pi}(n) \tag{11.5}$$

$$z_i(n) = \begin{cases} 1 & f_i(n) > \theta \\ 0 & f_i(n) \leq \theta \end{cases} \tag{11.6}$$

where w is a positive weight and $\theta = \frac{3}{2}w$ the threshold. Other combinations are also possible. For this implementation, w was set to 1 and θ to 1.5.

Similarly, the second type of logical operation, $z_s(n) = z_t(n-M)\text{AND}z_{p0}(n-M)\text{ANDNOT}z'_t(n-M)$ may be realized in an equivalent way by the summation $f_s(n) = z_t(n-M) + z_{p0}(n-M) - z'_t(n-M)$ followed by thresholding at 1.5. Again the operation is invariant to scaling.

If the above operations are implemented as part of the MLP structure and training is to be employed, the hard threshold function would have to be replaced by the softer sigmoidal function. It would then be simpler to include the threshold values in the form of the bias weight in the processing elements. Then, following the analysis of the previous section,

scaling up the weights (and thresholds) would make the sigmoidal behaviour very similar to the threshold function. Because, however, the signals at the output of the output layer are generally expected to be binary, it is possible to replace the sigmoidal with hard thresholding after training, or to employ an additional mapping on the outputs of the network by replacing output values that are above 0.5 by 1 and those below 0.5 by 0.

With these final notes, it is evident that the system of the previous Chapter is equivalent to an MLP. It may be postulated that the problems encountered by the former are associated with the non-optimality of its parameters for the particular application. Hence, by properly training it, the performance is expected to improve.

11.6 Alternative system configurations

The equivalence between the intuitive system and the MLP ceased by the replacement of the heuristic ad hoc selection of the weights with more formal criteria. Although it is possible to use the same structure as the former system and apply the learning algorithm to determine the optimal set of weights, a procedure that was initially followed, the backpropagation neural network offered other alternatives, which were tested also. A point of concern was the form of input conditioning. Squaring and normalization was inherited from the earlier system, where it had some significance. With the use of the MLP, the significance is lost along with the individuality of the processing elements. The purpose, and indeed the necessity of conditioning the inputs, were questioned. This led to different realizations and network configurations to cater for the peculiarities of each, which will be presented here.

11.6.1 Input preconditioning: Is it necessary?

Some artificial neural network enthusiasts state that the inputs may be applied to the network without any preprocessing (see for example [83]). This may be true in the majority of cases, where the mapping approximated by the network is not scale-invariant. When dealing with natural signals, like the EEG, the system is expected to tolerate changes in the dynamic range of the recorded signal, since these are usually related to changes in the impedance between the source of the signal and the recording apparatus. This argument is better supported by a simple example.

Let a signal $x(n)$ with variance $\sigma^2 = 1$ input to a single-input processing element that performs weighting by $w_1 = 10$ thresholding at $15 = -w_0$, with a sigmoidal non-linearity. If $x(n) = 1$, the linear output of the element would be $f(n) = 10 - 15 = -5$ and $z(n) = \frac{1}{1+e^5} = 0.006 \approx 0$. If the dynamic range of the signal was doubled (amplified by a factor of

2) its variance would be $\sigma^2 = 4$ and the value of the signal would be $x(n) = 2$. This would produce a linear sum $f(n) = 20 - 15 = 5$ and after the non-linearity, $z(n) = 0.993 \approx 1$. In both cases, of course, the system ought to have behaved in a similar manner, since the input signal is essentially the same, apart from a scaling factor, which in many signal processing applications should have had no effect. Without preprocessing, the difference in the output caused by the change in the range of the input was rather dramatic. If, however, the input to the MLP was normalized by dividing it by the standard deviation of the signal ($\sigma = 1$ in the first case and 2 in the second) then the effective input would have been $\hat{x}(n) = \frac{x(n)}{\sigma} = 1$ in both cases, producing 'consistent' results. It is, therefore, quite important to condition the inputs to the MLP if the consistency, and hence the reliability, of the system is to be maintained at run-time.

Sensible preconditioning may also result in a simpler MLP structure, with fewer elements or layers, which may be interpreted as performing an intermediate mapping of the inputs prior to their application to the network. If this was not done the MLP would have had to perform this mapping itself, a process requiring extra processing elements or even a few additional hidden layers. Of course such a mapping cannot always be determined, since the network is normally treated as a 'black box'. In the case of the decision-detection unit, however, some insight into the function of the MLP is available from the earlier intuitive system and what is required for the network to perform. If, for instance large amplitudes are to be detected, irrespective of polarity, the network is simplified if the corresponding input is squared (thus eliminating polarity). In this case, a single element with one weight and a negative bias (i.e. positive threshold value) would be sufficient. If the 'raw' input was applied, three elements would have been required. One to detect large positive amplitudes, using a positive weight and threshold, one to detect large negative amplitudes (with negative weight and threshold) and one to combine these outputs, having an amplitude of 1 if either of its input is 1 (an OR operation, with two equal positive weights, $w_1 = w_2 = w$ and a threshold, $\frac{w}{2}$).

From these arguments one may conclude that, in general, input preconditioning is often vital for the correct operation of the MLP and could lead to a simpler system without compromising its performance, if done carefully.

11.6.2 Treatment of the inputs to the MLP

The arguments on the use of preprocessing put forward in the previous paragraph were put forward in the realization of the MLP decision-detection unit. Normalization of the inputs to maintain the function of the system through changes in the dynamic range of the input

will be examined first. This is the only necessary operation but it will be demonstrated that some additional transformations of the inputs may lead to simpler network configurations.

Normalization

The elimination of the dependence of the output of the network on the strength of the input signal is vital for the success of the decision-detection system, especially if it is to operate with minimal supervision.

Even in this seemingly straightforward matter a number of questions arose. Should each signal be normalized by a measure (such as the variance), derived from individual signals, or should a common normalizing factor be used for all? If the second approach is adopted, it is not clear which signal measure to use in normalization.

In the system of section 10.4 each signal was normalized by its own variance. This had its origins in the statistical origins of the detection procedures employed then, where normalization of the corresponding probability distributions was required. Although this is justified for the random uncorrelated signals $r(n)$ and $r'(n)$, the same cannot be said about any of the $y_i(n)$, which are far from uncorrelated random sequences. Hence, although the same techniques were applied to these signals, the arguments used then were not strictly valid.

In the case of the MLP, normalization was viewed away from its significance in the earlier system, just as a means of reducing the effects of changes in the signal level on the operation of the network. It is, of course, possible to use measures derived from each signal separately, like before. Therefore an input signal to the MLP, $x(n)$, could be divided by its own measure of spread, like the ones that follow:

$$\begin{aligned} \sigma_x &= \{E[x(n)^2]\}^{\frac{1}{2}} && \text{standard deviation} \\ \delta_x &= E[|x(n)|] && \text{mean deviation} \\ m_x &= x(n);\, P[X(n) < x(n)] = \tfrac{1}{2} && \text{median} \\ p_x &= \max_n[x(n)] - \min_n[x(n)] && \text{peak} - \text{to} - \text{peak} \end{aligned} \tag{11.7}$$

where $x(n)$ represents $r(n)$, $r'(n)$ or any one of the $y_i(n)$. It should be noted that these signals have zero mean, a fact that has been used to simplify some of the above expressions. The standard and the mean deviation, are the most favoured measures, because they are linear functions of the input signal and may be estimated with relatively simple operations.

The normalization of a signal by its own measure of spread is a valid procedure when standard probability distributions are used for statistical testing. This is not so meaningful, however when comparisons are to be carried out on the various signals, with common origin, like in the proposed analysis system. Again, the reasons are better established by an example,

with reference to the extended system of Figure 10.1. Let a segment of the EEG, $e(n)$, consist of a sinusoidal background activity signal of amplitude A_0 and frequency ω_0 with negligible noise and a transient generated from the excitation by an impulse of amplitude B_i of a system $F_i(z)$, whose PSD has a single peak at ω_1, away from ω_0 and a gain such that the peak of its output is 1. It will be assumed that another transient has been modelled, whose transfer function, $F_j(z)$ has a spectral peak near ω_0. After the action of the inverse of $F_i(z)$ there would be an impulse of amplitude B_i in its output, $y(n)$, but this will be buried in the large sinusoidal due to $e(n)$, which would not have been suppressed by $F_i^{-1}(z)$ (which reduces frequencies around ω_1). On the contrary, the output of $F_j(z)$, $y_j(n)$ will reduce the background activity to a level ρA_0 ($\rho \ll 1$) and maintain the transient.

The standard deviation, or indeed any measure of spread, depends on the long-term variability in the signal. Hence in the case of $y_i(n)$ this will be essentially equal to the standard deviation of $e(n)$, $\sigma_e = \sigma_{y_i} = \frac{A_0}{\sqrt{2}}$, whereas in the second it will have the reduced magnitude, $\sigma_{y_j} = \frac{\rho A_0}{\sqrt{2}}$. Therefore, after normalization, the amplitude of the transient in $y_i(n)$ would have been reduced to $\hat{B}_i = \sqrt{2}\frac{B_i}{A_0}$, which is smaller than the equivalent peak amplitude in $y_j(n)$, $\sqrt{2}\frac{B_i}{\rho A_0} = \frac{\hat{B}_i}{\rho}$. Because the normalized transient in $y_j(n)$ has a higher amplitude than that in $y_i(n)$, a system that depends solely on amplitudes for the detection of its source would be misled into believing that the transient was of type j, when in fact it was of type i. Testing $r'(n)$ would, of course reveal the mistake, but there would be no means of correcting it. This is an artificial example and depicts a rather extreme event, but demonstrates that by scaling all transient signals by the same factor would have prevented the occurrence of such mistakes.

Such normalizing factors could be derived from a number of sources within the system, independent of the transfer functions of individual transients. One candidate is the standard or mean deviation of the output of the background activity, $r(n)$, or of the input EEG signal, $e(n)$. Although the first was used successfully[131], it has one potential disadvantage. If the EEG signal becomes essentially deterministic, the 'random' element may almost vanish, reflecting the predictability of the signal. Therefore, its variance (standard deviation, mean deviation, etc.) would be small, whereas this may not be the case in the $\{y_i(n)\}$, thus contradicting the purpose of normalization. On the contrary, the input signal $e(n)$ provided a better source for the scaling factor, since it can only be small if the outputs of the $F_i(z)$ are small also. Besides, it offers the possibility of scaling $e(n)$ directly, before any processing takes place, thus avoiding individual scaling of many samples of the signals which constitute the inputs to the MLP, saving a considerable amount of computation as well.

Estimating the variance or the mean deviation recursively could be performed by recursive

means, like the ones suggested earlier for the computation of the power of the signal in Chapter 8 (see equation (8.52)):

$$\sigma_e^2(n) = \alpha\sigma_e^2(n-1) + (1-\alpha)e(n)^2 \tag{11.8}$$

$$\delta_e(n) = \alpha\delta_e(n-1) + (1-\alpha)|e(n)| \tag{11.9}$$

As before, α is a filtering factor, sometimes related to the length of the signal stationarity (section 8.4.3). In this case, however, it was made larger, thus taking into account a longer sequence of samples, because the changes in the signal power which normalization attempts to counteract occur on a much longer time scale than changes in the signal behaviour and often take effect over several minutes or even hours. For a sampling rate of 160 s^{-1} catering for changes slower than 60 seconds, $\alpha = 0.9999278$ (see equation (8.52)). Because this is very close to unity and quantization may deteriorate its filtering operation, a practical solution to this problem might be a two stage process, with squaring and band-limiting up to a frequency, f_s', being $\frac{1}{n}$ times smaller than the sampling frequency, f_s, followed by re-sampling at the lower rate f_s'. This permits a lower value of α, corresponding to the new sampling rate, to be subsequently applied.

Although normalization is the only *essential* input preprocessing operation, other data transformations have been investigated. These were carried out on the already normalized signals and may be considered as attempts to simplify the task performed by the MLP, by removing information believed to be redundant. Hence, the network has a less complex mapping to learn and can 'concentrate' on learning those hidden properties in the structure of the input signals rather than waste its time and processing elements modelling a data transformation known beforehand. As a result the number of processing elements and/or layers of the MLP may be reduced, leading to a simpler structure to perform the required task.

Numerous such operations are possible. Some are described next, emphasizing for each the main reason for their selection, any other properties that are significant to this application and the network configuration used in each case. It should be noted that the more elaborate network corresponds to a system whose inputs are unprocessed (but normalized) samples of the corresponding signals. For this reason it will be presented last.

Squaring

This is the one more reminiscent of the intuitive system (section 10.4), whose input (preprocessing) elements performed squaring and normalization prior to thresholding. The difference from that system lies in the fact that the weights of the earlier system were selected infor-

mally, whereas for the MLP implementation these are selected by optimization. The basic operation on the input signals is equivalent to squaring and scaling by the variance of the input signal, $\sigma_e^2(n)$:

$$\hat{x}_i(n) = \frac{x_i(n)^2}{\sigma_e(n)^2} \tag{11.10}$$

where $x_i(n)$ represents any of the signals $r(n-k)$, $r'(n-k)$ or $\{y_j(n-k); j = 0, 1, \ldots, N\}$ and the delay factor k ranges over the window of samples applied to the network, $[-M, M]$.

Squaring has two essential properties, concerning the detection of large atypical values in a sequence. First of all, $\hat{x}(n)$ is always positive. In other words, what the subsequent preprocessing elements can concentrate on is the magnitude of a sample, and not its direction (positive or negative), which had already been removed. The second characteristic of squaring is the enhancement of large amplitudes with respect to small ones, stretching them over a greater dynamic range. Hence it is easier to estimate the threshold for the distinction of large magnitudes, especially if the line between large and normal is not very clearly drawn owing to the two populations being closely clustered.

Using the intuitive system as a reference, the elements in the first hidden layer must perform weighting and thresholding of their inputs, although with different parameters than that system. This is suggestive of the number of elements necessary in that layer, which, should be equal to $N + 5$ for the detection of N transients (one for every transient plus one for spikes, one for $r(n)$ and one for delayed versions of $r(n)$, $r'(n)$ and $y_0(n)$, as explained in the end of section 10.4).

The outputs of these (pseudo-binary) elements are then used in the output layer directly for the production of the N transient indications, one for spike indications plus one more for spike detections ($N + 2$ outputs in total). It was assumed, like in the intuitive system, that one processing element is capable of performing the logic operations necessary. If this was not the case, extra hidden layers would have been required.

This network is perhaps not the smallest possible structure, since it may, in theory, be possible to combine more than one of the preprocessing operations in a single element. This, however would largely depend on the nature of the signals involved. The proposed rule for selecting the number of elements knowing the number of transients to be detected permits the system to be generalized in a simple and systematic way. This selection criterion is a practical guideline, based on a structure that seems intuitively adequate and not on any formal performance criteria. It has worked well in this application and is believed to perform equally well, should the system be extended to more transients.

Using the moduli of the inputs

Akin to squaring is the modulus operation, in the sense that both eliminate the sign information from the signals they are applied to. In conjunction with normalization, this is described by the following operation performed on every input to the MLP:

$$\hat{x}_i(n) = \frac{|x_i(n)|}{\sigma_e(n)} \tag{11.11}$$

The normalizing factor now is the standard deviation of the input signal, not the variance. Alternatively, the mean deviation could have been used.

The modulus operation has the same basic property as squaring, but does not enhance the amplitudes. It is, however, simpler and faster to compute, even on a small computer.

The structure of the required MLP for this type of normalization is effectively the same as that for squaring, already described, with $N+5$ elements in its only hidden layer, byt the values of the individual weights would, in general be different.

A hybrid realization

Squaring and normalization are sensible if the direction (positive or negative) of all transients considered is irrelevant in their detection, which sometimes is not the case. Epileptic spikes, for instance, are always electro-negative, hence are registered in a unipolar montage with always the same polarity (pointing up). If a spike of the reverse polarity is encountered, it does not constitute a valid occurrence. In a working system, should polarity be ignored, a number of false detections may be registered, an undesirable phenomenon that could have been avoided. This would be possible by maintaining some sign information, especially in the central samples for every signal, where transients are detected. Constituting a natural extension to the existing system, it is also linked to the structure of the system without preprocessing, described under the next heading. It deserves some elaboration because it requires modifications to the MLP configuration to work properly.

For this function the inputs are squared and normalized as described earlier with the additional re-introduction of the polarity for the samples in the centre of the window from each signal. For example for $y_i(n)$, the corresponding inputs to the network would be as follows:

$$\hat{y}_i(n+k-N) = \begin{cases} \frac{y_i(n+k-N)^2}{\sigma_e(n)^2} & -N < k \leq -Q \\ \frac{|y_i(n+k-N)|y_i(n+k-N)}{\sigma_e(n)^2} & -Q \leq k \leq Q \\ \frac{y_i(n+k-N)^2}{\sigma_e(n)^2} & Q < k \leq N \end{cases} \tag{11.12}$$

where N and Q have the same significance as in the earlier system (section 10.4). An equivalent implementation could be developed for the modulus function.

The price to pay for the discrimination of polarity is an extra layer and an increased number of elements on the first hidden layer. The latter is necessary to cater for transients which may occur with either polarity. This is established by considering what operations are required. A single processing element consisting of a weighted linear sum and thresholding can only distinguish linearly separable input conditions. This has been the reason why the XOR logic function could not be implemented by the perceptron[124, 166]. For a given set of weights and threshold, it is not possible to have a node produce the same output for both a positive and a negative central sample. Therefore it is necessary to have two elements, one detecting atypically positive (having output $z^+(n)$) and the other atypically negative ($z^-(n)$) samples as well as a further one to combine the two using a logical OR operation ($f(n) = wz^+(n) + wz^-(n) - \frac{1}{2}w$, $z(n) = \sigma[f(n)]$, Allowing for all transients to be bipolar, and maintaining the layered structure of the MLP where each layer feeds to the one directly above it, $2N + 10$ elements are necessary in the first hidden (preprocessing) layer, but this may be reduced to $2N + 6$, since spike detection and initiation only deal with positive transients (signals $y_0(n)$, $y_0(n - M)$, $r(n - M)$ and $r'(n - M)$) and require one, instead of two, preprocessing elements each. A further $N + 5$ OR units form the second hidden layer, with the $N + 2$ output elements. Although the elements in the second hidden layer have been referred to as OR units, it is understood that this is a simplification that may not reflect their function after optimization has been completed.

A System without preprocessing

Evidently, from the three types of preprocessing already described, the network complexity increases with the amount of information included in the input. To investigate whether preprocessing was really necessary, the an MLP with its inputs normalized, but *without any other processing* was constructed:

$$\hat{x}_i(n) = \frac{x_i(n)}{\sigma_e(n)} \tag{11.13}$$

Normalization is a scaling operation and does not modify the shape of the applied signal, thus preserving virtually all information contained in the waveforms. This is expected to result in a more accurate system, since the network can 'choose' an appropriate mapping of the inputs using its ability to approximate arbitrary functions. Perhaps different inputs need to be preprocessed in different ways and their inter-dependencies exploited from the first hidden layer, rather than by units that can only combine already processed information.

This configuration, like the hybrid preprocessing structure of the previous section preserves the polarity of the signals and requires at least another hidden layer. The first hidden layer must have $2N + 10$ elements, the second $N + 5$ and the output layer $N + 2$ elements.

Again, these are guidelines to deal with N transients and give a near minimal system. Because the processing elements can only perform simple functions, it may be necessary, in practice to have more elements or layers than the ones specified by the given rule, depending on the complexity of the embedded mapping. For the system tested, however, this did not prove necessary, as demonstrated by the results presented later.

11.6.3 A non-intuitive realization

In most applications of neural networks there is very little or no information as to which configuration is appropriate. The selection of a network topology is a task that its designer has to complete normally by trial and error. Effectively, if a structure does not work, more processing elements are introduced in the hidden layers, whose number may also be increased. It has been a general observation that the introduction of more layers often reduces the total number of nodes to achieve a desirable performance level[136, pp. 132–133].

Although in this application there are definite clues about the structure and size of the MLP, which were exploited to suggest the topology of the network for each type of preprocessing in the preceding section, the 'traditional' approach could not be ignored, at least for comparative purposes. By constructing an MLP without using the earlier system to suggest the topology, it is possible to verify whether the cues derived from that system can be used as satisfactory guidelines for constructing and generalizing the system.

The performance of the earlier networks, which were based on the structure of the intuitive system (section 11.6.2) may be evaluated explicitly, as the MSE after training is completed, or in terms of the performance of the spike and transient detection system. Performance determines whether a given MLP configuration can approximate a specified mapping well, but it provides no indication whether the same performance could not be achieved by a simpler network. Starting from a small network, the best approximation to the required mapping is expected to be poor and the minimum MSE after optimization large. As more processing elements and layers are added, the approximation improves and there is a point when the minimum MSE 'levels off' and further increase in the network size is accompanied by marginal, if any, improvement in its performance. Having redundancy in the network is undesirable, since it is accompanied by unnecessary processing and waste of computing power and could lead to over-training, since the extraneous processing elements may model detailed features specific to the training set and not to the underlying mapping (see Appendix F),

The usual approach of constructing a neural network system consists of building and training a number of different configurations and selecting the one having the best performance[136

p. 119].

For the system considered, a number of configurations with one or two hidden layers were tested. The only preprocessing performed on the inputs was normalization, as in the corresponding system of section 11.6.2. Initially one hidden layer was considered. The performance of the system for an increasing number of processing elements was evaluated, by recording how the MSE dropped with the progressive increase in the number of processing elements in the hidden layer, until the error function stopped decreasing significantly. The same was repeated for two hidden layers, although matters here have been complicated by the fact that the number of processing elements in either layer may be increased independently, requiring long hours to train each one of them. The main obstacle was connected with configurations whose MSE was high, even after it seemed to have levelled off. In these cases it is uncertain whether a local minimum or flat region of the error surface was the cause, or whether the global minimum was at a high error level. Hence it was necessary to re-train the network using different parameters to increase the confidence of the results.

Results for these system configurations are reported separately in section 11.8.3.

11.7 Implementation problems and other issues

The previous section focused on the interface of the MLP with the rest of the system, with emphasis on preprocessing. The structure of the required network and how to select it was also elaborated for every type of preprocessing, but little was mentioned on its implementation. Having decided how many layers and how many elements should exist in each, it is necessary to optimize the performance of the network for this particular application, as an untrained MLP is of little use. This requires the construction of a training set of input-output example pairs which define the mapping the network is required to approximate (see section 11.4.2). Then, before training (optimization) may commence, the weights of the processing elements must be initialized to some value, preferably close to the optimal, if possible, to minimize the number of iterations necessary to achieve a low MSE. These implementation issues and problems encountered during training are examined in this section.

11.7.1 Defining a training set

The mapping to be approximated by the MLP is specified at a discrete set of points in its input space, paired with the corresponding output points, on which optimization is based. The inability to form a comprehensive set of such pairs is one of the reasons why this network cannot be used in a number of cases. A suitable training set should consist of a

large set of examples of the desired mapping and should cover a representative sample of all input conditions that may arise in practice, selected according to a uniform probability distribution, if possible.

It became apparent from the beginning that defining such a training set would not have been a trivial task. Firstly, the mapping to be approximated is not known exactly for real EEG records, due to the limited number of transients present, most of which are spikes. Hence, other transients are grossly under-represented and would have a small effect on the MSE, deteriorating their detection. It is desirable for all transients to be represented well, if the network is to be successful. Under-representation is not the only concern here. Because the models for the transients used in the system are only approximate, there are differences between their theoretical outputs and those actually produced when a real signal is applied. Thus a practical 'impulse' generated by one of the $\{F_i(z)\}$ (see section 10.3) may spread over more than one sample and its exact location, with respect to the input transient, may be ambiguous. Besides, the only reference available for the required mapping is the behaviour of the earlier system (section 8.5.2), which is not 100% accurate. Therefore, a different approach was adopted.

Using the signal model of section 7.3 synthetic data records were generated, containing large numbers of transients, with approximately similar probability distributions for each, ensuring that large, similar numbers for all the modelled transients are present. This would eliminate the problems of under-representation already stated.

The use of artificial data has other advantages. The excitation functions for the various transients, particularly their location in time, is known exactly and errors associated with their uncertainty in real records are automatically eliminated. Moreover, for synthetic data records all the elements of the model are known, since they are used for the construction of the signal (section 7.3).

The MLP, being a versatile non-linear modelling tool is expected to learn the inverse of the synthesis procedure, based on the linear modelling elements (section 7.4). This is only possible if a set of examples of the input excitation functions are known during optimization. A training set may then be formed so that the indications of the transients in the various outputs of the inverse systems are the inputs to the system and the instances of occurrence of the transients, become the desired response.

For the artificial signal, the desired mapping is known, since at a given EEG sample, $e(n)$ the values of the excitation functions for the various transients, $c_0(n), c_1(n), \ldots$ are known. A training set could then be constructed by processing $e(n)$ through the inverse linear systems $H^{-1}(z)$, $\{F_i^{-1}(z)\}$, to obtain the inputs to the MLP, $r(n)$, $r'(n)$ and $y_0(n), y_1(n), \ldots, y_N(n)$.

The desired signals are derived from the $\{c_i(n)\}$, having the value of 1 when the corresponding $c_i(n)$ are non-zero and 0 otherwise. A delayed indication for spikes was also included, as this is the desired spike detection signal.

Although in general the desired responses consist of long sequences of zeros with isolated unity-valued samples, in order to cater for deviations from the precise model and the possibility that in real life an impulse could occur in the period between two sampling instants with its energy 'leaking' in the two adjacent samples, each time an impulse was recorded at the nth instant in the excitation sequence, $c_i(n)$, corresponding desired signal, $d_{c_i}(n)$ was constructed by extending the indication of the impulse in the two adjacent sample values. Thus, if $c_i(n) \neq 0$ then $d_{c_i}(n-1) = d_{c_i}(n) = d_{c_i}(n+1) = 1$.

Another advantage of this approach lies in the ability to generate as many artificial samples as required to train the MLP, without depending on the limited supply of real EEG records, which were used primarily as a *validation set* for testing the integrity of the trained network[136, pp. 118–120].

It should be noted that a real EEG record was also used at the final stages of training, in order to observe how large was the deviation of the network parameters for real and artificial data. The changes in the weight values, were, however, insignificant, demonstrating that the MLP was close to the desired optimal (see section 11.8.1). The desired response for the real EEG data was formed from the response of the earlier system (section 10.4) by correcting the signal when the system was in error (removing overdetections and introducing missed transients).

11.7.2 Weight initialization

It is necessary for the backpropagation training algorithm to have an initial value for the required weight vector, from which optimization may commence. In the absence of a better alternative, a small random weight vector is often used. It is, however, clear that this may be far from the basin of attraction of a global minimum of the error surface and in its course towards it, the weight vector may wander around the surface having to find a direction towards the minimum, possibly encountering flat regions and local minima as it moves around the error surface. If, however, there is some more 'sensible' way to select this initial value, so that it is close to one of the global minima of the error surface, many of these problems could be avoided, since once in its basin of attraction, the weight vector would inevitably converge to the optimum.

Such an initial condition is not available, in general, but for the present application the earlier fixed-weight system may be useful, as its performance was better than pure chance and

the weight values are consistent with the function each processing element performed[131].

Initialization with intuitive weights

The generalized fixed-weight system, described in Chapter 10, corresponds to only one type of preprocessing, namely squaring and normalization. For the other types of preprocessing the weights of this system would be of no great use. The task of the decision-detection system may be divided in two functional entities: A preprocessing stage, performed by the hidden layers whose outputs indicate the presence or absence of certain phenomena in the inputs and the detection stage, which combines these to form the outputs of the system.

The preprocessing stage depends on the operation performed on the inputs prior to their application to the MLP. The detection stage is generally independent of the inputs, because it is performed on the output of the preprocessing stage. Hence it is possible to define a set of weights for the output stage to perform the task of detection for all tested configurations and then define the weights of the preprocessing elements for each input treatment independently.

The detection stage consists of a single layer of $N+2$ elements, one for each output of the system. Each one of these combines the outputs of the previous layer. With reference to section 11.5.1 these signals may be interpreted as indications of whether a transient (signals $z_t(n)$, $z_t(n-K)$ and $z'_t(n-K)$) or an impulse (signals $z_{p0}(n), z_{p1}(n), \ldots, z_{pN}(n)$ and $z_{p0}(n-K)$) occurred in one of the inputs. The weights of the processing elements in this stage were set to zero (to ignore the inputs) apart from the ones more relevant to each output, which were set so that the logic functions of the intuitive system were realized, as described in section 11.5.1.

The outputs of the preprocessing stage were generated by weighting the inputs to the network. The initial weights will now be examined separately for every type of treatment of the inputs. Because all elements of this stage perform the same task the operation of one such element will be examined. Each one of these depended on delayed samples of one of the input signals to produce its output. Therefore the weights for all other signal samples are set to zero. The same operation applies to any one of the input signals (normalized versions of $r(n)$, $r'(n-K)$, $y_0(n-K)$ or $\{y_i(n)\}$) which will be referred to as $x(n)$.

For squared inputs (section 11.6.2), there is a single hidden layer with $N+5$ elements. Each element deals with a window of a signal, $\frac{x^2(n)}{\sigma_x^2(n)}$. The weights of such an element were related to those of the earlier system (equation (10.2)), where the values of the weights were set to -1 for all samples in the window of size $2M+1$, apart from the central $2Q+1$ samples where the weights were set to $\frac{2(M-Q)}{t_0^2}$. Hence the bias weight $w_{1pi}=0$.

When the moduli of the inputs are used (section 11.6.2), the weight distribution for a

processing element is similar, but the amplitudes of the central samples are equal to $\frac{2(M-Q)}{t_0}$ (scaled by t_0, instead of t_0^2, for the same statistical significance).

For the hybrid realization, which incorporates the sign of the central samples the above weights are not directly relevant, because every processing element in the previous realizations would be replaced by three, two on in first and one in the second hidden layer, (section 11.6.2). The first two are complementary, one indicating positive and the other negative transients in the centre of one of the input signals. One is identical to the processing element used with squared inputs, with the same central sample weights of $\frac{2(M-Q)}{t_0^2}$ and the rest equal to -1 and can detect large positive amplitudes in the central $2Q+1$ samples. If they are negative, or some positive and some negative but with large magnitudes there may not be a detection, since their weighted combination may be small. Like before, the bias is zero. The second element has a similar weight distribution (outer weights are equal to -1), but with the central ones have negative sign, $(-\frac{2(M-Q)}{t_0^2})$ to detect negative transients.

The $N+5$ elements in the second hidden layer combine the indications for positive and negative transients from the first hidden layer ($z^+(n)$ and $z^-(n)$) to form a single indication of a transient. Their represents an OR operation, defined by the following linear combination $f(n) = wz^+(n) + wz^-(n) - \frac{w}{2}$, where w is a positive weight value (set to 1 in this case), followed by range-limiting. These also receive inputs from all units of the first layer, even for other transients, but these are ignored by initially setting the appropriate weights to zero. For some transients, like spikes, which only occur with one of the two polarities, the OR units would a single one non-zero weight, corresponding to the valid polarity and they are effectively redundant.

Finally for the structure where normalization is the only treatment of the inputs (section 11.6.2), the structure with the two hidden layers is similar to the one for the hybrid realization. The elements of the second layer are still OR operators and the function of those in the first layer is the same as the ones of the previous system. The corresponding weights, though, are not the same. In fact no simple set of weights could be defined for this system, since the inputs may have any polarity and if they are random, as they are in some signals, any trivial weighted sum could be close to zero. In the absence of a better alternative it was decided to set all weights to zero, apart from those in the central $2Q+1$ samples, which were set to 1 for the detection of positive transients and to -1 for negative. The threshold was set to t_0, hence when any of the central $\frac{x(n)}{\sigma_e(n)}$ is above this value the output of the corresponding element would be close to 1. A problem could arise in theory, if the central samples have large amplitudes of different sign, as the resulting sum may be small, but this was not encountered for any valid transients.

One possible drawback, perhaps the only one, of this weight initialization scheme is the possibility that the intuitively derived weights are near a local minimum of the performance function. Such an event would result in a network whose performance is lower than that which could be theoretically achieved. Although this did not appear to occur in practice, it should be noted, as there is no guarantee that this form of initialization is generally better than random weight selection. Even if the initial condition is near a local minimum, it is one at a low error level and could serve well in practice, arguing that a random initial weight could also be near a local minimum and lead to a sub-optimal solution.

Small random initial weights

The above scheme was compared to the traditional initialization of the weights of the network to small random values. Hence this alternative approach was also put to the test. Typically, the weights are made small, so that the outputs of the processing elements are close to zero and the sigmoidal non-linearities operate in their 'linear' region away from saturation.

The approximation of the mapping by this network would be a nearly linear one, an almost flat surface. During optimization, the magnitudes of some of the weights is expected to increase, driving the corresponding elements' outputs into saturation and making the mapping more 'creased'.

The magnitude of the initial random weights (including the bias units) was selected so that if all inputs have their maximum achievable values (close to unity) and weights happened to have polarities such that their application would contribute to the linear output (the summation) of the element in the same way, then the corresponding output would not be driven in one of the narrow tails of the non-linearity. This range was chosen to be from 0.25 to 0.75, which translates in values of the linear sum in the range of approximately [-1.1,1.1]. These values were used to define the range of the weights so that their linear output when all are combined is never above this value. Thus for an element with N inputs, the magnitude of the weights should be such that $N|w| = 1.1$. Thus the initial weights were selected separately for each element from a uniform distribution having a probability density function:

$$p(w) = \begin{cases} \frac{N}{2.2} & -\frac{1.1}{N} < w < \frac{1.1}{N} \\ 0 & \text{otherwise} \end{cases} \tag{11.14}$$

Random initialization of the weight vector was the only means of starting the algorithm for the structure described in section 11.6.3, which is completely disconnected from the earlier system.

11.7.3 Software simulation of the MLP

A neural network may be implemented either as a purely hardware system, or on a separate dedicated neurocomputer. In both these cases the user communicates with the network through a conventional computer via a suitable interface.

In recent years, however, a number of ANN simulators have appeared. These are packages that run on a general-purpose computer usually having a graphic user interface to enable the interactive definition, construction, training and testing of a network. They typically offer a comprehensive repertoire of standard network topologies and learning rules. The progress of the network during training may be monitored by inspecting the MSE and the weights on graphical displays. Some ANN simulators offer the option to save a trained network as a program module which may be linked with a user's program and called as a routine, when necessary, thus providing a cheaper option (at a price of reduced speed of operation), than purchasing a dedicated neurocomputer.

Initially, the latter approach seemed the most appropriate and efforts were directed into finding an appropriate ANN simulator. Further consideration of the system as a whole revealed that the network could not be trained in isolation, as it is an integral part of the system. Referring to the block diagram of the system (Figure 10.1), some of the inputs to the MLP are generated from the EEG signal after processing through the inverse modelling systems. Most of the outputs of the MLP control the operation of other elements of the system, namely the transient modelling systems. But their outputs are subtracted from the EEG and $e'(n)$ is processed to generate $r'(n)$, which is also an input to the network. In other words, there is *feedback*. This is undesirable, since incorrect operation of the MLP, while training, would almost certainly affect the response of the systems it controls. As a result, a different sequence $r'(n)$ may be generated during and after training, invalidating the learning process. Recurrent backpropagation neural networks, where feedback is intentional, have been developed[136, pp. 183–191]. Their theoretical basis, however, is still weak and a lot of ground has to be covered before they are understood well enough to be used with confidence. Another problem had to do with the details of many of the packages,

Even if this problem was overcome, ANN simulators, in general, do not provide enough information, especially on the exact algorithms used for optimization and weight initialization schemes. Hence, although these may be excellent for the fast development of neural network solutions to problems, they tend to hide many of the implementation details. These may be unimportant in most cases but ignorance of their operation is rather prohibiting for the replication of the work presented.

An alternative approach consisted of writing a library of routines, which may be linked

directly with the existing system and facilitate the construction and training of a neural network. A brief description of this library appears in Appendix G.

11.8 Results

The design of a Multi-Layer Perceptron system to overcome some of the limitations of the earlier decision-detection unit raised numerous new questions. Because of the complexity and the non-linearity of the system and the function it attempts to approximate, these could not be answered by considering the system in general. Therefore, any attempts to address these questions were based on building and testing the various configurations.

Perhaps there is no end in the number of issues raised when using the system in question. The tests carried out concentrated on the following points:

- Whether a system trained with artificial data is consistent with one trained with real EEG signals
- The effect on the performance of the system with various types of input treatments
- How weight initialization affected the convergence properties of the training algorithm and the performance of the system
- How the systems based on the earlier structure compared with those whose architecture was chosen experimentally
- The effects on the performance in terms of size and accuracy for the experimentally selected configurations
- Whether the systems based on the earlier structure are sufficiently large to perform the required task and, if they are, how does their size compare with the smallest system having the same performance.

The investigation of all possible system configurations for all of the listed aspects would have been impossible as well as unnecessary. Hence only a sample of all possible combinations were investigated, but even this proved a formidable task. The conclusions from their results may be easily generalized.

Tests were grouped into three sections, covering training with real and synthetic data, systems with various types of treatment of the inputs and weight initialization and tests on networks having different sizes.

All systems tested were concerned with the detection of epileptic spikes and one more transient, as described in section 10.5, with the earlier decision-detection unit replaced by an MLP. Due to the considerable computational load for both training and running the MLP, it was necessary to reduce the length of the windows of its input signals. The signal buffers used remained unaffected, but only a fraction of their centred values were used. The value of Q was reduced from 2 to 1, resulting in 3 central samples, whereas M was reduced from 32 to 5, making the effective length of the buffer equal to 11. As demonstrated later, despite these restrictions, the MLP outperformed the earlier system.

11.8.1 Training with real and synthetic EEG records

Some of the primary concerns when using the MLP in the current system was the unavailability of large numbers of transients, their rarity and their short duration. This had a double impact on the issue of the adequacy of a training set. As explained in Appendix F.2 an inappropriate training set could result in an inferior performance of the system. Under-representation of transients is a fact, because these occupy only a small fraction of the signal used during training and also the formation of a training set for a real data record (section 11.7.1). Moreover, due to the limited number of real EEG records, it would not have been possible to reserve any for validation if training was based entirely on real data records. Training using artificial data records was an option that solved these problems, but the consistency in the behaviour of a system so trained when used on real records was questioned.

To investigate the last issue, the system was optimized using two schemes. The first consisted of training the system with artificial data only. These were generated from the model of the EEG signal, according to the prescription given in Appendix D. Because this provided an unlimited supply of input and desired signals, different data were used for training and testing the MLP. Five records were constructed, each consisting of 10000 samples, containing approximately 25 spikes and 20 exponential decay transients each. Four of those were used for training, by recycling them in a random order through the network. An adaptation rate of $\mu = 0.1$ was employed throughout the learning procedure in conjunction with a momentum term of $\alpha = 0.9$ (section 11.4.2). Training was initialized with random weights, following the scheme of section 11.7.2 and was terminated when the difference between the estimated MSE between two successive iterations was less than 0.1% of the sum of powers of the three desired output signals. After that, the trained system was tested with the remaining artificial EEG record and the available real records. The results of the test appear in table 11.1.

Record			A				B		C		D		E	
Duration in seconds			62				210		130		123		242	
Number of spikes present			25				52		8		10		0	
Number of transients present			20				0		0		0		0	
Muscle artifacts present?			no				some		yes		yes		many	
MLP			Detections											
Configuration			Spikes		Transients		Spikes							
#	Input Treatment	Topology§	√	×	√	×	√	×	√	×	√	×	√	×
1†	$\frac{x^2}{\sigma_e^2}$	49-6-3	25	0	20	0	52	2	8	4	10	8	0	34
2‡	$\frac{x^2}{\sigma_e^2}$	49-6-3	25	0	20	0	52	2	8	4	10	6	0	33
3†	$\frac{x}{\sigma_e}$	49-12-6-3	25	0	20	0	52	2	8	2	10	5	0	28
4‡	$\frac{x}{\sigma_e}$	49-12-6-3	25	0	20	0	52	2	8	2	10	3	0	30

Notes:

§: This indicates the number of processing elements in each layer, separated by dashes. The first is equal to the number of inputs (49) and the last to the number of outputs (3)

†: Trained with artificial data only

‡: Trained with artificial and real data

Table 11.1: Results for simulated and real EEG records for the extended system structure

The alternative scheme used the same learning parameters and random initialization of the weights. Initially the artificial set was used. Training was interrupted when the error level fell at 5% of the desired signal power and the training set of four artificial records was replaced by a single real EEG record, containing 10 epileptic spikes. The desired outputs of the MLP were constructed from the response of the earlier system (section 10.4) with small amendments performed manually to eliminate false detections and to introduce the ones missed by the system. Testing this system used all real data records and the remaining artificial EEG (Table 11.1).

Two MLP configurations were considered. The first used squaring, as part of the preprocessing, whereas the second employed only normalization.

The weights of the trained networks with artificial and real data were also recorded. These were almost indistinguishable by visual inspection after being plotted. To quantify their similarity, a summary measure was computed. The weights sequences may be regarded as vectors in the 'weight space' $\mathbf{w}_r$, for training with real and $\mathbf{w}_a$ for training with artificial

data. The modulus of their difference was first computed (the square root of the sum of squared differences) and then divided by the modulus of their mean, $\bar{\mathbf{w}} = \frac{1}{2}(\mathbf{w}_a + \mathbf{w}_r)$:

$$\delta_{ar} = \frac{\|\mathbf{w}_r - \mathbf{w}_a\|}{\|\bar{\mathbf{w}}\|} \tag{11.15}$$

The value of this measure was approximately 0.001 for squared and normalized and 0.034 for normalized inputs.

11.8.2 Performance of systems with different preprocessing and weight initialization

Most of the remaining issues related to the use of the backpropagation decision-detection unit are related to the fundamental question whether the inputs to the MLP should be processed beforehand, and the subsequent selection of an appropriate input mapping. In order to investigate this, a number of network configurations were tested, covering the operations on the inputs described in section 11.6.2. Four types of preprocessing were performed (squaring and normalization, modulus and normalization, a hybrid normalized structure and normalization only). The network configuration corresponding to each type of input treatment was related to the earlier intuitive structure, as explained in section 11.6.2.

Training was carried out with artificial data only, following the scheme used in the preceding section, using the same parameters for accomplishing optimization as before ($\mu = 0.1$ and $\alpha = 0.9$).

Two initialization schemes were tried for each configuration. In the first, the initial values of the weights corresponed to the earlier system, following the reasoning of section 11.7.2. The second was the more traditional random weight initialization, whose details appear in section the same section.

The eight input processing and weight initialization combinations were trained and tested with both real and artificial data, as before. The number of correct and incorrect spike and transient detections were recorded as well as the number of iterations required for the network to converge, so that the change between two successive iterations was less than 0.1% of the average power of the three desired output signals, as in the previous section. This number is an integer multiple of 10000, which has been the length of every sequence of samples used for training. The results of these tests have been summarized in Table 11.2.

Typical waveforms during the test of one of the configurations appear in Figures 11.4 and 11.5. These show the input EEG signal, the outputs of the inverse modelling systems, the estimated spikes and transients, the transient-free signal and the spike detections. The

Record				A				B		C		D		E	
Duration in seconds				62				210		130		123		242	
Number of spikes present				25				52		8		10		0	
Number of transients present				20				0		0		0		0	
Muscle artifacts present?				no				some		yes		yes		many	
MLP Configuration			No. of iter-	Detections											
				Spikes		Tran.		Spikes							
#	Preproc.	Topology§	ations	√	×	√	×	√	×	√	×	√	×	√	×
1†	$\frac{x^2}{\sigma_e^2}$	49-6-3	10000	25	2	20	0	52	2	8	4	10	8	0	34
2‡	$\frac{x^2}{\sigma_e^2}$	49-6-3	60000	25	2	20	0	52	2	8	4	10	7	0	34
3†	$\frac{\lvert x\rvert}{\sigma_e}$	49-6-3	10000	25	2	20	0	52	2	8	6	10	11	0	42
4‡	$\frac{\lvert x\rvert}{\sigma_e}$	49-6-3	50000	25	2	20	0	52	2	8	6	10	11	0	45
5†	hybrid	49-12-6-3	30000	25	0	20	0	52	2	8	4	10	4	0	31
6‡	hybrid	49-12-6-3	50000	25	0	20	0	52	2	8	4	10	4	0	33
7†	$\frac{x}{\sigma_e}$	49-12-6-3	60000	25	0	20	0	52	2	8	2	10	5	0	28
8‡	$\frac{x}{\sigma_e}$	49-12-6-3	80000	25	0	20	0	52	2	8	2	10	5	0	29

Notes:

§: This indicates the number of processing elements in each layer, separated by dashes. The first is equal to the number of inputs (49) and the last to the number of outputs (3)

†: Initial weights corresponding to the earlier system, as prescribed in section 11.7.2

‡: Random initial weights

Table 11.2: Results for different input treatments and MLP weight initializations

system corresponds to the seventh configuration of table 11.2. These are similar to the waveforms for the extended system, presented in section 10.5 as well as those for the earlier spike detection system, described in Chapters 8 and 9.

11.8.3 Results for the different MLP sizes

The final set of tests had a different aim. To reveal whether a working system based on the assumed functionality of its units, described in sections 11.5.1 and 11.6.2, is actually adequate to perform the required task and whether the intuitive configuration is close to the minimal in terms of number of processing elements and layers. To investigate this issue, different backpropagation neural network configurations were constructed, according to section 11.6.3.

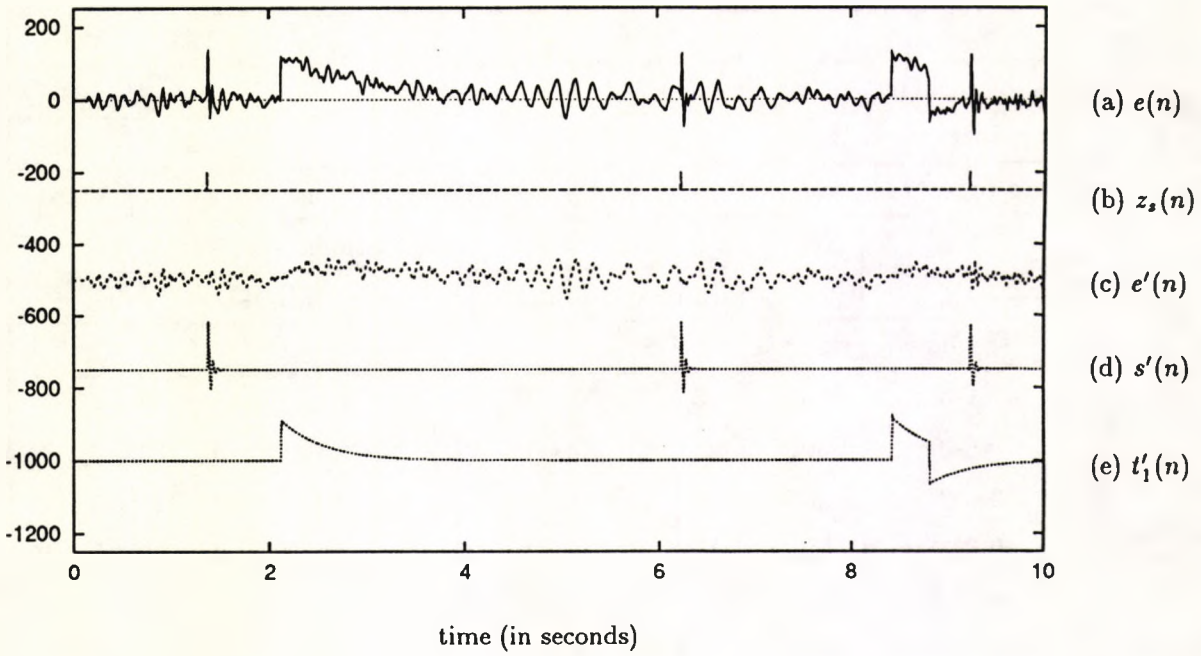


Figure 11.4: Behaviour of the system with artificial input data

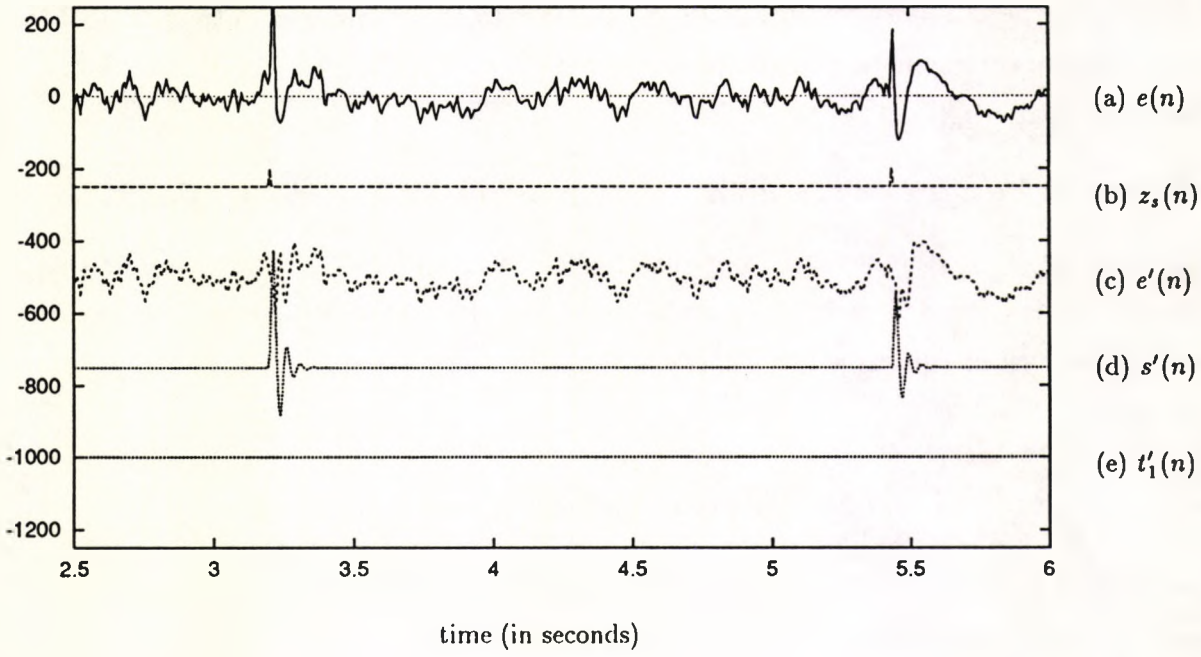


Figure 11.5: Behaviour of the system with real EEG input

Layer 2 / Layer 1	n/a†	2	3	4	5	6	7	8	10	20
2	19.21†	24.92	22.55	23.98	22.64	21.59	22.06	21.96	22.05	21.82
4	11.45†	17.59	16.46	19.24	18.56	18.23	18.09	17.33	17.16	17.19
6	7.32†	12.70	11.53	9.96	9.74	9.78	9.73	9.69	9.47	9.45
8	5.77†	10.67	8.22	8.53	8.45	8.33	8.29	8.05	7.78	7.43
10	2.36†	8.50	4.19	3.29	1.97	0.06	0.03	0.03	0.01	0.01
12	1.24†	7.61	3.65	0.75	0.02	0.01‡	0.01	0.01	0.01	0.00
14	0.96†	7.20	2.21	0.36	0.01	0.01	0.01	0.01	0.01	0.00
16	0.84†	7.13	0.98	0.10	0.01	0.01	0.01	0.01	0.00	0.00
18	0.32†	6.68	0.50	0.07	0.01	0.01	0.01	0.01	0.00	0.00
20	0.27†	5.97	0.32	0.06	0.01	0.01	0.01	0.00	0.00	0.00
30	0.09†	5.66	0.11	0.06	0.01	0.01	0.01	0.00	0.00	0.00

Notes:

All configurations have 49 inputs and 3 outputs

†These configurations have only one hidden layer

‡This setting corresponds to the system derived from the earlier intuitive system

All entries have been rounded to 2 decimal places. The zero entries on the bottom-right corner of the table correspond to MSE values of less than 0.005

Table 11.3: The final MSE for various network configurations

Although all types of preprocessing could have been tested, this is unnecessary, since it would not have provided any further indications to the suitability of the intuitive systems. Besides, the tests carried out in this section are not conclusive, since the general case of the detection of any number of transients is not covered here. Hence only one case was investigated, the network whose inputs are simply normalized by the standard deviation of the input signal, $\frac{x(n)}{\sigma_e(n)}$. The intuitive configuration had two hidden layers with 12 and 6 elements each, 49 inputs and 3 outputs. The object of the test was to vary the number of layers and processing elements, train the resulting configuration and record the final MSE at the end of training. The configurations considered had either one or two hidden layers having up to 30 processing elements each. Training was performed with the momentum variant of the backpropagation algorithm, as in the previous sections. The initial weights were made random.

Training followed the procedure of the previous section, using artificial records only. The resulting normalized MSE for every configuration is recorded in table 11.3.

11.9 Discussion, comments and conclusions

The objectives of the tests whose results have been listed in the preceding section were connected with the performance of the system, compared with the previous intuitive structure and with various issues related to the selection of the appropriate size, weight initialization scheme and treatment of the inputs for the MLP structure. These results are analysed here.

As with the results of the previous Chapters, the small number of EEG records used in the tests limited the level of confidence of analysis. They give, however a good indication of how the system would operate in practice. Besides, the proposed analysis procedures constitute a rather flexible methodology, whose parameters are derived from the available signals. Hence, as new EEG records become available, these could be applied to modify the parameters of the system and hence improve the performance of the suggested system.

11.9.1 Comparison of training schemes

The first test, described in section 11.8.1 attempted to demonstrate whether training with real EEG records had a significant impact on the performance of the system, compared to training with artificial records generated by the model of the EEG signal. Two configurations were tested separately with artificial EEG data and with a mixture of real and artificial records.

The results, of the application of the trained systems to a number of records, summarized in table 11.1 revealed that there was little difference between the systems trained with artificial data records and with real data. Configurations 1 and 3, which were trained only with artificial data performed equally well compared to configurations 2 and 4, which were trained using real EEG records. There were no noticeable differences in the detection of spikes and transients for the four combinations when tested on artificial data. A similar observation was made for real records containing spikes. All spikes present in records B, C and D were detected. A small number of false detections occurred in each record, related to some muscle artifacts which resembled spikes. Both systems trained with either real or artificial data produced the same number of overdetections for records B and C. Small differences were observed in records D and E, which contained more profound artifact activity. The results for record D favoured the systems trained with real data but those for record E were inconclusive, as for the system with normalized inputs training with real EEG signals appeared to have worsened the performance slightly.

The differences in the response of the two types of training have not been significant. The disagreement is in a small number (1 or 2) false detections and there is no evidence that one

training scheme was better than the other. Variations are believed to be circumstantial and more extensive testing with a larger set of records is expected to confirm their equivalence. Of course this may reflect the similarity in the model for the spikes and real epileptic spikes, which was evaluated from real spikes using the methods of section 8.4.4. Therefore artificial spikes have a good resemblance to real spikes and hence training with artificial data is indirectly related to the real ones, a fact that could explain the observed similarity in the results.

The likeness in the two systems is also apparent from the closeness of the weight vectors for the systems trained with real and artificial data. The comparative measure, δ_{ar}, given by equation (11.15) is a characteristic indication of their similarity. For the MLP with normalized and squared inputs δ_{ar} had a value of 0.01. This value is small, indicating that the weights after the completion of training with artificial data were close to those obtained when real EEG spikes were included. δ_{ar} quantifies the proximity of the two weight sets in weight space. In this case it shows that they lie on sphere whose radius is only 1% of the distance between its centre and the origin, hence they are close to each other. For the normalized input case, $\delta_{ar} = 0.03$, which is higher than the in previous case, but is still relatively small.

In conclusion, no significant differences were observed between the weights of the MLP trained with either purely artificial data or a mixture of artificial and real EEG signals. Thus, when real data are scarce, it is still possible to attain a reasonable performance if the system is trained with synthetic data, based on a signal model for the transients in question. This, of course, requires a reasonable model for the signal, especially the transients to be detected.

11.9.2 The effects of input preprocessing

Having established the suitability of the training set, the effects of input preprocessing on the performance of the system were assessed. Four types of preprocessing were used, all described in section 11.6.2. The test was described in section 11.8.2 and the results summarized in table 11.2. Although this table contained other pieces of information, which are analysed later, this section is concerned with the performance of the various configurations.

All spikes and transients of the synthetic record A were detected correctly by all systems, irrespective of the treatment of the inputs. Two false spike detections were, however, registered by the system with normalized and squared inputs (table entries 1 and 2) and by the configuration which used the moduli of the normalized inputs (entries 3 and 4). These detections are associated with two spikes, which were introduced with reverse polarity in the

record, to verify the response of the four configurations to signals with a shape similar to one of the modelled transients, but a direction contradicting its definition.

The inability of the systems mentioned to distinguish polarity is inherently linked with the type of preprocessing performed. Squaring and the modulus operations removed the sign of the signals prior to their application to the MLP, which is then unable to distinguish between signals with opposite polarity. False detections may then be observed, when it is known that a particular transient always occurs with the same polarity. This limitation was not observed for the other modelled transient, because occurrences in either direction constitute valid transient events. Similarly, no difference was present in the first real record, B, since this did not contain any spikes with the incorrect polarity. It was evident in the rest of them, especially in the results for record E, which contained artifacts, some of which imitate epileptic spikes. The smaller number of false detections for the systems which use polarity (5 to 8 in table 11.2) is clearly linked to their ability to reject signals which do not have the correct polarity. The improvement in artifact rejection was small but noticeable.

Evidently, the systems with no preprocessing (7 and 8) performed better than the rest. This is believed to be linked with the maintenance of all information in the inputs of the MLP which enabled the network to form its own internal operations, rather than those imposed externally. Nonetheless, the differences from the hybrid structure were small. The latter provided the network with most of the information in the central samples of each signal window, on which the detection of transients is primarily based. As there is no improvement in the performance of the hybrid structure compared to one with no preprocessing and both use a network of the same size, one may conclude that the latter is preferable. The hybrid structure may prove useful if the system is extended to more transients, when this type of preprocessing may simplify the task of the MLP, especially if correlations exist between the modelled transients. Squaring the inputs came third in the rank of performance, with the modulus preprocessing being the last. The extraneous false detections by the modulus-preprocessed system were associated with some boundary cases, like artifacts causing transient disturbances registered with small amplitudes in the inputs to the MLP. These have caused detections, because they were possibly near the threshold value for the nodes of the first hidden layer, although this is difficult to verify. In other words, the clusters of transients and non-transients in input space are closer for the modulus operation. Perhaps squaring expanded the range of the inputs, stretching these clusters and, thus, making their discrimination easier.

11.9.3 Convergence of network initialization schemes

One of the elements recorded in Table 11.2 during the tests for the systems with different preprocessing was the number of iterations until convergence of the MLP was achieved. This has been recorded at the end of every record used for training. It represents the number of samples from initialization until the estimated MSE for two consecutive iterations changed by less than a preset value (see section 11.8.2).

The objective then was to observe the convergence characteristics of the backpropagation learning algorithm for a given network architecture with two initialization schemes. The four tested networks were initialized either using the hints derived from the earlier intuitive system (section 11.7.2) or using random values. The tested networks with the first type of initialization appear as odd-numbered entries in Table 11.2, whereas the second as even-numbered entries.

When consecutive entries in the aforementioned table (representing the same setup with different initialization) were compared in terms of the number of iterations to obtain a satisfactory performance, some of interesting observations were made. In general, the systems initialized using the concepts of the earlier system required fewer iterations to achieve an acceptable error level than their randomly initialized counterparts. This supported the view that the considerations that led to the development of the earlier system (section 10.4) were 'in the right direction', although the system was sub-optimal. Hence, training acted as 'fine tuning' of the MLP parameters, rather than as a blind search for an appropriate set, which is the case with random initial weights.

The improvement in convergence, when using the preset instead of the random initial weights, was more apparent for the system whose inputs were normalized and squared ($\frac{x^2}{\sigma_e^2}$). Here, a speed-up by a factor of 6 was recorded, which shows that for the given system the initial values were close to the optimal. In fact the weight vector was in the basin of attraction for one of the minima of the performance surface. The improvement was somewhat lesser for the system with $\frac{|x^2|}{\sigma_e}$ (a reduction of the number of iterations by a factor of 5). The improvements for the hybrid input treatment was less than 2 and for $\frac{x}{\sigma_e}$ it had the marginal value of $\frac{3}{2}$.

In terms of performance, the two initialization schemes were comparable. Small differences were observed, which were more profound in the results for record E, for which the number of false spike detections appeared to depend on the initialization scheme employed. These differences were, however, small and did not favour one or the other initialization methods, since for the first system, the number of false detections was the same irrespective of initialization, for the next two the number was higher with random initial weights, whereas

for the last it was lower.

In general, initializing the network to the preset weights, suggested by the earlier system, was more favourable than the conventional random initialization procedure, not only because it accelerated convergence without any apparent loss in performance, but also because it reduced the risk of encountering local minima and, possibly, flat regions which could prevent the generalized delta rule from finding an optimal solution. This cannot be guaranteed, but it is evident that there is a higher probability that the algorithm may encounter such a pathological condition during its progress, if the initial weight vector is far from one of the global minima of $\xi(\mathbf{w})$. In general, for the systems tested, the intuitively derived weights appear to be near an optimal, something that cannot be claimed for a weight vector selected at random.

Comparison between the final weights for the two initialization schemes was not possible. Despite their equivalent performance, the two weight sets for all tested configurations were in three out of four cases very different. This is not unexpected, since gradient-based algorithms attempt to find the nearest point of zero derivative, irrespective of the level of the MSE it represents. Thus, different initial weights may lead to either local minima, or to different global minima. As Chen and Nielsen[42] demonstrated there are multiple global minima in the performance surface of a given MLP application. Hence, starting from different initial weights is likely to lead to different, but equivalent solutions. In the tests carried out there is a small possibility that some of the solutions correspond to local minima. If this is the case they must be at an error level close to the theoretical minimum, otherwise the observed performance level would not have been possible.

11.9.4 Comparison with the earlier detector

The primary reason for the development of the MLP decision-detection unit was the poor performance of the earlier detector, which was based on the apparently successful spike detector of the previous Chapters. Hence it was considered important to compare the performance of the MLP-based system with both earlier systems.

The results of the earlier transient detector were listed in Table 10.1. Comparing these with the results of any of the tested combinations of treatment of the inputs and MLP configurations (Table 11.2), there is a clear advantage in favour of the MLP-based systems. Even the worst performance, achieved by the MLP system with modulus preprocessing (entries 3 and 4 in the aforementioned table) was markedly better than that for the earlier system, which failed to detect all spikes and transients even in artificial record A. For real EEG records the response of the earlier system was clearly not very successful, since it missed

about a fifth of the spikes present in records B, C and D and produced a large number of false detections. On the contrary, the MLP-based systems detected all spikes in these records with a smaller number of false detections. The reduction in the number of false detections in record E, which contained artifacts, was also noticeable for the optimized backpropagation systems, being about 40% less than that for the intuitively constructed system. In conclusion, the introduction of the MLP in the system made the difference between a moderate and a fairly good performance, justifying the extra processing for off-line training and on-line operation of the system. The MLP systems are considerably more flexible than the earlier system and are expected to be more versatile in possible future extensions of the system.

The response of the system to spikes was comparable with that of the successful spike detectors of Chapter 9. Those were only concerned with the detection of epileptic spikes and disregarded other transients. Its results for the four real EEG records appear in table 9.1. Only the first system in that table is comparable with the MLP-based structures, since all other configurations contained adaptive sub-systems not corresponding to the configuration used in conjunction with the MLP decision-detection system. Both systems detected all spikes present in the records they were applied to. In terms of false detections, they were also comparable. Some of the MLP-based systems produced a higher number of false detections than the earlier structure, but this was compensated by other configurations, which had a lower number of false detections. On average the MLP structures had a higher number of false detections, especially for the artifact-contaminated record E, but the increase has not been significant.

11.9.5 Sufficiency and 'minimality' of the intuitively defined configurations

The final two questions raised at the beginning of section 11.8 were connected with the size and configuration of the neural network. The number of layers and elements in each layer of the MLP were selected using the earlier intuitive system as a guideline. This method of selection has not been based on any formal criteria (these do not exist anyway) and raised two questions: Whether a system so selected has a sufficient number of parameters to perform the designated task satisfactorily and whether there exists a smaller MLP that can do the same job, in which case there is redundancy in the proposed system. To investigate these issues, MLPs with different numbers of processing elements and layers were tested and the MSE, after training was completed, was recorded for one of the input conditioning schemes (see section 11.8.3). Table 11.3 shows the normalized mean-squared error function for one or two hidden layers with various processing elements in each.

The question of sufficiency of the network proposed by the intuitive system was answered positively by its performance, but the results of Table 11.3 support this claim by further demonstrating that the MSE achieved by this system (corresponding to a 2-layer structure with 12 and 6 elements respectively) is quite close to that with larger numbers of elements. Indeed, an improvement by 1% required almost double the number of elements in each layer to take effect.

The question of minimality was more difficult to answer, but the reported MSE values suggested that the intuitive structure is close to the minimal system with acceptable performance. Looking into the rows of Table 11.3 it is not difficult to establish that for a given number of elements in the first layer, the MSE decreases with the introduction of more elements in the second layer. The error level, however did not drop significantly, until the number of elements in the first layer reached 10. For configurations having more elements in the first layer there is a marked drop in the MSE from an unacceptable level to a small fraction. This is associated with the transition from a network too small for the task to one larger than necessary, for which there is no marked improvement in the MSE. Somewhere close to this transition point there exists system configurations which perform quite well, yet do not have an excess of processing elements.

A similar observation can be made for the second layer (looking at the columns of Table 11.3). In this case, acceptable performance required at least 5 processing elements.

As a result, for the error level to be small there was a minimum requirement of about 10 elements in the first hidden layer and 5 elements in the second. The configuration based on the intuitive system had 12 and 6 nodes respectively, which was quite close to the minimum requirement.

These conclusions about adequacy and minimality only apply to the system tested (for the detection of spikes and one more transient). If this is extended to detect many transients, it is possible for the intuitively derived system to be excessively large. For example, a smaller network could exploit possible differences and redundancies which would enable one node to perform clustering or separation of more than one input signal. Similarly, there is a possibility that inter-correlations between the inputs may lead to more complicated boundaries between the different classes of input conditions, thus requiring a larger network. In any case, the system suggested by the intuitive system may be used as a guideline, which can be trimmed, or extended to make an adequately large working system.

11.9.6 Concluding remarks

Generally speaking, all MLP-based configurations tested performed well with both synthetic as well as with real EEG data. It is interesting to note that in spite of training the network mainly with synthetic data, its behaviour with real EEG signals was quite good. This may be due to the small number of data records considered in this study, but it demonstrates that the neural network approach has the capabilities to perform in a reasonable manner, even when there are differences between the training and testing sets. This is a general characteristic of mapping neural networks, which behave as generalized regression or interpolation systems[136, pp. 120–121].

Chapter 12

Comments and conclusions

This thesis was concerned with the analysis of EEG signals suspected of containing specific patterns (spikes) connected with epilepsy. In the first two chapters it was shown that although there is a connection between the recorded signal and the activity of underlying neural activity, this is by no means straightforward and is influenced by many factors. Nonetheless, the presence of specific patterns in the EEG, which are characteristic of the general operation of the brain and its disorders, like epilepsy have been identified. Despite its importance in the diagnosis of epilepsy, as shown in Chapter 3, brief routine EEG recordings sometimes prove inadequate for the reliable assessment of patients and for controlling anticonvulsant medication, where reaching a therapeutic dosage while minimizing the possibility of side-effects of the drugs used are equally important for the successful treatment of constitutional epileptic attacks.

To overcome practical problems associated with the large volume of data produced by the more effective long-term EEG monitoring, automatic analysis was proposed. From a review of this subject appearing in Chapter 4, it became obvious that authors do not all follow a specific methodology, but agree that to achieve reliability of an automatic system, it is necessary to work closely with medical experts and to find a way to use the experience of the latter. This is not an easy task, as medical experiece cannot be expressed exactly and medical definitions are often vague and are based on subjective, visual observations, rather than on objective criteria, based on understanding in detail the underlying processes.

With these considerations in mind, two main problems were identified, producing accurate definitions and finding a way to utilize medical experience in an automatic system. As both problems are linked to the internal description, the model, of the signal, used by human analysts, producing a suitable model for the EEG signal became the initial target of this work. Such a model could suggest a way for the analysis of the signal without considering

specific examples of EEG data records.

12.1 Discussion of methods used

In this Thesis there have been two approaches for the solution of the problem, both withing the common framework of signal modelling, but with distinct characteristics. The generalized modelling approach is preferred over the early ad hoc methods, but both methods will be discussed briefly here concentrating on their usefulness, effectiveness, elegance, clarity and simplicity.

12.1.1 The early methods

In the beginning, some models were proposed, based on the appearance of the signal, especially the spike-and-wave epileptic precursor. Some of the early models, presented in Chapter 5 did not lead to an analysis procedure. They motivated the study of certain attributes of the signal that led to the development of a class of simple single-wave models, on which two useful spike detectors were based.

These were based on the time derivatives of the signal, as many of their predecessors, but have one fundamental difference from them. They were based on a rigorous analysis procedure, founded on a model rather than on heuristics extracted from specific test signals. As a result, their functionality could be predicted and even when they did not perform as expected, it was possible to give a detailed and accurate explanation, based on the proposed models. Some of limitations of the initial detector were solved by the state-space approach of the second attempt, which was less sensitive to muscle artifacts than the first.

The main advantage of these methods, apart from the rigor of the approach is their simplicity, and their capability of operating in real-time on relatively modest hardware.

Among their disadvantages is the absense of neurological significance in the proposed models making the intervention of medical experts difficult. The early specialization of the analysis, the simplicity of the model and the use of application-specific techniques made these techniques inflexible and difficult to assimilate to the function of a human expert. As a result, the approach is not easy to describe in a general way to medical experts, and its performance cannot be enhanced without changing its fundamental concept.

12.1.2 The generalized modelling approach

It became obvious that ad hoc models, constructed without considering the generating mechanisms of the signal and neurological definitions would not be a suitable starting point for collaboration with human experts from the medical field. Therefore, a composite model was proposed. This was based on existing definitions used by neurophysiologists to describe the EEG, like background activity, spikes, transients and artifacts. These terms were examined both in terms of their generating processes and their nature (amplitude, persistence, shape, etc.). From this examination these terms were formulated using general signal processing techniques. Specifically, these were considered as components of the (composite) EEG signal. Each one was described in terms of a system with (generally) constant or rarely changing characteristics and a simple input signal. Hence, the 'uncertainty' of a signal (time of occurrence, amplitude or power) was condensed in the input and the known characteristics in the system.

This approach is general, covering all the components of the EEG and not specific to neither epileptic patterns, nor to any particular signal processing method. The significance of the latter is its ability to describe the signal in a way familiar to the EEG analysts while maintaining their definition for the benefit of the signal processing analysts.

12.2 Conclusions

Evidently, the detailed signal modelling approach is far superior to the earlier method. Not only has it accomplished its primary function, the detection of epileptic spikes in the EEG but has also proved its potential for improvement and generalization. These issues will be elaborated in this section, demonstrating that the methodology suggested may form the basis of a continuously improving system, capable of analyzing the EEG in general, useful to individual medical analysts, with plenty of scope for application to other signals.

12.2.1 Reliability of the detector

First of all, it should be noted that in terms of spike detection the results of the system, based on an elaborate signal modelling approach were not considerably better than those for the simpler system, based on derivatives, described in section 6.5. In fact the performance-to-complexity ratio of the earlier detector was much better than the signal modelling approach. The latter, however, has other desirable properties making it a more versatile analysis tool, applicable not only for the analysis of the EEG, but also for other signals, not necessarily of

biological origin.

12.2.2 Generality

This system does not depend on the shapes of the signals involved, nor on heuristics derived by making specific assumptions about the nature of transients. It rather describes a *procedure* to follow for the modelling and detection of transients, whose approximate characteristics are known a priori, when these are embedded in a continuous signal which may have possibly unknown characteristics. The transients in question, spikes in the case of the EEG, are modelled in a consistent way and, unlike the earlier system, where the model was used only implicitly to develop the method, here it constitutes an *integral part* of the analysis procedure.

12.2.3 Useful by-products of the rigorous system

Signal analysis

At the centre of the suggested EEG analysis method is signal modelling. The background activity $b(n)$ is modelled by the prediction error system $H^{-1}(z)$ and the error sequence, $\hat{u}(n)$. Spikes are described by the all-pole system $G(z)$ and the impulse excitation sequence, $\hat{d}(n)$. In this manner, the power signal component, $b(n)$ is separated from the energy components, spikes and other transients[128, 129, 130].

The system offers the ability to study these signals *separately*, using their parametric representations in terms of a transfer function and an input sequence. This is related to more conventional methods of treatment of the EEG[133].

Synthetic spectrum evaluation

It is often advantageous to visualize the EEG in the frequency domain, where dominant frequency components and other characteristics are easier to detect[130]. The presence of transients makes the interpretation of the spectrum difficult, because they are wide-band signals, contaminating a broad area of frequencies. This has been solved by the proposed system, because spikes and the background activity are characterized by separate input sequences and transfer functions. Hence, their corresponding frequency characteristics may then be studied independently. In fact, the power spectral density, $S_{bb}(\omega)$ of the background activity may be computed from $H(z)$ and the power $\sigma_{\hat{u}}^2$ in the white sequence $\hat{u}(n)$ by

evaluating $H(z)$ around the unit circle in z as suggested by equation 7.8:

$$S_{bb}(\omega) = \sigma_{\hat{u}}^2 H_i(z) H_i(z^{-1})]_{z=e^{j\omega T}} \tag{12.1}$$

This is much simpler to evaluate than the FFT and it may be computed at any desirable frequency resolution. The PSD of individual spikes may be evaluated in an equivalent way, using $G(z)$, although the usefulness of this may be limited.

Capability for data compression

In long-term monitoring, there is a requirement to store the recorded signal for visual inspection or for additional off-line analysis. Although the proposed method performs analysis and spike detection on-line, it may be required to verify that it performed as expected by visually assessing the signal. Besides, maintaining some record of the progression of events in the EEG apart from spikes may be of clinical use. Storing the signal directly is inefficient, requiring a large amount of extra memory or even a mass storage device, thus increasing the complexity, size and cost of the system. Therefore some form of data compression is required.

The spike detection system was not designed for signal compression, but it may be possible to use it directly for this purpose, although the propositions that follow need to be considered in more detail during the development of such a system.

A scheme for signal compression based on the parametric form of the EEG will be outlined briefly. Assuming that the background activity is piecewise stationary, small segments of it may be modelled by a random input sequence and a time invariant system $H(z)$ (section 7.3.1). Therefore, by ignoring small changes in $H(z)$ during adaptation, it is possible to define segments in the signal, as sets of consecutive samples between two large transitions in the parameters of $H(z)$. Each segment may be characterized by the parameters of $H(z)$ before the transition or some average over the whole segment and the error sequence, $\hat{u}(n)$, which has a smaller dynamic range than the input EEG signal. Assuming that a segment has an average duration of one second (160 samples for the sampling rate used), $H(z)$ should only be saved at the end of every segment. For a 16th order system with every parameter stored as a 16-bit value, its storage requires 256 bits per second. If the sampled signal is a 12-bit quantity and the error signal an 8-bit value, then with the parametric form, saving the number of parameters, the length of the segment as a 6-bit value and the error signal requires $256 + 8 + 6 \times 160 = 1224$ bits per second, whereas the unprocessed EEG requires $12 \times 160 = 1920$, that is a saving of about 36%.

If the shape and the phase of the signal is not important, (for example during visual inspection, where a general 'feel' of the signal is required), the error sequence may not be

stored, only its power (16 bits, say), which leads to a compression rate of 85%! Spikes may be incorporated in the signal during reconstruction, by saving their time of occurrence from the beginning of the segment (8 bits), the amplitude of the generating impulse (12 bits) and the difference vector from the typical spike (optional). Since no more than three spikes can occur in a second, and they are infrequent, the overhead in their storage is small.

Further compression may be achieved if individual parameters are not stored, but on-line spectral analysis is carried out, to detect the type of activity occurring (e.g θ, δ, α or β). Reporting only this information for every segment, as well as the instants of spike occurrences may be sufficient.

Although this would not allow visual inspection of the record directly, it would provide a general impression of the behaviour of the signal during the period of monitoring, while being extremely economical in terms of storage, as the memory required for twenty-four hours of monitoring does not exceed 300k bytes.

12.2.4 Improvements, generalization, future prospects

Generalization for the detection of many transients

One of the key issues in the modelling approach for EEG analysis was the possibility of generalization. This may be viewed in the context of the EEG signal or in the more general frame of signals consisting of a structured (correlated) sequence with superimposed transient components whose detection is required.

As far as EEG analysis is concerned, the suggested system may be generalized to detect any number of transients, as demonstrated in Chapters 10 and 11. This would require the isolation of the transients manually and the application of the off-line estimation procedure of section 8.4.4 to derive a model for the new transients and their incorporation in the model along with the existing spike and other transient models. The decision elements (Chapters 10 and 11) must be changed also, but this modification is almost trivial, consisting of the introduction of a few more processing elements in the artificial neural network, as already explained in section 11.6, preferably with their weights initialized according to the scheme of section 11.7.2. Although the system is expected to operate as it is, retraining the network to finalize the details of the new transient and to integrate it fully with the remaining system may be necessary.

Introduction of secondary features

Other minor improvements on the existing system include the realization of the computation of the synthetic spectrum and the subsequent segmentation of the signal, as suggested in section 12.2.3. It would then be possible to construct the EEG data compression method outlined in the same section and by detecting the type of activity that occurred during every segment to produce the encoded summary of the activity, which constituted the proposed high-compression method. Ths could lead to an off-line reporting system, which would read and interpret that information, providing a report of the findings during the period of monitoring in a concise form using standard EEG terminology. This could be of better value for clinical purposes as it may be directly interpreted by human analysts and medical staff.

Provision of a user-friendly interface

The developed prototype system consists of a number of software parts: The main, on-line EEG analsysis system, a small program reading its output and isolating segments corresponding to transients and spikes, the spiking filter for extracting the parameters of new transients and converting them to an average and a covariance matrix and the neural network implementation.

Combining them under a unified user interface would result in a tool suitable for use by medical experts. The expert may then view existing sampled EEG records on a graphic display (a primitive form of this is currently available), mark and label the transients of interest, or correct the existing annotation by the automatic procedure. The system may then use this information to improve detection, based on the cues provided by the expert, by re-training the existing MLP. The user may also define new transients to be incorporated by the system, by marking them and applying the spiking filter (non-interactively). The output of the spiking filter may be put automatically as a new set of parameters and the system may be automatically modified to include it in the future.

Artifact detection and reduction

Artifacts, signals that have their origin outside the cerebrum and contaminate the scalp EEG recordings, often to a degree beyond recognition, are a serious problem of many EEG analysis procedures. Although it has been shown in section 7.4 that these may fall in the category of the background activity (thus modelled by $H(z)$) or they may be considered as transients, they are still a major concern in any long-term EEG analysis system. The most serious source of artifacts are muscle potentials, whose shape is variable and amplitude high.

It was demonstrated by testing that the rigorous system has an inherent ability to reject many artifacts as unrecognized transients, but this was not always possible, as some muscle artifacts resembled epileptic spikes. A more rigorous approach is therefore required. Artifacts may be reduced by some separate pre-processing, or by using an additional artifact rejection procedure fed with the parameters of the system, like the transfer function $H(z)$ and the power of the noise sequence output from its inverse after the occurrence of a transient.

Multichannel analsysis

Closely related to artifact reduction is the use of more than one EEG channel during analysis. Evidently, muscle artifacts would vary between electrodes, whereas genuine spikes due to generalized epileptic attacks are expected to be present simultaneously in more than one channel. The proposed modelling concepts need not be modified for the multichannel case, only the scalar signals have to be converted to vector quantities. The analysis of such a system is a generalization of the existing system. Although the case may be trivialized by treating each channel independently and combining their results, a more appropriate concept would be the integration of all signals in one system from the beginning, so that existing inter-correlations may be utilised for more accurate results.

Generalization to other signals

Finally, the proposed method is not confined to the EEG. Other biological signals, like the ECG may be treated in a similar way. The QRS complexes in the ECG may substitute the rôle of spikes, although the equivalent to 'background activity' may only consist of occasional interfering signals. The blood pressure waveform is another possibility. Non-biological signals consisting of some background signal on which transients are superimposed may be analyzed in a similar way. These could include images, where edges are to be detected in a noisy signal, containing texture information. Some initial work on the use of adaptive filtering for the purpose of edge detection has been reported[127], but the realization of a system equivalent to the one for EEG analysis has been hindered by the two-dimensional nature of images and hence the absense of a natural, unique, direction for the application of the algorithm, provided by 'time' in the case of the EEG.

Non-linear modelling

The modelling elements of the constructed system have been linear. This was a limitation imposed by existing theory. Work in non-linear signal processing systems is still experimental but new developments are anticipated in the near future. The concepts on which the

model was constructed do not have linearity as a necessary condition. Therefore elements of the analysis procedure may be converted to non-linear if required (e.g. for more accurate modelling). This would not require the modification of the rest of the system. The first elements that might be modelled by non-linear processes are the spikes and perhaps some other transients. Although small perturbations of a low-order linear system were often adequate to describe spikes, a non-linear model may provide a more accurate and 'neat' solution. A method that appears to be promising for this application is Volterra Networks[113, 173]. These are a special class of artificial neural networks, whose process consists of a non-linear process, $\mathbf{y}(n) = P_m(\mathbf{x}(n))$ of its input vector followed by linear weighting to produce its output, $z(n) = \mathbf{w}^T\mathbf{y}(n)$. The non-linear operator is a polynomial expansion of the input vector, so that the intermediate vector $\mathbf{y}(n)$ consists of the components of $\mathbf{x}(n)$, their cross-terms and all terms of their polynomial expansion up to and including those of order m. The linear weighting is effectively a Wiener Filter which was the basis of linear adaptive filters. In this sense Volterra networks are a non-linear extension to conventional adaptive filters and subject to similar theory.

12.3 Closing remarks

In this thesis a method for the detection of epileptic spikes in the interictal EEG has been analyzed. Based on a rigorous model, derived from neurophysiological descriptions and the nature of the processes that generate the EEG, this is a generalized approach, and provides a framework for collaboration of experts from both the medical and the signal processing fields. The method is versatile and with trivial amendments it may be applied to many other transients, and to non-EEG signals. The main contribution lies in the concept of a detailed model of the EEG, that includes transients, on which to base the analysis method, so that its performance may be predicted and its limitations understood. The developed model is rather general and there is no unique method to realize the implied analysis system. The proposed implementation, was based on essentially linear systems, but with advances in signal processing some of its elements may change in the future. Being largely a procedure to follow, rather than a rigid system, it may be improved by interaction from expert human analysts, without the supervision of a 'designer'. New transients may be included and the overall systems performance may be tailored to the performance of one analyst, or it may act as a reference for many analysts, depending on how it was trained.

The ability of the system to 'learn' by example is very important because it may be the only means to incorporate medical experience into an automatic system.

APPENDICES

Appendix A

Orthogonality and triangular waveforms

For the set of triangular functions mentioned in section 5.4 to be orthogonal it is necessary to prove that the following hold true:

$$\int_{-\infty}^{+\infty} \phi_{A_{T_1}}(t)\phi_{B_{T_2}}(t)dt = 0 \tag{A.1}$$

$$\int_{-\infty}^{+\infty} \phi_{A_{T_1}}(t)\phi_{A_{T_2}}(t)dt = 0 \quad T_1 \neq T_2 \tag{A.2}$$

$$\int_{-\infty}^{+\infty} \phi_{B_{T_1}}(t)\phi_{B_{T_2}}(t)dt = 0 \quad T_1 \neq T_2 \tag{A.3}$$

These integrals may be computed easily using the Fourier expansions for $\phi_{A_T}(t)$ and $\phi_{B_T}(t)$ to simplify and generalize the operations. Direct evaluation is tedious because triangular functions are piecewise-linear and so are their integrals. It can be shown (see for example [121][pp. 96–97]) that the corresponding Fourier series are:

$$\phi_{A_T}(t) = \frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\cos\left[\frac{2\pi}{T}(2k-1)t\right] \tag{A.4}$$

$$\phi_{B_T}(t) = \frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{(2k-1)^2}\sin\left[\frac{2\pi}{T}(2k-1)t\right] \tag{A.5}$$

Note that the Fourier coefficients of the expansion for $\phi A_T(t)$ are all positive, whereas those of $\phi B_T(t)$ have alternating sign. Both expansions contain only odd harmonics. Using the left-hand side of equation (A.1),

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\frac{8}{\pi^2}\sum_{m=1}^{\infty}\frac{1}{(2m-1)^2}\cos\left[\frac{2\pi}{T_1}(2m-1)t\right]\frac{8}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{(2n-1)^2}\sin\left[\frac{2\pi}{T_2}(2n-1)t\right]dt \\
&= \frac{16}{\pi^4}\sum_{m=1}^{\infty}\frac{1}{(2m-1)^2}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{(2n-1)^2}\int_{-\infty}^{+\infty}\cos\left[\frac{2\pi}{T_1}(2m-1)t\right]\sin\left[\frac{2\pi}{T_2}(2n-1)t\right]dt \\
&= 0
\end{aligned}$$

since $\int_{-\infty}^{+\infty} \cos at \sin btdt = 0$ for all a and b. Hence the $\phi_{A_{T_1}}(t)$ and the $\phi_{B_{T_1}}(t)$ are orthogonal.

Orthogonality is also necessary for functions belonging to the same group. Using the left-hand side of equation (A.2),

$$\int_{-\infty}^{+\infty} \frac{8}{\pi^2} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^2} \cos\left[\frac{2\pi}{T_1}(2m-1)t\right] \frac{8}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos\left[\frac{2\pi}{T_2}(2n-1)t\right] dt$$

$$= \frac{16}{\pi^4} \sum_{m=1}^{\infty} \frac{1}{(2m-1)^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \int_{-\infty}^{+\infty} \cos\left[\frac{2\pi}{T_1}(2m-1)t\right] \cos\left[\frac{2\pi}{T_2}(2n-1)t\right] dt \quad \text{(A.6)}$$

Since $\int_{-\infty}^{+\infty} \cos at \cos btdt = 0$ the above integrals would cancel for

$$\frac{2m-1}{T_1} \neq \frac{2n-1}{T_2} \quad \text{(A.7)}$$

However, there are values of n, m, for which in this relation the left- and right-hand sides are equal although $T_1 \neq T_2$. For example, if $T_1 = 3T_2$ Then for $m = 3n - 1$ the integrals would not cancel and the result of the expression in equation (A.6) would be non-zero. In other words, the $\{\phi_{A_T}(t)\}$ functions are not orthogonal. It can be shown, using similar arguments, that the $\{\phi_{B_T}(t)\}$ are not orthogonal either.

Note that this argument also holds for a a finite period of integration, T when the basis is limited to the functions that are harmonically related (i.e. have periods $\frac{T}{n}$, $n = 1, 2, \ldots$).

□

Appendix B

Spike models with special derivative properties

B.1 A modification of the Hanning window for continuous derivatives

The Hanning window of duration τ_0 and amplitude A_0 is expressed by the following function:

$$\begin{aligned} W_h(t) &= 0 && t < 0 \\ &= \tfrac{A_0}{2}(1 - \cos\omega_0 t) && 0 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \tag{B.1}$$

where $\frac{\omega_0 = 2\pi}{\tau_0}$. The main concern is the behaviour of the function and its first two derivatives at the points where $t = 0$ and $t = \tau_0$. For $tin[0, \tau_0]$, the derivatives are:

$$W_h'(t) = \frac{A_0\omega_0}{2}\sin\omega_0 t \tag{B.2}$$

and (B.3)

$$W_h''(t) = \frac{A_0\omega_0^2}{2}\cos\omega_0 t \tag{B.4}$$

Obviously, $W_h(0) = W_h\tau_0 = 0$ and $W_h'(0) = W_h'\tau_0 = 0$ but $W_h''(0) = W_h''\tau_0 = \frac{A_0\omega_0^2}{2} \neq 0$. Hence, $W_h(t)$ has to be modified to form a new window, W_s, such that $W_s''(0) = W_s''(\tau_0) = 0$ as well.

A simple way to accomplish this is by windowing the derivative of the signal, using a second Hanning window:

$$W_s'(t) = \frac{A_0\omega_0}{4}\sin\omega_0 t(1 - \cos\omega_0 t) \tag{B.5}$$

$$= \frac{A_0\omega_0}{4}(\sin\omega_0 t - \frac{1}{2}\sin 2\omega_0 t) \tag{B.6}$$

Then, the second derivative will be

$$W_s''(t) = \frac{A_0\omega_0^2}{4}(\cos\omega_0 t - \cos 2\omega_0 t) \tag{B.7}$$

which is zero at the specified two points. The actual window can be easily determined by integration and setting $W_s(0) = 0$ to be as follows:

$$W_s(t) = A_0\,(\frac{1}{16}\cos 2\omega_0 t - \frac{1}{4}\cos\omega_0 t + \frac{3}{16}) \tag{B.8}$$

This function is zero at the end-points of the range $[0, \tau_0]$ but in the centre of its this range it has a value of $\frac{1}{2}$, instead of 1. This can be easily corrected by multiplying all terms by 2, which gives the required function:

$$\begin{aligned} A_s(t) &= 0 && t < 0 \\ &= A_0\,(\tfrac{3}{8} - \tfrac{1}{2}\cos\omega_0 t + \tfrac{1}{8}\cos 2\omega_0 t) && 0 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \tag{B.9}$$

B.2 A further modification to achieve half-wave symmetry

W_s satisfies the requirement for the end-points. At $t = \frac{\tau_0}{2}$, $W_s(\frac{tau_0}{2}) = 1$ and $W_s(\frac{\tau_0}{2}) = 0$. But the second derivative, $W_s''(\frac{\tau_0}{2}) = -A_0\omega_0^2 \neq 0$. This is a problem in case the window is used in conjunction with a constant amplitude, like in the case of the Tukey window, where the first half of a hanning window is used for a smooth transition form 0 to some value A_0, which is maintained for a period of time, before it is smoothly returned to zero by the second half of the Hanning window. In this case the transition curve will need to maintain continuity at the end-points of the "constant" segment, which has zero first and second derivatives. This means that at $t = \frac{\tau_0}{2}$ both derivatives have to be zero. This has been achieved, as before, by windowing the first derivative of the function by a window which is zero at that critical point. A cosine window with double the frequency of the one used in the development of $W_s(t)$ would suffice. Hence, the new function has a derivative

$$W_t'(t) = \frac{A_0\omega_0}{4}\sin\omega_0 t(1 - \cos 2\omega_0 t) \tag{B.10}$$

$$= \frac{A_0\omega_0}{4}\sin\omega_0 t(2\sin^2\omega_0 t) \tag{B.11}$$

$$= \frac{A_0\omega_0}{2}\sin^3\omega_0 t \tag{B.12}$$

which gives a second derivative

$$W_t''(t) = \frac{3A_0\omega_0^2}{2}\sin^2\omega_0 t\ \cos\omega_0 t \tag{B.13}$$

which is zero when $t = \frac{\tau_0}{2}$. Integrating $W_t'(t)$ like before and using the fact that $W_t(0) = 0$ to determine the value of the constant produces the following function:

$$W_t(t) = A_0(\frac{1}{3} - \frac{1}{2}\cos\omega_0 t + \frac{1}{6}\cos^3\omega_0 t) \tag{B.14}$$

Of course, the amplitude needs to be normalized, since $W_t(\frac{\tau_0}{2}) = \frac{2}{3}$. Hence dividing by this value gives the final version of the function:

$$\begin{aligned} A_t(t) &= 0 && t < 0 \\ &= A_0\,(\tfrac{1}{2} - \tfrac{3}{4}\cos\omega_0 t + \tfrac{1}{4}\cos^3\omega_0 t) && 0 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \tag{B.15}$$

B.3 A function for non-linear variations in $\theta(t)$

$\theta(t)$ is the angular argument in the oscillatory part of the model suggested in section 6.5. Since it should be such that as t varies from 0 to τ_0 exactly one oscillation is completed, then $\theta(0) = 0$ and $\theta(\tau_0) = 2\pi$, and $\theta'(t) > 0$ for $t \in [0, \tau]$. The requirement for continuity and differentiability of the derivative puts additional restrictions on the functions to use. It was decided to use a model that allows θ to increase slowly near 0 maintain a constant rate of increase between some time limits $[t_1, t_2]$ after which its rate of increase should drop gradually, until $t = \tau_0$. The behaviour resembles that of an epileptic spike, which starts from a slowly varying frequency, acqyires a maximum and then gradually decreases its frequency after the trough is reached. Since the specifications are mainly on the "frequency", $\theta'(t)$, modelling began with this function. The slow variations were modelled by Hanning window sections with a constant in the middle:

$$\begin{aligned} \theta'(t) &= 0 && t < 0 \\ &= K\,(1 - \cos\omega_1 t) && 0 \geq t \geq t_1 \\ &= 2K && t_1 < t < t_2 \\ &= K\,[1 - \cos\omega_2(t - \tau_0)] && t_2 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \tag{B.16}$$

where $\omega_1 = \frac{\pi}{t_1}$ and $\omega_2 = \frac{\pi}{\tau - t_2}$ and K is a constant scaling factor.

The second derivative of this function,

$$\begin{aligned} \theta''(t) &= 0 && t < 0 \\ &= K\,\omega_1\,\sin\omega_1 t) && 0 \geq t \geq t_1 \\ &= 0 && t_1 < t < t_2 \\ &= K\,\omega_2 \sin\omega_2(t - \tau_0) && t_2 \geq t \geq \tau_0 \\ &= 0 && t > \tau_0 \end{aligned} \tag{B.17}$$

is continuous, since $\theta''(0) = \theta''(\tau_0) = 0$ and $\theta''(t_1) = \theta''(t_2) = 0$.

To find $\theta(t)$ one has to integrate $\theta'(t)$, taking into account the initial condition $\theta(0) = 0$. Hence, integrating the first section gives:

$$\theta(t) = K \int 1 - \cos \omega_1 t \, dt \tag{B.18}$$

$$= K \, (t - \frac{\sin \omega_1 t}{\omega_1}) \tag{B.19}$$

At the end of this segment, $\theta(t_1) = K \, t_1$, which is the initial condition for the next segment:

$$\theta(t) = 2K \int 1 \, dt \tag{B.20}$$

$$= 2K \, t - K \, t_1 \tag{B.21}$$

$$= K \, (2t - t_1) \tag{B.22}$$

Which ends up when $t = t_2$ with a value of $K \, (2t_2 - t_1)$. The final segment is similar to the first apart from a constant:

$$\theta(t) = K \int 1 - \cos \omega_2 (t - \tau_0) \, dt \tag{B.23}$$

$$= K \, [t - \frac{\sin \omega_2 (t - \tau_0)}{\omega_2}] + K \, (t_2 - t_0) \tag{B.24}$$

$$\tag{B.25}$$

At the end of this segment the value of the angle should be equal to 2π, thus $\theta(\tau_0) = K \, \tau_0 + K \, (t_2 - t_1) = 2\pi$, which gives the value of the scaling factor $K = \frac{2\pi}{\tau_0 + t_2 - t_1}$.

Appendix C

Properties of the Wiener filter

This Appendix contains material to complement the brief summary on Wiener filter theory, described in Chapter 8. Although it is not by any means complete, there are proofs of most of the properties of the Wiener estimator stated in section 8.1 and enough material to support both the concept of system modelling and its adaptive implelemtation, covered in section 8.4.2 and in Chapter 9.

Most of the properties of the estimator are associated with the Mean Squared Error optimization criterion, which may be transformed using the linear properties of statistical expectation, $E[\cdot]$, from its defining equation to the as follows:

$$\begin{aligned} \xi &= E[y(n)^2 - 2y(n)\mathbf{w}^T\mathbf{x}(n) + \mathbf{w}^T\mathbf{x}(n)\mathbf{x}(n)^T\mathbf{w}] \\ &= E[y(n)^2] - 2\mathbf{w}^T E[y(n)\mathbf{x}(n)] + \mathbf{w}^T E[\mathbf{x}(n)\mathbf{x}(n)^T]\mathbf{w} \\ &= r_{yy}(0) - 2\mathbf{w}^T\mathbf{r}_{xy} + \mathbf{w}^T R_{xx}\mathbf{w} \end{aligned} \tag{C.1}$$

This expression was given in section 8.1 as equation (8.6) and was employed to derive the Wiener-Hopf solution to the optimal filtering problem:

$$\mathbf{w}^* = R_{xx}^{-1}\mathbf{r}_{xy} \tag{C.2}$$

The shape of the Mean Squared Error function, which is a scalar function of the weight vector, $\xi(\mathbf{w})$, will now be considered, because most of the properties of the filter and its implementation stem directly from it. Because ξ, a function of $\mathbf{w}$, is non-negative it must possess at least one point of minimum MSE. In fact this minimum is unique, given by equation (C.2), occurring at $\mathbf{w} = \mathbf{w}^*$. Its value, $\xi_{min} = \xi(\mathbf{w}^*)$, may be found by substituting equation (C.2) into (C.1):

$$\xi_{min} = r_{yy}(0) - 2\mathbf{w}^{*T}\mathbf{r}_{xy} + \mathbf{w}^{*T}R_{xx}\mathbf{w}^*$$

$$
\begin{aligned}
&= r_{yy}(0) - 2[R_{xx}^{-1}\mathbf{r}_{xy}]^T\mathbf{r}_{xy} + [R_{xx}^{-1}\mathbf{r}_{xy}]^T R_{xx}[R_{xx}^{-1}\mathbf{r}_{xy}] \\
&= r_{yy}(0) - 2\mathbf{r}_{xy}^T R_{xx}^{-1}\mathbf{r}_{xy} + \mathbf{r}_{xy}^T R_{xx}^{-1}[R_{xx}R_{xx}^{-1}]\mathbf{r}_{xy} \\
&= r_{yy}(0) - \mathbf{r}_{xy}^T R_{xx}^{-1}\mathbf{r}_{xy} = r_{yy}(0) - \mathbf{r}_{xy}^T\mathbf{w}^* \qquad \text{(C.3)}
\end{aligned}
$$

where the symmetry of the autocorrelation matrix $R_{xx}^{-1T} = R_{xx}^{-1}$ was used.

Starting from equation C.1, substituting $I = R_{xx}R_{xx}^{-1}$ and adding and subtracting the term $\mathbf{r}_{xy}^T R_{xx}^{-1}\mathbf{r}_{xy}$ results in the following expressions:

$$
\begin{aligned}
\xi(\mathbf{w}) &= r_{yy}(0) - 2\mathbf{w}^T R_{xx}R_{xx}^{-1}\mathbf{r}_{xy} + \mathbf{r}_{xy}^T R_{xx}^{-1}\mathbf{r}_{xy} - \mathbf{r}_{xy}^T R_{xx}^{-1}\mathbf{r}_{xy} + \mathbf{w}^T R_{xx}\mathbf{w} \\
&= r_{yy}(0) - 2\mathbf{w}^T R_{xx}[R_{xx}^{-1}\mathbf{r}_{xy}] + [\mathbf{r}_{xy}^T R_{xx}^{-1}]R_{xx}[R_{xx}^{-1}\mathbf{r}_{xy}] - \mathbf{r}_{xy}^T[R_{xx}^{-1}\mathbf{r}_{xy}] + \mathbf{w}^T R_{xx}\mathbf{w} \\
&= [r_{yy}(0) - \mathbf{r}_{xy}^T\mathbf{w}^*] - 2\mathbf{w}^T R_{xx}\mathbf{w}^* + \mathbf{w}^{*T} R_{xx}\mathbf{w}^* + \mathbf{w}^T R_{xx}\mathbf{w} \qquad \text{(C.4)}
\end{aligned}
$$

where the symmetry of R_{xx} has been used as well as Wiener-Hopf equation. Identifying the term in square brackets as ξ_{min} and using the fact that scalar terms like $\mathbf{w}^T R_{xx}\mathbf{w}^*$ are equal to their transpose, the above equation may be rewritten as

$$
\begin{aligned}
\xi(\mathbf{w}) - \xi_{min} &= \mathbf{w}^{*T} R_{xx}\mathbf{w}^* - [\mathbf{w}^T R_{xx}\mathbf{w}^*]^T + \mathbf{w}^T R_{xx}\mathbf{w} - \mathbf{w}^T R_{xx}\mathbf{w}^* \\
&= \mathbf{w}^T R_{xx}\mathbf{w} - \mathbf{w}^T R_{xx}\mathbf{w}^* + \mathbf{w}^{*T} R_{xx}\mathbf{w}^* - \mathbf{w}^{*T} R_{xx}\mathbf{w} \\
&= \mathbf{w}^T R_{xx}(\mathbf{w} - \mathbf{w}^*) + \mathbf{w}^{*T} R_{xx}(\mathbf{w}^* - \mathbf{w}) \\
&= (\mathbf{w} - \mathbf{w}^*)^T R_{xx}(\mathbf{w} - \mathbf{w}^*) \\
\text{or} \quad \xi\mathbf{v}) - \xi_{min} &= \xi(\mathbf{w}) - \xi_{min}) = \mathbf{v}^T R_{xx}\mathbf{v} \qquad \text{(C.5)}
\end{aligned}
$$

where $\mathbf{v} = \mathbf{w} - \mathbf{w}^*$. The expression above has a clear form of a quadratic (hyperparaboloid) in $\mathbf{v}$ (hence in $\mathbf{w}$ too, which is translated by a factor of $\mathbf{w}^*$). Because R_{xx} is positive (semi)definite, it is concave upwards[202][p. 20–25]. For these reasons the unique point of zero gradient is a minimum. Also, the locus of $\mathbf{w}$ having the same MSE, ξ_c lie on a hyperellipse[202][p. 38].

These statements will be substantiated using equation (C.5), where the origin of the weight vector space has been translated to $\mathbf{w}^*$ and ξ_{min} to zero. The eigenvalues of the correlation matrix, $\lambda_1, \lambda_2, \ldots$ are all real and positive (since R_{xx} is positive definite)[202][p. 34–43]. The eigenvetors corresponding to each one of the eigenvalues, $\mathbf{q}_1, \mathbf{q}_2, \ldots$ are linearly independent and may be constructed to be orthonormal. Then the *modal matrix* $Q = (\mathbf{q}_1\mathbf{q}_2\ldots)$ whose columns are the eigenvectors is such that $QQ^T = I$ or $Q^{-1} = Q^T$, a direct

consequence of the orthonormality of its columns. Using the diagonal eigenvalue matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & \cdots \\ 0 & \lambda_1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & & \lambda_k & \cdots \\ \vdots & \vdots & \cdots & \vdots & \ddots \end{bmatrix} \tag{C.6}$$

and the characteristic equation of the correlation matrix, R_{xx} may be factored as follows:

$$\begin{aligned} R_{xx}Q &= \Lambda Q = Q\Lambda \\ \Rightarrow R_{xx} &= Q\Lambda Q^{-1} \\ \Rightarrow R_{xx} &= Q\Lambda Q^{T} \end{aligned} \tag{C.7}$$

The matrix Q corresponds to a rotation[202][p. 34–43] and when equation C.7 is substituted into C.5 it corresponds to a change of basis from $\mathbf{v}$ to a new one $\mathbf{u}$:

$$\begin{aligned} \xi\mathbf{v}) - \xi_{min} &= [\mathbf{v}^T Q]\Lambda[Q^T\mathbf{v}] \\ &= [Q^T\mathbf{v}]^T\Lambda[Q^T\mathbf{v}] \\ \xi\mathbf{u}) - \xi_{min} &= \mathbf{u}^T\Lambda\mathbf{u} \end{aligned} \tag{C.8}$$

Since Λ is a diagonal matrix with positive entries, the above quadratic expression may be epressed in the more recognizable form

$$\xi(u_1, u2, \ldots) - \xi_{min} = \lambda_1 u_1^2 + \lambda_2 u_2^2 + \ldots \tag{C.9}$$

In this form, the parabolic shape in any direction u_i is revealed by setting all other variables to constants, giving the parabola $\xi - \xi_{min} = \lambda_i u_i^2 + const.$ Because all the eigenvalues of the positive-definite matrix R_{xx} are positive, the coefficient of u_i is positive, resulting in an upwards-concave parabola.

Similarly, for constant $\xi - \xi_{min} = \xi_c$ the above expression gives the locus of the weight vectors having the same MSE output level. This has the form of a (hyper)ellipse. Different values of ξ_c give concentric hyperellipses (centred at $\mathbf{w}^*$) having the same orientation. In this case the eigenvalues and eigenvectors have a geometrical sigificance. The ith eigenvector, $\mathbf{q}_i$ gives the direction of the ith principal axis of the hyperellipse. The corresponding eigenvalue, λ_i is the reciprogal of the square of the radial distance from the centre of the hyperellipses to the one corresponding to $\xi_c = 1$ in that direction. Widrow[202][p. 34–43] interprets the eigenvalues as the second derivatives in the direction of their corresponding eigenvectors. It

should be noted that the smallest eigenvalue, λ_{min} corresponds to the longest principal axis of any hyperellipse whereas the largest, λ_{max} to the smallest.

One of the most important properties of the Wiener filter arising from the use of the MSE function is the decorrelation of the input and the error. Algebraically, the desired response vector, $\mathbf{y} = (\ldots, y(n-1), y(n), y(n+1), \ldots)$ may be represented in a vector space, $\mathcal{V}$. The vector space, $\mathcal{U}$, spanned by the input vectors to the weights, $\{\mathbf{x}(n)\}$ is a subspace of $\mathcal{V}$. The estimate produced by the Wiener filter, $\hat{\mathbf{y}}$, may then be regarded as a *projection* of $\mathbf{y}$ on this subspace, *calU*. $\hat{\mathbf{y}}$ is a linear combination of the $\{x(n)\}$, scaled by the weight vector, $\mathbf{w}$. The error vector is the difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Hence, $\varepsilon(n) = y(n) - \mathbf{x}(n)^T\mathbf{w}$ and the expected value of the product of $\varepsilon(n)$ and $\mathbf{x}(n)$ when $\mathbf{w} = \mathbf{w}^*$ is zero:

$$\begin{aligned} E[\varepsilon(n)\mathbf{x}(n)] &= E[y(n)\mathbf{x}(n)] - E[\mathbf{x}(n)\mathbf{x}(n)^T\mathbf{w}^*] \\ &= \mathbf{r}_{xy} - R_{xx}\mathbf{w}^* \\ &= \mathbf{r}_{xy} - R_{xx}R_{xx}^{-1}\mathbf{r}_{xy} \qquad \text{(C.10)} \\ &= \mathbf{r}_{xy} - \mathbf{r}_{xy} = 0 \qquad \text{(C.11)} \end{aligned}$$

In other words, the error is orthogonal to the input vector space when the optimum weight vector is used. This has a geometric interpretation, because the MSE criterion is equivalent to the square of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Since the perpendicular distance to a the (hyper)plane spanned by the $\{\mathbf{x}(n)\}$ is shorter than that in any other direction, the length of the error vector is minimized when ε is perpendicular to $\mathcal{U}$. Consequently, the output of the filter $\hat{\mathbf{y}} \in \mathcal{U}$, and ε are orthogonal (decorrelated) as well.

Appendix D

Details of the simulated EEG record used for testing the adaptive spike detection system(s)

The background activity $b(n)$ (see section 7.3.1) was generated by processing a white sequence, $u(n)$, by a piecewise linear system, H. The input sequence was a uniform random, uncorrelated sequence with zero mean and constant variance (power), σ_{uu}^2. This was generated from a standardized uniform random number generating function whose output $w(n) \sim \text{Uni}(0,1)$ was shifted by $-\frac{1}{2}$ and then scaled by the standard deviation of the required signal to produce $u(n) = \sigma_{uu}\,[z(n) - \frac{1}{2}]$. The piecewise-linear system H consisted of a collection of transfer functions, $\{H_i(z)\}$. For practical reasons this was restricted to a set of five. Normally one of these is active at any time, but there is a transition period between consecutive segments, during which both are active. According to the model, all transfer functions are continuously active (to prevent the transients of switching their input on and off) and hence the selection of one of the six outputs to be $b(n)$ determines which transfer function is active at any time. The duration N_i of activity of the current $H_i(z)$ was made variable (random perturbations around a mean duration). N_i was conveniently assumed to be an exponential random variable having a mean of μ_n samples (corresponding to approximately 2 s of real time. For a sampling rate of 160 s^{-1}, $N_i \sim \text{Exp}(320)$. Selecting which transfer function was active for this duration was adone in a random manner, using the uniform random number generator, with output between 0 and 1, scaled by a factor of 6 and truncated into an integer between 0 and 5, with equal probability, used to index the active transfer function. At the end of the activation period a new period and transfer function was selected in the same manner. Transition periods were considered at the end of

every segment. These were short, their duration being $M \sim \text{Exp}(20)$ facilitating the smooth transition between segments. In this period, the output of the expiring transfer function was scaled down to zero using a $\cos^2(\frac{\pi}{2M}n)$, while the new transfer function was scaled up from zero to its maximum value using $\sin^2(\frac{\pi}{2M}n)$. The selection of these two functions was based on the fact that their sum is always 1. Therefore, if both transfer functions have the same amplitude levels in their output, they are mixed so that the amplitude level is maintained throughout the transition period, although the composition will change gradually.

Spikes, $s(n)$, were generated independently using a software implementation of the generating process described in section 7.3.2. According to this definition, spikes are the output of a LTI system with transfer function $G(z)$, under the excitation of infrequently occurring impulses with scaled amplitudes, $d(n)$. The input sequence was easily generated, as it is normally zero, apart from certain rare discrete instances. The time difference, ΔM (in samples) between two consecutive spike occurrences was varying randomly about some mean value μ_m following an Exponential distribution, $\Delta M \sim \text{Exp}(\mu_m)$. Although spikes occur quite infrequently in the interictal EEG, μ_m was taken as 3 s, (or 480 samples), so that a reasonable number of spikes were present in a record of 10000 samples. The amplitudes of the excitation impulses in $d(n)$ were made variable, but at a level well above that for the backgrond activity, since spikes should be clearly distinguished from this. It was found that by setting the mean amplitude of the excitation impulses to about six times above the RMS value of the background activity input process ($\mu_d = 6\sigma_{uu}$) produced spikes which were at the same level above the background activity as the real signals available. Variability was introduced on spike amplitudes by random variations on this mean amplitude with a variance $\sigma_{dd}^2 = \frac{1}{2}\sigma_{uu}^2$, following a Normal distribution. This ensured the presence of spikes with a wide range of amplitudes. The transfer function, $G(z)$ contained the average coefficients for real spikes, g^*, which were estimated by the method suggested in section 8.4.4. The order of $G(z)$ was chosen to be $L = 10$, although the spikes out of a second order system, were still reasonably similar to real spikes. Variability on the shape of spikes was introduced by small perturbations on the coefficients of $G(z)$ every time a new spike was generated. These were in the range described in section 8.4.4 and were used to test on-line adaptation to spikes described in that section. The resulting coefficient vectors were Normally distributed with mean $\mathbf{g}^*$. This is a multivariate distribution with equal entries in the diagonal of the covariance marix and zero everywhere else (each coefficient is independent of all others and are all identically distributed). With this model no two spikes are exactly the same and tests for the behaviour of the detector with variable spikes are more objective.

The process of generating any other transient (section 7.3.3) is similar to the one just

described for spikes. It differs only in the values of its parameters, especially the transfer function. Because there are many transients, there should actually be a collection of such processes, each one with its own parameters and transfer function. Due to the small number of EEG records available, most of which contained essentially spikes, constructing models for other transients was impossible because of unavailability of data. To test how the method behaved in the presence of at least one non-spike transient an artificial simple model was constructed. This consisted of a second-order system, $F(z)$, with two real poles. These were typically set to 0.75 and 0.40, but were allowed to vary randomly by 20% of their typical values in either direction from one transient to the next. The impulse response of $F(z)$ was an exponential decay. This could be taken as a model of some electrode movement artifacts or a simplified model of eye movement arifacts. Small variations were introduced on the filter parameters, which changed the shape of its output slightly. $F(z)$ was excited by a sequence of impulses occurring randomly, $c(n)$, following the same distribution as with the spike model, but with a mean duration of 320 samples between occurrences (corresponding to 2 s for the sampling rate used). The amplitudes of the impulses in $c(n)$ were made random, following a normal distribution around a mean $\mu_c = 4\sigma_{uu}$ (higher than that for spikes) having a variance $\sigma_{cc}^2 = \sigma_{uu}^2$, which produced more dispersed amplitudes than for spikes. There is no particular reason for the selection of these statistical parameters, since the signal is, unlike spikes, artificial. An additional polarity parameter taking the value $+1$ or -1 with equal probability determined whether the excitation impulse and hence the generated transient was positive or negative.

Finally, random, Gaussian, uncorrelated noise $v(n)$, as described in section 7.3.4, was generated by the Normal random value generator, used in all previous signal components. This had zero mean and a variance (power) depending on a variable user-defined signal-to-noise ratio, κ. The variance for the noise was computed based on that of the 'signal', $b(n)$ (or rather $u(n)$), $\sigma_{vv}^2 = \frac{1}{\kappa}\sigma_{uu}^2$. The actual signal-to-noise ratio is normally larger than κ, because $H(z)$ tends to amplify the signal power in $b(n)$. It does serve as a guideline, though.

These signals were then added together to generate an artificial EEG signal $e(n)$ on which the spike detection method was tested.

Appendix E

An introduction to Artificial Neural Networks (ANN)

E.1 Historical Overview

This possibly misleading name was adopted to describe a wide class of systems and is rather suggestive of the neurological origins of these systems. They emerged in the mid 1940's along with the first digital computers[136][p. 14–20]. Then the idea of computing machines was quite novel and digital computer technology was still in its infancy. In those early days the primary concern was the development of a machine exhibiting intelligence. There were different theories on how this could be achieved. On one extreme were the supporters of a 'programmed' machine, having a powerful processing unit executing instructions explicitly. This led to the development of today's computers. On the opposite extreme were the 'connectionists', who supported a distributed system, consisting of massively interconnected processing elements. Characteristic of that period was the work of W. McCulloch and W. Pitts, who demonstrated that simple models of neurons could compute basic logic functions. Although the structure and function of these elements was simple, the supporters of this theory believed that the system as a whole could exhibit complex behaviour. The latter were the predecessors of todays Artificial Neural Networks. Another importance contribution was the work of D. Hebb, who presented in 1949 a 'learning law', by which individual neurons could change their response depending on their input stimuli.

In the beginning the supporters of the two philosophies overestimated the capabilities of their corresponding systems. It is quite clear nowadays that the first approach, which led to the development of the digital computer, was not the way to intelligence, although it comprises a powerful machine for executing predefined sequences of mathematical and

data manipulations. The second approach was also pursued and the first 'neurocomputers' were built in the 1950's. The most successful ones were Rosenblatt's *Perceptron* in 1958 and the ADALINE by Widrow and Hoff, which has been the predecessor of the LMS algorithm discussed in a section 8.4.2. The apparent success of these early methods in solving simple problems had its shortcomings. Many of the researchers approached the field from an experimental standpoint and overstated the capabilities of the early methods they used. Consequently, the field ran out of new ideas and was also discredited, because of the lack of a sound theoretical basis. The unrealistic speculations of some about the future of the subject were criticised and eventually attention was drawn away from neurocomputing in favour of the more reliable and consistent digital technology. The book *Perceptrons* by Minsky and Papert (1969)[124] contained extensive criticisms on the field, as it emphasized that the perceptron was unable to compute the logical exclusive-or function, concluding that all neural networks suffered from similar limitations. Some serious artificial neural network enthusiasts continued their work in associated fields (for example Widrow in signal processing) but a few new 'recruits' managed to maintain the interest in the field until the 1980's, when the field re-emerged, with stronger foundations laid by a handful of enthusiasts with analytical skills, like J. Hopfield, T. Kohonen, S. Grossberg, D. Rumelhard and J. McClelland. The last two were the editors of two volumes on Parallel Distributed Processing (PDP) in 1986[166], containing a number of key papers on neurocomputing, including some on the Multi-Layer Perceptron (MLP) and its training algorithm.

E.2 Structure and Function of ANNs

The internal structure (organization) of the network may have, in theory, any form. The only restrictions that apply are that every processing element has only one output, all memory is local and the connections between the elements are unidirectional and instantaneous[136, pp. 21–44]. Every processing element has a transfer function, which describes how the output may be generated from the inputs and the local memory. This may be continuous- or discrete-time. In the latter case a scheduler may be required to synchronize the activity of the various processing elements. The inputs and the outputs of the network need not all be of the same mathematical type, a rule that applies to the inputs and outputs of the processing elements, which may differ in their structure and function from one another.

Despite the generality of the definition of what constitutes an ANN, most network structures used in practice are much more restrictive in their architecture. Typically, one or two data types are used and the topology of the interconnections has a great deal of symmetry

and regularity. Moreover, although the definition does not rule out feedback, this is a property of only a few ANN structures. The structure of a network is often related to its general function. Hecht-Nielsen in his book "*Neurocomputing*"[136] classified ANNs into *associative*, *mapping*, *spatiotemporal*, *stochastic* and *hierarchical networks*.

Associative networks, produce a mapping from an input vector to an output vector. This is effected by one array (layer) of processing elements. ANNs belonging to this category may have feedback (recurrent) or not (feed-forward). Typical examples of associative networks include the linear associator and its binary equivalent, the *learnmatrix*, both of which are feed-forward networks. The most famous recurrent associative ANNs are the *Hopfield* network, which is essentially binary, and its real-number equivalent *brain state in a box* (BSB) network. Associative networks have been reviewed in an article by Lippman[112] and described in some detail by Hecht-Nielsen[136, pp. 79–109].

Mapping networks attempt to form an approximation to the relation of input vectors to output vectors by means of examples of the required mapping. Both the input and the output spaces must be bounded. The approximation is normally effected by minimizing a performance measure, such as the mean squared error of the approximation for all examples presented to the network. Mapping networks are related to statistical linear regression of which they are a generalization. Examples of mapping networks include the *backpropagation neural network* (a form of the MLP), Kohonen's *self organizing map*, the *counterpropagation network* and the *group method of data handling* (GMDH)[136, pp. 110–162].

Spatiotemporal networks, as the name suggests are ANN structures, which can incorporate time in their input patterns and hence can deal with temporal sequences. The time information may be incorporated as external feedback using network connections like in the *recurrent backpropagation network* or as part of the transfer function of individual nodes, like in the *spatiotemporal pattern recognizer*. Stochastic neural networks employ probabilistic means of 'learning', like the *Boltzmann Machine*. Finally, hierarchical networks use sparse connections between groups (layers) of processing elements with rules to direct training towards specific intermediate patterns. The *neocognitron* is a typical example of this class of ANNs.

Appendix F

The MLP and its training algorithm

F.1 The generalized delta rule

Because of the layered structure of the MLP there are differences in the expressions for the derivatives depending on where a weight is located in the network and there is need to consider them individually.

The ith weight of the pth processing element of the lth layer of the network, is adjusted using the following scalar equivalent of equation 11.3:

$$w_{lpi}(n+1) = w_{lpi}(n) - \mu \frac{\partial \xi(\mathbf{w}(n))}{\partial w_{lpi}} \tag{F.1}$$

It is necessary to compute $\frac{\partial \xi(\mathbf{w}(n))}{\partial w_{lpi}}$. Using the expressions for the transfer function of the element (equation 11.1) and the chain rule for differentiation, the partial derivative may be decomposed as follows:

$$\begin{aligned}
\frac{\partial \xi(\mathbf{w}(n))}{\partial w_{lpi}} &= \frac{\partial \xi(\mathbf{w}(n))}{\partial z_{lp}} \frac{\partial z_{lp}(n)}{\partial f_{lp}(n)} \frac{\partial f_{lp}(n)}{\partial w_{lpi}} \\
&= \frac{\partial \xi(\mathbf{w}(n))}{\partial z_{lp}} \frac{\partial \sigma[f_{lp}(n)]}{\partial f_{lp}(n)} \frac{\partial \sum_{k=0}^{N_{l-1}} w_{lpk}(n) z_{(l-1),k}(n)}{\partial w_{lpi}} \\
&= \frac{\partial \xi(\mathbf{w}(n))}{\partial z_{lp}} \sigma'[f_{lp}(n)] z_{(L-1),i}(n)
\end{aligned} \tag{F.2}$$

where $\sigma'[\cdot]$ denotes the first derivative of the non-linearity in the output of every element. In some applications where this is absent, as explained earlier, it may be substituted by unity.

For the algorithm to be complete $\frac{\partial \xi(\mathbf{w}(n))}{\partial z_{lp}}$ must be evaluated. A general step is its estimation has been due to Widrow and Hoff and is known as the *delta rule*, although it is more well known in signal processing circles as the LMS algorithm, the basis of adaptive filtering described previously in section 8.4.2. This is based on the assumption that the MSE may be

estimated without bias from a finite sum of squared errors, $\hat{\xi}(\mathbf{w}(n)) = \frac{1}{N}\sum_{n=1}^{N}\varepsilon^2(n)$, where $\varepsilon^2(n) = \sum_{j=1}^{N_L}[d_j(n) - z_{Lj}(n)]^2$. Since this is true for any N, it is also true for $N = 1$, which further simplifies estimation to $\hat{\xi}(\mathbf{w}(n)) = \varepsilon^2(n)$.

If the weight considered belongs to a processing element in the output layer ($l = L$), then the required derivative is

$$\begin{aligned}\frac{\partial\hat{\xi}(\mathbf{w}(n))}{\partial z_{Lp}} &= \frac{\partial\varepsilon^2(n)}{\partial z_{Lp}} \\ &= \frac{\partial\{\sum_{j=1}^{N_L}[d_j(n) - z_{Lj}(n)]^2\}}{\partial z_{Lp}} \\ &= \frac{\partial[d_p(n) - z_{Lp}(n)]^2}{\partial z_{Lp}} \\ &= -2[d_p(n) - z_{Lp}(n)] \end{aligned} \tag{F.3}$$

For weights in the intermediate (hidden) layers, the expressions for the partial derivatives are somewhat more complex, because each output from the layer affects every element of the subsequent layers and there are multiple paths to the network outputs. Considering a node in the hidden layer just before the output layer,

$$\begin{aligned}\frac{\partial\hat{\xi}(\mathbf{w}(n))}{\partial z_{(L-1),p}} &= \frac{\partial\varepsilon^2(n)}{\partial z_{(L-1),p}} \\ &= \frac{\partial\{\sum_{j=1}^{N_L}[d_j(n) - z_{Lj}(n)]^2\}}{\partial z_{(L-1),p}} \\ &= \sum_{j=1}^{N_L}\frac{\partial[d_j(n) - z_{Lj}(n)]^2}{\partial z_{(L-1),p}} \\ &= -2\sum_{j=1}^{N_L}[d_j(n) - z_{Lj}(n)]\frac{\partial z_{Lj}(n)}{\partial z_{(L-1),p}} \\ &= -2\sum_{j=1}^{N_L}[d_j(n) - z_{Lj}(n)]\frac{\partial z_{Lj}(n)}{\partial z_{(L-1),p}} \end{aligned} \tag{F.4}$$

Since $z_{Lj}(n) = \sigma[f_{lj}(n)]$ and $f_{lj}(n) = \sum_{k=0}^{N_{L-1}} w_{Ljk}(n)z_{(L-1),k}$, by a further application of the chain rule, $\frac{\partial z_{Lj}(n)}{\partial z_{(L-1),p}} = \sigma'[f_{Lj}(n)]w_{Ljp}$. Substituting into equation F.4 and noting that $-2[d_j(n) - z_{Lj}(n)]\sigma'[f_{Lj}(n)] = \frac{\partial\hat{\xi}(\mathbf{w}(n))}{\partial f_{Lj}}$ the above expression may be written as:

$$\frac{\partial\hat{\xi}(\mathbf{w}(n))}{\partial z_{(L-1),p}} = \sum_{j=1}^{N_L}\frac{\partial\hat{\xi}(\mathbf{w})}{\partial f_{Lj}}w_{Ljp}(n) \tag{F.5}$$

This expression may be generalized to any layer, binding the required partial derivatives to those of the layer above the one considered. By denoting $\delta_{lp}(n) = \frac{\partial\hat{\xi}(\mathbf{w}(n))}{\partial f_{lp}}$ it is possible to express the weight updating algorithm in terms of the $\{z_{lp}(n)\}$, the $\{f_{lp}(n)\}$ and the

$\{\delta_{lp}(n)\}$, which can be evaluated recursively starting from the output layer and propagating backwards through the hidden layers, an action that gave the network its peculiar name. Summarizing the weight updating procedure, usually called 'network training' in neural network nomenclature:

- Initially, the MLP with L layers, has random weights. There also exist a training set, $\{[\mathbf{x}(n), \mathbf{d}(n)]; n = 0, 1, 2, \ldots\}$, and a learning rate, μ
- For every pair $[\mathbf{x}(n), \mathbf{d}(n)]$ in the training set repeat the following:
- Apply $\mathbf{x}(n)$ to the input layer and run the network to produce the output vector $\mathbf{z}_L(n)$
- For every layer in the network $l = L, L-1, \ldots, 1$ do the following:
  - For every processing element, $p = 1, 2, \ldots, N_l$ in the current layer perform the following:
    * Compute the the partial derivative, $\delta_{lp}(n) = \frac{\partial \xi(\mathbf{w}(n))}{\partial f_{lp}}$, where

$$\delta_{lp}(n) = \begin{cases} -2\sigma'[f_{Lp}(n)][d_p(n) - z_{Lp}(n)] & \text{for the output layer, } L \\ \sigma'[f_{lp}(n)] \sum_{j=1}^{N_{l+1}} \zeta_{l+1,j,p}(n) & \text{for all other layers} \end{cases} \tag{F.6}$$

    * For every input $i = 0, 1, \ldots, N_{l-1}$ of the processing element do the following:
      · propagate the errors to the previous layer, by computing $\zeta_{lpi}(n) = \delta_{lp}(n) w_{lpi}(n)$
      · Update the weight using the update relation $w_{lpi}(n+1) = w_{lpi}(n) - \mu \delta_{lp}(n) z_{(l-1),i}(n-k)$

It should be noted that if $z(n) = \sigma[f(n)]$ is the logistic function $z(n) = \frac{1}{1+e^{-f(n)}}$ then its derivative is given by $\sigma'[f(n)] = z(n)[1 - z(n)]$, which may be computed without the expensive operation of exponentiation.

Although the procedure outlined is the one generally employed, small differences that appear to improve the performance of training have been suggested. These normally affect the weight updating stage of the algorithm only. The two most widely used modifications of the updating equation result in the *batch* and the *momentum* variants[136][p. 59–63, 133–137].

The first does not update the weights every time a new input vector is applied to the network. The errors are still propagated every time, and the errors collected and averaged in 'batches' of K trials. Updating occurs at the end of every batch, according to the following

relation:

$$w_{lpi}(n+1) = w_{lpi}(n-K) - \mu\frac{1}{K}\sum_{k=0}^{K-1}\delta_{lp}(n-k)z_{(l-1),i}(n-k) \qquad \text{(F.7)}$$

This approach produces more reasonable (less 'noisy') estimates of the propagated errors than the original delta rule. This is advantageous in some applications.

The momentum version of the algorithm utilises the previous value of the weight, $w_{lpi}(n-1)$. The change in the weight is not only based on the 'correction' term $-\mu\delta_{lp}(n)z_{(l-1),i}(n)$, but on a 'prediction' term, based on the 'trend' of the weight, which is approximated by the difference from the previous weight value. In physical terms this is equivalent to giving the weight resistance to instantaneous changes in the direction of adaptation, which is analogous to momentum. The weight updating equation is normally expressed as follows:

$$w_{lpi}(n+1) = w_{lpi}(n) + \alpha[w_{lpi}(n) - w_{lpi}(n-1)] - (1-\alpha)\mu\delta_{lp}(n-k)z_{(l-1),i}(n) \qquad \text{(F.8)}$$

Where the new factor $\alpha \in (0,1)$ defines the contribution of the prediction and the correction terms. A value of α close to unity pays more attention to the trend of adaptation rather than the new information and requires a large number of iterations with $\delta_{lp}(n-k)z_{(l-1),i}(n)$ pointing in a specific direction to effect a change in the direction of adaptation. A value of α close to 0 makes the algorithm similar to the original delta rule.

The momentum algorithm may help the weight to descend to a minimum when a flat area or a shallow trough of the error surface is encountered. In both these cases the traditional algorithm will proceed very slowly or even stop descending, since the gradients will be small. Smoothing of erroneous oscillations of the weight in a direction lateral to the one of minimum derivative due to noise in gradient estimation is another desired property of this variant.

Other gradient-based algorithms have also been adapted for the layered backpropagation algorithm with some success, like the conjugate gradient algorithm[40].

F.2 Some notes on the use of the MLP

From the description of the backpropagation neural network it may appear that this it is a universal tool, a panacea to all modelling problems. This is, in fact far from truth. It should be used with care and the some scepticism when initial results appear to be 'too good'. There are a number of issues that should be kept in mind, when developing a mapping neural network. Some of these are inherent to the MLP and its training algorithm and anyone who has attempted to train a network must have come across at least one of them.

One of the primary issues is the selection of the architecture of the network. With the number of inputs and outputs known in advance, it remains to determine the number of

hidden layers and the number of processing elements in each layer. Unfortunately, there are no general rules on how to achieve this. In theory[136][p. 122–124], it is possible to solve a problem with a single hidden layer, but the required number of processing elements may be unacceptably high. The introduction of multiple hidden layers may yield a more economical procedure, but the determination of how many is still ambiguous. Fortunately, in the current application a study of the earlier intuitive system provided some guidelines which worked well in practice (see section 11.5.1).

The remaining issues are related to training. Unlike adaptive filtering, where range for the adaptation rate, μ for which the algorithm converged was known, the range of the learning rate for the backpropagation algorithm is largely unknown. This is associated with the non-quadratic shape of the performance surface, $\xi(\mathbf{w})$. Although this is not known exactly, it may be shown that it is highly degenerate, with plenty of global minima, which in some cases lie at infinity. The shape between these minima is unknown, but flat regions as well as local minima are known to exist and their presence has been verified experimentally [136][p. 111–125]. Consequently, there are many basins of attraction where a gradient-based algorithm may converge. Although locally these may 'look' quadratic, their parameters are not known and hence it is not possible to determine the upper limit for the adaptation rate a priori. It is therefore necessary to find it by 'trial and error' and sometimes to change it to navigate around the error surface until a minimum is reached. The shape of the error surface is often the reason gradient-based algorithms fail to reach the point of minimum achievable error. Because the search for the minimum proceeds in a direction opposite to that of the gradient of the surface, it fails to proceed if the gradient is small, like on a flat region of the surface or zero, like a local minimum. In the first case the descent may become extremely slow, whereas in the second it may be stopped because the local minimum may act as a basin of attraction for the algorithm[136][p. 111–115]. Things may become worse because of the noisy estimates of the gradient. Both problems are often overcome by employing one of the modified algorithms. For example the noise level is reduced by the batch process, whereas overcoming flat surfaces and shallow basins of attraction due to local minima are often solved by the momentum term. There are, however, cases when the modifications of the algorithm may make things worse.

Finally there are two issues, *under-representation* of some input conditions and *overtraining*, which are specific to the MLP. These are related to the way the training set is selected and applied to the network[136][p. 115–121].

Under-representation may arise if the selection of the training set is far from uniform over the domain of the mapping. In this case there are more examples drawn from one

region and less from another. Because training schemes are based on the mean-squared-error criterion, which is more sensitive to large frequently occurring errors, poorly represented regions will have a lesser effect on the MSE than those represented by large numbers of examples[136][p. 111–114]. As a consequence, the final approximation may be inadequate for regions represented by few examples.

Overtraining arises in situations where the training set is small and there is need to apply it many times before the network parameters converge to the optimal (training often requires a considerable numbers of iterations). If this occurs, the following peculiar effect is often observed. The MSE for the training set reduces and levels off. At this point the network has 'learned' all information it could and further training produces little improvement. When testing with new data, not present in the training set, though, the MSE for this set drops and then starts increasing again. In other words, the ability of the network to interpolate has been lost, although the fit to the training set may be almost perfect. This is often linked in the way the network forms the approximation. It originally starts by constructing a smooth, continuous input/output mapping surface, which becomes more 'crinkled', as the network learns the required mapping. If training continues past that stage, the surface becomes convoluted and attempts to fit the points presented as a training set, ignoring what occurs between them. To remedy this problem, noise is often introduced in the input vectors during successive applications of the training set. This generates is equivalent to an expansion of the training set by the introduction of a number of 'fake' pairs. This is not expected to be a problem in the application considered, since the training set is large (thousands of samples).

Appendix G

Artificial Neural Network Library

A portable library of routines to construct, train and use Artificial Neural Networks is described here. It was written in the C programming language under the UNIX operating system but it may be ported to other systems without modifications. The library provides a small number of subroutines which may be grouped into three categories:

Network Constructors: These are routines that allow the user to define the topology of the network, how many inputs, outputs and layers it has, how many processing elements in each layer and how these are interconnected. Because of the complexity of this information, these were specified in a file, which is read by the constructor routine, called `creatNet()` and allocates the required network in memory, forming the necessary connections between the various elements. The current system only permits the construction of layered structures, like the MLP, but the data structures were defined in anticipation of its generalization later. Initial weights may also be specified along with the structure. Thus the system may be set up as explained in section 11.6.2. An example configuration file is shown at the end of this Appendix.

Network Operators: These include three main routines, which deal with the operation of the network. `initNet()` initializes all weights of the network to small random values, as already explained in section 11.7.2. `runNet()` performs the forward pass of the network, thus for the current input it produces the output vector. This is repeatedly called both for training and afterwards, during the operation of the MLP. Finally, `trainNet()` performs backward propagation of errors and weight adjustment. There are also some auxiliary routines, related to the operation of the network. These provide a means of selecting between a number of sigmoidal non-linearities (see section 11.4.1) and variants of the backpropagation learning algorithm (section 11.4.2).

Interface Routines These include a routine `showNetStat()`, which reports the status (MSE, weight values, changes in the MSE from the previous adaptation of weights, etc.) so that the progress of the network during training may be monitored. It also includes two more routines, which write the network configuration and the weight values after training, either on the screen (`showNetInfo()`) or in a specified file, (`saveNet()`). The format of the written information is the same as that used by `creatNet()` and may be readily used for further training. The format of this is also convenient for visual inspection, which proved helpful for drawing some conclusions at the end of Chapter 11.

A file used by the library to read or save an ANN contains both the configuration (geometry of the network) as well the weights of all its processing elements. An example of such a file is shown here:

```
lambda 0.100000
nu     0.100000
alpha  0.000000
initW  0.300000
traceR 288.866753
inputs 6
layers 2 {
nodes 6[7 2] {
{ 0.131312 -0.160101 0.042230 0.081605 0.151219 0.058248 0.012904 }
{ 0.029806 -0.001687 -0.025264 0.156314 -0.042600 -0.102633 -0.022871 }
{ 0.132936 -0.090213 0.017072 0.015283 -0.031051 -0.122041 0.061764 }
{ 0.096641 -0.136355 0.039885 -0.106059 -0.005646 0.006175 -0.051060 }
{ 0.022416 0.153271 -0.100071 -0.115765 -0.095419 -0.002896 0.126090 }
{ 0.127473 0.100500 0.108526 0.065083 0.127340 -0.013964 -0.064449 }
}
nodes 2[7 1] {
{ 0.165319 0.055438 -0.085959 0.178045 0.078292 -0.538782 0.101768 }
{ -0.186715 0.108167 0.248357 0.001600 0.297533 0.149091 -0.213862 }
}
}
outputs 2
```

Appendix H

Publications

- MYLONAS, S. A.; COMLEY, R. A.; (July, 1991): **On-Line Detection of the Epileptic Precursor**, Proceedings, First Cyprus International Conference on Computer Applications to Engineering Systems, pp. 178–183.

- OMAROUAYACHE, S.; MYLONAS, S. A.; ELLIS, T. J.; COMLEY, R. A.; (1992): **Transputer Implementation of Adaptive Noise Cancelling In Digital Images**, Proceedings, PACTA '92, Barcelona, Spain, pp. 964–974.

- MYLONAS, S. A.; COMLEY, R. A.; (November 1992): **Detection of Epileptic Spikes in the EEG Using Adaptive Filters**, Proceedings, I Fórum Nacional de Ciência e Tecnologia em Saúde—XIII Congresso Brasileiro de Engenharia Biomedica, Caxambu (MG), Brazil.,

- MYLONAS, S. A.; COMLEY, R. A.; (December 1992): **Adaptive Predictive Modelling for the Analysis of the Epileptic EEG**, Proceedings, ICCS/ISITA '92, Singapore, *Vol. 3*, pp. 1214–1218.

- MYLONAS, S. A.; COMLEY, R. A.; (July 1993): **Linear Prediction, Neural Networks and the Analysis of EEG signals**, Proceedings, International Conference on DSP and Second Cyprus International Conf. on Computer Appl. to Eng. Sys., pp. 286–291.

- MYLONAS, S. A.; COMLEY, R. A.; (September 1994): **EEG Analysis Using a Multi-Layer Perceptron With Linear Preprocessing**, Proceedings, IV International IEEE Workshop on Artificial Neural Networks In Signal Processing, pp. 671–680.

ON-LINE DETECTION OF THE EPILEPTIC PRECURSOR

S. A. Mylonas† R. A. Comley†

ABSTRACT:- The use of a portable microcomputer-based programmable medical instrument for long-term EEG monitoring is outlined. An algorithm for real-time detection of a feature associated with the onset of petit mal epilepsy is described. Early results on the evaluation of the performance of the method are presented. The degree of success in meeting the objectives is discussed.

1.0 INTRODUCTION

The human electroencephalogram (EEG) has been used as a medical diagnostic aid for many years. Taking an EEG is a simple, harmless and non-invasive procedure. Electrodes are placed on the surface of the scalp, in a standard montage, with eight, sixteen, or more channels recorded on paper for subsequent human interpretation, based on characteristic features related to disorders of brain function [1]. This is often a tedious, time-consuming, repetitive and error-prone activity, leading to subjective results, as Barlow [2] explains. The need for standardization and more time-effective use of human analysts has led to numerous attempts to automate the analysis of the EEG. However, despite several decades of research, human interpretation remains the major means of analysis to this day.

This paper is concerned with the automatic detection of the Spike and Wave (SAW) feature (Fig. 1.1), which is associated with a form of idiopathic epilepsy, known as petit mal (absence). The absence attacks usually start in childhood. They last for a few seconds and are characterized by transient loss of consciousness and momentary inhibition of activity [3]. Several attacks may occur every day but due to their nature and duration they cannot be identified unambiguously from non-absence incidents with a short, traditional, EEG recording. Long-term monitoring has been suggested as a better alternative, monitoring the patients in their normal working environment. A method, leading to the development of a fully portable unit, based on microprocessor technology, has been proposed[4]. The EEG is analyzed on-line, in real-time, discarding hours of clinically insignificant recording and isolating the few seconds of petit mal seizures to be recorded for confirmation by an experienced human analyst. Results of automatic analysis, also performed in real-time, could be useful for cross referencing.

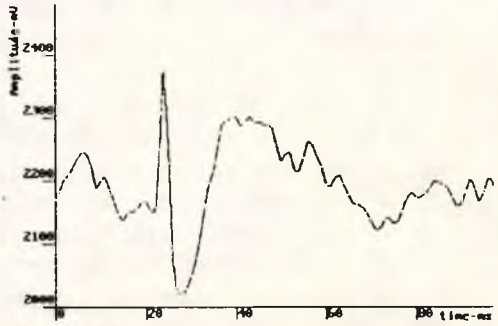


Fig. 1.1 Typical Spike-and-Wave (SAW) feature

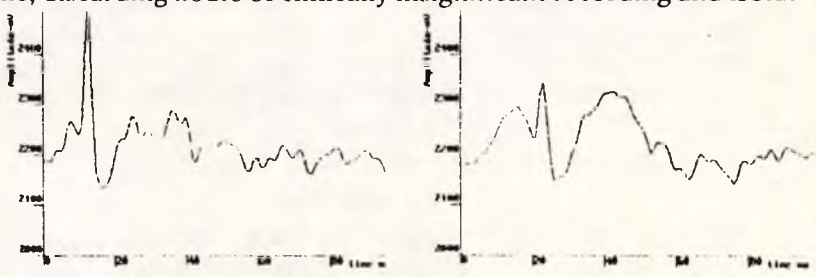


Fig. 1.2 Atypical SAW features

An algorithm to meet these requirements has been developed and is the main subject area of this paper. The proposal of a single SAW detection method is inhibited by factors as fundamental as its definition, since the medical language is often qualitative and informal, not suitable for direct computer

† Centre for Information Engineering, City University, London, EC1V 0HB, United Kingdom.

implementation. Moreover, the experience on which human analysts so often rely is not possible to incorporate as software on a computer. These two factors account for the successful identification of variable and distorted SAWs (Fig. 1.2) which has not been matched by automatic detectors yet. Finally the contamination of the EEG by spiky muscle artifact potentials (Fig.1.3) often causes false detections. It should be stressed that automatic analysis of EEG signals is not aiming to replace the traditional manual analysis but rather to complement it.

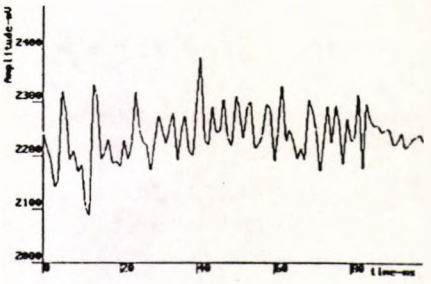


Fig 1.3 A Muscle Artifact

2.0 BRIEF REVIEW OF COMPUTERIZED EEG ANALYSIS

The possibility of computerizing EEG analysis has long been appreciated. With objectives varying from noise cancellation to the production of full reports, a number of workers from medical, electronics and computer disciplines have produced numerous systems to meet these objectives with varying degrees of success. Some techniques used for the detection of epileptic events will be briefly outlined here. The interested reader is referred to the review by Barlow [2] and the papers by Gevins [5] and Jansen [6] for a detailed account of these and other techniques. These methods, may be broadly classified into five categories.

Baseline crossing and selective filtering are used to detect specific features that are related to specific frequency bands [Ehrenberg and Penry (1976)].

Matched Filtering and its variants can be used to detect predefined features in a signal, using convolution and thresholding. It finds most use in the detection of known signals in random noise [Saltzberg et al. (1971), Herolf (1971), Saltzberg (1975) and Barlow and Dubinsky (1976)]. Similar to this is inverse filtering [Barlow (1978)].

Signal parametrization consists of representing the EEG signal by various amplitudes and duration measures and using pattern recognition or statistical techniques to classify sections of the EEG [Bickford (1959), Kooi (1966), Brazier (1967), Salzberg (1967), Buckley (1968), Smith and Ktonas (1974), Gotman and Gloor (1976)].

Various measures based on the time derivatives of the EEG signal have been used to enhance and detect spikes. Fast, simple and relatively effective spike detectors, suitable for real-time have been reported [Carrie (1972), Walter (1973), Hill and Townsend (1973), Gevins et al. (1975), Comley and Brignell [4](1981), Stelle and Comley [8](1989)].

Autoregressive filtering has been used for signal and system modelling, calculation of spectra etc.. An autoregressive filter is estimated, whose output resembles the required signal, using error minimization criteria. Large error levels may indicate transients, such as spikes. The method is generally unsuitable for real-time and it has been used with variations by known research groups [Lopes Da Silva et al. (1977), Praetorius et al. (1977), Birkner et al. (1978), Pfurtscheller and Fischer (1978)].

3.0 THEORETICAL BASIS OF OUR METHOD

3.1 Modelling the EEG waveform

Several mathematical models of the SAW have been suggested. These are often employed for formal description of detection algotithms. The model discussed here was selected not for its accuracy but for the foundation of the theoretical basis of our work. The EEG waveform was assumed to be oscillating with both its frequency, Ω, and amplitude, A, varying with time:

$$e(t) = A(t) \sin\{\Omega(t)t\}$$

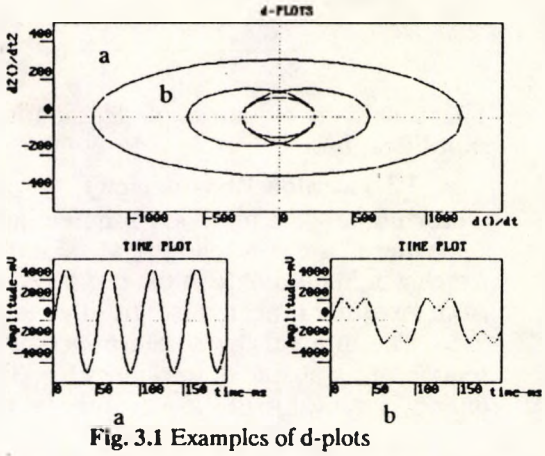
It was further assumed that both A(t) and Ω(t) do not suffer sudden jumps (discontinuities).

If A(t) and Ω(t) are constants A_0 and Ω_0 repectively, the signal is a sinusoidal and its first two derivatives are a cosinusoid and a sinusoid scaled according to its frequency. Therefore, plotting the second as a function of the first derivative, let us call this the d-plot, will give an ellipse centred at the origin (Fig.3.1a):

$$\frac{[e(t)']^2}{a^2} + \frac{[e(t)'']^2}{b^2} = 1$$

where $a = A_0\Omega_0$ and $b = A_0\Omega_0^2$

Obviously, the shape of the d-plot with varying A(t) and W(t) will not be an ellipse any more (Fig. 3.1b).



a b

Fig. 3.1 Examples of d-plots

3.2 The Spike Detector

According to the model, if the range of both A(t) and Ω(t) is limited, the case for the normal EEG signal, the d-plot will occupy a limited area, bound by a closed curve c. Spikes, on the other hand, are characterized by values of Ω(t) where amplitudes A(t) are higher than normal. If "normal" is related to the behaviour of the signal excluding spikes, it is expressed as the interior of the curve c. Therefore, assuming that the spike was sinusoidal, with $A(t) = A_s$ and $\Omega(t) = \Omega_s$ respectively, the d-plot would be an ellipse of larger size than the normal d-plots. This fact is used to distinguish spikes from normal EEG signals, using the description of a spike d-plot as:

- passing from the interior to the exterior of c and
- describing an ellipse with $a_s = A_s\,\Omega_s$ and $b_s = A_s\Omega_s^2$

In practice, though, c is not generally known and real spikes are not sinusoidal. Additionally, real spikes vary one from another. For these reasons some tolerance limits on spike d-plots were specified, as outlined later in section 4.1. The problem of defining the separation surface c was solved by statistical means. Analysis of existing EEG records showed that d-plots were concentrated at the origin, spreading in both horizontal and vertical directions, the number of points reducing with increasing distance from the origin. Under the assumption that the distribution of such points is bivariate normal with zero mean, the decision surface, c, is an ellipse, whose shape and orientation depends on the covariance matrix:

$$\Phi = \begin{pmatrix} \sigma_{xx}^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_{yy}^2 \end{pmatrix} \qquad \text{with } \Phi^{-1} = \frac{1}{\Delta}\begin{pmatrix} \sigma_{yy}^2 & -\sigma_{xy}^2 \\ -\sigma_{xy}^2 & \sigma_{xx}^2 \end{pmatrix}$$

where $\Delta = \sigma_{xx}^2\sigma_{yy}^2 - \sigma_{xy}^2\sigma_{xy}^2$ and x and y are the first and second time derivatives of the EEG signal, respectively. Its size depends on a preset significance level, d_0 and is defined by the (Mahalanobis) distance measure:

c: $\underline{y}^T\Phi^{-1}\underline{y} = d_0$, where $\underline{y}^T = (x \;\; y)$, giving the ellipse

$$\frac{\sigma_{yy}^2}{d0\Delta}x^2 - 2\frac{\sigma_{xy}^2}{d0\Delta}xy + \frac{\sigma_{xx}^2}{d0\Delta}y^2 = 1 \qquad (1)$$

To determine the location of a point with respect to this ellipse, it is substituted in the LHS of equation (1). If the result, d, is between 0 and 1 the point is inside the ellipse, if it is 1 it is on, if greater than 1 it is outside the ellipse. The result, d, is a measure of how far away from the ellipse a point lies. It was found, during testing, that σ_{xy}^2 was small and changed randomly about

zero. This was assumed to be due to sampling uncertainty. Therefore the equation for c was further simplified:

$$\frac{x^2}{d0\,\sigma_{xx}^2} + \frac{y^2}{d0\,\sigma_{yy}^2} = 1 \qquad (2)$$

There was no evidence of degradation in the performance of the detector due to this simplification.

3.3 The Slow Wave detector

Unfortunately the method for detecting spikes is unsuitable for the detection of slow waves, since these have very low derivatives and are often contaminated by higher frequency sinusoids which would dominate in the d-plot. A method using the energy difference of the EEG signal from an initial time, t_0, just after the negative peak of the spike, to some later time, $t_0+\tau$, was used. The interval τ was determined as the duration of a typical slow wave, assumed to be a section of a sinusoid of frequency $f_s = 3.5$ Hz, starting at t_0. The baseline for calculation of the energy is the value of the waveform at $t = t_0$, $e(t_0)$. The integral

$$E_c = \int_{t_0}^{t_0+\tau} \{e(t-t_0)-e(t_0)\}\, dt$$

is compared to the energy of a typical slow wave

$$E_s = \int_{t_0}^{t_0+\tau} \{A_s \sin[2\pi \tfrac{1}{T}(t-t_0)]\, dt = \frac{A_s}{2\pi} T\, (1-\cos 2\pi \tfrac{\tau}{T})$$

In practice a relative measure, $\hat{E}_s = E_s/E_c$, is evaluated, where the normalizing factor $E_s = A_s\tau$ is the energy of a constant signal of value A_s.

A similar relative measure, $\hat{E}_c = E_c/E_c$ is computed on-line, and compared to $\hat{E}_s$ for the detection of the slow wave.The unknown amplitude A_s estimated by the maximum deviation of the low-pass filtered e(t) from the baseline, $e(t_0)$, for the duration of integration.

4.0 IMPLEMENTATION AND PRACTICAL CONSIDERATIONS

Although the theoretical basis of the detection technique was developed using continuous signals, the implementation was in discrete-time. A scheduler, called on the arrival of a new sample, organizes the execution of all scheduled tasks (Fig. 4.1) within one sampling period, T (i.e. in real-time). A more detailed description of these tasks follows.

The analogue-to-digital converter reader obtains a new sample and stores it in a circular buffer, e().

The low-pass filter removes any frequencies above 5Hz in the EEG, with a simple second order IIR filter. Its output is used by the slow wave detector.

The spike detector computes the first two derivatives of the EEG signal, calculates the distance from the ellipse bounding the background (normal) EEG activity (see d-plots in section 3.2) and determines if a spike has occurred. A discrete-time, differentiator is represented by a filter $G(\omega) = j\omega$ for $(k-\frac{1}{2})\omega_s < \omega \le (k+\frac{1}{2})\omega_s$, $k \in Z$ (Fig. 4.2), where ω_s is the sampling frequency. The

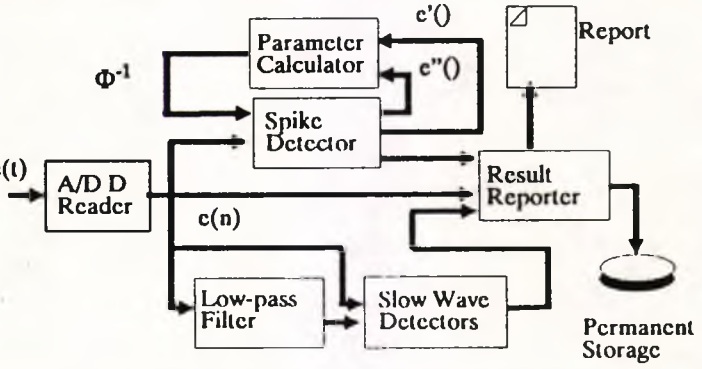


Fig. 4.1 A block diagram of the implemented algorithm.

corresponding impulse response, $g(nT) = \{\frac{(-1)^n}{2nT} x[(n-i)T]\}$, $i \in Z$ was approximated by the simpler difference equation $h(n) = \frac{1}{2T}\{x[(n+1)T]-x[(n-1)T]\}$ (the constant T dropped in the practical realization). The declining performance, as frequency increases (Fig. 4.2), proved to be advantageous since it reduces high-frequency interference causing false alarms [3][8]. Circular buffers are used to store the first and second derivatives, e'() and e''(). The values corresponding to sample e(n), x = e'(n) and y = e''(n), are used to calculate the distance from the ellipse c (section 3.2),

$$d = \frac{x^2}{d0\,\sigma_{xx}^2} + \frac{y^2}{d0\,\sigma_{yy}^2}$$

If d is less than 1, the current sample is considered part of the background EEG activity. Otherwise a possible spike is in progress. Upper limits on the value of d were established, to guard against muscle artifacts and other high-derivative interference, using definitions by other authors [4][7] and the spikes from sample EEG recordings. Instead of a single optimal maximum d, four were used, one per quadrant, increasing the immunity of the detector against non-epileptic spikes. Lower limits were also specified in the same way. Four classes of waveforms are distinguished: True background activity ($d \le 1$), extended background activity ($d > 1$ but less than the lower spike limits), spikes (as defined above) and artifacts (d exceeds the upper spike limits). If a spike is detected, a new slow wave detector is initialized. Many slow wave detectors could be running concurrently since a new spike may appear before a slow wave terminates.

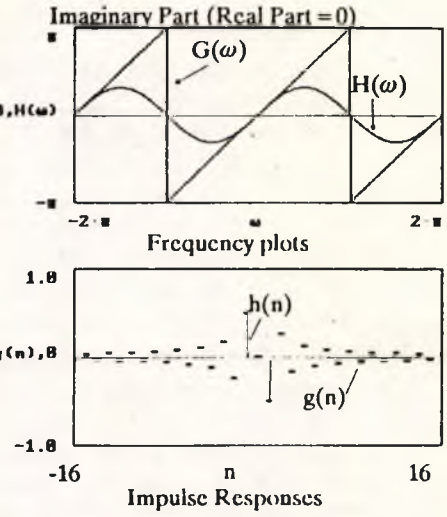


Fig. 4.2 Digital differentiators

The slow-wave detector is almost a direct implementation of the method described in section 3.3. It is initialized by the spike detector which provides the initial value e(0) and the time of integration in samples, n, so that $\tau = NT$. The energy integral is approximated by a summation:

$$E_c = \sum_{n=0}^{N-1} [e(nT)-e(0)]T$$

with the constant T dropped for simplicity. This is calculated by updating E_c with each new sample, e(nT), by adding the value [e(nT)-e(0)] to the previous sum. The maximum of the low-pass filtered EEG for the period of integration is recorded as an estimate of A_s. Low pass filtering was necessary to avoid miscalculation of A_s in the presence of high-frequency components. When the integration time is over, $E_c = A_s.n$ is computed, and $\hat{E}_c = E_c/E_c$ is compared to the constant $\hat{E}_s$ with some tolerance threshold, for an optimal false alarm and missed detection rate.

The parameter calculator estimates parameters of the ellipse, bounding the background activity (see section 3.2), used by the spike detector. The initial parameter values are not very critical, as they are quickly discarded. Only samples classified as true or extended background activity are used for updating. Estimates of the covariance matrix elements for the first (x) and second (y) derivatives of the signal for a sample of size N are calculated:

$$\sigma_{xx}^2 = \frac{1}{N-1}\Sigma x^2, \qquad \sigma_{xy}^2 = \frac{1}{N-1}\Sigma xy \qquad \text{and } \sigma_{yy}^2 = \frac{1}{N-1}\Sigma y^2$$

Circular buffers are used here as well, to store the squares of the derivatives and the cross products. The ellipse parameters are then computed as indicated in section 3.2.

Finally, the result reporter keeps track of the numbers of the spike and wave detections and records clinically important sections of the EEG signal on non-volatile storage media for human assessment later.

EEG record Number	Spikes found by Human EEGer	Automatic Detector	Missed Features	False Alarms
1	39	38(97.4%)	2(5.1%)	1(2.6%)
2	52	45(86.5%)	7(13.4%)	1(1.9%)
3	0	4	0	4
Total	91	87(95.6%)	9(9.9%)	6(6.6%)

Table 5.1 Experimental results

5.0 EXPERIMENTAL RESULTS

The method was tested using a number of FM tape recordings of EEG signals, bandlimited to 70Hz and sampled by a 12-bit data acquisition unit at a rate of 160 samples per second. The method is currently tested with higher sampling rates. Normal and abnormal records were tested by an experienced human analyst and the abnormal features were indicated. The automatic system was used to classify the same recordings by detecting the SAW features. The number of correctly classified and missed spikes and waves and the number of false alarms were recorded. These are presented in table 5.1.

6.0 DISCUSSION AND CONCLUSIONS

In general the method performed well. The muscle artifacts that have caused the false detections proved to be visually similar to spikes and were rejected by the human analyst because they were part of an artifact burst. The missed SAW features were caused by almost non-existent slow waves. Human observers have the ability to change their "thresholding" criteria, using experience, to deal with such cases; this is very difficult emulate using an automatic system.

Some further improvements are planned. These include more artifact rejection, based on the frequency of occurrence of spikes and the signal behaviour before and after the spike, in order to identify artifact bursts. Reduction of missed features is also planned by revising the multiple hard-thresholding on the spike and the slow wave. Thus the detection levels for the slow wave will be conditioned on the quality of the preceding spike. Slow wave detection seems to be a general problem in SAW detection. This could be the reason why many workers often concentrate on spikes, avoiding any attempts to recognize the slow wave feature.

ACKNOWLEDGEMENTS

One of the authors, Mr. S. A. Mylonas, would like to thank the Committee of Vice Chancellors and Principals of the Universities of the United Kingdom (CVCP-ORS award schemes), for the provision of financial support.

REFERENCES

[1] R. COOPER; J. W. OSSELTON; J. C. SHAW:"EEG Technology", 3rd ed. (1980); Butterworths.

[2] J. H. BARLOW:"Computerized Clinical Electroencephalography in Perspective"; IEEE Trans. on Biomed. Eng., Vol. BME-26, No. 7, Jul. 1979; pp. 377-391.

[3] J. N. WALTON:I"Brain Diseases of the Nervous System"; 8th Ed.-Oxford Medical Publishers Company Ltd; (Chapter 22 - Paroxysmal and Convulsive Disorders- Epilepsy), pp. 1093-1132.

[4] R. A. COMLEY; J. E. BRIGNELL:"Real-Time Detection of the Epileptic Precursor"; J. Phys. E:Sci. Instrum., Vol. 14., 1981; pp. 963-967.

[5] A. S. GEVINS:"Analysis of the Electromagnetic Signals of the Human Brain- Milestones, Obstacles and Goals"; IEEE Trans. on Biomed. Eng., Vol. BME-31, No. 12, Dec. 1984; pp. 833-850.

[6] B. H. JANSEN:"Qualitative analysis of Electroencephalograms; Is there Chaos in the Future?"; Int J. Biomed. Comput., 27, 1991; pp. 95-124.

[7] J. R. G. CARRIE:"A Technique for Analyzing Transient EEG Abnormalities"; Electroenceph. Clin. Neurophys., 1972, 32; pp. 199-201.

[8] A. L. STELLE; R. A. COMLEY:"Portable Analyzer for Real-Time Detection of the Epileptic Precursor"; XI Brazilian Congress on Biomedical Eng., Sept. 1989, Proceedings; pp. 101-107.

TRANSPUTER IMPLEMENTATION OF ADAPTIVE NOISE CANCELLING IN DIGITAL IMAGES

S. Omarouayache[1], T.J. Ellis[1], S.A. Mylonas[2] and R.A. Comley[2]
Dep. of Electrical, Electronic and Information Engineering
City University, Northampton Square
London EC1V 0HB
U.K.

SUMMARY

A method for removing noise from digital images using adaptive filtering is introduced. Noise cancelling is extended to a related edge detection method, based on the derivative of the adaptive filters, simple non-maximum suppression and thresholding. It will be demonstrated that this process can be partitioned for data and functional parallelism, suitable for implementation on a transputer network. Some network configurations and the corresponding algorithms are presented. The performance of these structures and the results of the method on some images are discussed.

1. INTRODUCTION

The work presented in this paper uses an adaptive approach to noise cancelling in digital images. This improves the performance of a subsequent edge operator. Noise filtering is necessary as a pre-processing stage for edge detection [1]. By using an adaptive filter, assumptions usually made about the nature of the noise in the image are relaxed. This leads to an improved performance over a wide range of input images.

The algorithm implemented consists of two linear noise cancelling filters cascaded with the transpose of their derivatives. The output, thus produced, is passed through a non-maximum suppression process which ensures a single response. The resulting image is thresholded using a smooth non-linear function.

Inherent parallelisms are identified and exploited through the use of a transputer network. Data parallelism leads to partitioning of the image into small segments that are processed independently. Further parallelism can be identified in the algorithm, since the horizontal and vertical kernel can be applied to the image separately.

---

[1] Machine Vision Group.
[2] Centre for Information Engineering.

2 SIGNAL PROCESSING AND DIGITAL IMAGES

One of the first stages in image processing is the extraction of information from an image, presented in digital form as a matrix of intensities in quantized (grey level) form. At this level the image may be regarded as a two-dimensional signal containing information such as objects on a background. These are often corrupted by other signals, such as noise.

Noise affects the reliable extraction of features, like edges from an image. Most edge detectors use gradient information to detect boundaries of objects and are particularly susceptible to "high-frequency" (spiky) interference. It is therefore necessary to reduce or, if possible, remove the noise before edge detection can take place.

2.1 Noise cancelling

Noise is often removed by applying a filter to the image. The filter coefficients are selected to remove as much of the noise as possible, by smoothing the image whilst maintaining the desired image features unchanged. Failure to select the appropriate coefficient values may result in either high levels of noise or excessive smoothing. Both effects can be hazardous for subsequent edge detection since they may lead to either false detections or missed edges.

If the characteristics of the noise are known beforehand, the filter can be tailored as described. This is often not the case. It is, however, possible to start from an initial guess for the coefficients and correct them when the image, and hence the noise characteristics, become available.

To determine the optimal noise cancelling filter, the study of the signal through a model is often useful.

2.2 A model for the image

Although an image is a two-dimensional signal, its horizontal and vertical components can be studied separately. For edge detection purposes the cross-section of an edge is almost constant in any direction, for small steps in the perpendicular direction [1].

Edges usually signify boundaries between regions of interest, such as objects. In a grey-level image, these are distinguished from their different intensities. Objects normally appear as transitions in the intensity of the image that last for several pixels. Noise, on the other hand, consists of random transitions in the intensity that flood the image and have no structure. The noise-free image signal S(i,j) and noise N(i,j) are added to form the recorded image I(i,j) (figure 3.1(a)):

$$I(i,j) = S(i,j) + N(i,j)$$

N(i,j) is normally treated as a stochastic process with fixed statistical properties throughout an image. The best noise cancelling filter should reflect these properties.

2.3 The adaptive solution to optimum filtering

If a signal d(n) is to be produced from another signal $x(n)$ by linear weighting,

$$y(n) = \sum_{i=-N}^{N} w x(n-i) = \vec{W}\vec{X}(n)$$

a cost function, such as the mean squared error

$$J = E\left[\left[d(n) - y(n)\right]^2\right]$$

must be minimised. The best weight (coefficients) values $\vec{W} = (w_{-N}, \ldots, w_0, \ldots w_N)^t$ satisfy the Wiener-Hopf Equation

$$R_{xx}\vec{W} = \vec{R}_{dx}$$

Where R_{xx} is the auto correlation matrix of x(n) and R_{dx} is the cross correlation vector between x(n) and d(n).

In a noise cancelling situation, x(n) is the linear combination of the wanted signal s(n) with uncorrelated additive noise n(n):

$$x(n) = s(n) + n(n)$$

The objective is to find W such that y(n) is the best approximation of s(n). If d(n) is not available separately, like in our case, x(n) can be used as a noisy estimate of it.

The Wiener-Hopf Equation can be solved recursively, e.g. using gradient search techniques[2][4]. A simple recursion equation for W was described by Widrow and Hoff [2]. It is known as the Least Mean Squares (LMS) algorithm:

$$\vec{W}(n+1) = \vec{W}(n) + 2\mu e(n)\vec{X}(n)$$

where

$e(n) = x(n) - y(n)$

$y(n) = \vec{W}(n)\vec{X}(n)$

$\mu = \lambda / tr(R_{xx}) \quad (0 < \lambda < 1)$

μ is the adaptation step and governs the rate of convergence of the filter. Small μ causes slower adaptation but smaller error in the final $\vec{W}$.

3. ADAPTIVE NOISE CANCELLING AND EDGE DETECTION

3.1 Noise cancelling in images

Most of the adaptive algorithms were initially developed for time sequences. The LMS algorithm is not linked with time-related data and is therefore suitable for applying to image data. There is however a fundamental difficulty in the adaptation procedure. Whereas in a time sequence there is a natural and compulsory direction for filter application and adaptation, (i.e. forward time), such a direction does not exist for a single image. A good procedure is to select randomly the next location in the image on which the filter is to be applied. This method, however, suffers from two practical limitations. First, the adaptation and filtering stages have to be separated, to guarantee that all image elements are processed. Second, localisation of filter coefficients could be lost. Last, noise characteristics may differ from one part of the image to another. Better results are to be expected from locally optimal filters than a globally optimised filter over a large area. For this reason the filter is applied in the horizontal and the vertical directions.

Initially, a two-dimensional (NXN) filter was used for smoothing. This required $2N^2$ operations and did smoothing in both directions simultaneously and resulted in rounded corners. Subsequent edge detection had to be completely separated from the filtering operation.

A different approach proved more effective both in terms of number of computations necessary and in terms of edge detection. It was based on a commonly used filer structure in image processing, the use of two one-dimensional filters to reduce noise in the both the horizontal and the vertical directions. Typically, these filters have fixed coefficients based on assumptions about the nature of the noise. Gaussian filters, are among the commonest. The variance of the Gaussian function, which determines the amount of smoothing is specified at the beginning of the filter process and is not modified afterwards. Each filter produces its own output, which is used for further processing.

Our proposed method of smoothing is similar, but the fixed-coefficient filters were replaced by adaptive ones. The coefficients of the two filters were initialised to either zeros or to those of an initial estimate of a Gaussian filter. Both methods performed almost identically. An example of a smoothed image profile is shown in figure 3.1(b).

3.2 Edge detection

Usually edge detectors are fixed-coefficient differential operators. The differentiator presented here uses the derivatives of the adaptive filters after adaptation is completed. Each of the differentiated filters was then applied to the smoothed image produced by the other filter, like in the Canny operator. The output of this operation is a set of two images, with sharp transitions emphasised and the rest of the image suppressed. Because the other source of large gradient values (i.e. noise) was already suppressed, the remaining large gradients are expected to indicate edges.

Although this operation lacks mathematical rigour, early results, presented in this paper, were encouraging.

3.2.1 Gradient detection

The two differential images produced in the previous section were then combined to detect edges in any direction. The horizontal and vertical differentials at every point may be regarded as a vector of two orthogonal components. The modulus of the vector was taken as a measure of "edgeness". The "direction" of the edge was also determined as angular deviation from the horizontal, although this feature was not used.

3.2.2 Non-maximum suppression

The differentiated image was then treated further to produce a binary image indicating whether a point was classified as an edge or not. A good edge detector should not only detect the existing edges, but perform good localisation too. Part of the second process is a single response to an edge.

A simple non-maximum suppression algorithm produced good results in terms of single response. A candidate edge is not suppressed if it is a local maximum in any direction. This test involves evaluating its gradient modulus against that of its eight neighbouring locations. A further test is then performed to determine if the gradient is high for the area where the point lies, by passing its gradient, g(i,j) through a sigmoidal non linearity where the threshold ϑ is the mean and its exponentiating factor σ is the standard deviation of the gradient in the neighbourhood:

$$f(g(i,j)) = \frac{1.0}{1 + e^{-\sigma(g(i,j) - \vartheta)}}$$

Only if the obtained value is over 1/2, the pixel at (i,j) is considered to be an edge (figure 3.1(c)).

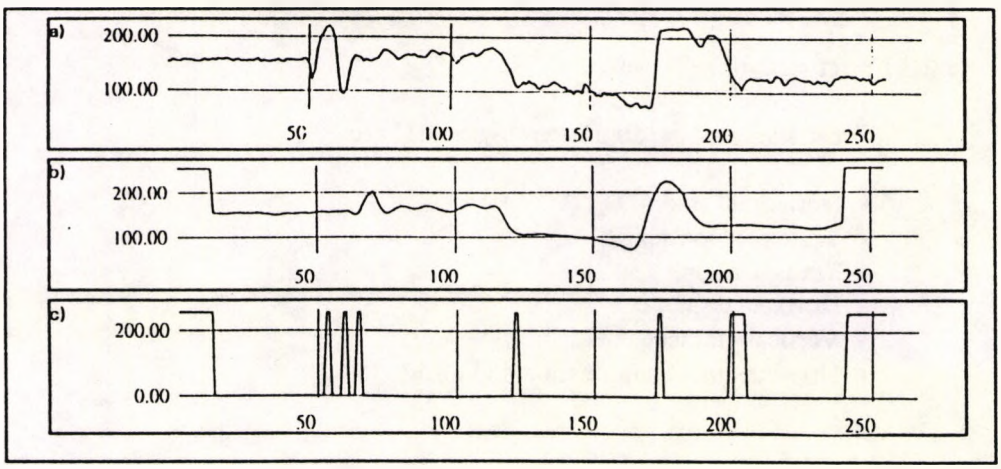


Figure 3.1 One scan line from the image

4. IMPLEMENTATION

In this section the algorithm is described and a transputer implementation is proposed. First, the potential parallelisms are identified; then a mapping of the algorithm on a network of processing elements is presented. The analysis is based on a message passing framework, suitable for transputer modelling.

4.1 Algorithm description

The ideas behind the algorithm can be derived from the definition of the procedure adopted. First, because operations in different areas of the image are independent, data can be shared amongst a population of processors. Next, the unit process is identified in order to take advantage of any inherent functional parallelism. The notion of identifying independence derives from the necessity to limit the amount of communication among processors, and try to achieve induction i.e. linear speedup.

4.2 Inherent parallelism

As mentioned above the system consists of an adaptive noise canceller, cascaded with a local gradient operator, obtained by differentiating the filter coefficients. The output of the latter is fed into a non-maximum suppression unit that enforces single response. Figure 4.1 shows a graphical representation of the processes that implement the system.

4.2.1 Data parallelism

Adaptive noise cancelling relies on the ability of a variable kernel evaluated from a neighbourhood of pixels around the sample to estimate the signal. This completely local dependency suggests the partition of the image into independent blocks that do not require external data. Because the noise canceller and the subsequent gradient operator are convolution operations, half a mask size overlap has to be packed with the data for each block. Although this presents an overhead, it is a major set back only for very small blocks. At this point, the important issue in mapping such a partition on a network of processors is to minimise the distances travelled by data packets, bearing in mind that the real distance is not only the number of hops that data packets makes but the delay incurred during routing as well.

4.2.2 Functional parallelism

The different stages of the algorithm (figure 4.1) are:

- Horizontal smoothing,
- Vertical smoothing,
- Transposition,
- Horizontal gradient,
- Vertical gradient,
- Non-maximum suppression and thresholding.

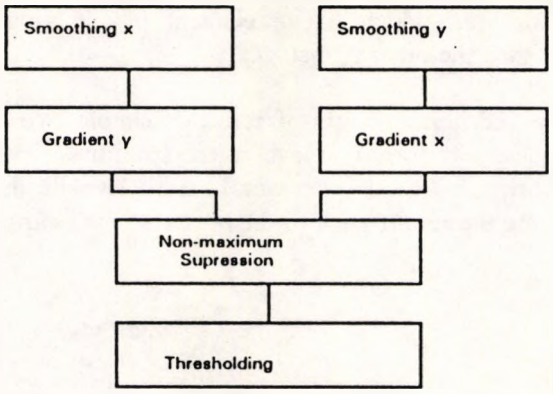


Figure 4.1 Block diagram of the system

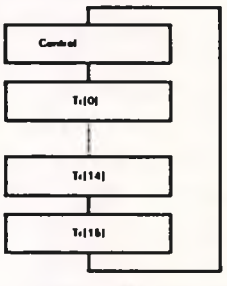
The horizontal and vertical smoothing operations are independent and can be carried out in parallel. The transposition operation refers to the fact that the gradient operator in one direction is applied to the output of the smoothing operator in the other. The outputs of the gradient operations are combined to produce the combined gradient map. This leaves the non-maximum suppression and thresholding as a final stage. This can be implemented as three separate stages.

4.3 Transputer Configurations

The networks described here two types of parallelisms mentioned above. First a simple linear structure is given, then the functional parallelism is incorporated to yield a more efficient architecture.

4.3.1 Simple linear structure

A simple linear structure of sixteen transputers was implemented to take advantage of data parallelism. More complex networks have not been considered because of the large processing to communication ratio. The further performance improvement that can be achieved for such a small number of transputers will be marginal. Figure 4.2a shows this simple structure, while figure 4.2b presents the processes inside each transputer in the network. The root transputer interfaces with the host system (a Sun Sparc station) and provides data partition services and overlap calculations for the rest of the network.



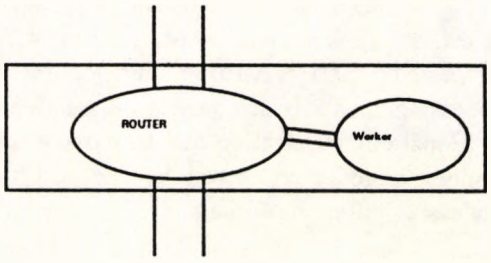


Figure 4.2 a) The simple linear structure, b) Inside T:[i] i = 0..15

4.3.2 Multi-level pipelines

The algorithm as described above decomposes into three separate stages. However, an efficient implementation cannot consider the third stage (i.e. non-maximum suppression and thresholding) as an element of the pipeline since this process is far less complex then the other two stages.

The structure proposed here consists of series of simple two stage pipelines combined with a third processor to form one level in the structure. Levels are stacked to form the overall network. The third processor in each level is used to run the simpler final stage. And route the results back to the host system (figure 4.3).

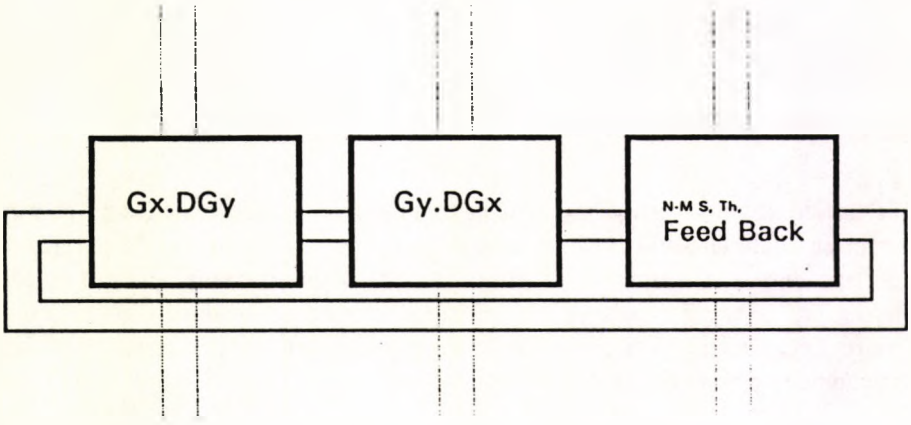


Figure 4.3 Unit structure for multi-level pipeline

5 RESULTS

Results of the edge detector are shown for two sample images in figure 5.1 (c,d), together with the original images (figure 5.1 a,b). The output shows single response to different types of edges and a good localisation.

Filter coefficients after adaptation (figure 5.2) seem to confirm the validity of using a gaussian model for additive noise in digital images. However, this noise is not stationary. The adaptive filter relaxes stationarity and 'zero-mean' conditions.

A speed-up of 12.5 was registered with a sixteen transputer network (see section 4.3.1). The network of section 4.3.2 was partly implemented and results are not available at this writing. Further improvements are anticipated since the main processing load which consists of the two convolution processes is partitioned. The additional communication due to passing the same block to the two processors in the pipeline stage is countered by the fact that only coefficients are passed between processors, after smoothing.

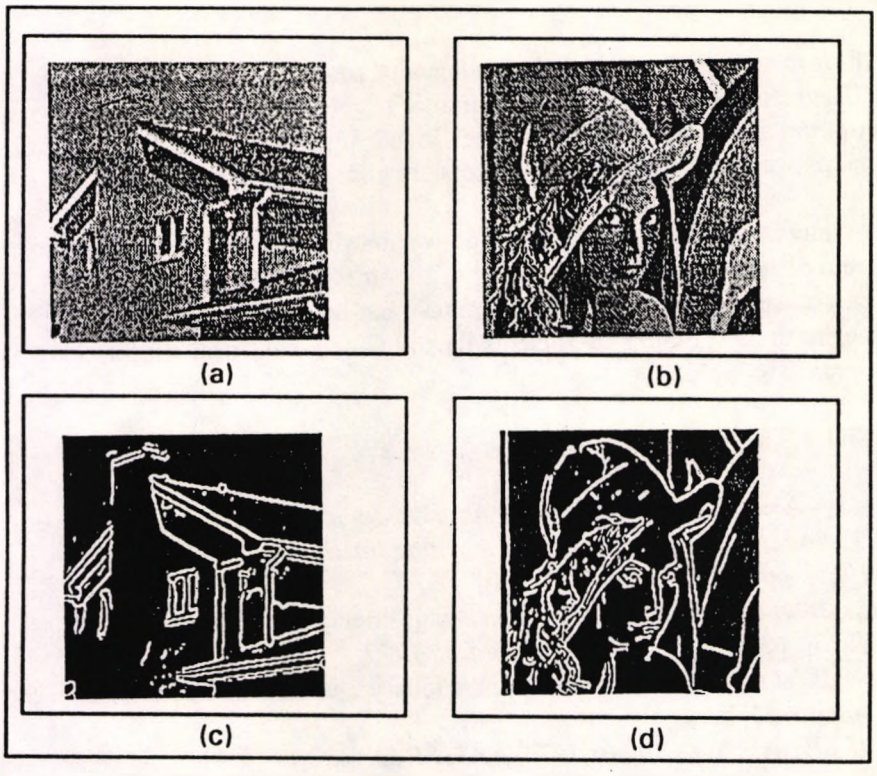


Figure 5.1 Original and processed images

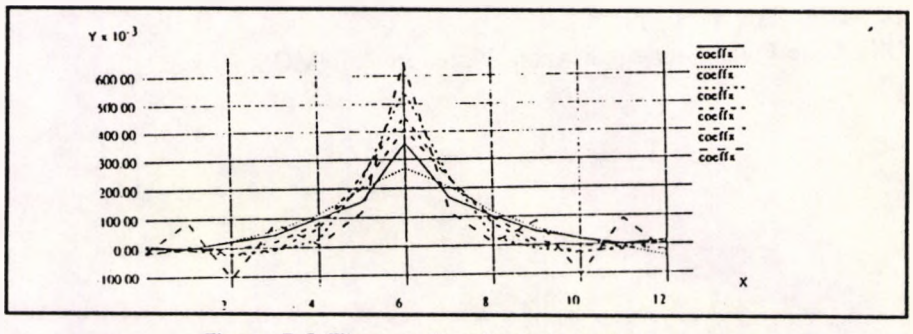


Figure 5.2 Filter profile after adaptation

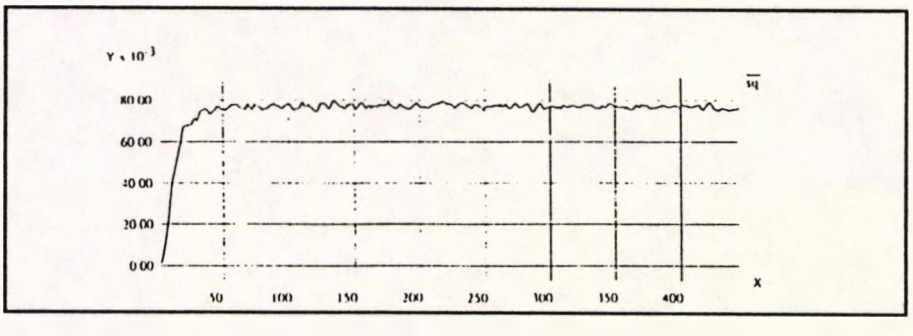


Figure 5.3 Modulus of weights vector during adaptation

6 COMMENTS AND CONCLUSIONS

The system presented here is implements adaptive noise canceller and edge detector. The results show a good localisation of edge elements and immunity to noise. The transputer implementation showed speed ups which suggest that further improvements are possible if more processors are added to the system.

An interesting observation was made when analysing the filter coefficients. The kernels seem to latch on to textured areas of the image. This, together with the good immunity to noise, suggests that adaptive filters can be used as a learning process for texture registration. The filter coefficients thus obtained can form cluster centres of a parameter space to search.

REFERENCES

1. Canny, J. - A computational Approach to edge detection, *IEEE Transact. Pattern Analysis and machine Intelligence: PAMG-8* **6**, *NOV 1986*, pp 679-697.
2. Widrow, B. *et al.* - Adaptive noise cancelling: Principles and Applications, *Proc. IEEE*, **63**, *No 12, pp 1692-1716, Dec 1975.*
3. Widrow, B. *et al.* - Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter, *Proc. IEEE*, **64**, No 8, *August 1976, pp 1151-62.*
4. Orfanidis, S. - "Optimum Signal Processing - An Introduction",2nd Ed., McGraw-Hill, pp 405-34.
5. Quinn, M.J. "Designing Efficient Algorithms for Parallel Computers". McGraw-Hill, New York, 1987.
6. INMOS, "The Transputer Applications Notebook". INMOS, Bristol, 1989

DETECTION OF EPILEPTIC SPIKES IN THE EEG USING ADAPTIVE FILTERS

S. A. Mylonas and R. A. Comley

Centre for Information Engineering
CITY UNIVERSITY
Northampton Square, London, EC1V 0HB
U.K.

Abstract

A method for automatic detection of epileptic spikes in the EEG is presented. It is based on adaptive linear prediction and filtering. The method is suitable for on-line in real-time operation on relatively inexpensive hardware. The theory behind the method is outlined and the implementation is described. Finally some results are presented and discussed. The technique also produces intermediate results which may be of clinical value and can be generalized for the detection of other EEG transients and the removal of artifacts.

INTRODUCTION

Electroencephalography is a simple diagnostic aid for the detection of a number of brain disorders. The electroencephalogram (EEG) has been used for many years for the detection of epilepsy and, despite advances in brain imaging, it is still invaluable for the screening of patients with disorders not related to structural damage of the brain. Petit mal (minor epilepsy) in children, a form of Idiopathic epilepsy[7], is a good example.

Minor epileptic attacks have the form of transient 'absencess', typically associated with the appearance of a characterisic pattern in the EEG record. This pattern, known as the spike-and-wave complex (SAWC), is often not detected in the brief period of an EEG recording session. It has been suggested that prolonged EEG monitoring in the patient's natural environment would improve the reliability of assessment[6]. The large volume of data produced by these methods makes human assessment impractical[1]. The need for mobility of the patient suggests the use of small protable monitoring devices[3]. The use of 'intelligent' monitoring devices, based on microcomputers has been used for some years now with varying degrees of success[1]. EEG assessment, however, is heavily based on the experience of human experts rather than formal criteria which are required by a computer. Therefore computerized analysis cannot substitute human interpretation; it can only assist and complement it.

This paper describes a computerized system made primarily for the detection of the precursor of minor epilepsy (SAWC) in the EEG. It is based on modelling of the time evolution of the EEG signal using adaptive linear prediction. It will be shown that this method is more general than a purpose-built SAWC detector and can be used for other signals as well. The detector is also capable of providing a spike-free EEG signal and other parameters which can be used by either a human or a computer system for subsequent processing.

THEORY

The EEG signal from a signal processing point of view has a non-stationary nature. Its frequency content changes with time. However, for short periods of time, the signal properties could be considered fairly constant. It is therefore possible to study the EEG signal using stochastic models[2, 4].

Model of the signal

A sampled EEG signal $e(n)$ was modelled by three components.

- The background activity, $b(k)$, which is considered as a stochastic process, generated by passing white random noise through an autoregressive filter. This component corresponds to the ongoing EEG activity. Changes in the properties of this component are associated with changes in the filter coefficients.
- Transient activity, $t(k)$, consisting of characteristic patterns of short duration, due to signals of either cerebral origin, like epileptic spikes and evoked potentials, or external sources, like ocular and muscle artifacts.
- Uncorrelated Noise, $n(k)$, from various sources.

Thus

$$e(k) = b(k) + t(k) + n(k) \quad (1)$$

The model described in this section was chosen for its simplicity. It is a signal processing rather than a neurophysiological model, suitable for developing the algorithm to be used and not for describing the underlying mechanisms of the EEG.

Adaptive filtering

According to the model, detecting epileptic spikes is associated with separating the transients, $t(k)$, from the stochastic component, $b(k)$, and the noise. Optimal linear prediction is a method often used for modelling stochastic signals, like $b(k)$, with a few parameters. It is based on estimating the best filter coefficients, w_i, that minimize the mean-squared error of the prediction process.

Then the inverse of the filter would correspond to the generating process of the signal model and the output of the filter is white noise[4].

An optimal linear predictor can be implemented as an adaptive filter, so that it continuously adjusts itself to match changing signal conditions. This can be achieved by one of the adaptive algorithms, like the Least Mean Squares (LMS), developed by Widrow and Hoff[9, 10]:

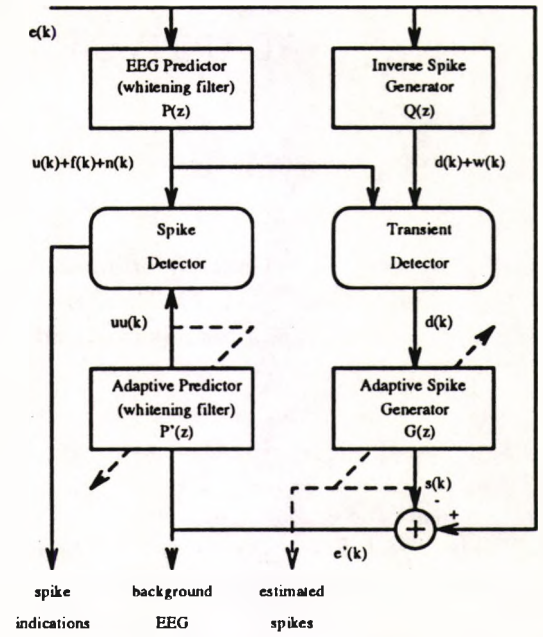


Figure 1: Block diagram of the system

$$e(k) = b(k) + \sum_{n=1}^{N} w_n b(k-n) \quad (2)$$

$$w_i(k+1) = w_i(k) + 2\mu e(k) b(k-i) \quad (3)$$

$$where\ 0 < \mu < \frac{1}{N(Power\ in\ b(k))}$$

Although signal tracking has not been mathematically proved, it is often achieved in practice, provided the signal characteristics change slowly with time[8].

Similarly, spikes and other transients can be modelled by adaptive filters, with impulse rather than noise excitation.

The mathematical basis of these processes can be found in any book on optimal filtering like [5].

IMPLEMENTATION

The main elements of the system

The two whitening filters (figure 1) are associated with the stochastic component of the EEG. They essentially perform the same operation on different signals. The input signal is processed by

one predictor to model the stochastic component, $b(k)$, as explained in section 2. Its output will contain white noise, $u(k)$, from modelling of $b(k)$, excess noise, $v(k)$, due to input noise and transients $f(k)$, which correspond to transient changes in the input, $t(k)$. The latter, is a reliable indication of transients but, being distorted by the prediction filter, it cannot be used for its reliable classification. The other predictor performs the same opeation on a signal $e'(k)$ which is identical to the input with estimated spikes removed. Both predictors have the same coefficients at any time instance but one of them is made adaptive so as to achieve and maintain a good model of $b(k)$.

The inverse spike generator filter acts in a similar way, but focuses on the transients, $t(k)$, instead. Its transfer function is the inverse of that of the adaptive spike generator (figure 1). If a pattern similar to the spike model is encountered, an impulse-like waveform, $d(k)$, is generated at the output (inverse filtering action). This can be used to regenerate the spike from the model using the adaptive spike generator but is an unreliable spike indicator, since non-spike signals may produce similar responses. The input also contains $b(k)$, $n(k)$ and non-spike transients, which cause a compound output component $w(k)$. The spike generator, when excited by an impulse, generates a model of the spike transient in its output. It is made adaptive to cater for variations between the theoretical model and real spikes and for variations among spikes.

The transient detector identifies transients in the output of the predictor by simple statistical significance testing. The spike detector performs a similar operation on both of its inputs.

How the system operates

If the input signal contains a transient, the transient detector will set the hypothesis that it is a spike by causing an impulse, based on $u(k)+f(k)$, to its output. This will produce a spike estimate at the spike generator output which is subtracted from the input signal to produce a hypothetical spike-free signal, $e'(k)$. If the hypothesis is correct, applying the same linear predictor as used on the input, should generate no transient in its output, $uu(k)$, since it was a spike and was removed. This is tested by the spike detector to produce a final decision.

RESULTS, DISCUSSION AND CONCLUSIONS

The adaptive system was tested with both simulated and real data. The method performed quite well on both types of data records. There were only four false alarms out of sixty two detections in a 10 minute record, and no spikes were missed. Muscle artifacts increase the number of false detections but extensions in the basic system to improve the versatility of the detector are currently being studied. The results for a typical input are shown in figure 2. The noise sequences in the error signals show that the selection of the linear prediction model has not been unrealistic. The transients which are apparent in the output of the EEG predictor but not in that of the adaptive predictor demonstarte that spike removal is performed well. This also supports the chosen model of the spike generator.

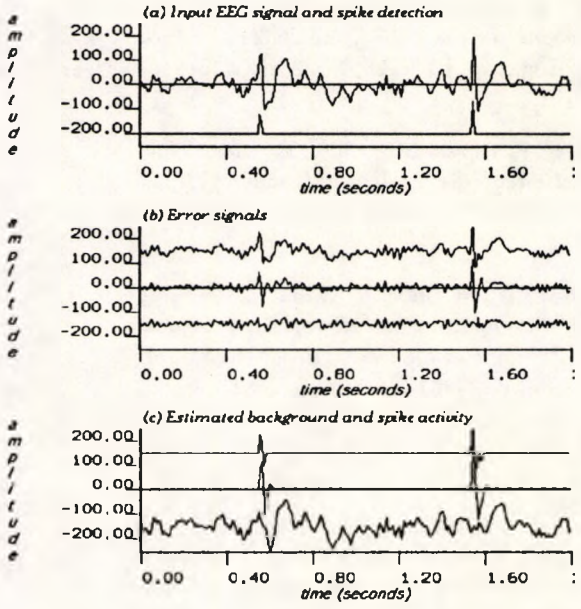


Figure 2: Results

The spike indicator, shown under the input waveform, is not the only useful output of the system. As an intermediate result, a spike-free EEG is generated. This may contain other information, which can be studied, without the obstruc-

tion caused by the spikes, for further assessment of the patient's condition. Moreover, the coefficients of the predictor and the spike filters can be used for the computation of the synthetic spectra of the spike-free EEG and the spikes separately or used directly as spectral parameters.

Finally, the methods described for spike detection may be extended to other transients by the inclusion of the necessary inverse filters. This could prove valuable for the detection and removal of artifacts and for the isolation of evoked potentials.

References

[1] Barlow, J. H. *"Computerized Clinical Electroencephalography in Perspective"*, IEEE Trans. Biomed. Eng., vol. BME-26, No. 7, pp. 377–391, Jul. 1979.

[2] Birkemeier, W. P.; Fontaine, A. B.; Celesia G. G. and Ma K. M. *"Pattern Recognition Techniques for the Detection of Epileptic Transients in the EEG"*, IEEE Trans. Biomed. Eng., vol. BME-25, No. 3, pp. 213–216, Jul. 1978.

[3] Comley, R. A. and Brignell, J. E. *"Real-Time Detection of the Epileptic Precursor"*, J. Phys. E: Sci. Instrum., vol. 14, pp. 963–967, 1981.

[4] Isaksson, A.; Wennberg, A. and Zetterberg, L. H. *"Computer Analysis of EEG Signals with Parametric Models"*, Proc. IEEE, vol. 69, No. 4, pp. 4151–4161, Apr. 1981.

[5] Orfanidis, S. J. *"Optimum Signal Processing-An Introduction" (2nd Ed.)*, McGraw-Hill, NY, 1990.

[6] Stelle, A. L. and Comley, R. A. *"Portable Analyser for Real-Time Detection of the Epileptic Precursor"*, Proc., XI Brazilian Conf. in Biomed. Eng., pp. 101–107, Sep. 1989.

[7] Walton, J. N. *"Brain Diseases of the Nervous System" (8th Ed.): Chapter 22, pp. 1093–1132*, Oxford Medical Publications.

[8] Widrow, B.; McCool, J. M.; Larimore, M. G. and Johnson Jr., C. R. *"Stationary and Non-stationary Learning Characteristics of the LMS Adaptive filter"*, Proc. IEEE, vol. 64, No. 8, pp. 1151–1162, Aug. 1976.

[9] Widrow, B. and Stearns, S. D. *"Adaptive Signal Processing"*, Prentice-Hall Inc., NJ, 1985.

[10] Widrow, B.; McCool, J. M.; Kaunitz, J.; Williams C. S.; Hearn, R. H.; Zeidlen J. R.; Dong, Jr. E. and Goodlin, R. C. *"Adaptive Noise Cancelling: Principles and Applications"*, Proc. IEEE, vol. 63: No. 12, pp 1692–1716, Dec. 1985.

ADAPTIVE PREDICTIVE MODELLING FOR THE ANALYSIS OF THE EPILEPTIC EEG

S. A. Mylonas R. A. Comley

Centre for Information Engineering
CITY UNIVERSITY
Northampton Square, London, EC1V 0HB
U.K.

Abstract

A signal processing model for the epileptic EEG is proposed. This is used to formulate an analysis model, based on linear prediction. This formulation is implemented as a number of adaptive filters and applied for the detection of epileptic spikes. The theory behind the method is explained and the implementation is described. Results are presented and compared for two adaptive filter realizations. The algorithm is computationally efficient and can be implemented in real-time on a small microcomputer system for on-line analysis. Intermediate results, produced by this method may be used for further analysis. Generalization for the detection of other EEG transients and the removal of artifacts can be easily achieved.

1 Introduction

Electroencephalography is a simple non-invasive technique for monitoring the electrical activity of the brain by placing electrodes on the surface of the scalp. The evolution of the electrical potentials of the brain over a period of time, may be manually inspected for specific patterns known to be associated with mental functions or mental disorders. Recent advances in Clinical neurophysiology, like magnetic resonance imaging (MRI) have undermined the usefulness of the EEG. The latter, though, is still a useful diagnostic aid for brain disorders not generally associated with structural damage of the brain. A major application is in the detection of idiopathic epilepsy[11], like petit mal in children.

Petit mal, or minor epilepsy, is evidenced in the form of brief 'absences' in children, thought to be associated with boredom[5]. Minor epilepsy is characterized by a specific pattern in the EEG, known to neurophysiologists as the spike-and-wave complex (SAWC) because of its shape on conventional EEG paper records. Although this pattern is quite characteristic of petit mal, it may not be present in short clinical recordings. The use of prolonged monitoring, has been suggested for more reliable assessment of patients[5, 10]. Human interpretation, however, would be impractical because of the large volume of data. Computer analysis has been employed[1] and with the increasing speed, accuracy and integration of microcomputer systems this can now be carried out on-line in real time. Unfortunately, human assessment is subjective and qualitative and does not rely on formal criteria. It varies according to the ongoing EEG activity and between observers.

In this paper a method for the analysis of the epileptic EEG using adaptive prediction and estimation will be described. This is believed to mimic more closely some aspects of the human analysis process than alternative techniques. Variations on the basic method using different adaptive filter structures are outlined and compared.

Although the work presented here concentrates on the detection of the precursor of petit mal, it can be easily generalized for other signals. It is also capable of separating the EEG into transient and ongoing activity as well as parametrizing it for further study or compression.

2 Theory

2.1 Real EEG signals and signal modelling

An EEG signal is a record of the time evolution of the electrical potential between a point on the scalp and a reference point. Neurons, on the surface of the brain, emit feeble electrical impulses When great numbers of neighbouring neurons operate simultaneously, the electric field produced becomes strong enough to be measured from the surface of the scalp, but is often contaminated by external signals, such as electrical interference, eye movement potentials and muscle potentials (artifacts).

The EEG is often studied isolated from its neurophysiological generating mechanisms, which are often incompletely understood and poorly modelled. Thus, numerous ad hoc signal models have been proposed [1].

Signal analysis shows that the EEG signal is non-stationary. That is short segments, extracted from it have different frequency content. This is partly due to (permanent) changes in the signal behaviour and partly because of transient events. In the absense of transients the non-stationarity would be due to permanent changes only. Assuming that these variations are slow, then the signal could be regarded as (quasi-)stationary for brief periods of time [3, 4, 6]. Under these conditions, stochastic methods can be employed.

A stationary signal, $b(n)$, can be produced by passing white noise, $u(n)$, through a shaping filter, $H(z)$, to obtain a specific power spectral density. Although $H(z)$ has, generally, both poles and zeros, all-pole (autoregressive) models (Equation 1 are still general enough and have some

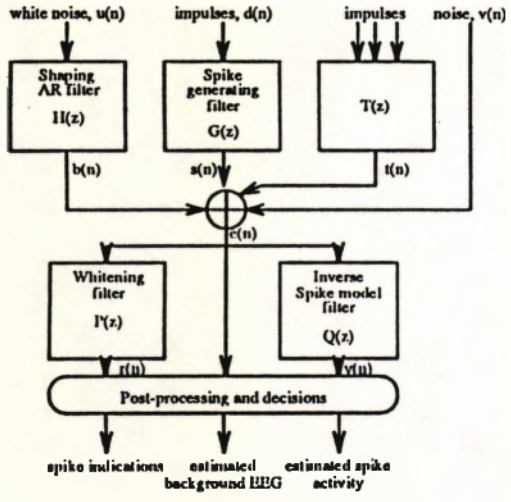


Figure 1: synthesis–analysis model

useful mathematical properties.

$$H(z) = \frac{1}{1 - \sum_{i=1}^{N} h_i z^{-i}} \tag{1}$$

This model, shown in figure 1, has been used as a reference for the analysis of the EEG in this paper. Nevertheless, transients are often more important in clinical applications. The SAWC complex was also modelled by an autoregressive filter, $G(z)$, with slowly varying parameters, excited by a sequence of impulses, $d(n)$, of varying amplitudes, occurring at random intervals following a Poisson distribution.

For completeness, other transients, $t(n)$, like muscle and ocular artifacts, were modelled by impulse excitation of a filter, $T(z)$, whose structure is not of particular interest in this paper. Finally, random, uncorrelated, zero mean normally distributed (Gaussian) noise, $v(n)$, was added to the signal to model measurement, quantization and environmental noise.

Therefore the recorded EEG signal, $e(n)$, is a composite signal:

$$e(n) = b(n) + s(n) + t(n) + v(n) \tag{2}$$

This agrees with the way neurophysiologists often qualify the EEG. Spikes, $s(n)$, and other transients, $t(n)$, are superimposed on the ongoing (background) activity, $b(n)$.

2.2 EEG analysis and inverse modelling

Analysis of the EEG may be regarded as the inverse of the synthesis operation just described. Although the detection of epileptic spikes, $s(n)$, is the aim of this paper, the extraction of $b(n)$ from $e(n)$ is part of the process developed.

For a real EEG, only the combined output, $e(n)$ is known. If the transfer function $H(z)$ was known or could be estimated, the effects of the application of its inverse, $H^{-1}(z)$ ($P(z)$ in figure 1) on $e(n)$ may be studied in the z-domain:

$$\begin{aligned} R(z) &= H^{-1}(z)E(z) \\ &= H^{-1}(z)H(z)U(z) + H^{-1}(z)G(z)D(z) \\ &+ H^{-1}(z)T(z) + H^{-1}(z)V(z) \\ &= U(z) + A(z)D(z) \\ &+ H^{-1}(z)T(z) + H^{-1}(z)V(z) \qquad (3) \\ \textit{where } A(z) &= H^{-1}(z)G(z) \end{aligned}$$

The output of the inverse filter, $r(n)$, contains the sequence $u(n)$ that has generated the ongoing activity and transients from the excitation of $A(z)$ and $H^{-1}(z)$ by the impulse sequences $d(n)$ and $t(n)$ respectively. The transients are apparent because their amplitudes are typically larger than the other elements of the signal and have very brief durations. These properties have been used by [3] and [7] for the detection of spikes.

Unfortunately non-spike transients, generated by T(z), often produce a signature in $r(n)$ similar to that of true spikes. Applying the inverse of the spike generating filter, $G^{-1}(z)$ ($Q(z)$ in figure 1), to $e(n)$,

$$\begin{aligned} Y(z) &= G^{-1}(z)E(z) \\ &= C(z)U(z) + D(z) \\ &+ G^{-1}(z)T(z) + G^{-1}(z)V(z) \qquad (4) \\ \textit{where } C(z) &= G^{-1}(z)H(z) \end{aligned}$$

assuming as before that $G(z)$ could be estimated. The spike generating impulse sequence, $d(n)$, is one of the output elements. The other components, are not impulse-like, but have power levels large enough to make the detection of impulses in $y(n)$ difficult.

In conclusion, $r(z)$ may be used for the *reliable* detection of transients but is unable to distinguish which of them are spikes, whereas $y(n)$ is capable of the *accurate* detection of spikes, albeit unreliable. The combination of the two can form a composite spike detector which may be both reliable and accurate. The proposed detector has the ability to model the spikes and produce separate outputs, for the estimated spikes and the signal with the spikes remove (figure 1 and section 3).

2.3 Optimal signal processing, linear prediction and adaptive filtering

The transfer functions of the filters $H(z)$ and $G(z)$ which are unknown, may be estimated using optimum signal processing. The theory, based on the work of Wiener and Kalman [2, 9, 14] is quite extensive. It deals with the problem of approximating a (time) sequence $\{\ldots, y(n-2), y(n-1), y(n), y(n+1), \ldots\}$ by linear weighting of another sequence, $\{\ldots, x(n-2), x(n-1), x(n), x(n+1), \ldots\}$:

$$\hat{y}(n) = \sum_{i=-\infty}^{\infty} w_i x(n-i) = \mathbf{w} \cdot \mathbf{x}_{(n)} = \mathbf{w}^{\mathbf{T}} \mathbf{x}_{(n)} \tag{5}$$

where $\hat{y}(n)$ is an approximation to $y(n)$ and $\mathbf{x}_{(n)}$ is a vector formed by delaying the $\{x(i)\}$ sequence by n samples. In practice the boundaries of the summation are limited to small integers, for obvious reasons. The (unknown) weights, $\{w_i\}$, are selected to optimize a cost function (criterion for goodness of fit). The mean squared error,

$$\bar{\epsilon^2} = E[y(n) - \hat{y}(n)]^2 = E\{[y(n) - \sum_{i=1}^{N} w_i x(n-i)]^2\} \tag{6}$$

is the one normally used. The optimal weights must minimize $\bar{\epsilon}^2$. This leads to the Wiener–Hopf equation

$$\begin{aligned} R\mathbf{w} &= \mathbf{q} \\ \mathbf{w} &= R^{-1}\mathbf{q} \qquad (7) \end{aligned}$$

where $R(i,j) = E[x(n-j)x(n-i)]$ and $q(j) = E[y(n)x(n-i)]$ are the autocorrelation matrix of x and the cross correlation vector between x and y respectively. Optimal linear prediction, [2, 9] is a special case of constrained Wiener filtering, where the current sample value, $x(n)$ is the desired output, to be estimated from N past samples of the same sequence.

$$\hat{x}(n) = \sum_{i=1}^{N} w_i x(n-i) \tag{8}$$

This gives

$$\epsilon(n) = [x(n) - \hat{x}(n)]^2 = \sum_{i=0}^{N} a_i x(n-i) \quad (9)$$

$$= \mathbf{a}^{\mathbf{T}}\mathbf{x}_{(n)} \quad (10)$$

assuming that $a_0 = 1$ and $a_i = -w_i \forall i \geq 1$. If $\{x(n)\}$ is an autoregressive sequence, like $b(n)$ in the previous section, the optimal linear predictor will be the inverse operation, whose error sequence is the uncorrelated random generating sequence. In other words, linear prediction could be used to find the required inverse of the modelled EEG processes.

Adaptive filters solve equation (7) using recursive techniques, hence achieving significant computational savings. As such, they can be implemented on small cost-effective hardware. Methods that estimate the inverse correlation matrix recursively, known as recursive least squares (RLS) techniques[2, 9, 14] are available, but other methods are often simpler.

Gradient search techniques perform numerical optimization on equation (6) directly by iteratively modifying the weight vector, w, in small steps to reduce the gradient of $\bar{\epsilon^2}$:

$$\mathbf{w}_{(p)} = \mathbf{w}_{(p-1)} - \mu \nabla_{\mathbf{w}} \bar{\epsilon^2}_{(p-1)} \quad (11)$$

where p is the number of iterations.

Widrow and Hoff made a significant step in the estimation of the unknown gradient vector by ignoring the expectation operator letting $\bar{\epsilon^2} = \epsilon^2$, which converts equation (11), to the Widrow-Hoff (least mean squares-LMS) algorithm[14, 12]:

$$\mathbf{w}_{(p)} = \mathbf{w}_{(p-1)} - \mu \nabla_{\mathbf{w}} \epsilon^2_{(p-1)}$$

$$\mathbf{w}_{(p)} = \mathbf{w}_{(p-1)} - 2\mu\epsilon_{(p-1)} \nabla_{\mathbf{w}} \epsilon_{(p-1)} \quad (12)$$

$$\mathbf{w}_{(p)} = \mathbf{w}_{(p-1)} + 2\mu\epsilon_{(p-1)}\mathbf{x}_{(p-1)} \quad (13)$$

where $\epsilon_{(p-1)} = y_{(p-1)} - \mathbf{w}_{(p-1)}{}^T \mathbf{x}_{(p-1)}$. It converges for $0 < \mu < \frac{1}{\lambda_{max}}$ where λ_{max} is the largest eigenvalue of the autocorrelation matrix of $\{x(n)\}$. In practice a more strict range is used which does not use the correlation matrix explicitly:

$$0 < \mu < \frac{1}{N(Power\ in\ \{x(n)\})}$$

Updating of w is done on-line, as every new sample $x(p)$ becomes available. Its simplicity is sometimes offset by its slow convergence, when the ratio of the largest to the smallest eigenvalue of the correlation matrix is large. In the case of the EEG signal convergence is achieved in a few seconds and is insignificant compared to the time of operation of the filter (several hours).

2.4 Comments

We have seen that transfer functions, initially assumed constant, can be estimated by optimal linear prediction. The constancy assumption may also be partly relaxed with adaptive filters, since they continuously adjust themselves trying to maintain a small error . Although they cannot follow all changes in a transfer function, tracking of time-varying coefficients is possible for changes slower than the time constant of adaptation[13].

3 System description

3.1 Structural description

The synthesis-analysis model presented in the previous section is only a theoretical framework. A working system needs to be more elaborate like the one shown in figure 2. The various elements of this will be described here.

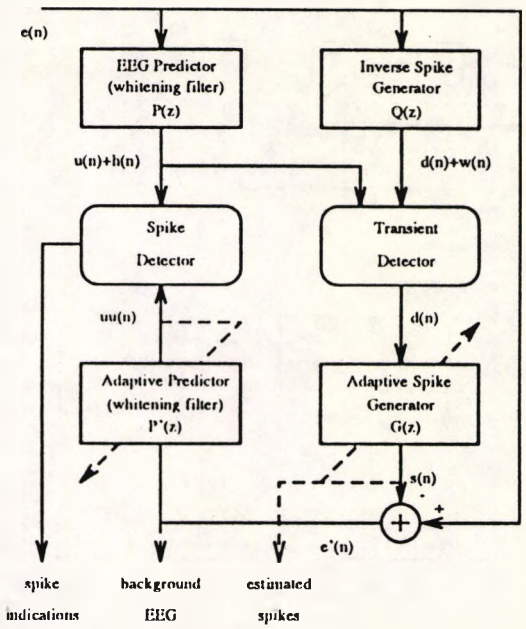


Figure 2: Structure of complete system

3.1.1 Linear predictor structures and algorithms

The linear predictor, $P(z)$, and its function have already been discussed. Two common realizations and the corresponding adaptation algorithms are presented here, the direct and the lattice gradient adaptive filters.

The direct realization (figure 3(a)) uses the transfer function of the linear predictor in its polynomial form:

$$H(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N} \quad (14)$$

Adaptation is done using Widrow's algorithm (equation (13)).

The lattice structure, shown in figure 3(b), is based on the decomposition of the input signal into orthogonal signals and it naturally arises in linear prediction. The (partial correlation) coefficient for stage i, γ_i, can be updated according to the following algorithm:

$$d_i(p) = \alpha d_i(p-1) + [\epsilon^2_{f_{i-1}}(p) + \epsilon^2_{b_{i-1}}(p-1)] \quad (15)$$

$$c_i(p) = [\epsilon_{f_i}(p)\epsilon_{b_{i-1}}(p-1) + \epsilon_{b_i}(p)\epsilon_{f_{i-1}}(p)] \quad (16)$$

$$\gamma_i(p+1) = \gamma_i(p) + \frac{2\lambda}{d_i(p)} c_i(p) \quad (17)$$

where α is a power estimating smoothing factor and λ is the normalized adaptation rate.

3.1.2 Spike detection and generation

In section 2.1, a general model for SAWC generation was given. Formally, a SAWC consists of a brief disturbance (a spike) with duration 20-70ms followed by a half sinusoidal component with frequency around 3 Hz (a slow wave). Earlier studies[8] have shown that the slow wave is not a reliable feature, since many spikes are not followed by distinct slow waves, and so was ignored in this study. The transfer function of the spike-generating autoregressive model, $G(z)$, was determined by optimization, experiment and earlier models. This lead to a heavily damped sinusoid,

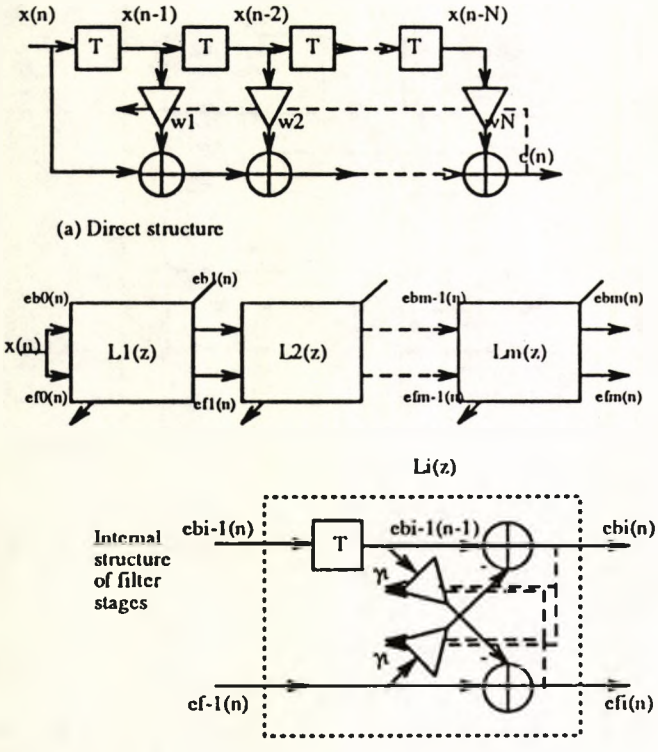


Figure 3: predictor realizations

$s(n) = r^n sin(2\pi f_o nT)$, which corresponds to the following autoregressive transfer function:

$$G(z) = \frac{1}{1 - 2rcos(2\pi f_o)z^{-1} + r^2 z^{-2}} \qquad (18)$$

whose inverse is linear prediction structure (section 2.1). This loosely relates to an earlier model of the spike as a single sinusoidal oscillation[8].

Initially $G(z)$ and $Q(z)$ were set to correspond to a typical spike. Detected spikes are used at run-time to adapt $Q(z)$ to maintain the efficiency of the detector for "non-ideal" spikes. The poles of $Q(z)$ are confined inside the unit circle in the z-domain, to ensure stability, and in the vicinity of the poles for ideal spikes, to limit the effects of biassing from possible false detections.

A gradient adaptive algorithm for IIR filters was implemented for $Q(z)$. This is similar to Widrow's algorithm for FIR filters [14]:

$$\mathbf{a}(p) = \mathbf{a}(p-1) + 2\mu\epsilon(p)\hat{\mathbf{y}}(p) \qquad (19)$$

3.1.3 Decision elements

In the analysis model, a decision block was necessary to detect spikes, using the outputs of the linear predictor and the inverse spike filter. This consists of the transient detector and the spike detector (figure 2).

The transient detector performs a statistical hypothesis test on the output of the prediction filter to find samples that have a high probability of being caused by transients. Its output is an excitation impulse for the spike generator, when a transient is detected.

The spike detector performs similar tests on the outputs of both the input and the output predictors. If a transient exists in the input but not in the output predictor, a spike indicator is activated.

3.2 Functional description

If the input , $e(n)$, contains a spike, $s(n)$, or other transient, $t(n)$, it will produce a transient disturbance, $h(n)$, at the output of $P(z)$ and if it is a true spike, an impulse $\hat{d}(n)$, at the output of $Q(z)$ (section 2.1). The transient detector will detect $h(n)$, and under the hypothesis that it is a spike, will output an estimate of $\hat{d}(n)$, to the spike generator, $G(z)$. This will produce an estimate of a spike, $\hat{s}(n)$, which will be subtracted from the input signal, $e(n)$, to produce a hypothetically spike-free signal, $e'(n)$. If the transient in $e(n)$ was really a spike, then $e'(n)$ would be identical to $e(n)$ with the spike removed. Hence applying $P'(z) = P(z)$ on $e'(n)$ should generate no transient disturbance in its output sequence, $u'(n)$. The spike detector performs the same operation as the transient detector on the outputs of both $P(z)$ and $P'(z)$. It is only the presence of a transient in the output of $P(z)$ and the absense of one from the output of $P'(z)$ that indicates a spike, otherwise the transient is ignored. The model of the spike, $\hat{s}(n)$ and the spike-free EEG signal, $e'(n)$, are also made available for further processing.

4 Results and discussion

The system was tested with both simulated and real data. The results of these tests are presented here for the direct and lattice adaptive predictor structures. Typical waveforms are shown in figure 4. The noise sequences in the error signals show that the selection of the linear prediction model has not been unrealistic. The transients which are apparent in the output of the EEG predictor (Figure 4(b), middle waveform) but not in that of the adaptive predictor (Figure4(b), bottom waveform) demonstrate that spike modelling is performed well.

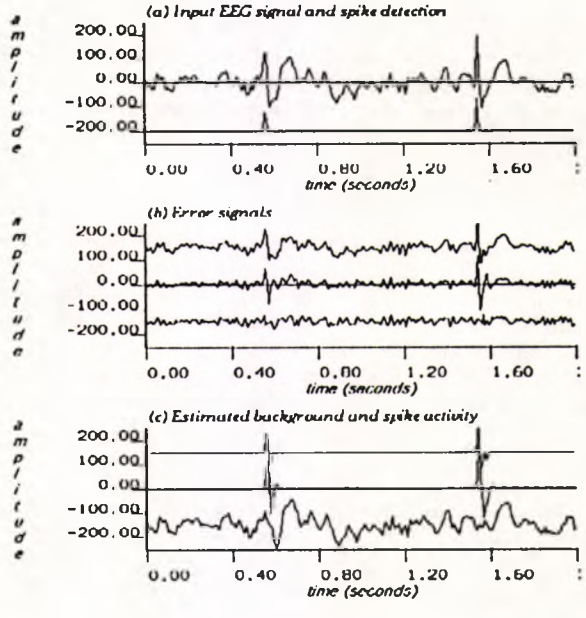


Figure 4: Example waveforms

The two structures are compared, based on a number of criteria.

4.1 Effectiveness

Both filter realizations performed equally well with simulated data, detecting all spikes with no false alarms. In the case of real data, small differences were present (see table

1). In record A, which contained genuine spikes, minor differences were observed. Although the direct realization performed slightly better than the lattice, there is no significant difference in the results. Sensitivity to muscle artifacts, giving a number of false detections, is evident for both realizations. This is primarily because some artifacts are indistinguishable from spikes. They can only be identified from their context (they occur in bursts, whereas genuine spikes are isolated), but this would not have affected the comparison.

Predictor structure	Spike detections					
	Record A (72 Spikes)			Record B (only artifacts)		
	Corr	Miss	False	Corr	Miss	False
Direct	72	0	7	0	0	25
Lattice	71	1	9	0	0	31

Table 1: Results

4.2 Speed of convergence

The lattice structure has the fastest initial convergence rate. This proved to be true for tracking permanent changes in the input signal behaviour (non-stationarity). The direct structure was slower. This is not a very objective result, since the normalized adaptation rate has different significance for each structure. The speed of adaptation of the lattice structure is probably responsible for the high false detection rate in record B, which contained muscle artifacts.

4.3 Modeling accuracy and stability

The accuracy of the filters can only be quoted for the simulated data, where the correct transfer functions are known. Both filters approximate the true transfer function with minor fluctuations particularly in coefficients with a large time lag.

Stability of the adaptation process is the only critical issue. The direct realization did not exhibit instability for λ inside its theoretical bounds. The lattice appears to be more sensitive to large initial transient signals, for adaptation rates close to 1, in which case coefficients may grow to large magnitudes and sometimes diverge. This could be due to the assumption of signal stationarity being more critical in this case. Stability is easy to check, though, since a stable lattice filter has reflection coefficients with magnitudes less than 1.

4.4 Special properties of each filter realization

Although both filters produced comparable results, the choice of which structure to use in a particular application depends on secondary properties of each realization.

The direct form is often used when simplicity and the number of computations are important. It requires $2N+4$ multiplications per sample processed. Its behaviour with non-stationary signals and many of its properties have been studied and there is a considerable theoretical background to support this type of filter.

Studies of the lattice structure indicate that the dynamic range of its coefficients to achieve a particular performance is smaller than other structures[2, 9]. The main advantage of the lattice form arises from the orthogonality of the signals at every stage, making the already calculated coefficients unaffected by changes changes in the order of the filter[9]. Thus, in principle, it is possible to increase the order of the predictor by adding more stages and adapting the new stages only.

5 Comments and conclusions

The results are comparable with those of purpose-built spike detectors [5, 8, 10]. The spike indicator, shown in figure 4(a) under the input waveform, is not the only useful output of the system. Extensions in the basic system to improve the versatility of the detector are currently being studied. Unlike other methods, the one described here not only detects epileptic spikes but also separates them from the ongoing EEG activity. The spike-free EEG produced as an intermediate result (bottom waveform, Figure 4(c)), may contain other information, which can be studied, without the obstruction caused by the spikes. Moreover, the coefficients of the predictor and the spike filters can be used for the computation of the synthetic spectra[9] of the spike-free EEG and the spikes, separately, or used directly as spectral parameters.

Finally, the methods described for spike detection may be extended to other transients by the inclusion of additional analysis-synthesis filter pairs. This could prove valuable for the detection and removal of artifacts and the isolation of other signal components with definite form, like evoked potentials.

References

[1] J. H. Barlow. *Computerized Clinical Electroencephalography in Perspective,* IEEE Trans. Biomed. Eng., vol. BME-26, No. 7, pp. 377–391 (Jul. 1979).

[2] M. G. Bellanger. *Adaptive Digital Filters and Signal Analysis,* Marcel Dekker, NY (1987).

[3] W. P. Birkemeier, et al. *Pattern Recognition Techniques for the Detection of Epileptic Transients in the EEG,* IEEE Trans. Biomed. Eng., vol. BME-25, No. 3, pp. 213–216 (Jul. 1978).

[4] G. Bodenstein and M. Praetorius. *Feature Extraction from the EEG by Adaptive Segmentation,* Proc. IEEE, vol. 65, No. 5, pp. 642–657 (May. 1977).

[5] R. A. Comley and J. E. Brignell *Real-Time Detection of the Epileptic Precursor,* J. Phys. E: Sci. Instrum., vol. 14, pp. 963–967 (1981).

[6] A. Isaksson, et al. *Computer Analysis of EEG Signals with Parametric Models,* Proc. IEEE, vol. 69, No. 4, pp. 4151–4161 (Apr. 1981).

[7] F. H. Lopes Da Silva, et al. *Automatic Detection and Localization of Epileptic Foci,* Electroenceph. Clin. Neurophys., No. 43, pp. 1–13 (1977).

[8] S. A. Mylonas and R. A. Comley *On-Line Detection of the Epileptic Precursor,* First Cyprus International Conf. on Computer Appl. to Eng. Sys. (Jul. 1991)

[9] S. J. Orfanidis *Optimum Signal Processing-An Introduction (2nd Ed.),* McGraw-Hill, NY (1990).

[10] A. L. Stelle and R. A. Comley. *Portable Analyser for Real-Time Detection of the Epileptic Precursor,* Proc., XI Brazilian Conf. in Biomed. Eng., pp. 101–107 (Sep. 1989).

[11] J. N. Walton. *Brain Diseases of the Nervous System (8th Ed.): Chapter 22, pp. 1093–1132,* Oxford Medical Publications.

[12] B. Widrow, et al. *Adaptive Noise Cancelling: Principles and Applications,* Proc. IEEE, vol. 64, No. 12, pp. 1692–1716 (Dec. 1975).

[13] B. Widrow, et al. *Stationary and Nonstationary Learning Characteristics of the LMS Adaptive filter,* Proc. IEEE, vol. 64, No. 8, pp. 1151–1162 (Aug. 1976).

[14] B. Widrow and S. D. Stearns. *Adaptive Signal Processing,* Prentice-Hall Inc., NJ (1985).

LINEAR PREDICTION, NEURAL NETWORKS AND THE ANALYSIS OF EEG SIGNALS

S. A. Mylonas R. A. Comley

City University, Centre for Information Engineering,
Northampton Square, London, EC1V 0HB, UK
e-mail: za300@uk.ac.city

Abstract

Automatic analysis of the electroencephalogram is important in clinical diagnosis of epilepsy and other brain disorders. A model of the EEG signal is proposed and used to derive an analysis method, based on adaptive linear prediction. An artificial neural network is used to distinguish epileptic spikes from other transients. Early results were encouraging and the generalization of the method to detect other transients is possible. The method is suitable for on-line analysis in real-time.

1 Introduction

The importance of the electroencephalogram (EEG) has been undermined by recent techniques, like computed tomography (CT) and magnetic resonance imaging (MRI). Despite this, the EEG is clinically indispensable because of its simplicity, low-cost and diagnostic value, particularly in some forms of idiopathic epilepsy, where there is no structural abnormality in the brain, only a transient functional disorder persisting for the duration of the seizure. The EEG provides a profile of the brain function over a period of time, and often contains indications of an abnormality. These have been identified and some correlate well with specific brain disorders. The "spike and wave" pattern, named after its shape on the EEG record, is characteristic of petit mal epilepsy. Petit mal attacks are characterized by transient loss of consciousness, lasting for a few seconds, during which the EEG is swamped by successions of spike and wave discharges, appearing synchronously over all areas of the scalp[15]. Occasionally isolated spike and wave patterns appear in the EEG between seizures and may or may not be recorded during the course of a usual clinical EEG recording, which only lasts for about 20 minutes. For the more reliable screening of patients, prolonged monitoring was recommended[5]. This makes human interpretation difficult because of the large volume of data produced.

Computer interpretation has been suggested as a viable alternative[5, 14]. In fact, computerized EEG analysis has been applied with varying degrees of success for many years[1]. With the increasing speed, accuracy and integration of microcomputer technology, on-line processing of EEG signals is possible on a fully portable pocket-sized system. The real challenge is the establishment of a reliable automatic assessment procedure. This is not an easy task at all, because human assessment is subjective and qualitative having no formal criteria. Often it is heavily dependent on the experience of the observer. A plethora of computerized methods of incredible diversity has hence developed[1].

The method presented in this paper attempts to address the EEG assessment problem as one of signal analysis. By using a theoretical model of the EEG signal, the problem becomes one of decomposition of the signal into its basic elements. It was achieved using adaptive linear prediction and estimation for preprocessing and a neural network for making decisions, and controlling the other elements of the system. This bears some resemblance to certain aspects of the analysis process as carried out by human analysts. Although the main objective was the detection of the petit mal epileptic pattern, the separation of the EEG in transient and ongoing activity and its parametrization are performed as part of the process.

2 Theory

2.1 EEG Signal modelling

The source of the EEG signal is, primarily, the cerebral cortex. External signals, such as electrical interference, potentials due to eye movement and muscle potentials are sometimes picked up by the electrodes. These non-cerebral potentials, usually referred to as artifacts, contaminate EEG and cause problems in its interpretation.

Most EEG analysis methods use ad hoc models[1], because the neurophysiology cannot provide an accurate and detailed model of the signal. Signal processing methodology, previously applied by a number of workers[1, 4, 7] forms the basis of the method presented here.

It is quite easy to demonstrate the non-stationarity of the EEG signal, since its frequency content changes over time. This is due to permanent changes in the signal behaviour, related to the changing mental state of the subject, and transient disturbances, making small EEG segments to have completely different spectral properties from longer ones. If one could ignore the transients, selected segments of the EEG signal could be regarded as quasi-stationary[4, 3, 7] enabling stochastic modelling methods to be used. A stationary signal, $b(n)$, with a specific power spectral density, is produced at the output of a shaping filter, $H(z)$, when excited by white noise, $u(n)$. $H(z)$ may have both poles and zeros, but an all-pole (autoregressive) model is adequate for many applications (see Figure 1):

$$H(z) = \frac{1}{1 - \sum_{i=1}^{N} h_i z^{-i}} \tag{1}$$

Nevertheless, it is transients like the spike and wave complex, that have more diagnostic value. In fact, spikes are characteristic of many forms of epilepsy. Isolated spikes, $s(n)$, are produced at the output of an autoregressive filter, $G(z)$, having an impulse response that resembles a spike, when excited by a sequence of impulses, $d(n)$, whose occurrence follows a Poisson distribution. Varying the amplitude of the impulses produces spikes of varying amplitude (Figure 1).

Other transients, $t(n)$, like muscle and ocular artifacts, may be modelled in a similar way. To avoid complications these have been grouped into a multiple input multiple output filter, $T(z)$, excited by impulses (see Figure 1). The whole process is also affected by quantization and environmental noise, $v(n)$, assumed to be additive, uncorrelated, following a Normal (Gaussian) distribution with zero mean.

The superposition of these signals produces the EEG model, $e(n)$, which agrees well with those assumed by neurophysiologists, where $b(n)$ is often called the "background" and $s(n)$ and $t(n)$, collectively, the "transient" activity:

$$e(n) = b(n) + s(n) + t(n) + v(n) \tag{2}$$

2.2 EEG Signal analysis

The advantage of having a signal model lies in the ability to derive an analysis method using the inverse of the model and some additional assumptions. In real EEG recordings only $e(n)$ is known. If the transfer functions $H(z)$ and $G(z)$ were known, the application of their inverses, $H^{-1}(z)$ and $G^{-1}(z)$ on $e(n)$ (shown in Figure 1 as $P(z)$ and $Q(z)$, respectively), would produce outputs $r(n)$ and $y(n)$, respectively.

The noise sequence $u(n)$ which generated $b(n)$, is regenerated in $r(n)$ along with a distorted noise signal due to $v(n)$ and transients due to $d(n)$ and $t(n)$ exciting $A(z)$ and $H^{-1}(z)$ respectively[11]. The transient responses have larger amplitudes, and short duration, making them visible on top of other elements of the signal, a fact used by Da Silva et al.[9] and Birkemeier et al.[3] for the detection of spikes in the EEG. Similarly, the spike generating sequence, $d(n)$, is one of the components of $y(n)$. The other components, are distorted versions of $u(n)$, $v(n)$ and $t(n)$. Their power is often large enough to make the detection of impulses in $y(n)$ difficult, especially the non-spike transients, generated by T(z)[11].

The simultaneous use of $r(z)$ and $y(n)$ has combined the reliable transient detection properties of $H^{-1}(z)$ with the accuracy of spike modelling of $G^{-1}(z)$ to produce a reliable spike detector. Post-processing is necessary for this combination (Figure 1), leading to the separation of spikes $s(n)$, from the ongoing EEG activity, $b(n)$, and the detection of spikes[11, 12].

The concepts described so far do not solve the problem since $H(z)$ and $G(z)$ are not known in practice, not to mention the ambiguous post-processing element. The unknown transfer functions of the filters $P(z)$ and $Q(z)$ (Figure 1) may be estimated using optimum signal processing methods. The supporting theory, based on the work of Wiener and Kalman, is well-known and can be found in signal processing texts[17, 2, 13]. In its general form Wiener theory deals with finding a set of linear weights, $\mathbf{w} = \{\ldots, w_{n-1}, w_n, w_{n+1}, \ldots\}$ that can produce the best approximation, $\hat{y}(n)$, of one (time) sequence, $\mathbf{y} = \{\ldots, y(n-1), y(n), y(n+1), \ldots\}$ using another, $\mathbf{x} = \{\ldots, x(n-1), x(n), x(n+1), \ldots\}$:

$$\hat{y}(n) = \sum_{i=-\infty}^{\infty} w_i x(n-i) \tag{3}$$

In practice, the boundaries of the summation are constrained, for obvious reasons. The (unknown) optimal weights, w_i, minimize the mean squared error $\epsilon^2 = E[y(n) - \hat{y}(n)]^2$ This leads to the Wiener-Hopf equation:

$$Rw = q \tag{4}$$

where $R(i,j) = E[x(n-j)x(n-i)]$ is the auto-correlation matrix of $\mathbf{x}$ and $q(i) = E[y(n)x(n-i)]$ the cross-correlation vector between $\mathbf{x}$ and $y(n)$.

A Wiener filter can be implemented by recursively optimizing the weight vector, $\mathbf{w}$, instead of solving equation 4 directly. This is called an adaptive filter and leads to significant computational savings. Recursive Least Squares (RLS) adaptive filters converge fast[2, 17, 13], but gradient-search methods are simpler and numerically stable[16], although they generally converge slowly. These repeatedly estimate the gradient of ϵ^2 and modify $\mathbf{w}$ in small steps to make the gradient close to zero. Widrow and Hoff made a significant step in the estimation of the unknown gradient vector by letting $\hat{\epsilon}^2 = \epsilon^2$ which leads to the Least Mean Squares (LMS) algorithm[17]:

$$\mathbf{w}_{(p)} = \mathbf{w}_{(p-1)} + 2\mu\epsilon(p-1)\mathbf{x}(p-1) \tag{5}$$

where $\epsilon(p-1) = y(p-1) - \mathbf{w}_{(}p-1)^T\mathbf{x}(p-1)$. $\mathbf{w}$ is updated as every new sample, $x(p)$, becomes available. Converge is guaranteed for when $0 < \mu < \frac{1}{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of R, but in practice the simpler upper limit $\frac{1}{N(Power\ in\ \{x(n)\})}$ is used.

When $\mathbf{y}$ is identical to $\mathbf{x}$, the special case of linear prediction arises[2, 13], equivalent to estimating the current value of this sequence using the previous N samples.

$$\epsilon(n) = \sum_{i=0}^{N} a_i x(n-i) \tag{6}$$

where $a_0 = 1$ and $a_i = -w_i \forall i \geq 1$. Adaptive linear prediction is the basis of the realization of the EEG signal analysis method presented here.

3 Implementation

3.1 Predicting the stationary part of the signal

The background EEG signal, $b(n)$ is produced by an autoregressive filter, $H(z)$ (section 2). The optimal linear predictor filter (equation 6) for this signal is the inverse of the generating filter, $H^{-1}(z)$, and $\epsilon(n)$ identical to the generating sequence, $u(n)$:

$$\begin{aligned} u(n) &= \sum_{i=0}^{N} h_i b(n-i) \\ where\ H^{-1}(z) &= 1 - \sum_{i=1}^{N} h_i z^{-i} \end{aligned} \tag{7}$$

Adaptation was performed by the LMS algorithm (equation 5). Slow convergence is not a real issue in our case, since it is achieved in the first few seconds of operation, an insignificant period of time compared to several hours of operation after convergence. Once it has converged, the filter can follow permanent changes in the spectral properties of the signal, provided these are slower than the time constant of adaptation[16]. The use of other structures and algorithms, that have faster convergence, like the lattice was investigated. The advantages were insignificant; fast convergence led to sensitivity to large transients[12].

3.2 Detecting and estimating spikes

Optimization and experiment lead to the description of the spike as a heavily damped sinusoid[12, 11]. This corresponds to the autoregressive transfer function

$$G(z) = \frac{1}{1 - 2rcos(1\pi f_o)z^{-1} + r^2 z^{-2}} \tag{8}$$

whose inverse is a linear prediction structure (equation 6). Initially $G(z)$ and its inverse, $Q(z)$ were set to correspond to a typical spike. $Q(z)$ is adapted every time a spike is encountered to follow small discrepancies from the ideal, using a gradient (LMS) adaptive algorithm for IIR filters[16], similar to equation 5. Divergence from the typical spike, due to false detections, was prevented by restricting adaptation so that the poles of $G(z)$ stay in the vicinity of those of a typical spike, using a non-linear ($tan^{-1}(\cdot)$) mapping. Stability is ensured since this restriction ensures that the poles stay inside the unit circle in the z-plane.

3.3 The working system

The analysis model (section 2.1) required some refinement and extensions to make a working system. This is shown in figure 2, where the decision making block has been refined and the transfer functions $P'(z)$ and $G(z)$ have been added. Their purpose is now explained.

Spikes, $s(n)$ and transients, $t(n)$ in the input, $e(n)$, when passed through $P'(z)$ cause a disturbance $h(n)$, at the output. A genuine spike would also generate an impulse $\tilde{d}(n)$, when passed through $Q(z)$ (section 2.1). These two are combined by the transient detector to generate an impulse, $\tilde{d}(n)$, assuming that the transient is a spike. The spike generator $G(z)$, excited by this, will generate a spike estimate, $\hat{s}(n)$. Subtracting this from the input signal, $e(n)$, should produce a, hypothetically, spike-free signal $e'(n)$. $P'(z)$ is identical to $P(z)$ but attempts to predict $e'(n)$. Hence, if the hypothesis that the transient was a spike is true, no transient should be observed in its output, $u'(n)$, a fact verified by the spike detector. A transient in $u'(n)$ would indicate that the detected transient was not a spike. The model of the spike, $\hat{s}(n)$ and the spike-free EEG signal, $e'(n)$, are also made available for further processing.

Adaptation of $P'(z)$ ensures the correct function of the predictor. Similarly adaptation of $G(z)$ ensures tracking of detected spikes. It should be noted that $P(z)$ and $P'(z)$ share the same coefficient values at all times, as do $G(z)$ and $Q(z)$.

3.4 Decision elements–the multi-layer perceptron

Little has been said so far about the structure of the transient and the spike detectors. Initially, these were represented by simple statistical hypothesis testing elements, detecting whether a sample input belongs or not to the same statistical distribution as the rest of the samples, thus detecting "outliers", assumed to be due to transients[11, 12]. This approach worked quite well, but required that the significance level of the statistical tests be tailored to minimize false alarms and false detections. A more flexible approach was found in the form of artificial neural networks. The multi-layer perceptron (MLP) structure was found quite suitable for this purpose[6, 8].

The MLP learns its function by example. It is "programmed" by presenting it with training patterns, consisting of pairs of input and desired output vector patterns, $\{(\mathbf{x}_0(p), \mathbf{d}(p)\}$. It is then trained so that its output, $\mathbf{y}$ is as close as possible to the desired response, much like an adaptive filter. This is an optimization problem, involving minimization of the mean-squared error and is traditionally done by the generalized delta rule on every node n of layer l:

$$\mathbf{w}_{ln}(p) = \alpha \mathbf{w}_{ln}(p-1) + (1-\alpha)\mathbf{w}_{ln}(p-2) + 2\mu\delta_{ln}(p-1)\mathbf{x}l(p-1) \tag{9}$$

where $\delta_{ln}(p) = d_n(p) - y_{ln}(p)$ for the nodes in the output layer and $\delta_{ln}(p) = \sigma'[v_{ln}(p)] \sum_j \delta_{l+1,n,j}(p) w_{l+1,n,j}$ for the nodes in the hidden layers and $\sigma'[\cdot]$ is the derivative of the sigmoidal non-linearity in the output $y_{l,n}(p) = \sigma[v_{l,n}(p)]$ of every node.

The inputs to the network are power-normalized past outputs of $P'(z)$, $Q(z)$ and $P'(z)$ (11 samples from each signal). It has three outputs, one for spike indications, one for producing $\tilde{d}(n)$ and one for the production of ocular artifact indications (still on experimental stage). It has 38 input, 20 hidden and 3 output nodes. It has been trained mainly with simulated data at this stage. Training required approximately 3000 iterations to produce a satisfactory output.

4 Results

The system was tested with both simulated and real EEG data. Simulated EEG records contained a stochastic component, produced by filtering white noise through a low-pass autoregressive filter, spikes, using impulse excitations of the modelling filter with random perturbations on its poles and ocular artifacts using a impulse

excitation of a first order filter with random perturbations of its only pole. White noise was also added to the system.

Two real records have been used to test the method. The first contained 72 spikes, which were all detected by the system. There were also 5 false detections. The second record contained muscle artifacts and no spikes. There were 28 false detections.

Figure 4 shows typical waveforms of the process. The error signals agree with the theoretical model. Spike modelling is not unrealistic, since the removed spikes give no apparent transients in the output spike-free signal.

5 Comments and conclusions

The method presented here may be regarded as superior to earlier special-purpose spike detectors[5, 14, 10], mainly because of its generality. Adaptation for the detection of other transients has already begun.

The ability of the system to separate transients from the ongoing EEG activity allows these signals to be studied independently. This permits spectral methods to operate on a spike-free EEG signal, without the obstruction caused by spikes and other transients. Parametrization of the EEG signal, performed by linear prediction, may be used for compression, storage, or further analysis.

Although the use of the neural network for decision making did not show any apparent improvements in the system performance, which is almost identical to earlier systems[11, 12], significant advantages are expected to be gained by the ability to train the system whilst in operation. An expert clinician, who need not know the internal details of the system, can continue training of the neural network by presenting it with an EEG signal where transients have been marked and labelled. This work is still at an early stage, but has worked well with simulated EEG records containing both spikes and ocular artifacts. It has been tested on real records containing spikes and showed good prospects. The advantage of having a learning element in the system is twofold. Individual experts could tailor its behaviour to closely resemble their own decisions and groups of them could collectively train it as a common reference for EEG analysis bringing together medical experience and engineering formalism.

References

[1] J. H. Barlow: *Computerized Clinical Electroencephalography in Perspective,* IEEE Trans. Biomed. Eng., vol. BME-26, No. 7, pp. 377–391 (Jul. 1979).

[2] M. G. Bellanger: *Adaptive Digital Filters and Signal Analysis,* Marcel Dekker, NY (1987).

[3] W. P. Birkemeier, et al.: *Pattern Recognition Techniques for the Detection of Epileptic Transients in the EEG,* IEEE Trans. Biomed. Eng., vol. BME-25, No. 3, pp. 213–216 (Jul. 1978).

[4] G. Bodenstein and M. Praetorius: *Feature Extraction from the EEG by Adaptive Segmentation,* Proc. IEEE, vol. 65, No. 5, pp. 642–657 (May. 1977).

[5] R. A. Comley and J. E. Brignell: *Real-Time Detection of the Epileptic Precursor,* J. Phys. E: Sci. Instrum., vol. 14, pp. 963–967 (1981).

[6] R. Hecht-Nielsen: *Neurocomputing,* Addison-Wesley (1990).

[7] A. Isaksson, et al.: *Computer Analysis of EEG Signals with Parametric Models,* Proc. IEEE, vol. 69, No. 4, pp. 4151–4161 (Apr. 1981).

[8] R. P. Lippmann: *An Introduction to Computing with Neural Nets,* IEEE ASSP magazine, Apr. 1987, pp. 4–22.

[9] F. H. Lopes Da Silva, et al.: *Automatic Detection and Localization of Epileptic Foci,* Electroenceph. Clin. Neurophys., No. 43, pp. 1–13 (1977).

[10] S. A. Mylonas and R. A. Comley: *On-Line Detection of the Epileptic Precursor,* First Cyprus International Conf. on Computer Appl. to Eng. Sys. (Jul. 1991)

[11] S. A. Mylonas and R. A. Comley: *Detection of Epileptic Spikes in the EEG Using Adaptive Filters,* I Fórum Nacional de Ciência e Tecnologia em Saúde – XIII Congresso Brasileiro de Engenharia Biomedica, Caxambu (MG), Brazil (Nov. 1992)

[12] S. A. Mylonas and R. A. Comley: *Adaptive Predictive Modelling for the Analysis of the Epileptic EEG,* Singapore ICCS/ISITA'92, vol. 3, pp. 1214-1218 (Nov. 1992)

[13] S. J. Orfanidis: *Optimum Signal Processing-An Introduction (2nd Ed.),* McGraw-Hill, NY (1990).

[14] A. L. Stelle and R. A. Comley: *Portable Analyser for Real-Time Detection of the Epileptic Precursor,* Proc., XI Brazilian Conf. in Biomed. Eng., pp. 101–107 (Sep. 1989).

[15] J. N. Walton: *Brain Diseases of the Nervous System (8th Ed.): Chapter 22, pp. 1093–1132,* Oxford Medical Publications.

[16] B. Widrow, et al.: *Stationary and Nonstationary Learning Characteristics of the LMS Adaptive filter,* Proc. IEEE, vol. 64, No. 8, pp. 1151–1162 (Aug. 1976).

[17] B. Widrow and S. D. Stearns.: *Adaptive Signal Processing,* Prentice-Hall Inc., NJ (1985).

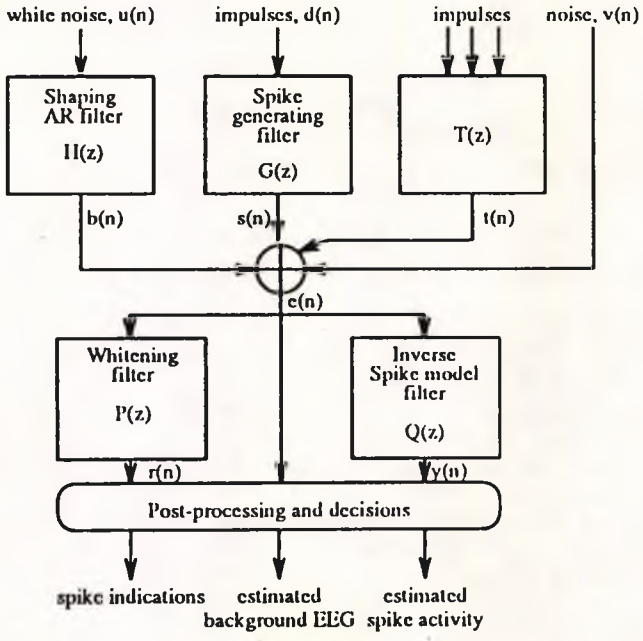


Figure 1: Synthesis-analysis model

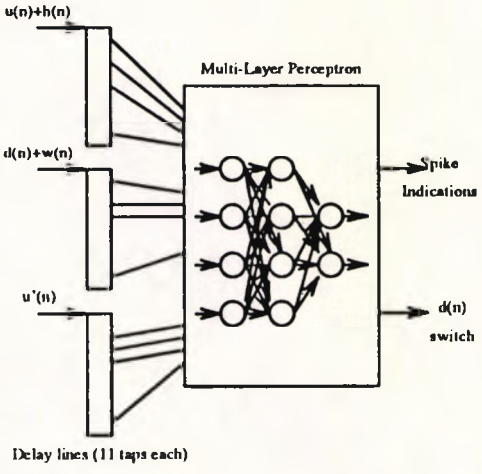


Figure 3: ANN Decision unit

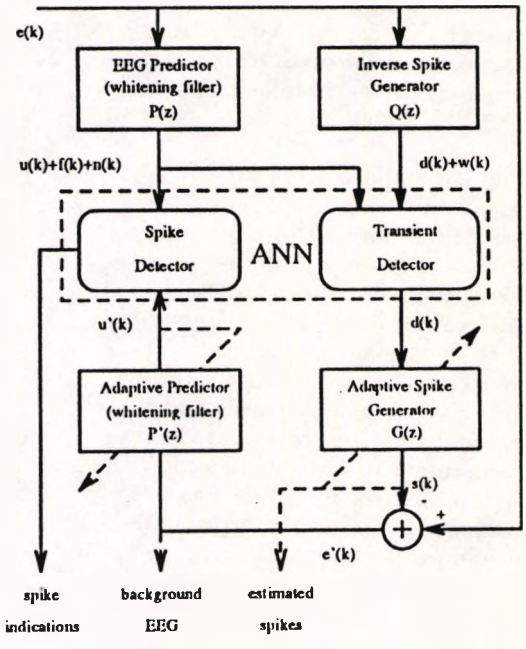


Figure 2: Structure of complete system

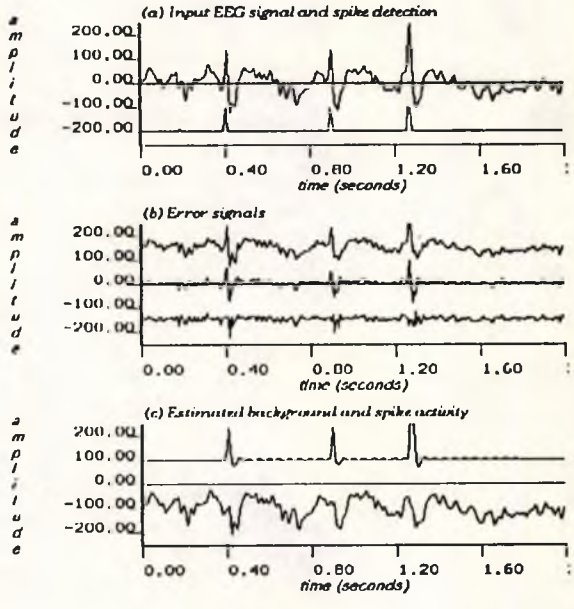


Figure 4: Example waveforms

EEG Signal Analysis Using a Multi-Layer Perceptron with Linear Preprocessing

S. A. Mylonas R.A. Comley
City University, Centre for Information Engineering
Northampton Square, London, EC1V 0HB, UK
e-mail za300@uk.ac.city

Abstract

A system for the detection of epileptic spikes and other transients in the EEG will be described. It consists of a number of adaptive linear filters combined with a non-linear detection unit to control their operation. This has been implemented as a multi-layer perceptron. Configurations using different input preprocessing, initialization and network sizes will be presented along with a discussion on their corresponding results.

1 Introduction

The generation of feeble electrical signals by the brain was known since the end of last century, although their study with the primitive equipment of that time was not easy and the restricted understanding of their origins made interpretation difficult. With the establishment of neurophysiology on a scientific basis and the advances in electronics in the 1950's it became possible to record these time-varying signals on paper by attaching electrodes on the surface of the scalp. Nowadays, such a recording, called the *electroencephalogram* (EEG) can be taken in many hospitals following a harmless as well as inexpensive procedure. Features that relate to the age and level of consciousness and general indications of the mental activity are registered in the EEG, giving a profile of mental health.

Despite recent advances in medical imaging, the EEG is still of great value in monitoring and screening patients suffering from neurological conditions or *idiopathic epilepsy*, where the abnormality is only functional and transient[17].

In epilepsy the abnormal EEG patterns that characterize seizures often occur isolated in "larval" form and are registered in the interictal clinical EEG. *Spikes*, so called because of their shape on a conventional recording are among the commonest patterns. Their presence, absense and frequency of occurrence are valuable clues in the diagnosis and the treatment of this condition[2]. Conventional EEG recordings are sometimes too brief to be trusted for the assessment of patients in order to prescribe some effective therapy. Hence prolonged recording has been suggested as a more reliable alternative[6, 16]. Human interpretation, however, is hindered by the enormous volume of data and the difficulty in their collection from several ambulatory patients. On-line analysis by

a portable microcomputer-based unit was considered as a feasible solution. Early methods were simple because of the limited capabilities of the computers of the time[1]. Despite the sophistication and speed of modern processors no method has been developed as yet to analyze the EEG with total success. This is mainly because the only reference for comparison are human experts, who learn by experience how to perform the highly qualitative and often subjective task of EEG analysis. No formal criteria seem to be followed and even the definitions of the various patterns[5] act as mere guidelines. Intra and inter-reader variability is not uncommon.

The method for automatic EEG analysis described here attempts to combine some attributes of human decision making with the formalism of conventional signal processing tools to form a flexible, but consistent automatic system.

2 Signal modelling and analysis

2.1 A simplified model for the signal

The behaviour of the EEG signal, varies with the level of consciousness, eye opening and closure, mental activity etc. Experts usually separate the EEG into *background activity*, which is the signal present at all times and on which *transients* are superimposed. These have been discriminated into epileptic spikes, which are of medical interest, and other transients. Noise is often present also. Therefore the recorded EEG may be represented as a composite signal:

$$e(n) = b(n) + s(n) + t(n) + v(n) \qquad (1)$$

where $b(n)$ is the background activity, $s(n)$ and $t(n)$ are spikes and other transients, respectively, and $v(n)$ is the noise component. Each one of these components was modelled as the output of an an all-pole system excited with either a sequence of impulses or a white uncorrelated sequence[12, 13, 14]. The components of the generating model of the EEG signal are shown in Figure 1.

2.2 EEG signal analysis

The analysis of the EEG signal was based on the inverse of the model. The inverses of $H(z)$, related to the background activity and $G(z)$, related to spike generation, are both transversal "linear predictor" filters[12, 13] with transfer functions of the form $P(z) = 1 + \sum_{i=1}^{L} h_i z^{-i}$. $H^{-1}(z)$ was estimated using on-line linear optimization (adaptive filtering)[19]. It can also track slow changes in the signal behaviour[12, 13], as explained in [18]. $G^{-1}(z)$ was estimated off-line from available spikes, but on-line adaptation of $G(z)$, in the neighbourhood of the optimal, was carried out for better modelling of individual spikes[13].

Other transients may be treated in a similar way. Only one transient was considered this, having the form of an exponential decay with arbitrary polarity, imitating interference potentials from movements of the recording electrodes on the scalp or of ocular origin.

When $H^{-1}(z)$ is applied to the (recorded) EEG signal $e(n)$ the output, $r(n)$, contains the white sequence, $u(n)$, transients due to $s(n)$ and $t(n)$ and additive distorted noise. Similarly, the application of $G^{-1}(z)$ on $e(n)$ would reproduce the sequence of impulses $d(n)$, and other components in the output $y(n)$. Likewise, application of the inverse transfer function of any other transient would produce its generating sequence in $y_i(n)$, among other components.

Spikes are registered in $r(n)$ a fact used by earlier systems for their detection[3, 4, 9, 11]. When other transients are present, they are also registered and the results are inconclusive. Impulses present in $y(n)$ and in any of the $y_i(n)$ are not very reliable indications, because they are burried in non-random signals, caused by $b(n)$.

The proposed method[12, 13], depicted in Figure 2, detected transients in $r(n)$. Making an initial assumption about the origin of the transient, by inspecting $y(n)$ and $y_i(n)$ for all modelled transients and deciding which one is the most likely to have occurred it was possible to generate an excitation impulse to the appropriate generating transfer function ($G(z)$ for spikes) whose output was subtracted from the recorded EEG signal, $e(n)$ to produce a supposedly transient-free signal, $e'(n)$. Processing this through $H^{-1}(z)$ and observing no disturbance in its output $r'(n)$ confirmed the presence of the suspected transient.

3 The need for decisions in signal classification

The reliability of the above scheme, is linked to the accuracy of the detection of transients in $r(n)$ and $r'(n)$ and the discrimination of their origin in either $y(n)$ or one of the $y_i(n)$.

Initially, simple statistical significance testing was employed for the detection of transients. Both $r(n)$ and $r'(n)$ are essentially random sequences, consisting mainly of $u(n)$ if the noise level is low. This follows a normal distribution with zero mean. Transients have amplitudes that are atypically large and may be detected with a certain degree of certainty, $p\%$ using the assumed probability distribution to derive a corresponding level of significance ν. Usually the standard normal distribution is employed, and hence the sample is normalized by the standard deviation of the signal.

This method, however, did not discriminate between *isolated* atypical samples, which are genuine transients and longer *bursts* which are related to certain extracerebral phenomena, like muscle artifacts. To overcome this problem, every time a new sample became available, a linear combination of the N most recent sample values was formed.

$$f(n) = \sum_{i=-N/2}^{N/2} w_i \frac{x(i)^2}{\sigma_x^2(n)} \tag{2}$$

where $x(n)$ is either $r(n)$ or $r'(n)$ and $\sigma_x^2(n)$ is an on-line estimate the power (variance) of $x(n)$, making $f(n)$ independent of the signal level. The weights, $\{w_i\}$ were positive for $i \in [-M/2, M/2]$ where M is a small

number of samples (1,3 or 5) and negative otherwise. The statistical significance level, $\frac{1}{\nu}$ is implicitly included in the weights. By applying a threshold θ on $f(n)$, a binary detection, $z(n)$ is formed.

$$z(n) = \begin{cases} 1 & f(n) > \theta \\ 0 & f(n) \leq \theta \end{cases} \tag{3}$$

Two such elements applied on $r(n)$ and $r'(n)$ verified whether transients were present ($z_1(n)$ and $z_2(n)$). Similar units were introduced to detect transients in $y(n)$ and $y_1(n)$ ($z_3(n)$ and $z_4(n)$). The outputs of these units were combined to produce activation signals for the spike ($z_s(n) = z_1(n)\text{AND}z_3(n)$) and for the other transient ($z_t(n) = z_1(n)\text{AND}z_4(n)$) as well as the "spike detected" output, $z_o(n) = z_1(n)\text{AND NOT}z_2(n)$. $z_s(n)$ and $z_t(n)$ were used as windows on $y(n)$ and $y_1(n)$ to generate the (presumed) impulse activation function of the spike or transient generating filters, for example $d(n) = z_s(n)y(n)$. The logical operations required may also be implemented as weighted sums (e.g. $f_s(n) = z_1(n) + z_2(n)$ and $f_o(n) = z_1(n) - z_2(n)$) followed by thresholding ($\theta_s = 1.5$ and $\theta_o = -0.5$). This scheme is an extension of an earlier method which considered only spikes[12]. Unlike its predecessor, which was quite successful, the performance of the extended system was only moderate. The problems associated with the generalized structure are believed to be associated with the selection of an appropriate set of weights for the elements that act on the inputs, rather than with the structure itself.

4 A Multi-Layer Perceptron structure as a decision unit

The layered fixed-weight structure of the decision unit had a logical interpretation. The function of any of its elements may be described in terms of a linear combination of its inputs (equation 2) followed by thresholding (equation 3).

This bears a strong resemblance to the Multi-Layer Perceptron (MLP) neural network, originally described by Rumelhart and McClelland[15], as every layer receives inputs from the previous layer only. The only difference is that the intuitive system does not have all outputs of one layer connected to the next, but with the introduction of these with zero weights the two became equivalent.

Because all the inputs to the MLP are all treated in the same manner, they loose their individual significance and they may be grouped to form a single input vector: $\mathbf{x}(n) = [1r(n + N/2) \ldots r(n - N/2)r'(n + N/2) \ldots r'(n - N/2)y(n + N/2) \ldots y(n - N/2)y_1(n + N/2) \ldots y_1(n - N/2) \ldots]^T$. Describing the weights of the kth element of the lth layer in a similar way $\mathbf{w}_{lk} = [w_{lk0}w_{lk1} \ldots]^T$ the weighting operation may be defined as an inner product, $f_{1k}(n) = \mathbf{w}_{1k}^T\mathbf{x}(n)$. The bias weight, w_{1k0} is multiplied by unity and plays the role of the threshold in the earlier system. The output of the element, $y_{1k}(n) = \sigma[f_{1k}(n)]$ is produced by applying a limiting function (non-linearity), $\sigma[\cdot]$ on the linear output $f_{lk}(n)$. The operation of the element of any layer may be described in terms of the

inner product $f_{lk}(n) = \mathbf{w}_{lk}^T \mathbf{y}_{l-1}(n)$ as described above, using the vector of the outputs of the previous layer, $\mathbf{y}_{l-1}(n) = [1 y_{l-1,1}(n) y_{l-1,2}(n) \ldots]^T$.

The function of the MLP description of the system is defined as a relation between its inputs and its desired output rather than by the behaviour of its individual elements, which was the basis of the earlier system. This is specified as a set of examples, consisting of pairs $[\mathbf{x}(n), \mathbf{d}(n)]$, the training set, where $\mathbf{d}(n)$ is the desired output vector. An optimal weight vector may be found by minimizing the mean-squared error between the desired and the actual output vectors of the output layer $\xi = E\{[\mathbf{d}(n) - \mathbf{y}_L(n)]^2\}$ according to the well-known generalized delta rule (backpropagation algorithm)[15]. Many variants of the basic algorithm exist[10, 7]. The one used in this application has a momentum term and updates the weight vector on every sample, according to the following recursive relation, where for notational convenience, $\mathbf{y}_0(n)$ has been used to denote the input vector:

$$\mathbf{w}_{lk}(n+1) = \alpha \mathbf{w}_{lk}(n) + (1-\alpha)\mathbf{w}_{lk}(n-1) + 2\mu\delta_{lk}(n)\mathbf{y}_{l-1}(n) \quad (4)$$

where

$$\delta_{lk} = \begin{cases} d_k(n) - y_{lk}(n) & \text{if } l = L \\ \sigma'[f_{lk}(n)] \sum_j \delta_{l+1,j}(n) w_{l+1,j,k}(n) & \text{otherwise} \end{cases} \quad (5)$$

where $\sigma'(f) = y(1-y)$ is the derivative of the non-linear logistic function $y = sigma(f) = \frac{1}{1+e^{-f}}$, α is a filtering factor defining the "momentum" and μ is the learning rate of the algorithm.

5 Implementation and other issues

Although the structure and the learning algorithm for the MLP are well-defined several details and problems in the implementation needed to be resolved. The main ones are discussed here.

5.1 Forming a training set

A training set consisting of input-output examples needed to be defined. This is not directly possible from the real EEG records available, because spikes and other transients considered are fairly rare and the exact location of their occurrence unknown. Some spikes that were identified by an earlier system were used but the training set consisted mainly of artificial data, generated according to the described model (section 2.1). For these the location of the excitation functions for the various transients is known exactly. These were used to derive the desired output signals indicating the points of application of the excitation impulses for spikes, $z_s(n)$ and the other type of transients, $z_t(n)$, as well as a separate indication for the occurrence of spikes $z_o(n)$ (Figure 3). The real EEG training patterns were used as well, but their effect on the final weights was not visible.

5.2 Implementation of the MLP

One of the problems encountered during the implementation was the fact that the MLP is embedded in the rest of the system and that its spike detection output depends indirectly on the spike excitation signal, which is also an output. In other words, there is feedback. Hence training the network cannot be done independently and as a result it was not possible to use many of the available tools for developing and training the network.

It was therefore necessary to develop a library of functions that deal with the construction of networks of different configurations as well as the implementation of the backpropagation algorithm. The library, which was realized in the C programming language, like the rest of the system, also provides the means to read a network configuration and learning parameters from a file and save them in a format that is easily read by computers and humans alike.

5.3 Preprocessing of the inputs to the MLP

An important issue when using neural networks with natural signals like the EEG is the format of the inputs to the network. Although it is sometimes claimed that there is no need for preprocessing, it is obvious that performance may be affected by changes in the dynamic range of the inputs. To ensure that the input levels to the network are not affected much by such fluctuations, they were normalized in a way similar to the one described for the earlier system (equation 2). All inputs to the network were divided by the RMS value of $r(n)$. This is a convenient measure, because it is also representative of $r'(n)$ and is not affected by modifications to the system by the extension or restriction of the number of transients considered.

Preprocessing of the inputs using other operations to assist the network to learn or to simplify its structure were also considered. The latter is important if the system is to be implemented on a small portable microcomputer with limited processing capabilities and resources.

Two types of preprocessing were considered. The first consists of simply scaling all the elements in the input vector by σ_r, as already explained in the previous paragraph. The inputs to the network, $\hat{x}_i(n)$ are related to the corresponding "raw" inputs, $x_i(n)$ by the simple relation $\hat{x}_i(n) = \frac{x_i(n)}{\sigma_r}$. No information is lost during this operation, but the size of the network may be larger than other alternatives.

Squaring the elements of the input vector prior to their application to the MLP, so that $\hat{x}_i(n) = \frac{x_i^2(n)}{\sigma_r^2}$ was another simple form of processing. This resembles the operation performed by equation 2, which has an intuitive interpretation (section 3).

5.4 Initial conditions of the learning algorithm

When training the MLP using the backpropagation algorithm the weights of the processing elements are usually initialized to small random values. In this application the weights of the intuitive system of section 3

for which the performance was not perfect, but not unreasonable, were considered as a possible alternative, as they may be closer to an optimal solution than a random weight vector. Strictly speaking, these are only valid if the inputs are squared and normalized.

The elements on the first (hidden) layer were made to detect whether a sample at a specific position in the sequences of the input signals, $\{r(n)\}$, $\{r'(n)\}$, $\{y(n)\}$ and $\{y_1(n)\}$ has an atypically large amplitude compared to the ones in its neighbourhood. The result was then fed into one more (hidden) layer, implementing logical AND operations between the outputs of the first layer, before they are combined by the logical OR elements of the output layer for the producton of $z_s(n)$, $z_t(n)$ and $z_o(n)$.

6 Results

The system was tested for a number of combinations of input preprocessing and weight initialization methods. Training was primarily carried out using synthetic data, because the available EEG signals did not contain substantial numbers of other transients. Tests, however, were carried out on real EEG records as well.

Different MLP configurations were tested, either with normalized or with normalized-squared inputs, as explained in section 5.3. The convergence of the backpropagation algorithm for random and preset initial weight vectors (section 5.4) was also investigated.

Tests were carried out on a simulated EEG record containing 25 spikes and 20 other transients and a real EEG record containing 52 spikes. The number of successful spike detections was noted for every network configuration, as well as the number of iterations required for the network performance to stop improving. Some networks were also tested with a different number of elements in the first hidden layer. The results are shown in Table 1.

Net Config.	Iterations to Converge	Detections		
		Simulated EEG		Real EEG
		Spikes (25)	Transients (20)	Spikes (52)
1	60000	25 (0)	20 (0)	51 (5)
2	80000	25 (0)	20 (0)	51 (5)
3	60000	25 (2)	20 (0)	50 (5)
4	10000	25 (2)	20 (0)	50 (5)
5	*	*	*	*
6	30000	25 (2)	20 (0)	50 (12)

Note: * indicates complete failure of the system

Table 1: Results

All tested configurations had 49 inputs (11 samples from each input signal, 16 from $r'(n)$) and 3 outputs. The first four had two hidden layers with 12 and 6 elements respectively, and the other two had two hidden layers with 6 elements each. Configurations 1, 2 and 5 were for normalized inputs and 3, 4 and 6 for squared and normalized inputs.

1, 3, 5 and 6 had random initial weights, whereas 2 and 4 had preset weights.

Figure 4 shows a typical set of waveforms from the simulated EEG data records. The indications of spike detections under the input signal show the proper operation of the system, whereas the estimated spikes and transients in the third set demonstrate that $z_s(n)$ and $z_t(n)$ are generated correctly as well.

7 Comments, observations and conclusions

All but one of the configurations presented in Table 1 performed reasonably well with both simulated and real EEG data. In spite of training the network mainly with synthetic data, its behaviour with real EEG signals was still quite good. Although the number of data records considered in this study was limited, results show that the MLP-based system is capable of performing well even when its inputs deviate from those used for its training.

The number of iterations to obtain a satisfactory performance was considerably smaller for squared-normalized inputs, when the weights were initialized to those of the earlier system than when random initial values were used. This indicated that the former were closer to an optimal, which was also apparent from the small distance between the initial and the final weight vectors. The final weight sets for the two different initialization procedures were different, but this is not surprising, as there are many combinations of weights that give an optimal performance for a given network architecture[8]. For the normalized only input vector, this was not the case. It appears that starting from random weights produces faster convergence than when the weights of the earlier system were used. This is not surprising, because the latter corresponds to the weights for squared inputs, which is a completely different case with the preset initialization being perhaps far from the optimal.

The loss of the polarity of the input signal with squaring as preprocessing was evidenced in the form of the two false spike detections, indicated in brackets for the simulated EEG record. These corresponded to spikes with negative polarity which were intentionally introduced. Real epileptic spikes always have positive polarity. These have not been detected by the systems 1 and 2, which maintain the polarity of their inputs.

Finally, for configuration 6 with a reduced number of elements and the inputs squared, the system was still quite successful. This structure required a simpler network to produce a satisfactory output, because some of the burden of preprocessing was shifted to the the input. This was not the case in configuration 5, where the lack of an adequate number of elements led to a poor performance of the MLP leading to the degradation of the performance of the system.

The main advantage in using the MLP in this application lies in its ability to learn by example. This permits the inclusion of medical expertise which cannot be expressed in a set of rules. Hence the EEG analysis system presented may be trained by an individual expert to reflect his/her experience or by a group of analysts to act like a less

subjective analysis tool combining both medical experience and formal engineering methodology.

References

[1] J. H. Barlow: *Computerized Clinical Electroencephalography in Perspective,* IEEE Trans. Biomed. Eng., vol. BME-26, No. 7, pp. 377–391 (Jul. 1979).

[2] C. D. Binnie: *What's the use of EEG in epilepsy?* British Journal of Hospital Medicine, February 1988, p. 98, 1978, pp. 575–585.

[3] W. P. Birkemeier, et al.: *Pattern Recognition Techniques for the Detection of Epileptic Transients in the EEG,* IEEE Trans. Biomed. Eng., vol. BME-25, No. 3, pp. 213–216 (Jul. 1978).

[4] G. Bodenstein and M. Praetorius: *Feature Extraction from the EEG by Adaptive Segmentation,* Proc. IEEE, vol. 65, No. 5, pp. 642–657 (May. 1977).

[5] G. E. Chatrian et al.: (IFSECN International Assembly) *A Glossary of Terms Most Commonly Used By Clinical Electroencephalographers* Electroenceph. Clin. Neurophysiol., 37, (1974), pp. 538–548.

[6] R. A. Comley and J. E. Brignell: *Real-Time Detection of the Epileptic Precursor,* J. Phys. E: Sci. Instrum., vol. 14, pp. 963–967 (1981).

[7] R. Hecht-Nielsen: *Neurocomputing,* Addison–Wesley (1990).

[8] A. M. Chen and R. Hecht-Nielsen: *On the Geometry of Feedforward Neural Network Weight Spaces,* Proceedings, Second International Conference on Artificial Neural Networks, Bournemouth, (Nov. 1992), pp. 1–4.

[9] A. Isaksson, et al.: *Computer Analysis of EEG Signals with Parametric Models,* Proc. IEEE, vol. 69, No. 4, pp. 4151–4161 (Apr. 1981).

[10] R. P. Lippmann: *An Introduction to Computing with Neural Nets,* IEEE ASSP magazine, Apr. 1987, pp. 4–22.

[11] F. H. Lopes Da Silva, et al.: *Automatic Detection and Localization of Epileptic Foci,* Electroenceph. Clin. Neurophys., No. 43, pp. 1–13 (1977).

[12] S. A. Mylonas and R. A. Comley: *Detection of Epileptic Spikes in the EEG Using Adaptive Filters,* I Fórum Nacional de Ciência e Tecnologia em Saúde – XIII Congresso Brasileiro de Engenharia Biomedica, Caxambu (MG), Brazil (Nov. 1992)

[13] S. A. Mylonas and R. A. Comley: *Adaptive Predictive Modelling for the Analysis of the Epileptic EEG,* Singapore ICCS/ISITA'92, vol. 3, pp. 1214-1218 (Nov. 1992)

[14] S. A. Mylonas and R. A. Comley: *Linear Prediction, Neural Networks and the Analysis of EEG Signals,* Cyprus, Int. Conf. on DSP/II Int. Conf. on Comput. Appl. to Eng. Sys. (Jul. 1993)

[15] D. E. Rumelhart et al.: *Learning internal representations by error propagation,* in D. E. Rumelhart and J. L. McCLelland (Eds.) *Parallel Distributed Processing: Explorations in the microstructure of Cognition,* 1, pp. 318–362, (1986), MIT Press.

[16] A. L. Stelle and R. A. Comley: *Portable Analyser for Real-Time Detection of the Epileptic Precursor,* Proc., XI Brazilian Conf. in Biomed. Eng., pp. 101–107 (Sep. 1989).

[17] J. N. Walton: *Brain Diseases of the Nervous System (8th Ed.): Chapter 22, pp. 1093–1132,* Oxford Medical Publications.

[18] B. Widrow, et al.: *Stationary and Nonstationary Learning Characteristics of the LMS Adaptive filter,* Proc. IEEE, vol. 64, No. 8, pp. 1151–1162 (Aug. 1976).

[19] B. Widrow and S. D. Stearns: *Adaptive Signal Processing,* Prentice-Hall Inc., NJ (1985).

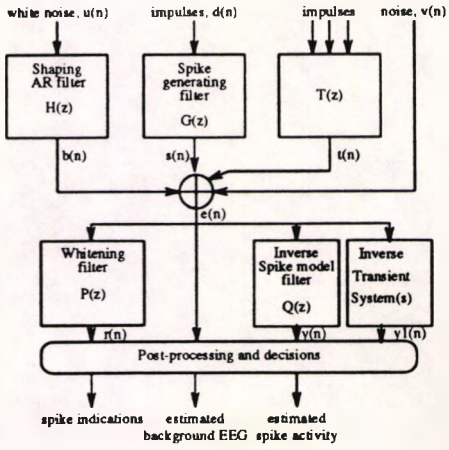


Figure 1: EEG modelling processes for synthesis and analysis

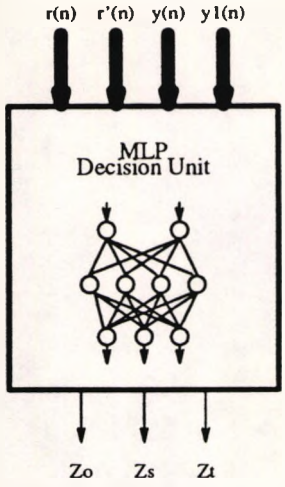


Figure 3: The MLP decision unit

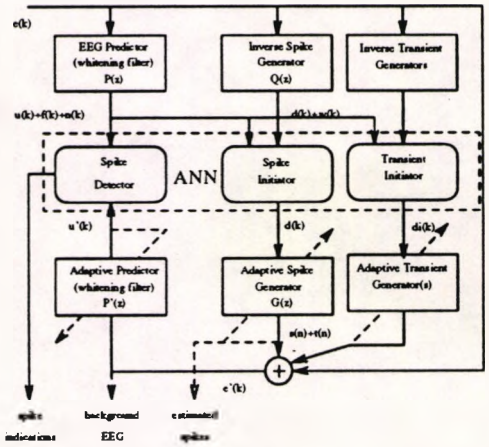


Figure 2: Block diagram of the proposed system

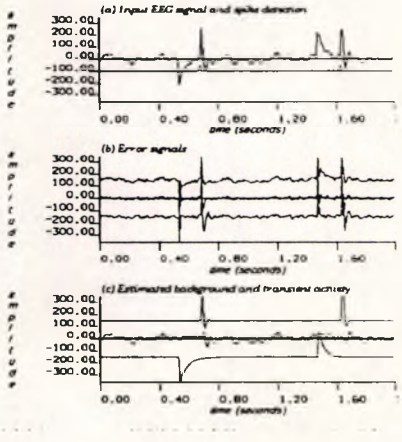


Figure 4: Waveforms at the various stages of the system

Bibliography

[1] ACCORNERO, N.; (1981): **Active Surface EMG Probe and Contour Follower**, Electroencephalography and Clinical Neurophysiology, *Vol. 51*, pp. 331–332.

[2] AICARDI, J.; (1988): **Clinical Approach to the Management of Intractable Epilepsy**, Developmental Medicine And Child Neurology, *Vol. 30*, pp. 429–440.

[3] APAK, A. G. S.; ÇALIŞKAN, A.; (1986): **Electrical Status Epilepticus Lasting For 17 Months Without Behavioural Changes**, Electroencephalography and Clinical Neurophysiology, *Vol. 63*, pp. 32–34.

[4] ARAKAWA, K.; FENDER, D. H.; HARASHIMA, H.; MIYAKAWA, H.; SAITOH, Y.; (1986): **A Non-Linear Digital Filter for Separating Non-Stationary Waves from Stationary Waves in The EEG**, Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, Tokyo, pp. 1809–1812

[5] ATTINGER, E. O.; (December 1984): **Impacts of Technological Revolution on Health Care**, IEEE Transactions on Biomedical Engineering, *Vol. BME-31*, **No. 12**, pp. 736–743.

[6] ATKINSON, K. E.; (1989): **An Introduction to Numerical Analysis**, (*2nd edition*), ISBN 0-47-150023-2, Wiley.

[7] BARLOW, J.; (July 1979): **Computerized Clinical Electroencephalography In Perspective,** , IEEE Transactions on Biomedical Engineering, *Vol. BME-26*, **No. 7**, pp. 377–391.

[8] BARLOW, J. S.; (1983): **Muscle Spike Artifact Minimization In EEGs By Time-Domain Filtering,,** Electroencephalography and Clinical Neurophysiology, *Vol. 55*, pp. 487–491.

[9] BARLOW, J. S.; (1984): **EMG Artifact Minimization During Clinical EEG Recordings By Special Analogue Filtering,,** Electroencephalography and Clinical Neurophysiology, *Vol. 58*, pp. 161–174.

[10] Barlow, J.; (1984): **Analysis of EEG Changes With Carotid Clamping by Selective Analogue Filtering, Matched Inverse Digital Filtering and Automatic Adaptive Segmentation: A Comparative Study**, Electroencephalography and Clinical Neurophysiology, *Vol. 58*, pp. 193–204.

[11] Barlow, J. S.; (May 1986): **Automatic Elimination of Electrode-Pop Artifacts in EEG's,,** IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 5**, pp. 517–521.

[12] Beauchamp, K. G.; (1987): **Transforms For Engineers—A Guide to Signal Processing**, ISBN 0-19-856174-1, Claredon Press, Oxford.

[13] Beghi, E.; Bollini, P.; Di Maschio, R.; Cerisola, N.; Merloni, T.; Manghi, E.; (1987): **Effects on Rationalizing Drug Treatment of Patients With Epilepsy and Mental Retardation**, Developmental Medicine and Child Neurology, *Vol. 29*, pp. 363–369.

[14] Bellanger, M. G.; (1987): **Adaptive Digital Filters and Signal Analysis**, ISBN 0-8247-7784-0, Marcel Dekker, New York.

[15] Bellanger, M.; (1989): **Digital Processing of Signals—Theory and Practice**, (*2nd edition*), ISBN 0-471-02101-7, Wiley–Interscience.

[16] Besag, F. M. C.; Mills, M.; Wardale, F.; Andrew, C. M.; Craggs, M. D.; (1989): **The Validation of a New Ambulatory Spike And Wave Monitor**, Electroencephalography and Clinical Neurophysiology, *Vol. 73*, pp. 157–161.

[17] Birkemeier, W. P.; Fontaine, A. B.; Celesia, G. G.; Ma, K. M.; (May 1978): **Pattern Recognition Techniques for the Detection of Epileptic Transients in the EEG**, IEEE Transactions on Biomedical Engineering, *Vol. BME-25*, **No. 3**, pp. 213–217.

[18] Binnie, C. D.; Batchelor, B. G.; Bowring, P. A.; Darby, C. E.; Herbert, L.; Lloyd, D. S. L.; Smith, D. M.; Smith, G. F.; Smith, M.; (1978): **Computer-Assisted Interpretation of Clinical EEGs**, Electroencephalography and Clinical Neurophysiology, *Vol. 44*, pp. 575–585.

[19] Binnie, C. D.; (February 1988): **What's the use of EEG in epilepsy?**, British Journal of Hospital Medicine, pp. 98.

[20] Binnie, C. D.; Aarts, J. H. P.; Houtkooper, M. A.; Laxminarayan, R.; Da Silva Martins, A.; Meinardi, H.; Nagelkerke, N.; Overweg, J.; (1984): **Temporal Characteristics of Seizures and Epileptiform Discharges**, Electroencephalography and Clinical Neurophysiology, *Vol. 58*, pp. 498–505.

[21] BLINOWSKA, K. J.; FRANASZCZUK, P. J.; MITRASZEWSKI, P.; (1988): **A New Metod of Presentation of the Average Properties of the EEG Time Series**, International Journal of Biomedical Computing, *Vol. 22*, pp. 97–106.

[22] BLUME, W. T.; YOUNG, G. B.; LEMIEUX, J, F; (1984): **EEG Morphology of Partial Epileptic Seizures**, Electroencephalography and Clinical Neurophysiology, *Vol. 57*, pp. 295–302.

[23] BOASHASH, B.; BLACK, P. J.; (November 1987): **An Efficient Real-Time Implementation of the Wigner-Ville Distribution**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-35*, **No. 11**, pp. 1611–1618.

[24] BOASHASH, B.; (September 1988): **Note on the Use of the Wigner Distribution for Time-Freqyency Signal Analysis**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-36*, **No. 9**, pp. 1518–1521.

[25] BODENSTEIN, G.; PRAETORIUS, H. M.; (May 1977): **Feature Extraction from the Electroencephalogram by Adaptive Segmentation**, IEEE Proceedings, *Vol. 65*, **No. 5**, pp. 642–657.

[26] BOUDREAUX-BARTELS, G. F.; (1985): **Time Varying Signal Processing Using the Wigner-Distribution Time-Frequency Signal Representation**, Advances in Geophysical Data Processing, *Vol. 2*, pp. 33–79.

[27] BRAZIER, M. A. B.; (1961): **A History of the Electrical Activity of the Brain**, Pitman Medical Publishing Company, Ltd.

[28] BRAZIER, M. A. B.; WALTER, D. O.; SCHNEIDER, D. (EDITORS); (1973): **Neural Modeling**, Brain Information Service, Research Report, **No. 1**, University of California, Los Angeles.

[29] BRENNER, R. P.; ULRICH, R. F.; SPIKER, D. G.; SCLABASSI, R. J.; REYNOLDS, C. F. III; MARIN, R. S.; BOLLER, F.; (1986): **Computerized EEG Spectral Analysis in Elderly Normal, Demented and Depressed Subjects**, Electroencephalography and Clinical Neurophysiology, *Vol. 64*, pp. 483–492.

[30] BROWN, K.; (1988): **'Epilepsy'—A New Disease?**, Developmental Medicine And Child Neurology, *Vol. 30*, pp. 427–428.

[31] BROWN, R. G.; (1983) **Introduction to Random Signal Analysis and Kalman Filtering**, ISBN 0-471-08732-7, Wiley & Sons Inc.

[32] CABELLO, D.; BARRO, S.; SALCEDA, J. M.; RUIZ, R.; MIRA, J.; (1991): **Fuzzy K-Nearest Neighbor Classifiers For Ventricular Arrhythmia Detection**, International Journal of Biomedical Computing, *Vol. 27*, pp. 77–93.

[33] CARAYANNIS, G; MANOLAKIS, D. G.; KALOUPTSIDIS, N; (December 1983): **A FAst Sequential Algorithm for Least-Squares Filtering and Prediction**, IEEE Transactions on Acoustics, Speech and Signal Procesing, *Vol. ASSP-31*, **No. 6**, pp. 1394–1402.

[34] CARRIE, J. R. G.; (1972): **A Technique for Analyzing Transient EEG Abnormalities**, Electroencephalography and Clinical Neurophysiology, *Vol. 32*, pp. 199–201.

[35] CARRIE, J. R. G.; (1972): **A Hybrid Computer Technique For Detecting Sharp EEG Transients**, Electroencephalography and Clinical Neurophysiology, *Vol. 33*, pp. 336–338.

[36] CARRIE, J. R. G.; (1972): **A Hybrid Computer System for Detecting and Quantifying Spike and Wave EEG Patterns**, Electroencephalography and Clinical Neurophysiology, *Vol. 33*, pp. 339–341.

[37] CARRIE, J. R. G.; FROST J. D., JR.; (1977): **Clinical Evaluation of a Method for Qualification of Generalized Spike-Wave EEG Patterns by Computer During Prolonged Recordings**, Computers and Biomedical Research, *Vol. 10*, pp. 449–457.

[38] CERUTTI, S.; BERSANI, V.; CARRARA, A.; LIBERATI, D.; (January 1987): **Analysis of Visual Evoked Potentials Through Wiener Filtering Applied to a Small Number of Sweeps**, Journal of Biomedical Engineering, *Vol. 9*, pp. 3–12.

[39] CHADWICK, D.; (February 1988): **Modern Treatment of Epilepsy**, British Journal of Hospital Medicine, pp. 104–111.

[40] CHARALAMBOUS, C; (July 1991): **Optimization and Neural Networks**, Proceedings, First Cyprus Interantional Conference on Computer Applications to Engineering Systems, pp. 154–159.

[41] CHATRIAN, G. E.; BERGAMINI, L.; DONDEY, M.; KLASS, D. W.; LENNOX-BUCHTHAL, M.; PETERS, N. I.; (1974): **A Glossary of Terms Most Commonly Used By Clinical Electroencephalographers**, (IFSECN International Assembly), Electroencephalography and Clinical Neurophysiology, *Vol. 37*, pp. 538–548.

[42] CHEN, A. M.; HECHT-NIELSEN, R.; (November 1991): **On the Geometry of Feedforward Neural Network Weight Spaces**, Proceedings, Second International Conference on Artificial Neural Networks, Bournemouth, pp. 1–4.

[43] CHILDERS, D. G.; (1989): **Biomedical Signal processing**, in **Selected Topics in Signal Processing**, HAYKIN, S. S. (EDITOR) pp. 194–250 Prentice Hall.

[44] CLAASEN, T. A. C. M.; MECKLENBRAUKER, W. F. G.; (June 1981): **Comparison of the Convergence of Two Algorithms for Adaptive FIR Digital Filters**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-29*, **No. 3**, pp. 670–678

[45] COMLEY, R. A.; BRIGNELL, J. E.; (1981): **Real-time Detection of the Epileptic Precursor**, Journal of Phys. E: Sci. Instrum., *Vol. 14*, pp. 963–967.

[46] COOLEY, J. N.;TUKEY, J.; (1965): **An Algorithm For The Machine Calculation Of Complex Fourier Series**, Journal of Mathemetical Computing, *Vol. 19*, pp. 297–301.

[47] COOPER, R.; OSSELTON, J. W.; SHAW, J. C.; (1980): **EEG Technology**, (*3rd Edition edition*), ISBN 0-407-16002-7, Butterworths & Co (Publishers) Ltd.

[48] COWEN, A.; HARTLEY, P.; WORKMAN, A.; (Autumn 1989): **Medical Archiving: A picture of Health**, Image Processing, pp. 44–46.

[49] COX, J. R., JR.; NOLLE, F. M.; ARTHUR, R. M.; (October 1972): **Digital Analysis of the Electroencephalogram, the Blood Pressure Wave and the Electrocardiogram**, IEEE Proceedings, *Vol. 60*, **No. 10**, pp. 1137–1164.

[50] CROWELL, D. H.; JONES, R. H.; KAPUNIAI, L. E.; LEUNG, P.; (1977): **Autoregressive Representation of Infant EEG for the Purpose of Hypothesis Testing and Classification**, Electroencephalography and Clinical Neurophysiology, *Vol. 43*, pp. 317–324.

[51] DARCEY, T. M.; WILLIAMSON, P. D.; (1985): **Spatiotemporal EEG Measures and their Application to Human Intracranially Recorded Epileptic Seizures**, Electroencephalography and Clinical Neurophysiology, *Vol. 61*, pp. 573–587.

[52] DASKALOVA, M. I.; (July 1988): **Wave Analysis of the Electroencephalogram**, Medical & Biological Engineering & Computing, *Vol. 26*, pp. 425–428.

[53] DAVID, R. A.; (1982): **A Cascade Structure for Equation Error Minimization**, Proceedings, IEEE International Signal Processing Conference, pp. 182–186.

[54] DAVEY, B. L. K.; FRIGHT, W. R.; CARROLL, G. J.; (July 1989): **Expert Systems Approach to Detection of Epileptiform Activity in the EEG**, Medical & Biological Engineering & Computing, *Vol. 27*, pp. 365–370.

[55] D'AZZO, J. J.; HOUPIS, C. H.; (1989): **Linear Control System Analysis and Design**, (*2nd edition*), ISBN 0-07-016183-6, McGraw-Hill.

[56] EHRENBERG, B. L.; PENRY, J. K.; (1976): **Computer Recognition of Generalized Spike-Wave Discharges**, Electroencephalography and Clinical Neurophysiology, *Vol. 41*, pp. 25–36.

[57] FITCH, Y. P.; WILLISON, R. G.; (1980): **High-Speed Automatic Analysis of EEG Spike and Wave Activity Using an Analogue Detection and Microcomputing Plotting System**, Electroencephalography and Clinical Neurophysiology, *Vol. 49*, pp. 187–189.

[58] FITZGERALD, K.; (January 1990): **Medical Electronics: 3-D Imaging Benefits from Faster Processing**, IEEE Spectrum, pp. 52–54.

[59] FORTGENS, C.; DE BRUIN, M. P.; (1983): **Removal of Eye Movement and ECG artifacts from the Non-Cephalic Reference EEG**, Electroencephalography and Clinical Neurophysiology, *Vol. 56*, pp. 90–96.

[60] FREUD, J. E.; WALPOLE, R. E. (1980): **Mathematical Statistics**, ISBN 0-13-562082-1, Prentice-Hall, Inc.

[61] FRIDMAN, J.; JOHN, E. R.; BERGELSON, M.; KAISER, J. B.; BAIRD, H. W.; (1982): **Application of Digital Filtering and Automatic Peak Detection to Brain Stem Auditory Evoked Potential**, Electroencephalography and Clinical Neurophysiology, *Vol. 53*, pp. 405–416.

[62] FROST, J. D., JR; (1970): **An Automatic Sleep Analyzer**, Electroencephalography and Clinical Neurophysiology, *Vol. 29*, pp. 88–92.

[63] GANONG, W. F.; (1979): **The Nervous System**, (*2nd edition*), ISBN 0-87041-241-8, Lange Medical Publications.

[64] GERSCH, W.; FOUTCH, D. A.; (December 1974): **Least Squares Estimates of Structural System Parameters Using Covariance Function Data**, IEEE Transactions on Automatic Control, *Vol. AC-19*, **No. 6**, pp. 898–903.

[65] GEVINS, A. S.; YEAGER, C. L.; DIAMOND, S. L.; SPIRE, J. P.; ZEITLIN, G. M.; GEVINS, A. H.; (October 1975): **Automated Analysis of the Electrical Activity of the Human Brain (EEG): A Progress Report**, IEEE Proceedings, *Vol. 63*, **No. 10**, pp. 1382–1399.

[66] GEVINS, A. S.; YEAGER, C. L.; ZEITLIN, G. M.; ANCOLI, S.; DEDON, M. F.; (1977): **On-Line Computer Rejection of EEG Artefacts**, Electroencephalography and Clinical Neurophysiology, *Vol. 42*, pp. 267–274.

[67] GEVINS, A. S.; ZEITLIN, G. M.; ANCOLI, S.; YEAGER, C. L.; (1977): **Computer Rejection of EEG Artifact, II. Contamination by Drowsiness**, Electroencephalography and Clinical Neurophysiology, *Vol. 42*, pp. 31–42.

[68] GEVINS, A. S.; (December 1984): **Analysis of the Electromagnetic Signals of the Human Brain: Milestones, Obstacles and Goals**, IEEE Transactions on Biomedical Engineering, *Vol. BME-31*, **No. 12**, pp. 833–850.

[69] GLOOR, P. (EDITOR); (1969): **Hans Berger on the Electroencephalogram of Man**, Supplement 28, Electroencephalography and Clinical Neurophysiology, ISBN 444-40739-1, Elsevier Publishing Company.

[70] GLOOR, P.; FARIELLO, R. G.; (1988): **Generalized Epilepsy: Some of Its Cellular Mechanisms Differ From Those of Focal Epilepsy**, Trends In Neurosciences, *Vol. 11*, **No. 2**, pp. 63–68.

[71] GLOVER, J. R., JR.; KTONAS, P. Y.; RAGHAVAN, N.; URUNUELA, J. M.; VELAMURI, S. S.; REILLY, E. L.; (December 1986): **A Multichannel Signal Processor For the Detection of Epileptogenic Sharp Transients in the EEG**, IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 12**, pp. 1121–1128.

[72] GLOVER, J. R., JR.; RAGHAVAN, N.; KTONAS, P. Y.; FROST, J. D., JR.; (May 1989): **Context-Based Automated Detection of Epileptogenic Sharp Transients in the EEG: Elimination of False Positives**, IEEE Transactions on Biomedical Engineering, *Vol. BME-36*, **No. 5**, pp. 519–527.

[73] GORDON, N.; (1988): **Intractable Epilepsy**, Letter, Developmental Medicine And Child Neurology, *Vol. 30*, pp. 830.

[74] GOTMAN, J.; SKUCE, D. R.; THOMPSON, C. J.; GLOOR, P.; IVES, J. R.; RAY, W. F.; (1973): **Clinical Applications of Spectral Analysis and Extraction of Features from Electroencephalograms with Slow Waves in Adult Patients**, Electroencephalography and Clinical Neurophysiology, *Vol. 35*, pp. 225–235.

[75] GOTMAN, J.; GLOOR, P.; (1976): **Automatic Recognition and Quantification of interictal Epileptic Activity in the Human Scalp EEG**, Electroencephalography and Clinical Neurophysiology, *Vol. 41*, pp. 513–529.

[76] GOTMAN, J.; GLOOR, P.; SCHAUL, N.; (1978): **Comparison of Traditional Reading of the EEG and Automatic Recognition of Interictal Epileptic Activity**, Electroencephalography and Clinical Neurophysiology, *Vol. 44*, pp. 48–60.

[77] GOTMAN, J.; (1980): **Quantitative Measurements of Epileptic Spike Morphology in the Human EEG**, Electroencephalography and Clinical Neurophysiology, *Vol. 48*, pp. 551–557.

[78] GOTMAN, J.; IVES, J. R.; GLOOR, P.; (1981): **Frequency Content of EEG and EMG at Seizure Onset: Possibility of Removal of EMG Artefact by Digital Filtering**, Electroencephalography and Clinical Neurophysiology, *Vol. 52*, pp. 626–639.

[79] GOTMAN, J.; (1989): **Automated Analysis of Ambulatory EEG Recordings**, in **Ambulatory EEG Monitoring**, EBERSOLE, S (EDITOR) pp. i97–110 Raven Press.

[80] GRATTON, G.; COLES, M. G. H.; DONCHIN, E.; (1983): **A New Method For Off-Line Removal Of Ocular Artifact**, Electroencephalography and Clinical Neurophysiology, *Vol. 55*, pp. 468–484

[81] GROCHULSKI, W.; PENCZEK, P.; POSIELSKI, J.; (1986): **Segmentation of the EEG by Means of a Finite State Automaton**, International Journal of Biomedical Computing, *Vol. 18*, pp. 35–44.

[82] GROSSMANN, S. I.; DERRICK, W. R.; (1988): **Advanced Engineering Mathematics**, ISBN 0-06-042534-2, Harper & Row.

[83] HARRIGAN, E.; KROH, J. R.; SANDHAM, W. A.; DURRANI, T. S.; (November 1992): **Seismic Wavelet Extraction Using Artificial Neural Networks, On the Geometry of Feedforward Neural Network Weight Spaces**, Proceedings, Second International Conference on Artificial Neural Networks, Bournemouth, pp. 95–99.

[84] HAUSTEIN, W.; PILCHER, J.; KLINK, J.; SCHULTZ, H.; (1986): **Automatic Analysis Overcomes Sleep Stage Scoring**, Electroencephalography and Clinical Neurophysiology, *Vol. 64*, pp. 364–374.

[85] HENRICI, P.; (1974): **Applied and Computational Complex Analysis**, *Vol. 1*, ISBN 0-471-37244-7, Wiley-Interscience.

[86] HILL, A. G.; TOWNSEND, H. R. A.; (1973): **The Automatic Estimation of Epileptic Spike Activity**, Biomedical Computing, *Vol. 4*, pp. 149–156.

[87] HJÖRTH, B.; 1970 **EEG Analysis Based on Time-Domain Properties**, Electroencephalography and Clinical Neurophysiology, *Vol. 29*, pp. 306–310.

[88] HJÖRTH, B.; 1973 **The Physical Significance of Time Domain Descriptors in EEG Analysis**, Electroencephalography and Clinical Neurophysiology, *Vol. 34*, pp. 321–325.

[89] HULTQUIST, P. F.; (1988): **Numerical Methods for Engineers and Computer Scientists**, ISBN 0-80534-652-X, The Benjamin/Cummings Publishing Company, Inc.

[90] IFEACHOR, E. C.; JERVIS, B. W.; MORRIS, E. L.; ALLEN, E. M.; HUDSON, N. R.; (1986): **New Online Method For Removing Ocular Artefacts From EEG Signals**, Medical & Biological Engineering & Computing, *Vol. 24*, pp. 356–364

[91] IFEACHOR, E. C.; JERVIS, B. W.; ALLEN, E.M.; MORRIS, E. L.; WRIGHT, D. E.; HUDSON, N. R.; (1988): **Investigation and Comparison of Some Models for Removing Ocular Artefacts from EEG signals—part 1: Review of models and data analysis**, Medical & Biological Engineering & Computing, *Vol. 26*, pp. 584–590.

[92] IFEACHOR, E. C.; JERVIS, B. W.; ALLEN, E. M.; MORRIS, E. L.; WRIGHT, D. E.; HUDSON, N. R.; (1988): **Investigation and Comparison of Some Models for Removing Ocular Artefacts from EEG signals—part 2: Quantitative and pictorial comparison of models**, Medical & Biological Engineering & Computing, *Vol. 26*, pp. 591–598.

[93] ISAKSSON, A.; WENNBERG, A.; ZETTERBERG, L. H.; (April 1981): **Computer Analysis of EEG Signals with Parametric Models**, IEEE Proceedings, *Vol. 69*, **No. 4**, pp. 451–461.

[94] JACKSON, L. B.; WOOD, S. L.; (December 1978): **Linear Prediction in Cascade Form**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-26*, **No. 6**, pp. 518–528.

[95] JANSEN, B. H.; (1991): **Quantitative Analysis of Electroencephalograms: Is There Chaos In The Future?**, International Journal of Biomedical Computing, *Vol. 27*, pp. 95–123.

[96] JERVIS, B. W.; IFEACHOR, E. C.; ALLEN, E. M.; (January 1988): **The Removal of Ocular Artefacts from the Electroencephalogram: A Review**, Medical & Biological Engineering & Computing, *Vol. 26*, pp. 2–12.

[97] JERVIS, B. W.; COHELLO, M.; MORGAN, G. W.; (May 1989): **Spectral Analysis of EEG Responses**, Medical & Biological Engineering & Computing, *Vol. 27*, pp. 230–238.

[98] JOHNSON, T. L.; WRIGHT, S. C.; SEGALL, A.; (October 1979): **Filtering of Muscle Artifact from the Electroencephalogram**, IEEE Transactions on Biomedical Engineering, *Vol. BME-26*, **No. 10**, pp. 556–563.

[99] JONES, R. H.; (December 1974): **Identification and Autoregressive Spectrum Estimation**, IEEE Transactions on Automatic Control, *Vol. AC-19*, **No. 6**, pp. 894–897.

[100] JONES, W. B.; STEINHARDT, A. O.; (October 1985): **Finding the Poles of the Lattice Filter**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-33*, **No. 4**, pp. 1329–1331.

[101] KAILATH, T. (EDITOR); (1977): **Linear Least-Squares Estimation**, ISBN 0-87933-098-8, Dowden, Hutchinson & Ross Inc.

[102] KATZ, M. K.; (1988): **Fractals and the Analysis of Waveforms**, Computers in Biology and Medicine, *Vol. 18*, **No. 3**, pp. 145–156.

[103] KILOH, L. G.; MCCOMAS, A. J.; OSSELTON, J. W; (1972): **Clinical Electroencephalography**, (*3rd edition*), ISBN 0-407-13602-9, Butterworth & Co (Publishers) Ltd.

[104] KLEINER, B.; (1973): **Some Comments on the Use of an EEG Autoregressive Model for the Time-Saving Calculation of Spectral Power Density Distributions with a Digital Computer**, Electroencephalography and Clinical Neurophysiology, *Vol. 35*, pp. 331–332.

[105] KOFFLER, D. J.; GOTMAN, J.; (1985): **Automatic Detection of Spike-and-Wave Bursts in Ambulatory EEG Recordings**, Electroencephalography and Clinical Neurophysiology, *Vol. 61*, pp. 165–180.

[106] KOHN, A. F.; (March 1987): **Phase Distortion in Biological Signal Analysis Caused by Linear Phase FIR Filters**, Medical & Biological Engineering & Computing, *Vol. 25*, pp. 231–238.

[107] KTONAS, P. Y.; SMITH, J. R.; (1974): **Quantification of Abnormal EEG Spike Characteristics**, Computers in Biology and Medicine, *Vol. 4*, pp. 157–163.

[108] KTONAS, P. Y.; LUOH, W. M.; KEJARIWAL, M. L.; REILLY, E. L.; SEWARD, M. A.; (1981): **Computer-Aided Quantification of EEG Spike and Sharp Wave Characteristics**, Electroencephalography and Clinical Neurophysiology, *Vol. 51*, pp. 237–243.

[109] LEADER, H. S.; COHN, R.; WEIHRER, A. L.; CACERES, C. A.; (1967): **Pattern Reading of the Clinical Electroencephalogram with a Digital Computer**, Electroencephalography and Clinical Neurophysiology, *Vol. 23*, pp. 566–570.

[110] LIM, A. J.; WINTERS, W. D.; (April 1980): **A Practical Method for Automatic Real-Time EEG Sleep Stage Analysis**, IEEE Transactions on Biomedical Engineering, *Vol. BME-27*, **No. 5**, pp. 212–220.

[111] LJUNG, S; LJUNG, L; (1985): **Error Propagation Properties of Recursive Least-Squares Adaptation Algorithms**, Automatica, *Vol. 21*, **No. 2**, pp. 236–246.

[112] LIPPMAN, R. P.; (April 1987): **An Introduction to Computing with Neural Nets**, IEEE Acoustics Speech and Signal Processing Magazine, pp. 4–22.

[113] LYNCH, M. R.; HOLDEN, S. B.; RAYNER, P. J.; (November 1991): **Complexity reduction in Volterra Connectionist Networks Using a Self-Structuring LMS Algorithm**, Proceedings, II International Conference on Artificial Neural Networks, Bournemouth, U.K., pp. 44–48.

[114] LYNN, P. A.; (September 1977): **Online Digital Filters for Biological Signals: Some Fast Designs for a Small Computer**, Medical & Biological Engineering & Computing, *Vol. 15*, pp. 534–540.

[115] LYNN, P. A.; (1982): **An Introduction to the Analysis and Processing of Signals**, (*2nd edition*), pp. 196–207 ISBN 0-333-34030-2, MacMillan Education.

[116] MAKHOUL, J.; (April 1975): **Linear Prediction: A Tutorial Review**, IEEE Proceedings, *Vol. 63*, **No. 4**, pp. 561–580.

[117] MANDELBROT, B. B.; (1983): **The Fractal Geometry of Nature**, ISBN 0-7167-1186-9, Freeman & Co., NY.

[118] MARIEB, E. N; (1989): **Human Anatomy and Physiology**, ISBN 0-8053-0122-4, The Benjamin/Cummings Publishing Company, Inc.

[119] MATOUŠEK, M.; PETERSÉN, I.; (1973): **Automatic Evaluation of EEG Background Activity by Means of Age-Dependent EEG Quotients**, Electroencephalography and Clinical Neurophysiology, *Vol. 35*, pp. 603–612.

[120] MATTHEWS, W. B.; (12 September 1964): **The Use and Abuse of Electroencephalography**, The Lancet, pp. 577–579.

[121] MCGILLEM, C. D.; COOPER, G. R.; (1986): **Continuous and Discrete Signal and System Analysis**, (*2nd edition*), ISBN 0-03-910770-1, CBS College Publishing.

[122] MCLEOD, J. G.; LANCE, J. W.; (1989): **Introductory Neurology**, (*2nd edition*), ISBN 0-867-930179, Blackwell Scientific Publications Ltd.

[123] MCLOCHLIN, C.; PRINCIPE, J. P.; SMITH, J. R.; (1988): **A Data Compression Algorithm for the Electroencephalogram**, International Journal of Biomedical Computing, *Vol. 22*, pp. 83–86.

[124] MINSKY, M; PAPERT, S; (1969): **Perceptrons**, MIT Press.

[125] MOSER, J. M.; AUNON, J. I.; (December 1986): **Classification and Detection of Single Evoked Brain Potentials Using Time-Frequency Amplitude Features**, IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 12**, pp. 1096–1106.

[126] MYLONAS, S. A.; COMLEY, R. A.; (July, 1991): **On-Line Detection of the Epileptic Precursor**, Proceedings, First Cyprus International Conference on Computer Applications to Engineering Systems, pp. 178–183.

[127] OMAROUAYACHE, S.; MYLONAS, S. A.; ELLIS, T. J.; COMLEY, R. A.; (1992): **Transputer Implementation of Adaptive Noise Cancelling In Digital Images**, Proceedings, PACTA '92, Barcelona, Spain, pp. 964–974.

[128] MYLONAS, S. A.; COMLEY, R. A.; (November 1992): **Detection of Epileptic Spikes in the EEG Using Adaptive Filters**, Proceedings, I Fórum Nacional de Ciência e Tecnologia em Saúde—XIII Congresso Brasileiro de Engenharia Biomedica, Caxambu (MG), Brazil, .

[129] MYLONAS, S. A.; COMLEY, R. A.; (December 1992): **Adaptive Predictive Modelling for the Analysis of the Epileptic EEG**, Proceedings, ICCS/ISITA '92, Singapore, *Vol. 3*, pp. 1214–1218.

[130] MYLONAS, S. A.; COMLEY, R. A.; (July 1993): **Linear Prediction, Neural Networks and the Analysis of EEG signals**, Proceedings, International Conference on DSP and Second Cyprus International Conf. on Computer Appl. to Eng. Sys., pp. 286–291.

[131] MYLONAS, S. A.; COMLEY, R. A.; (September 1994): **EEG Analysis Using a Multi-Layer Perceptron With Linear Preprocessing**, Proceedings, IV IEEE Workshop on Artificial Neural Networks In Signal Processing, pp. 671–680.

[132] MYLONAS, S. A.; (1993): **The The nervous system**, Internal Report, Centre for Information Engineering, City University.

[133] MYLONAS, S. A.; (1993): **The electroencephalogram**, Internal Report, Centre for Information Engineering, City University.

[134] MYLONAS, S. A.; (1994): **Epilepsy and the EEG**, Internal Report, Centre for Information Engineering, City University.

[135] MYLONAS, S. A.; (1994): **Internal Report, Review of Automatic EEG Quantification and Monitoring**, Centre for Information Engineering, City University.

[136] HECHT-NIELSEN, R.: (1990): **Neurocomputing,** , ISBN 0–201–09355–1, Addison–Wesley.

[137] OFFNER, F. F.; (December 1984): **Bioelectric Potentials—Their Source, Recording and Significance**, IEEE Transactions on Biomedical Engineering, *Vol. BME-31*, **No. 12**, pp. 563–568.

[138] OKEN, B. S.; CHIAPPA, K. H.; (1988): **Short-Term Variability in EEG Frequency Analysis**, Electroencephalography and Clinical Neurophysiology, *Vol. 69*, pp. 197–198.

[139] DE OLIVIERA, P. G.; QUEIROZ, C.; LOPES DA SILVA, F.; (1983): **Spike Detection Based on a Pattern Recognition Approach Using a Microcomputer**, Electroencephalography and Clinical Neurophysiology, *Vol. 56*, pp. 97–103.

[140] ORFANIDIS, S. J.; (1990): **Optimum Signal Processing—An Introduction**, (*2nd edition*), ISBN 0-07-100834-9, McGraw-Hill, New York.

[141] PANYCH, L. P.; WADA, J. A.; BEDDOES, M. P.; (1985): **Automation of the Seizure Investigation Unit at the University of British Columbia Health Sciences Centre Hospital**, (short Communication), Electroencephalography and Clinical Neurophysiology, *Vol. 61*, pp. 588–591.

[142] PANYCH, L. P.; WADA, J. A.; BEDDOES, M. P.; (1989): **Practical Digital Filters for Reducing EMG artefact in EEG Seizure Recordings**, Electroencephalography and Clinical Neurophysiology, *Vol. 72*, pp. 268–276.

[143] PAPP, Z.; PECELI, G.; BAGO, B.; PATAKI, B.; (June 1988): **Intelligent Medical Instruments**, IEEE Engineering in Medicine Magazine, pp. 18–23.

[144] PENCZEK, P.; GROCHULSKI, W.; GRYZB, J.; KOWALCZYK, M.; (1987): **The Use of Multichannel Kalman Filter Algorithm in Structural Analysis of the Epileptic EEG**, International Journal of Biomedical Computing, *Vol. 20*, pp. 135–151.

[145] PENCZEK, P.; GROCHULSKI, W.; (1989): **Analysis of Multi-Channel Epileptiform EEG Using the Markov Chains Formalism**, Methods of Information in Medicine, *Vol. 28*, pp. 160–167.

[146] PEPER, A.; GRIMBERGEN, C. A.; (April 1983): **EEG Measurement During Electrical Stimulation**, IEEE Transactions on Biomedical Engineering, *Vol. BME-30*, **No. 4**, pp. 231–233.

[147] PFURTSCHELLER, G.; HARING, G.; (1972): **The Use of an EEG Autoregressive model for the Time-Saving Calculation of Spectral Power Density Distributions With a Digital Computer**, Electroencephalography and Clinical Neurophysiology, *Vol. 33*, pp. 113–115.

[148] PFURTSCHELLER, G.; FISCHER, G.; **A New Approach to Spike Detection Using a Combination of Inverse and Matched Filter Techniques**, Electroencephalography and Clinical Neurophysiology, *Vol. 44*, (1978): pp. 243–247.

[149] POLA, P.; ROMAGNOLI, O.; (1979): **Automatic Analysis of Interictal Epileptic Activity Related to its Morphological Aspects**, Electroencephalography and Clinical Neurophysiology, *Vol. 46*, pp. 227–231.

[150] POULARIKAS, A. D.; SEELY, S.; (1985): **Signals and Systems**, ISBN 0-534-03402-0, PWS-Engineering

[151] PROAKIS, J. G.; MANOLAKIS, D. G.; (1989): **An Introduction to Digital Signal Processing**, ISBN 0-02-946253-3, Macmillan International.

[152] PROAKIS, J. G.; RADER, C. M.; LING, F.; NIKIAS, C. L.; (1992): **Advanced Digital Signal Processing**, ISBN 0-02-396841-9, Macmillan International.

[153] PRAETORIUS, H. M.; BODENSTEIN, G.; CREUTZFELDT, O. D.; (1977): **Adaptive Segmentation of EEG records: A New Approach to Automatic EEG analysis**, Electroencephalography and Clinical Neurophysiology, *Vol. 42*, pp. 84–94.

[154] PRINCIPE, J. C.; SMITH, J. R.; BALAKRISHNAN, S. K.; PAIGE, A.; (December 1979): **Microcomputer-Based Digital Filters for EEG Processing**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-27*, **No. 6**, pp. 697–705.

[155] PRINCIPE, J. C.; SMITH, J. R.; (1982): **Microcomputer-Based System for the Detection and Quantification of Petit Mal Epilepsy**, Computers In Biology and Medicine, *Vol. 12*, **No. 2**, pp. 87–95.

[156] PRINCIPE, J. C.; SMITH, J. R.; (June 1986): **Design and Implementation of Linear Phase FIR Filters for Biological Signal Processing**, IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 6**, pp. 550–559.

[157] QUIAN, J.; BARLOW, J. S.; BEDDOES, M. P.; (January 1988): **A Simplified Arithmetic Detector For EEG Sharp Transients—Preliminary Results**, IEEE Transactions on Biomedical Engineering, *Vol. BME-35*, **No. 1**, pp. 11–17.

[158] RABINER, L. R.; GOLD, B.; (1975): **Theory and Application of Digital Signal Processing**, ISBN 0-13-914101-4, Prentice-Hall/ Englewood Cliffs.

[159] RICCI, D.; FIORE, L.; COLOMBETTI, G.; (July 1988): **Time-Domain Cross-Correlation Assembly Language Routine for the IBM PC and an Application in Neural Signal Analysis**, Medical & Biological Engineering & Computing, *Vol. 26*, pp. 454–456.

[160] RICHENS, A.; (February 1988): **Framework of Medical Care For Epilepsy**, British Journal of Hospital Medicine, pp. 97.

[161] ROBERTS, S.; TARRASSENCO, L.; (January 1991): **A New Method of Automated Sleep Quantification**, Report No. OUEL 1875/91, University of Oxford, Department of Engineering Science.

[162] ROBERTS, S.; TARRASSENCO, L.; (November 1991): **A New Method of Automated Sleep Quantification**, Proceedings: II International Conference on Artificial Neural Networks, pp. 210–213.

[163] ROMANO, J. M. T.; BELLANGER, M.; (1988): **Fast Adaptive Notch Filtering in Cascade Form**, in **Signal Processing IV: Theories And Applications** , LACOUME, J. L.; CHEHICKIAN, A.; MARTIN N.; MALBOS, J. (EDITORS) EURASIP, pp. 567–570 Elesevier Science Publishers.

[164] ROMANO, J. M. T.; BELLANGER, M.; CORADINE, L. C.; (1990): **Least Squares Adaptive Filter in Cascade Form for Line Pair Spectrum Modelling**, Proceedings, EURASIP Conference, pp. 120-122.

[165] ROWLAND, L. P. (EDITOR); (1979): **Merritt's Textbook of Neurology**, (*8th edition*), ISBN 0-8121-1148-6, Lea & Febiger Publishers.

[166] RUMELHART, D. E; HINTON, G. E.; WILLIAMS, R. J.; (1986): **Learning internal representations by error propagation**, in **Parallel Distributed Processing: Explorations in the microstructure of Cognition,,** RUMELHART, D. E.; MCLELLAND, J. L. (EDITORS); *Vol. 1*, pp. 318–362 ISBN 0-262-18120-7, MIT Press.

[167] SALTZBERG, B.; BURCH, N. R.; (1971): **Period Analytic Estimates of Moments of the Power Spectrum: A Simplified Time Domain Procedure**, Electroencephalography and Clinical Neurophysiology, *Vol. 30*, pp. 568–570.

[168] SALTZBERG, B.; (July 1976): **A Model for Relating Ripples in the EEG Power Spectral Density to Transient Patterns of Brain Electrical Activity Induced by Subcortical Spiking**, IEEE Transactions on Biomedical Engineering, *Vol. BME-23*, **No. 7**, pp. 355–356.

[169] SALTZBERG, B.; BURTON, W. D., JR.; BARLOW, J. S.; BURCH, N. R.; (1985): **Moments of the Power Spectral Density Estimated from Samples of the Autocorrelation Function (A Robust Procedure for Monitoring Changes in the Statistical Properties of Lengthy Non-Stationary Time-Series Such as the EEG)**, Electroencephalography and Clinical Neurophysiology, *Vol. 61*, pp. 89–93.

[170] SALTZBERG, B.; (December 1986): **An Efficient Formula for Estimating Generalized Moments of The Power Spectal Density (PSD) without Computing the Fourier Transform**, IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 12**, pp. 1134–1136.

[171] SANDMAN, A.; SAPIR, B.; (1988): **Third Order Polynomial—Its Use in Data Compression**, Signal processing, **No. 15**, pp. 405–418.

[172] SCAMBLER, G.; (1989): **Epilepsy—The Experience of Illness Series**, ISBN 0-41501758-0, Tavistock/Routledge.

[173] (1980): **The Volterra and Wiener Theories of Non-Linear Systems**, ISBN 0-471-4455-5, Wiley.

[174] SCHWANN, H. P.; (December 1984): **The Development of Biomedical Engineering: Historical Comments and Personal Observations**, IEEE Transactions on Biomedical Engineering, *Vol. BME-31*, **No. 12**, pp. 730–736.

[175] SCHWARTZ, M.; SHAW, L.; (1976): **Signal Processing**, ISBN 0-07-055662-8, McGraw-Hill Inc.

[176] SCHWARTZ, M. D.; (December 1984): **The Emerging Field of Clinical Engineering and its Accomplishments**, IEEE Transactions on Biomedical Engineering, *Vol. BME-31*, **No. 12**, pp. 743–747.

[177] SCOTT, D.; (1976): **Understanding EEG—An Introduction to Electroencephalography**, ISBN 0-7156-0939-4, Gerald Duckworth & Co. Ltd.

[178] SCOTT, D. F.; (July 1987): **Indications For The EEG**, editorial, British Journal of Hospital Medicine, pp. 111.

[179] LOPES DA SILVA, F. H.; VAN HULTEN, K.; LOMMEN, J. G.; STORM VAN LEEUWEN, W.; VAN VEELEN, C. W. M.; VILIEGENTHART, W.; (1977): **Automatic Detection and Localization of Epileptic Foci**, Electroencephalography and Clinical Neurophysiology, *Vol. 43*, pp. 1–13.

[180] SIU, S.;COWAN; C. F. N; (November 1991): **Adaptive Equalization Using the l_p Back Propagation Algorithm**, Proceedings, Second International Conference on Artificial Neural Networks, Bournemouth, U.K., pp. 10–13.

[181] SMITH, J. R.; (January 1974): **Automatic Analysis and Detection of EEG Spikes**, IEEE Transactions on biomedical Engineering, *Vol. BME-21*, **No. 1**, pp. 1–7.

[182] SMITH, W. D.; LAGER, D. L.; (March 1986): **Evaluation of Simple Algorithms for Spectral Parameter Analysis of the Electroencephalogram**, IEEE Transactions on Biomedical Engineering, *Vol. BME-33*, **No. 4**, pp. 352–358.

[183] STELLE, A. L.; COMLEY, R. A.; (September 1989): **Portable Analyser for Real-Time Detection of the Epileptic Precursor**, Proceedings, XI Brazilian Conference For Biomedical Engineering, pp. 101–107.

[184] STELLE, A. L.; COMLEY, R. A.; (October 1990): **The Use of the Wigner Distribution in the analysis of EEG signals**, Proceedings, XII Brazilian Conference For Biomedical Engineering, .

[185] STELLE, A. L.; (1991): **On-line Analysis of the EEG: A Two-Dimensional Approach**, PhD Thesis, Centre for Information Engineering, City University.

[186] SUN, M.; LI, C. C.; SEKHAR, L. N.; SCLABASSI, R. J.; (November 1989): **Efficient Computation of the Discrete Pseudo-Wigner Distribution**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-37*, **No. 11**, pp. 1735–1741.

[187] THAKOR, N. V.; WEBSTER, J. G.; TOMPKINS, W. J.; **Design, Implementation and Evaluation of a Microcomputer-Based Arrhythmia Monitor**, Medical & Biological Engineering & Computing, *Vol. 22*, (March 1984): pp. 151–159.

[188] TURIN, G. L. (Jume 1960): **An Introduction to Matched Filters**, IRE Transactions on Information Theory, *Vol. IT-6*, pp. 311–329

[189] VAN DEN BERG-LENSSEN, M. M. C.; BRUNIA, C. H. M.; BLOM, J. A.; (1989): **Correction of Ocular Artifacts in EEGs Using an Autoregressive model to describe the EEG; a pilot Study**, Electroencephalography and Clinical Neurophysiology, *Vol. 73*, pp. 72–83.

[190] VIDAL, J. J.; (May 1977): **Real-Time Detection of Brain Events in the EEG**, IEEE Proceedings, *Vol. 65*, **No. 5**, pp. 633-641.

[191] DE VRIES, J.; WISMAN, T.; BINNIE, C. D.; (1981): **Evaluation of a Simple Spike-Wave Recognition System**, Electroencephalography and Clinical Neurophysiology, *Vol. 51*, pp. 328–330.

[192] WALTER, D. O.; MULLER, H. F.; JELL, R. M.; (January 1973): **Semiautomatic Quantification of Sharpness of EEG Phenomena**, IEEE Transactions on Biomedical Engineering, *Vol. BME-20*, **No. 1**, pp. 53–55.

[193] WALTON, J. N.; (1981): **Brain's Diseases of the Nervous System**, [Chapter 22: **Paroxysmal and Convulsive Disorders—Epilepsy**,], (*8th edition*), ISBN 0-19-261309-X, Oxford Medical Publications.

[194] WAX, M.; KAILATH, T.; (1985): **Detection of Signals by Information-Theoretic Criteria**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-33*, **No. 4**, pp. 387–392.

[195] WEIDE, B. W.; ANDREWS, L. T.; IANNONE, A. M.; (1978): **Real-Time Analysis of EEG using Walsh Transforms**, Computers in Biology and Medicine, *Vol. 8*, pp. 255–263.

[196] WEINBERG, H.; COOPER, R.; (August 1972): **The Use of Correlational Analysis for Pattern Recognition**, Nature, *Vol. 238*, pp. 292.

[197] WEINBERG, H.; COOPER, R.; (1972): **The Recognition Index: A Pattern Recognition Technique for Noisy Signals**, Electroencephalography and Clinical Neurophysiology, *Vol. 33*, pp. 608–613.

[198] WENNBERG, A.; ZETTERBERG, L. H.; (1971): **Application of a Computer-Based Model for EEG analysis**, Electroencephalography and Clinical Neurophysiology, *Vol. 31*, pp. 457–468.

[199] WHEELER, B. C.; VALESANO, W. R.; (1985): **Real-Time Digital-Filter-Based Data-Acquisition System for the Detection of Neural Signals**, Medical & Biological Engineering & Computing, *Vol. 23*, pp. 243–248.

[200] WIDROW, B.; GLOVER, J. R., JR.; MCCOOL, J. M.; KAUNITZ, J.; WILLIAMS, C. S.; HEARN, R. H.; ZEIDLEN, J. R.; DONG, E. JR.; GOODLIN, R. C.; (December 1975): **Adaptive Noise Cancelling: Principles and Applications**, IEEE Proceedings, *Vol. 63*, **No. 12**, pp. 1692–1716.

[201] WIDROW, B.; MCCOOL, J. M.; LARIMORE, M. G.; JOHNSON, C. R., JR.; (August 1976): **Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter**, IEEE Proceedings, *Vol. 64*, **No. 8**, pp. 1151–1162.

[202] WIDROW, B.; STEARNS, S. D.; (1985): **Adaptive Signal Processing**, ISBN 0-13-004029-03, Prentice-Hall Inc., NJ.

[203] WITTE, H.; GLASER, S.; ROTHER, M.; (March 1987): **New Spectral Detection and Elimination Test Algorithms of ECG and EOG Artefacts in Neonatal EEG Recordings**, Medical & Biological Engineering & Computing, *Vol. 25*, pp. 127–130.

[204] ZHU, Y. M.; PEYRIN, F.; GOUTTE, R.; (October 1989): **Equivalence Between Two-Dimensional Analytic and Real Signal Wigner Distributions**, IEEE Transactions on Acoustics, Speech and Signal Processing, *Vol. ASSP-37*, **No. 10**, pp. 1631–1633

Index