APPROXIMATELY PERIODIC TIME SERIES AND NONLINEAR STRUCTURES

THÈSE N^{O} 3113 (2004)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Institut des systèmes de communication

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES TECHNIQUES

PAR

Norman Urs BAIER

ingénieur électricien diplômé EPF
de nationalité suisse et originaire de Termen (VS)

acceptée sur proposition du jury:

Prof. M. Hasler, directeur de thèse
Dr O. De Feo, rapporteur
Prof. A. Ijspeert, rapporteur
Prof. G. Kubin, rapporteur
Dr V. Makarov, rapporteur

Lausanne, EPFL
2005

ABSTRACT

In this thesis a previously developed framework for modelling diversity of approximately periodic time series is considered. In this framework the diversity is modelled deterministically, exploiting the irregularity of chaos. This is an alternative to other well established frameworks which use probability distributions and other stochastic tools to describe diversity. The diversity which is to be modelled, on the other hand, is not assumed to be of chaotic nature, but can stem out from a stochastic process, though it has never been verified before, whether or not purely stochastic patterns can be modelled that way.

The main application of such a modelling technique would be pattern recognition; once a model for a learning set of approximately periodic time series is found, synchronisation-like phenomena could be used to determine if a novel time series is similar to the members of the learning set.

The most crucial step of the classification procedure outlined before is the automatic generation of a chaotic model from data, called identification. Originally, this was done using a simple low dimensional reference model. Here, on the other hand, a biologically inspired approach is taken. This has the advantage that the identification and classification procedure could be greatly simplified and the computational power involved significantly reduced.

The biologically inspired model used for identification was announced several time in literature under the name “Echo State Network”. The articles available on it consisted mainly of examples were it performed remarkably well, though a thorough analysis was still missing to the scientific community. Here the model is analysed using a measure that had appeared already in similar contexts and with help of this measure good settings of the models’ parameter were determined.

Finally, the model was used to assess if stochastic patterns can be modelled by chaotic signals. Indeed, it has been shown that, for the biologically inspired modelling technique considered, chaotic behaviour appears to implicitly model diversity and randomness of the learnt patterns whenever these are sufficiently structured; whilst chaos does not appear when the patterns are remarkably unstructured. In other words, deterministic chaos or strongly coloured noise lead to the chaos emergence as opposed to white-like noise which does not.

With this result in mind, the classification of gait signals was attempted, as no signs of chaoticity could be found in them and the previously available modelling technique seemed to have difficulties to model their diversity. The identification and classification results with the biologically inspired model turned out to be very good.

VERSION ABRÉGÉE

Dans cette thèse un cadre développé préalablement a été considéré, il a pour but de modéliser diversité de séries temporelles approximativement périodiques. Dans ce cadre la diversité est modélisée déterministiquement, en exploitant l'irrégularité du chaos. Ceci présente une alternative envers d'autres cadres bien connus qui utilisent des distributions de probabilités et d'autres outils pour décrire la diversité. La diversité à modéliser, d'autre part, n'est pas supposée être de nature chaotique. Elle peut provenir d'un processus stochastique, bien qu'il n'a pas été vérifié auparavant si des patterns purement stochastiques peuvent être modélisés ainsi.

L'application principale d'une telle technique de modélisation est la reconnaissance de patterns. Une fois un modèle trouvé pour un ensemble d'apprentissage, des phénomènes de synchronisation peuvent être utilisés pour déterminer si une nouvelle série temporelle est similaire à celles de l'ensemble d'apprentissage.

L'étape critique dans la classification esquissée juste avant est la génération automatique d'un modèle chaotique à partir de données, un processus appelé identification. Dans l'œuvre original, ceci a a été fait avec un modèle de référence simple, bas dimensionnel. Ici, au contraire, une approche inspirée biologiquement a été prise. Ceci avait l'avantage que l'identification et la classification pouvaient être largement simplifiées et la puissance de calcul nécessaire significativement réduite.

Le modèle biologiquement inspiré a été apparu plusieurs fois dans la littérature sous le nom "Echo State Network". Les articles disponibles donnaient surtout des exemples pour lesquelles le modèle se tenait remarquablement bien, une analyse approfondi manquait par contre. Ici le modèle est analysé à l'aide d'une mesure qui était apparue déjà plusieurs fois dans des contextes similaires. A l'aide de celle-ci les bonnes paramètres du modèle ont pu être déterminés.

Finalement, le modèle a été utilisé pour estimer si les patterns stochastiques peuvent être modélisés avec des signaux chaotiques. Effectivement, il a été montré que, pour l'approche inspirée biologiquement, le comportement chaotique semble modéliser implicitement la diversité et l'aléatoirité des patterns appris toujours quand ceux-ci sont assez structurés; tandis que le chaos n'apparaît pas quand les patterns sont remarquablement instructurés. En d'autres termes, chaos déterministique ou bruit fortement coloré mènent à l'apparition du chaos lors de l'identification, contrairement à du bruit blanc (ou presque blanc) qui ne le fait pas.

Se rappelant de ce résultat, la classification de signaux de marche a été abordée. Les signaux de marche ne montraient pas de signes apparent de chaos et la méthode de modélisation disponible préalablement avait des difficultés à modéliser leurs diversité. Les résultats de l'identification et de la classification avec le modèle biologiquement inspiré se sont avérés très bons.

ACKNOWLEDGEMENTS

I would like to thank god and the universe. God for not having interfered with physical laws during the last 4 years and the universe for being stable like it is. This facilitates scientific work a lot, thanks.

CONTENTS

1 INTRODUCTION 1
1.1 State of Current Research 1
1.1.1 Nonlinear Modelling 5
1.1.2 Biologically Inspired Modelling 6
1.2 Motivations 8
1.3 Outline of the Thesis 8

2 FUNDAMENTALS 11
2.1 Approximately Periodic Time Series (APTS) 11
2.1.1 APTS Used for Training and Classification 14
2.1.2 Analysis of APTS 18
2.2 Modelling Diversity of APTS 22
2.2.1 Iterative Identification of Systems in Lur'e Form 23
2.2.2 Drawbacks 29

3 DETERMINISTIC MODELLING OF GAIT SIGNALS 31

4 BIOLOGICALLY INSPIRED MODELLING OF DIVERSITY 37
4.1 Presentation of the Model 37
4.1.1 Design Parameters 39
4.1.2 The Memory of the Nodes 40
4.1.3 Diversity Pool Interpretation 41
4.1.4 Identification Process 41
4.2 Entropy as a Measure for Diversity 42
4.3 Optimising the System's Entropy 43
4.3.1 Nonuniform Bias 44
4.3.2 Training Length 46
4.3.3 Dimension 47
4.3.4 Connection Probability and Memory 48
4.3.5 Matrices with Components 49
4.4 Performance 51
4.4.1 Generalisation Ability 52
4.4.2 Classification Results 59

5 CONCLUSIONS & FINAL DISCUSSION 67
5.1 Future Work 68

A ASYMPTOTICALLY UNIQUE BEHAVIOUR 71
A.1 Necessary Conditions 71
A.2 Sufficient Conditions 72

CHAPTER 1

INTRODUCTION

Brief — In this chapter, the context of this work is introduced. The state of current research and the motivations that led to this work are given. At the end the outline of the thesis can be found.

In one of the later chapters of this thesis some classification tables can be found, it would therefore be natural to put it on the shelve where all the other material from the field "machine learning" stand. Having classified this thesis that way, the question emerges how the material presented here is different from what could have been found already on the shelve before. First of all, the tables at the end may be misleading: the emphasis of this work does not lie on classification, but on the modelling which takes place before, and this is exactly where the books already present on the shelve most probably differ. To make this point clearer, a brief overview, which estimates what already is on the shelve, follows.

1.1 STATE OF CURRENT RESEARCH

In connection with classification and pattern recognition Platon and Aristotle are often cited. Whereas it can be doubted that either of them had a modern digital signal processor (DSP) in mind when establishing their theory, it most certainly shows how diverse the motivations can be for entering the domain of cognition. The ones relevant here are rather of engineering nature and, in this respect, the centre of interest is a classification machine, or a pattern recognition system, which takes sensory data as input and makes a decision on what the sensory data represent. According to Duda et al. [2001], such a pattern recognition system can in general be divided into components, as shown in Figure 1.1.

1. The first component is responsible for sensing, it has to convert physical inputs available to the pattern recognition system into signal data.
2. During segmentation the data is preprocessed in a way to isolate the relevant part of the data from any other information like background noise.
3. A feature extractor measures object properties that are useful for classification.

4. The classifier uses the features extracted to assign the sensed object to a category.
5. Optionally, in the end other rules can be applied to the proposed assignment of the classifier. Rules which can take into account the contextual situation or the cost of errors.

Usually, the information flows from lower to higher levels only, but more complex pattern recognition systems may also employ feedback.

Of the different components possibly present in a pattern recognition system, most are problem or domain dependent. In fact, the aim of the blocks from "sensing" up to "feature extraction" is to make an abstraction of the the sensible world such that it can be represented as a point in a feature space, which is in general a subspace of $\mathbb{R}^n$, where n is the number of the features considered. The classifier is then required to divide the feature space in a way that every point has a category associated. For this task the classifier has a certain number of samples, called training data, available.

With respect to the scheme of Figure 1.1, the contents of this thesis has to be situated in the stage "feature extraction". Nothing contained herein will change the view on how "classification" is done, on the contrary, feature spaces used up to now could be extended with the features proposed here and then a well known classification algorithm could be used to do the final classification. "Classification" and "feature extraction" are linked by their interface; therefore, to understand better the role of "feature extraction" it is necessary to know how classification works, a review follows. There are a number of different strategies to do the classification, *i.e.* division of the feature space. In case the underlying distributions are known, or their parameters can easily be estimated from the training data, Bayes' decision theory can be used to do the division [Bernardo and Smith, 1996]; in case the distributions are not known, they can be estimated using Parzen windows, or some nearest neighbourhood rule can be used

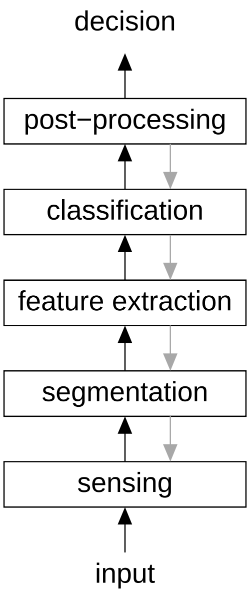


FIGURE 1.1: Block scheme showing the components of which most of pattern recognition systems are composed. The grey arrows indicate a possible feedback from higher to lower levels, which need not to be present [Duda et al., 2001].

[Parzen, 1962, Cover and Hart, 1967]. However, the need of calculating the underlying distribution can be avoided by estimating directly the bounds separating the training data.

Historically, the first approach to solve this separation problem has been to use linear decision functions, leading to the perceptron of Rosenblatt [Rosenblatt, 1958, Highleyman, 1962]. The main drawback with linear decision functions is, of course, that the categories have to be linearly separable, which unfortunately is rarely the case [Minsky and Papert, 1969]. There are two ways out of the dilemma: abandoning the analytically nice case of linear decision functions or transforming the input space such that the data become linearly separable. The efforts in abandoning linear decision functions led to multilayer nonlinear neural networks, which can be seen as an extension of the Rosenblatt perceptron. In early times the nonlinearity, or activation function, was a simple function like `AND`, `OR` or sigmoids, but with more computational power at hand they turned into Gaussian like functions, giving birth to the radial basis function networks [Duda et al., 2001, Hastie et al., 2001]. The second strategy to overcome the limitations of linear discriminant functions is to transform the feature space. The general idea is to augment the feature vector by nonlinear functions of its elements, the result is a generalised linear discriminant function [Duda et al., 2001]. In an attempt to overcome the computational issues associated with calculations in very high dimensional augmented feature spaces, the Support Vector Machine was created [Vapnik, 1995, Cristianini and Shawe-Taylor, 2004]. Its advantage over the generalised linear discriminant functions is that it does not need to specify the transformation functions explicitly, but only implicitly through inner products, its kernels [Hastie et al., 2001, Schölkopf and Smola, 2002].

This concludes the overview on currently used classification methods for static samples, it has been 300 years since Reverend Thomas Bayes has written his essay, and meanwhile the quantity of publications on classification has grown to a number where an overview of this size cannot be claimed complete, but all methods possibly of interest in this context are named. What follows from the overview is that “classification” is the act of dividing the feature space, a hypercube, into categories.

On the other hand, some problems encountered in classification cannot be solved by classifying statical samples, rather the dynamic evolution of a series of samples is important. For such problems, the hypercube of features has to be extended by the dimension along which the evolution takes place, and finally not a point but a path is considered for classification. One of the most successful frameworks within whose this is done are the Hidden Markov Models (HMMs), which are well known for their successful application to the speech recognition task, in which case evolution takes place along the time [Rabiner and Juang, 1993]. Other problems addressed by HMMs are sequences of bases in DNA molecules [Krogh et al., 1994]. In such a case the hypercube would have to be extended with the position of the basis in the DNA molecule. What is important in either case is the evolution that is modelled by the HMM. Throughout this work such a stochastic dependence between samples has not been considered and consequently HMMs have not been applied.

As stated in the beginning of this section, “classification” is the only component of a classifier which is not problem specific (Figure 1.1). This means that, as soon as some of the other components are discussed, generality is lost and specific examples have to be considered. The main interest lies here in features of approximately periodic time series.

A choice which may appear academic, but a lot of interesting applications lie in that domain and the next subsection (Section 1.1.1) gives further justification of that choice.

The most famous approximately periodic time series are probably voiced speech signals. Although their classification usually involves Hidden Markov Models, and HMMs will not be considered here, it may be worthwhile to look at the feature space chosen in the context of speech signals. Rabiner [1989] uses a 24-dimensional feature vector. The vector holds 12 coefficients of the cepstral analysis and an estimate of the time derivative of the cepstral coefficients, both determined from $45\,ms$ long windows of the speech signal. The cepstral analysis is a spectral analysis, which separates the effects of the vocal tract from the excitation coming from the larynx [Deller et al., 1993]. Two different strategies exist to handle the original $\mathbb{R}^{24}$ feature space. On one hand it can be reduced to a quantised set with a finite number of elements [Makhoul et al., 1985]. The quantised vectors can be interpreted as the stereotypes of verbal communication. The quantisation permits to associate to each member in the set an emission probability, which gives the probability of observing this particular feature vector [Makhoul et al., 1985]. Alternatively, the emission probabilities can be modelled with a continuous probability density function, which is usually a mixture of Gaussians [Rabiner and Juang, 1993]. The particularity with HMMs is that the emission probabilities depend on a hidden state. With help of transition probabilities between the states the lexical constraints of verbal communication can be interpreted. Summarising, the features of voiced speech signals are the spectral characteristics of the generating elements (vocal tract plus larynx). Diversity is modelled probabilistically. *This is exactly an issue for which the work presented here differs*, in so far as the diversity is to be modelled deterministically here. However, it is important to stress at this moment that if the method of deterministically modelling the diversity of approximately periodic time series were to be applied to speech recognition, still HMMs or similar techniques could be applied to interpret the lexical constraints. Therefore, what is presented here cannot be considered as replacement to HMMs but rather, beside spectral methods, as a complementary way of extracting features out of approximately periodic time series. The next example, where no evident constraints link the samples, will underline this point.

Walking of human beings is a highly automated, rhythmic behaviour that is mostly controlled by sub-cortical locomotor brain regions [Beauchet et al., 2003, Ijspeert, 2003]. Therefore, signals obtained from a walking subject, usually called *gait signals*, have strong chances to be approximately periodic. Applications for these signals come from pathology, where they help in diagnosing diseases, or detecting whether the subject walks on flat ground or on stairs [Sekine et al., 2000]. Means of obtaining gait signals depend on the application. Here gait signals holding accelerations of different parts of the body are mainly considered. For processing, likewise to the case of speech recognition, the gait signals are divided into frames, but in the case of gait signals all information is assumed to be contained in one frame, therefore the use of HMMs is not necessary. Section 2.1.1 discusses in greater detail gait signals, for the time being it can be stated that their analysis involves usually wavelets, which in turn are some kind of spectral analysis [Sekine et al., 2002, Coley et al., 2005]. Selected coefficients of the wavelet analysis are used to build up the feature space. Wavelet analysis, in turn, is looking for the presence of particular temporal patterns. The less the signal changes from period to period, the better works the classification based on wavelet analysis. This means that the method is sensible to inter-periodic diversity.

Both of these methods (speech recognition and classification of gait signals) have a rather precise view on how the measured signal should look like, more precisely they have a prototype of it in mind. Any deviation from this prototype deteriorates the reliability of the classification. Alternatively, the decision could be founded on an implicit ruling. For this, a model needs to be set up describing how the time series evolve, rather than describing them explicitly. This can be done, for instance, by using stochastic differential equations, which find their applications mainly in finance, where they are used to model asset prices [Mao, 1997]. With a similar approach the diversity would be modelled inherently, as part of the process. Still, the origin of the diversity would be explained with randomness.

There are two reasons to consider another approach, namely a deterministic modelling of diversity. First, there are time series that are strongly suspected to show chaotic behaviour. This is particularly true for voiced speech signals, but also for other signals like electrocardiograms [Herzel, 1993, Liebert, 1991]. By using a chaotic model to describe the temporal evolution of the speech signal, the diversity present in that same signal could be used to identify it even better, whereas in the case of a spectral analysis the diversity tends to deteriorate the classification. Second, even though it is still disputed how exactly information processing is done in the brain, and what its ingredients are, one possibility is that it involves chaotic phenomena to represent the diversity of the perceptible world. Whether or not this is the case is one of the key questions of the European project "APEREST"[1], under the hood of which most of this research was carried out. In case chaotic phenomena play a role for modelling diversity in the brain, a preliminary condition has to be fulfilled. Chaos would need to appear in biological neuronal networks to represent patterns even when they do not bear specific signs of chaoticity. Within this thesis research, and in the context of classification, this translates to the requirement that in artificial (dynamic) neural networks chaos should appear to model diversity independently from chaotic origins of the learnt patterns. This is exactly the point addressed in this thesis, and in the next two sections the "mathematical foundations" and the reference artificial neural network considered are briefly reviewed.

1.1.1 NONLINEAR MODELLING

The idea of using chaos to model diversity was pioneered by De Feo [2001]. Fundamental for that is the observation of a parallel between chaotic systems and categories of approximately periodic time series. Indeed, a chaotic system produces a whole family of trajectories that are all different, but still similar to each-other [Ott, 1993, Kuznetsov, 1998]. Equivalently, a class of approximately periodic time series contains a multitude of time series, all similar, but none identical to another one. This parallel between the perceptible world and dynamic systems can be used for modelling purposes, thereby replacing stochastic processes with deterministic equations in the modelling process [De Feo, 2001].

Adopting this new approach, the act of modelling becomes the automatic generation of a chaotic dynamical system representing the class of approximately periodic time series, where the class to be modelled is represented by examples of its members. The procedure of generating dynamical system from data is called identification, which would then become the crucial step toward classification [Bittanti and Picci, 1996]. Then, as

---

[1] `http://aperest.epfl.ch`

feature for the classification the synchronisation error can be employed; this reduces the time series to a 1-dimensional feature vector with simply connected decision regions making the classification trivial.

The method showed remarkable success when used in conjunction with a simple reference model for identification. Indeed, when identifying vowels and electrocardiograms, chaos did emerge during identification. However, the emerging chaotic behaviour was subharmonic, *i.e.* period doubling-like, and therefore not suitable to be exploited in a classification framework [De Feo, 2003]. In fact, the synchronisation-like phenomenon used for classification, named *qualitative resonance* by its inventor, needs homoclinic chaos [De Feo, 2004a, De Feo, 2004b]. So after identification the system has to be modified in a way to no longer exhibit subharmonic but homoclinic chaos. Making the whole procedure rather laborious.

As an alternative to this particular reference model, models in a biologically inspired form can be considered. Choosing such a model could mean trading in some of the system's transparency against a simpler identification and model tuning procedure.

1.1.2 BIOLOGICALLY INSPIRED MODELLING

Originally, the motivation behind biologically inspired modelling was to gain more insight in how information processing in the brain could work. The models that were derived thereby were called neural networks. Because they revealed a certain number of desirable properties for signal processing, they found application also in engineering, where they are often called basis function networks [Haykin, 1998]. As mentioned above with respect to the perceptron, one of the possible applications of these networks is classification, where the networks are used statically. Alternatively, when each neuron is a dynamic model, the case considered here, they may also exhibit self sustained oscillations. Though, the topology of the network remains similar and it is illustrated in Figure 1.2. Generally, a neural network is composed by an input layer, which usually holds one or several nodes that are connected with a weight to a first hidden layer of neurons (or basis functions). Through the input layer input functions (signals) are fed in. The sum of all inputs to a neuron is given as argument to its activation or basis function $a_{i1}(\cdot)$ in the static case, in case of dynamic neurons the same sum is given as input to the governing

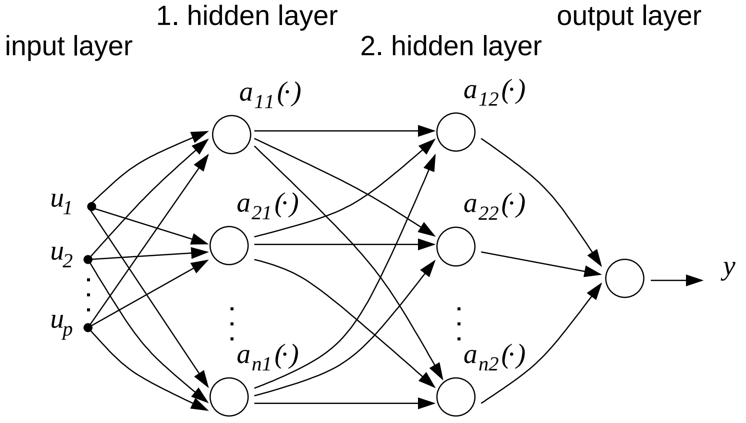


FIGURE 1.2: The general topology of a neural network. The arrows are weighted connections and the $a_{ij}(\cdot)$ are the activation functions.

difference / differential equation. The output of the neurons can then be input to another layer and so forth, until finally the output of the last layer is fed into the output neuron. Most often the activation functions $a_{ij}(\cdot)$ are all equal apart from a bias, and the output neuron is simply a linear combiner [Nelles, 2001]. The resulting network is finally called a feed-forward neural network with a certain number of hidden layers. The network can also be augmented with connections on the same level, which would result in a recurrent neural network.

The free parameters of the network – those adjusted during identification – are the weights connecting the different neurons. The most commonly used algorithms for training neural networks foresee to adapt all weights in the network [Arbib, 1995, Puskorius and Feldkamp, 1994]. To avoid the need for large computational power, recently another approach has been introduced, namely to choose randomly the weights connecting the neurons and to adjust solely the weights to the output connections. Jaeger [2001] calls the resulting model the *Echo State Network* (ESN). Essentially the same idea was also suggested by Maass et al. [2002,2003], there with the name *Liquid State Machine* (LSM). The different names reflect also that the corresponding authors motivated their work by different reasons.

Maass et al. give at first position “perspectives for the interpretation of neural coding, design of experiments and data analysis in neurophysiology”. Therefore, they use for their neurons a “realistic” model, namely the integrate & fire model, and consequently the applications are computational neuroscience oriented, like modelling motion detection in cortical tissue [Bülthoff et al., 2003]. The LSM owes its name to an analogy with shape recognition based on transient analysis in fluid dynamics [Natschläger et al., 2002, Holden et al., 1991]. The main idea is that, like the surface of a liquid, the LSM has only one stable state, the rest state. Indeed, the metaphor has already been used to perform speech recognition using a water surface as a “liquid brain” [Fernando and Sojakka, 2003]. The ESN, on the other hand, are not composed of integrate & fire neurons, but use dynamical neurons with a simple sigmoid saturation function. The purpose of the ESN is seen as a black box model for engineering dynamical systems rather than biological models, but the fundamental idea of fixing the internal weights in beforehand and adjusting only the output weights remains. Because the application considered here lies in the domain of engineering, I have considered this second paradigm; hence, I will review it in more detail.

In the first report on ESNs they were used for learning periodic sequences as well as the Mackey-Glass chaotic attractor [Mackey and Glass, 1977], mainly with output feedback [Jaeger, 2001]. The principal content are examples where the network performed either in line or better than existing modelling techniques. For the given examples, the parameters that were used for network generation are given. Unfortunately, none of the parameters is justified in any way, a problem which seems to be recognised also by the author, as a number of open questions are given in the final section of the report. Questions which include how the parameters have to be set exactly or whether an exogenous source of noise really is necessary in the identification process.

A second report on ESN from the same author introduces MC, the Memory Capacity [Jaeger, 2002a]. This measure estimates the maximal time interval τ for which the network can behave as an ideal delay. More precisely, the maximal τ for which, given an input $u(t)$, the network gives as output $\hat{u}(t) \approx u(t-\tau)$. In the second part of the report examples are given that make use of the memory capability, but no direct link is made on how ESNs perform with respect to this measure. Furthermore, none of the questions raised in the first report were taken up again in the second report. Finally, a third report can be seen as survey paper on recurrent neural networks and on how they are trained [Jaeger, 2002b]. The report gives an introduction to backpropagation through time, real-time recurrent learning, extended Kalman filtering and finishes with taking up what has appeared already in the first two reports. At the very end, hints for a good design of the ESN are given, although they stem mostly from observations that had already appeared in the other reports, without any further justification.

Some new information is given in an article entitled "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Telecommunication" [Jaeger and Haas, 2004]. The central element in this report is a graph showing how ESN outperform other methods of channel equalisation in a wireless communication model. The performance of the ESN is given with two graphs, a best and a mean graph for which performance differs by roughly one order of magnitude. This is the main drawback with ESNs, already there are examples for which working parameter settings exist, but it is not known if they are optimal. Furthermore, the parameters include random setting of weights and therefore their performance is not deterministic.

Despite the poor documentation available on them, ESNs appear to be a very promising biologically oriented modelling tool, hence they have been considered for deterministically modelling the diversity in approximately periodic time series.

1.2 MOTIVATIONS

As it can be guessed from the introduction so far, three elements did motivate the work presented in this thesis. The first, in order of appearance, is the theory of deterministically modelling the diversity in approximately periodic time series. As this theory is novel it has not yet been widely applied, consequently it was striven for to apply it further and improve it.

The second element are the aforementioned gait signals. A laboratory on the Campus, the Laboratory of Movement Analysis and Measurement (LMAM), had developed a device, which through its minor dimensions is perfectly adapted for the surveillance of gait and posture of elderly people. One application of the new device in which the LMAM was particularly interested was the classification of these signals, *i.e.* to detect whether the subject goes on flat ground or on stairs. This was of course a perfect occasion, where the deterministically modelling technique could prove useful in a real world example.

Finally, the ESNs themselves stand as a motivation. There is an apparent gap on how well they perform on certain examples and how well they are described in literature. The approach was judged interesting and therefore the model was scrutinised.

1.3 OUTLINE OF THE THESIS

The organisation of this thesis is as follows.

Chapter 1: Introduction. The purpose of this chapter is to cite the foundations on which the work is based, it defines the context of it and names the reasons, that led to it.

Chapter 2: Fundamentals. This chapter formally defines "approximately periodic time series". It also introduces the tools that are used for their analysis in the following chapters.

Chapter 3: Deterministic Modelling of Diversity. The theory presented in the second chapter is applied to data in this chapter.

Chapter 4: Biologically Inspired Modelling of Diversity. First, the biologically inspired model used throughout the chapter is presented, then a technique to analyse the fitness is introduced, which is used afterwards to tune the models parameters. At the end of the chapter, the tuned model is applied for identification and classification tasks.

Chapter 5: Conclusions & Final Discussion. In this chapter the different results of the thesis are summarised.

Chapter 2

Fundamentals

Brief — This chapter gives an overview on "approximately periodic time series" and introduces some terminology. After a formal definition of approximately periodic time series, a few examples are given, which will be used within the different modelling frameworks in the following chapters. The general properties of approximately periodic time series and tools for analysing them are given here as well.

Personal Contribution — Almost all the material presented is known. It can be found in any good book on nonlinear time series analysis. It is included here to support the choices of notation and representation of the following chapters.

The basic idea of deterministically modelling the diversity is to exploit somehow the irregular behaviour of chaos. Consequently, only patterns resembling chaos can be covered by this modelling technique. Considering dynamical systems, this means that exclusively approximately periodic time series can be modelled. This restriction is not so severe as it may seem in the first consideration; many interesting time series are approximately periodic. The next section presents some of them and also gives a simple, but in this context adequate, view on them.

2.1 Approximately Periodic Time Series

Approximately periodic time series are usually modelled within the framework of cyclostationarity [Gardner, 1994]. Cyclostationarity collects, practically, the mathematical tools framing linear periodic dynamical systems theory. As for linear dynamical systems theory, it considers two approaches; these are, the frequency (also known as external or spectral), approach, which includes, for instance, the spectrograms and the spectral-correlation density function analysis [Gardner, 1991], and the state space (also known as internal) approach, very much considered in linear periodic filtering and control [Bittanti and Colaneri, 1999]. However, the view on approximately periodic time series, which will be adopted here, is rather observational. The aim is not so much to describe

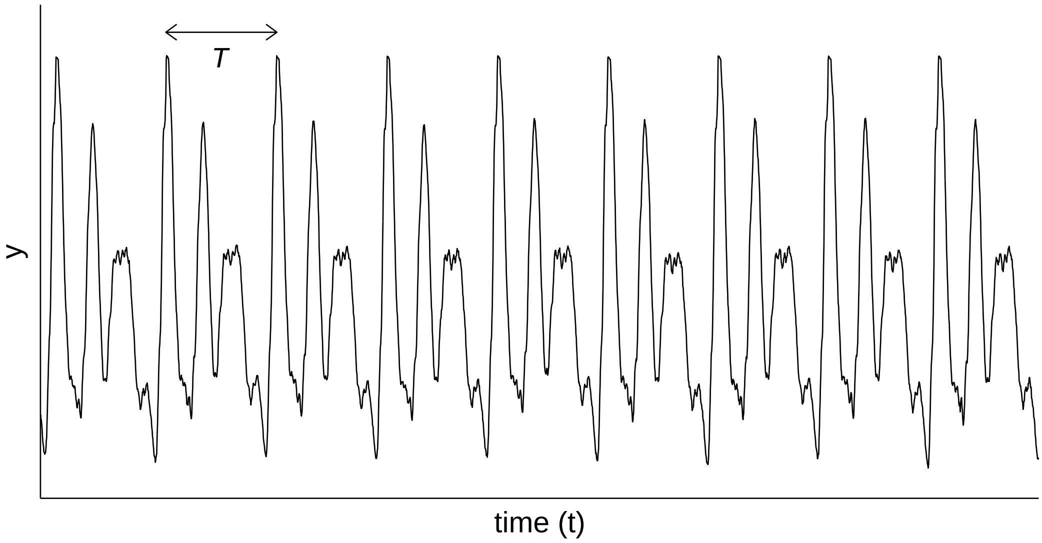


FIGURE 2.1: Example of an approximately periodic time series. In this normal representation in the time domain, it cannot be said on first sight, it is not periodic.

the time series dynamically, but to define the subject of the whole thesis and to give a simple tool for their synthesis, which will be used in a later Chapter (4.4.1).

To begin with, the definition of a periodic time series $y[t]$ is recalled. A time series is periodic if it is entirely defined by the values it takes during one period $t \in [t_0, t_0 + T)$:

$$y(t + nT) = y(t), \quad t \in [0, T), n \in \mathbb{Z} \tag{2.1}$$

Figure 2.1 shows an example of a time series that appears to fulfil the requirements for a periodic time series, the abscissa shows the time and the ordinate shows the amplitude. On the plot the period T is indicated. The same time series, with some additional periods is shown again in Figure 2.2 (grey lines); this time, on the other hand, the periods are not drawn one after the other, but rather on top the other. Such a plot can be called *stroboscopic plot.* In the case of a periodic time series all lines are perfectly congruent and the resulting plot would show only one line. For the time series in question, however, this is not the case. Clearly the different periods are similar, but not identical. So, strictly speaking the T indicated in Figure 2.1 is not the period of the time series, but its *pseudo period.*

The black line in the plot is the *mean stereotype* of the time series. The mean stereotype is the mean value the time series has at a given time after beginning of the time series. Formally this is written

$$y_s(t) = \frac{1}{P} \sum_{p=1}^{P} y(t + pT), \tag{2.2}$$

where P is the number of periods of the time series. An approximately periodic time series can then be written as the sum of its mean stereotype and a difference function $\delta(t)$

$$y(t) = y_s(t \mod T) + \delta(t). \tag{2.3}$$

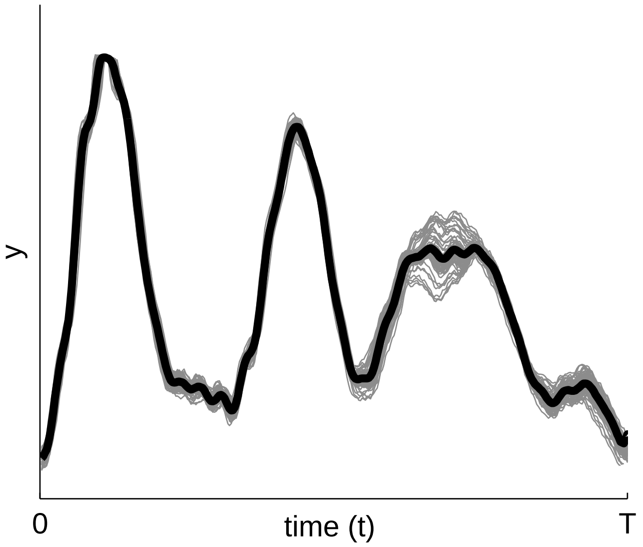


Figure 2.2: Stroboscopic plot of an approximately periodic time series. A stroboscopic plot is obtained by plotting the different pseudo periods of a time series on top of each-other. In this advanced representation, the realisation variation (see text) becomes visible.

Such a construction would of course be possible with any time series, not only with an approximately periodic one, but only when the time series bears inherent periodicity it is convenient. Here inherent periodicity means that the amplitude or the energy in $\delta(t)$ should be less than the one of $y(t)$. Other, more precise, criteria for inherent periodicity have been given in literature already, often in the context of speech processing [Deller et al., 1993], for the scope of this thesis the criteria depending on the energy fits best the needs.

The function $\delta(t)$ is the *realisation variation.* It can, but must not necessarily, give information on the state of the generating system and will be discussed in more detail in Section 2.1.2, but before some examples of approximately periodic time series will be given.

In contrast to a cyclostationary description of the same time series $y(t)$, here, the phase noise, *i.e.* the deviation of the period from its nominal value T, is not modelled explicitly. If the phase noise were to become important (in order of $\frac{T}{2}$), the variation term $\delta(t)$ would become similar in energy to the stereotype $y_s(t \mod T)$, thereby reducing the sense of the modelling. Before proceeding further, it should be noted that here (defined) time series have been considered, which show inherent approximately periodicity *in time.* This is not always the case; a better way of modelling approximately periodicity is with respect to a Poincare' section in a suitable state space, as done implicitly in cyclostationary signal modelling, especially when considering linear periodic models [Bittanti and Colaneri, 1999], or explicitly in reconstructed (embedded) state spaces [Kantz and Schreiber, 1999]. However, for the biologically inspired modelling approach considered in the following, any direct or indirect modelling of the time series in a suitable state space should be avoided, because the considered modelling method is supposed to do this implicitly. Nonetheless, as it will be clearer in a later Chapter (4), considering a pre-embedding of the signal, if appropriate, would not change the modelling method in its general outline, but will only augment the dimension of the input to consider. Hence, without any loss of generality, in the rest of this document I am going to consider only simple approximately periodic time series, which fit the definition given here.

2.1.1 APTS USED FOR TRAINING AND CLASSIFICATION

The time series that are considered later for identification will be presented here. Their origin and the reason why they are considered are given.

TIME SERIES FROM COLPITTS OSCILLATOR: SYNTHETIC SIGNALS FOR PERFORMANCE ASSESSMENT OF ESN

Time series obtained from the Colpitts oscillator have been used for a large part of the analysis. They were generated using a simplified, but realistic model

$$\begin{aligned} \dot{x}_1 &= \tfrac{g}{Q(1-k)}(-e^{-x_2}+1+x_3) \\ \dot{x}_2 &= \tfrac{g}{Q\,k}x_3 \\ \dot{x}_3 &= -\tfrac{Q(1-k)}{g}(x_1+x_2)-\tfrac{1}{Q}x_3, \end{aligned} \tag{2.4}$$

where the parameter values where chosen as

$$g = 3.5775, \quad Q = 1.5970, \text{and} \quad k = 0.5;$$

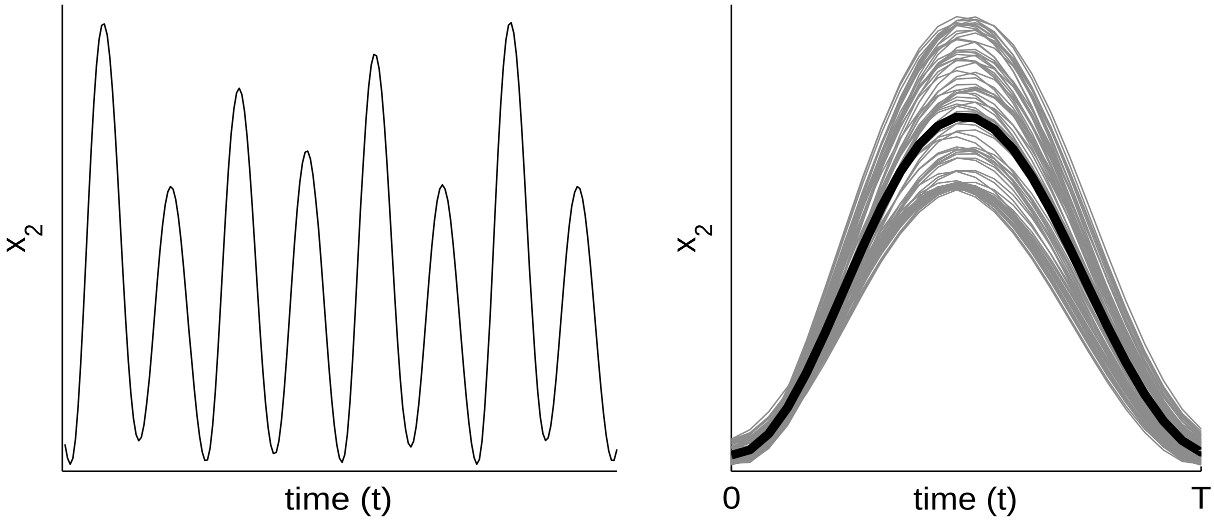
for these values the system operates in its chaotic region [Maggio et al., 1999, De Feo et al., 2000]. As output signal from the Colpitts oscillator, the second state variable was taken and sampled appropriately. A plot of a sample time series can be seen in figure 2.3.

The time series describe the chaotic behaviour of a system of ordinary differential equations and consequently the generating system has to be at least of order three, therefore it is not possible to construct another system producing the same output signal with fewer state variables [Strogatz, 1994]. More interesting is the fact that, by construction, a three-dimensional system of equations is enough to describe the dynamics in the time series. Applied to identification problems, these observations can be helpful to design the model, which is to be identified. The model must be at least of order three, on the other hand, if an order substantially higher than 3 is necessary to correct identification, then the model is not suited for the problem.

Summarised, the main criteria for choosing time series from the Colpitts oscillator were,

1. it produces chaotic time series,
2. they are produced by a 3-dimensional system of ODE.

Three-dimensional here means low dimensional with respect to the model used for identification which is high dimensional. The systems obeying to these criteria are numerous, to cite a few: the Rössler system, the Lorenz system and Chua's equations. All perfectly known to the scientific world. Such a stiff choice needs an explanation. The cited systems all contain symmetries. In the case of the Rössler system it is less evident, but still the equations contain a malicious term. As apparent symmetry is obviously a special case, which comes with some phenomenology associated to it – like anti-synchronisation – it should not be considered for testing purposes. Another property found in the Colpitts



(a) time domain

(b) stroboscopic plot

Figure 2.3: Example of a time series from the Colpitts oscillator. The time series were obtained using a simplified model of the well known electronic oscillator. In the plot the second out of three state variables is shown.

oscillator, but not in all of the other named systems is the stability, *i.e.* boundedness of solutions. Because the equations model a real physical circuit it would be bad modelling if for some parameter sets the state variables could reach infinity, this may not be true for the other systems, although this is not a decisive property. Finally, a reason also in favour of the Colpitts is the fact of personal experience [Baier et al., 2000].

Instead of generating approximately periodic time series by means of a chaotic model, they could be generated by somehow directly realising Equation (2.3), repeated here for convenience

$$y(t + nT) = y_s(t) + \delta(t + nT), \quad n \in \mathbb{Z}.$$

An important member of the category of synthetic time series is the *surrogate* [Schreiber and Schmitz, 2000] of another time series. For the scope of this thesis this is a time series that has the same spectral properties as the original one, but with arbitrary phase noise. The surrogate of the Colpitts time series would be constructed as follows: y_s would be the mean stereotype of the Colpitts, as defined in equation (2.2). Once calculated the mean stereotype, the realisation variation $\delta(t)$ can be calculated. Once the spectral properties of $\delta(t)$ are known, they can be realised by another process, generating the realisation variation of the surrogate ($\delta'(t)$) with the same spectral properties, but with arbitrary phase noise [Cohen et al., 1999]. Adding the realisation variation again to the stereotype calculated before, the surrogate of the original time series is obtained:

$$y'(t + nT) = y_s(t) + \delta'(t + nT), \quad n \in \mathbb{Z}. \tag{2.5}$$

The surrogate is introduced to test whether an identification algorithm focuses on identifying the very time series, meaning learns what sample comes after a number of previous samples, or if it focuses on identifying the spectral properties.

Another category within synthetic time series, would be time series that have freely designed spectral properties. The motivations for testing an identification algorithm with these time series are mainly the same as for surrogate time series. An example for this category are time series with coloured noise. These are time series with a spectrum of the form

$$\mathcal{F}(\delta(t)) = \begin{cases} 0 & \omega < \omega_0 \\ \frac{1}{(\omega-\omega_0+1)^{\frac{\gamma}{2}}} & \omega \geq \omega_0, \end{cases} \tag{2.6}$$

where γ was swept through $[0, 2]$, ω_0 was slightly lower than the radial frequency used for the stereotype, in order to avoid noise that appears be modulated by the stereotype (in the case the stereotype is almost a sinusoidal function).

GAIT SIGNALS: APPLYING ESN ON A REAL CLASSIFICATION PROBLEM

It has been stated in the introduction (page 4), that the walking of human beings bears strong periodic characteristics. Consequently, depending on how the walking is recorded, the resulting signals are approximately periodic time series. The way these signals, called gait signals, considered here are recorded has been proposed by the Laboratory of Movement Analysis and Measurement (LMAM). There, a device they call "Physilog" has been developed. It makes use of accelerometers and gyroscopes to capture any physical activity of the subject. The main advantage of using accelerometers and gyroscopes to capture gait signals is the transportability of the resulting device. Alternatively used capture components include sonic, magnetic and optical motion capture components, which often depend on a stationary installed unit.

The Physilog is a highly flexible device, which can record up to 8 channels of data from different sensors [Najafi et al., 2003]. The LMAM uses this device for different purposes. Among the different applications they consider, there is for instance, the evaluation of falling risk in elderly people [Najafi et al., 2002, Aminian and Najafi, 2004]. Because, compared to walking on flat ground, walking on stairs asks more energy from the subject, knowing whether the subject went on stairs or not can be useful in determining how exhausted the subject is. An information which finally influences the falling risk of the subject.

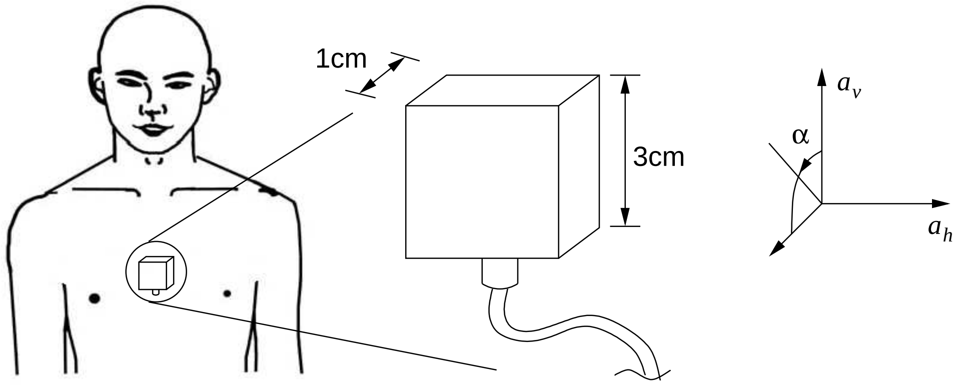


FIGURE 2.4: The sensor of the Physilog.

Finally, LMAM was interested in classifying gait signals, captured with a minimal setup of the Physilog, in samples of walking upstairs, downstairs and on flat ground. Because the problem formulated that way is exactly about classification of approximately periodic time series, this was an excellent opportunity to apply the technique of deterministically modelling approximately periodic time series to a real world example.

The Physilog, kindly offered by the LMAM to do the measures for this work, was composed of the recording unit and one sensor. The sensor was $3\,cm$ in height and width and about $1\,cm$ in depth, like drawn in Figure 2.4. The recording unit is small enough to hide in a pocket. Clearly, this presents a subtle and mobile way of measuring the human gait. The sensor measured 3 channels, the vertical acceleration a_v, the horizontal acceleration a_h and the angular acceleration α, like shown in Figure 2.4. The angular acceleration is meant vertical to the chest, the rotation axis leads through the hips; for measuring, the sensor is attached with the backside to the trunk of the subject. In this way, the accelerations along the spinal cord, along the shoulders and the angular acceleration of the trunk can be measured. The signals are measured with $8\,bits$ of precision and the sampling frequency can be set by the user. Here, it was set to $200\,Hz$. This corresponds roughly to 100samples per stride (or step), a similar value can be found in literature [Sekine et al., 2000, Coley et al., 2005].

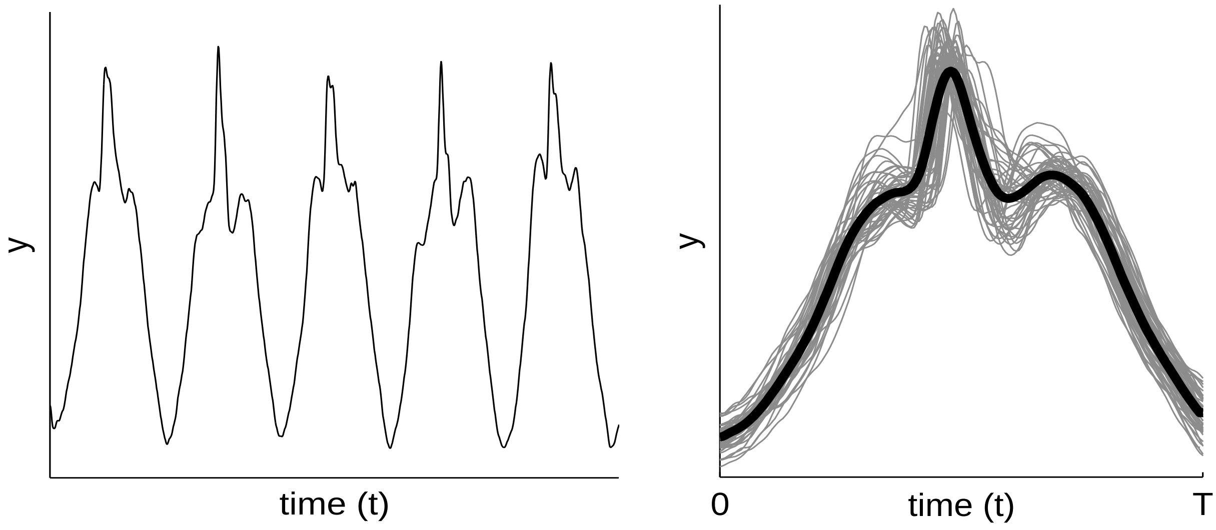
A similar setup was considered for classification by Sekine et al. [2000]. The main difference is that they mount their sensor on the backside of the trunk (*i.e.* the back) and not on the frontside (*i.e.* the chest). To do the classification they apply a wavelet transform to the vertical and the anteroposterior acceleration. The Physilog does not measure this acceleration, instead it measures the angular accelereration α. With help of the transform, they are able to detect, in the low frequency band, posture changes and significant peaks in the vertical acceleration. According to them, when classifying in walking upstairs, downstairs and walking on flat ground with this technique a 98.8% classification rate for all data sets they had at disposition is achieved.

The same authors propose in another publication the fractal dimension as features for classification. They were able to show a significant change in the fractal dimension according to the walking style, though no classification results are given [Sekine et al., 2002].

The only published result on classification done in LMAM considers another setup of the Physilog; the accelerometer was attached to the shank. In this study only classification between upstairs walking and flat walking is addressed. Classification rates are between 94% and 100% [Coley et al., 2005].

Other known attempts to classify gait signals obtained with wearable equipment include the results of Kern and Schiele [2003]. They used only one subject for testing their method and obtained classification rates contained in the band of 80% to 90%.

When trying to adapt the deterministically modelling method to the gait signals obtained with the Physilog, it was found that, depending on the subject, little or no information is contained in the horizontal and angular acceleration. Hence, also because the modelling method in its original form foresees only the modelling of a one dimensional time series, in the following only the vertical acceleration is used for the analysis. A typical sample of a gait signal considered for classification can be seen in Figure 2.5.



(a) time domain

(b) stroboscopic plot

FIGURE 2.5: Example of a gait signal obtained with the Physilog. Here represented is the vertical acceleration during 5 strides walking on flat ground.

2.1.2 ANALYSIS OF APTS

All that characterises an approximately periodic time series is of course contained in the realisation variation, because if it were not for this variance, the time series would be simply periodic. The different information that can be gained from the realisation variation will be described in this section. First a general reflection about it, is made, then a tool, the attractor reconstruction, is introduced. The attractor reconstruction is needed to estimate two properties of nonlinear time series, the correlation dimension and the Lyapunov exponent, which will be presented thereafter.

REALISATION VARIATION, DIVERSITY AND STRUCTURED NOISE

The realisation variation has been defined in equation (2.3) as the difference between the mean stereotype and the realisation, however it has not been said, where it stems from. This can be either the variedness, which is present in nature and in that case the realisation variation would stem from the *diversity*. On the other hand, realisation variation does not necessarily need to stem from diversity and it may be more indicated to refer to it as *structured noise*, above all for synthetic time series.

In both cases, the origin of the realisation variation is twofold. Once it appears within one realisation in the strict sense of the term, but it also appears between two realisations in a larger sense. In case both realisations were generated by exactly the same system, there might be no systematic difference between the two, however a difference in the generating system or its parameters cannot be excluded and a difference exists between the two time series, that cannot be explained by dynamic evolution.

Taking vowels as an example, the strict meaning of realisation variation describes the variation between some pseudo periods of the same vowel. In the larger sense it also includes the variation between a pseudo period of a different pronunciation of the same speaker, where the parameters of the generating systems might have slightly changed. It also includes the difference between the first pronunciation and a third pronunciation by a different speaker, where the parameters of the generating system surely have changed.

To sum it up, it cannot be assumed that a dynamic law forms the basis of the realisation variation. Though, it may be possible for an identification algorithm to define a model, which describes the realisation variation well. In that way the identification algorithm would have created an object, which has not been present before, a dynamic law, that ties up the different realisations. An algorithm, which is able to do so, is said here to have *generalisation ability.* This term comes originally from learning theory, where it refers to the ability to "learn characteristics of a general category of objects based on a series of specific examples from that category" [Caudill and Butler, 1990].

ATTRACTOR RECONSTRUCTION

Like it is stated aptly in the title, this thesis is about time series. Often time series are generated by a dynamical system with a state space of a few dimensions and observed through a function, which gives usually a one-dimensional output. Finally, by sampling the continuous time function a time series is obtained. During the observation all or some information on the state of the system got lost. The reasons for this loss of information is mostly because it is impossible or inconvenient to retrieve the desired information. Taking voiced speech signals as an example, the state space is made up of positions of the vocal cords and pressures along the throat, which are clearly less evident to measure than simply the acoustic wave once outside of the mouth [Ishizaka and Flanagan, 1972, Story and Titze, 1995].

Even though the purpose of the time series can be perfectly fulfilled, despite the loss of information, the analysis of the same time series is harder without knowing the state and its evolution in time, the trajectory. Among the tasks that are impossible or harder in the time series representation are prediction, it is impossible to predict accurately $y(t+1)$ knowing only $y(t)$, and comparison, two time series can look very different, but comparing the first time series to a delayed version of the second time series can make them look very similar. One possible solution for the two problems is attractor reconstruction, which aims at replacing the information stored in the state by the information contained in the evolution of time series [Sauer and Yorke, 1993, Kantz and Schreiber, 1999]. Different methods exist, but most commonly practised is the *time delay embedding.*

With time delay embedding a reconstructed attractor is built, whose samples are gradually delayed samples of the time series. Formally, for a reconstruction in n dimensions, this would give

$$\mathbf{x}_{\mathrm{rec},t} = \mathbf{x}_{\mathrm{rec}}(t) = \begin{bmatrix} y(t) \\ y(t-\Delta t) \\ \vdots \\ y(t-(n-1)\Delta t) \end{bmatrix} \tag{2.7}$$

Δt needs to be chosen appropriately. A good choice can be the first zero of the autocorrelation function [Kantz and Schreiber, 1999].

As illustration, the theory of delay embedding is applied to the Colpitts time series given above, the result is visible in Figure 2.6. The object represented in this Figure

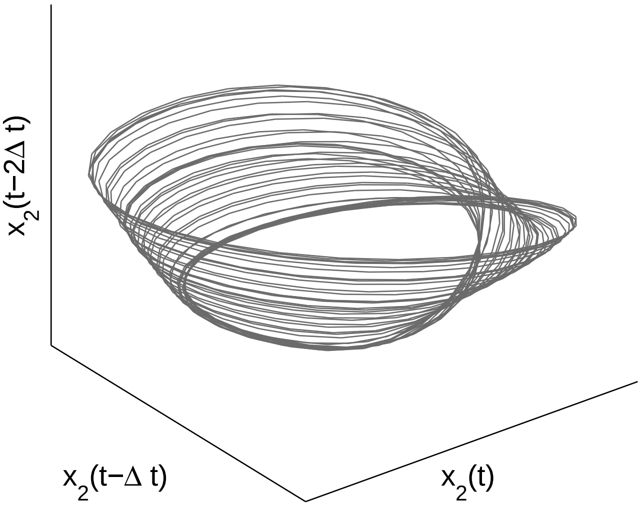


FIGURE 2.6: Attractor reconstruction of the Colpitts time series. The reconstruction was done using time delay embedding. The first dimension holds the time series, the others a successively delayed variant. Attractor reconstruction is used mathematical analysis and visualisation of the time series.

shows the same data as in Figure 2.3 (b). The delay used for the reconstruction was about $\Delta t = \frac{1}{3}T$, which was slightly more than the rule of thumb given above, just because the visualisation was better.

Finally, attractor reconstruction does not only lead to beautiful pictures, but it has to be used to estimate different properties of the time series and the generating system, among them the correlation dimension and the maximal Lyapunov exponent.

CORRELATION DIMENSION

Still unknown, up to now, is the dimension n that appeared in the last equation (2.7). When introducing the Colpitts time series in the last section, it has been emphasised, this dimension is three. In a general case, however, it has to be estimated. One way to do this is by estimating the correlation dimension; it estimates the dimension in which an object given by a set of data points $\mathbf{x}_i$ ($i = [1 \ldots N]$) lives. This can be applied to a set of points describing a reconstructed attractor $x_{\mathrm{rec},t}$ (2.7). If the estimation worked well and a valid correlation dimension could be obtained, this dimension gives an inferior bound for the dimension of the system that produced the time series.

The *correlation sum* is used to estimate the correlation dimension. It is defined (for simplicity without Theiler correction [Diks, 1999]) as

$$C(\epsilon, N) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} H(\epsilon - ||\mathbf{x}_i - \mathbf{x}_j||) \qquad (2.8)$$

where N is the length of the time series and $H(\cdot)$ is the Heaviside step function

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

ϵ is a small radius around $\mathbf{x}_i$ which is a point on the (reconstructed) attractor. The sum counts all points of the attractor that are nearer than ϵ. A fast reflection reveals that in the case of uniformly distributed points in a n-dimensional space this sum scales like a power law with respect to ϵ, $C(\epsilon) \propto \epsilon^n$. This leads to the following definition of the correlation dimension

$$D = \lim_{\epsilon \to 0} \lim_{N \to \infty} d(\epsilon, N) = \lim_{\epsilon \to 0} \lim_{N \to \infty} \frac{\partial \ln C(\epsilon, N)}{\partial \ln(\epsilon)} \tag{2.9}$$

Of course the estimated correlation dimension D cannot be greater than the embedding dimension n of the reconstruction, therefore the choice of a sufficiently high embedding dimension has to be assured, in fact the estimation should hold for several choices of the embedding dimension and a large range of ϵ.

Noise on the data tends to increase the value of $d(\epsilon, N)$ for small ϵ, this is intuitive, noise destroys the structure of the attractor and puts the data points unordered in the whole space of the reconstruction.

MAXIMAL LYAPUNOV EXPONENT

The maximal Lyapunov exponent measures quantitatively one of most apparent properties of a chaotic system, the sensitivity on initial conditions. It is sensitivity on initial conditions, which makes it in practise impossible to predict a chaotic system a long time into the future, despite the deterministic nature of the system. In a chaotic system any two trajectory starting arbitrarily close diverge exponentially from each other. More precisely, given two states of the system $\mathbf{x}_1(t_0)$ and $\mathbf{x}_2(t_0)$, that have a very small distance $d(t_0) = ||\mathbf{x}_1(t_0) - \mathbf{x}_2(t_0)|| \to 0$, their distance d would grow on the long run. In average the distance would grow exponentially, $d \propto e^{\lambda t}$ [Kantz and Schreiber, 1999, Ott, 1993]. Once the distance has reached the size of the attractor, it cannot grow anymore. In average means that the trajectories of two particular states may diverge and then for some time converge again, but the expectation for $d(t)$ grows exponentially.

A clear exponential divergence of nearby trajectories can therefore be taken as a sign for chaos. If on the other hand no stable exponential increase can be found within the time series, an eventual hypothesis of a chaotic time series has to be rejected, the time series was more likely generated by a stochastic process.

To actually estimate the Lyapunov exponent from a time series, the following equation can used [Kantz, 1994, Rosenstein et al., 1993]:

$$S(\Delta n) = \frac{1}{R} \sum_{r=1}^{R} \ln\Big(\frac{1}{|\mathcal{U}(\mathbf{x}_r)|} \sum_{\mathbf{x}_n \in \mathcal{U}(\mathbf{x}_r)} |x_{r+\Delta n} - x_{n+\Delta n}| \Big). \tag{2.10}$$

In this equation $\mathbf{x_r}$ are R reference points in a reconstructed attractor. $\mathcal{U}(\mathbf{x}_r)$ denotes the neighbourhood around the reference point $\mathbf{x_r}$ with diameter ε, which ideally is very small. The inner summation has as argument the distance at time Δn of the reference point $\mathbf{x_r}$ to one of its neighbours $\mathbf{x}_n$, $|x_{n_0+\Delta n} - x_{n+\Delta n}|$. The euclidean distance is avoided, in order to facilitate the comparison of results with different embedding dimensions [Kantz, 1994], instead the supremum out of the difference vector is used as distance. The outer sum has, consequently, as argument the logarithm of the average distance from a reference point to another point, given that Δn before they were at most ε far from each-other. The outer sum simply calculates the average of this logarithm.

If, finally, for some range of Δn the function $S(\Delta n)$ the function exhibits robust linear increase, its slope is an estimate of the maximal Lyapunov exponent in units of 1 over sampling periods [Kantz and Schreiber, 1999]. In case the increase is not robust over a large range of Δn and different values for embedding dimension and neighbourhood radii, a confident value cannot be determined and the realisation variation is likely to be due to noise [Rosenstein et al., 1993].

The realisation variation $\delta(t)$ may have systematically a higher amplitude for some values of t than for others, which means that for those t the standard deviation of the realisation variation is higher. In a stroboscopic plot this can be seen by the grey band, which has not the same width for all values of t. It has been found that identification can be negatively influenced by such a uneven distributed realisation variation, the small variation being masked by the large variation in a specific case [Baier, 2003]. A solution to this problem can be, depending on the time series, the observation of the time series through a nonlinear observation function $\chi(x)$. This applies a distortion to the time series, but does not affect its dynamical properties, furthermore if the function is invertible, the original time series can be recovered at any moment.

OBSERVATION FUNCTION

To equalise the standard deviation of the realisation variation for all t, the parts of the time series where the standard deviation is small should be more amplified than the other parts. Therefore, the observation function $\chi(x)$ should be inversely proportional to the standard deviation of the time series. Figure 2.7 illustrates this. The original time series at the bottom of the Figure is observed through the function $\chi(x)$, and consequently the realisation variation of the observed time series is distributed more evenly. The figure also shows the limits of this approach. When the stereotype $y_s(t)$ has the same value for two different values of t (t_1 and t_2 in occurrence), the realisation variation for those values cannot be amplified independently. For the time series shown in the figure this is not a problem, but clearly, the method is not applicable for the vowel presented in Figure 2.2 on page 13. In such a case a more elaborate method, like Kalman filtering, may lead to success [Grewal and Andrews, 1993]. However. in the scope of this thesis, this was not necessary.

2.2 MODELLING DIVERSITY OF APTS

Identification of approximately periodic time series and, accordingly, modelling the diversity of them is not a new subject, the use of a closed loop system in Lur'e form has given good results [De Feo, 2003, Bagni, 2004]. A system in Lur'e form consists of a closed loop of a dynamic linear filter $G(z)$ with a nonlinear static feedback $f_{\boldsymbol{\theta}}(x)$, like shown in Figure 2.8 (a), $\boldsymbol{\theta}$ indicates the function is parametrised and the vector $\boldsymbol{\theta}$ holds the parameters. The use of the z-transform indicates discrete time, but the formalism is also applicable to continuous time. The separation of the static nonlinearity from the

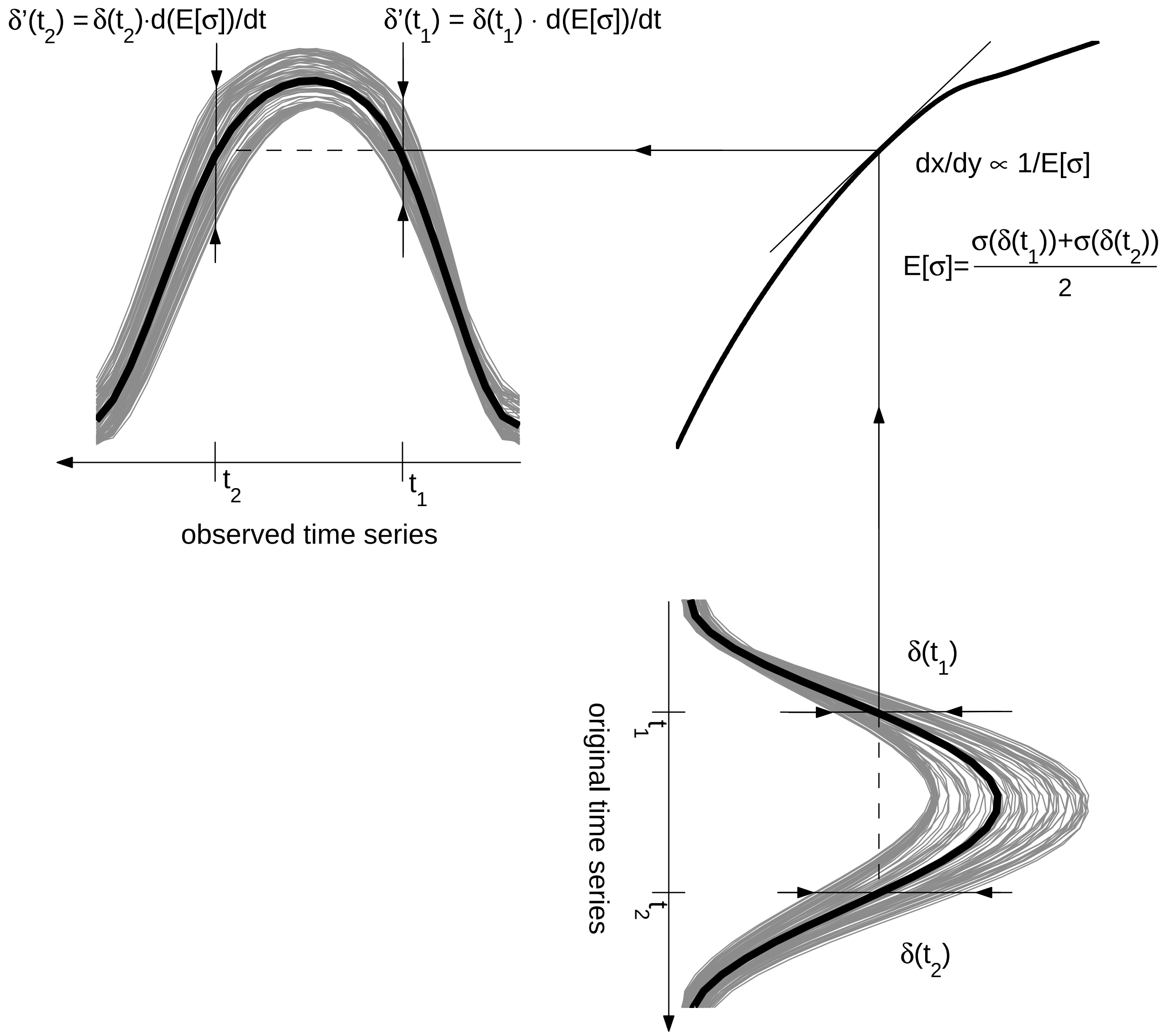

δ'(t₂) = δ(t₂)·d(E[σ])/dt
δ'(t₁) = δ(t₁) · d(E[σ])/dt
dx/dy ∝ 1/E[σ]
E[σ]= (σ(δ(t₁))+σ(δ(t₂)))/2
t₂
t₁
observed time series
original time series
δ(t₁)
δ(t₂)

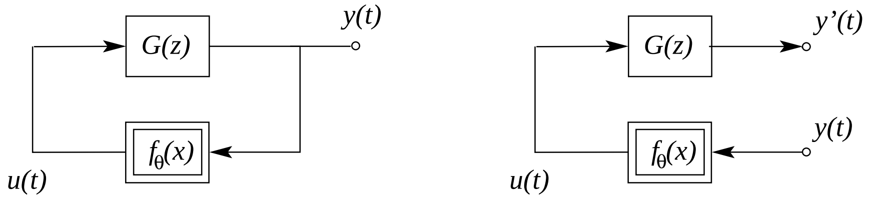
FIGURE 2.7: Observation of a time series through a nonlinear static function. The observation function $\chi(x)$ can be used to equalise the realisation variation.

linear dynamic part permits to apply different tools on both of them for identification. This is the main reason for choosing such a model structure. To identify the model an iterative method is proposed in literature, which is presented here.

2.2.1 ITERATIVE IDENTIFICATION OF SYSTEMS IN LUR'E FORM

To identify a system in Lur'e form, the closed loop is opened at the output of the linear filter $G(z)$, like shown in Figure 2.8 (b). Opened in this way, the model reminds of the Hammerstein (or after permutation the Wiener) model structure, unfortunately the corresponding identification algorithms are not applicable to closed loop systems, therefore the use of the iterative identification algorithm presented here [Bai, 2003].

The main idea behind the iterative approach is to split up the identification of the



(a) Closed loop. (b) Open loop during identification.

FIGURE 2.8: System in Lur'e form. The reference model used to model the diversity of APTS using the iterative approach. Nonlinear static function $f_{\boldsymbol{\theta}}(x)$ is separated from linear dynamic filter $G(z)$. This permits to use different tools for their respective identification.

whole system in two parts, namely a linear, dynamic identification part and a nonlinear, static optimisation part. Proceeding like this permits to use well known tools for the identification. The linear, dynamic subsystem $G(z)$ can be identified using standard linear techniques [Ljung, 1999] and for the the nonlinear, static function $f_{\boldsymbol{\theta}}(x)$ a standard algorithm can be used to find the optimal parameter vector $\boldsymbol{\theta}$.

Once the loop opened, the main point of the iterative algorithm is to note, that if a time series $y(t)$ were a solution of the closed system and if it would be fed into the open system, then the output of the open loop system $y'(t)$ should be identical to $y(t)$, its input. Given a nonlinear function $f_{\boldsymbol{\theta}}(x)$ and the solution $y(t)$, it is possible to calculate $u(t)$ and, hence, identify a linear model $G(z)$ using standard linear identification techniques. To find the right nonlinearity, a *cost function* that indicates how well $y(t)$ and $y'(t)$ correspond is defined. Any global optimisation algorithm can then be used to find the optimal parameter vector $\boldsymbol{\theta}$ of the nonlinearity by converging to the global minimum of the cost function.

The following subsections deal with the details of the nonlinear identification. Both parts, linear identification and nonlinear optimisation are treated separately. First the nonlinear optimisation is discussed and afterwards the linear identification.

NONLINEAR OPTIMISATION

The goal of the optimisation is to find the nonlinearity $f_{\boldsymbol{\theta}}(x)$, in Figure 2.8, such that the linear identification of $G(z)$ gives the best possible result. During the the optimisation, three characters interact with each-other: The **optimisation algorithm** is acting upon the parameter space of the **nonlinearity** and receives feedback through the **cost function**.

Optimisation Algorithm Optimisation is a topic with a broad field of applications and several algorithms are accessible to the scientific world. An important difference between the existing algorithms is their ability to find global optima or focusing only on local ones. Cost functions considered during iterative identification are of complicated nature and show many local optima. Figure 2.9 shows a typical example of a cost function

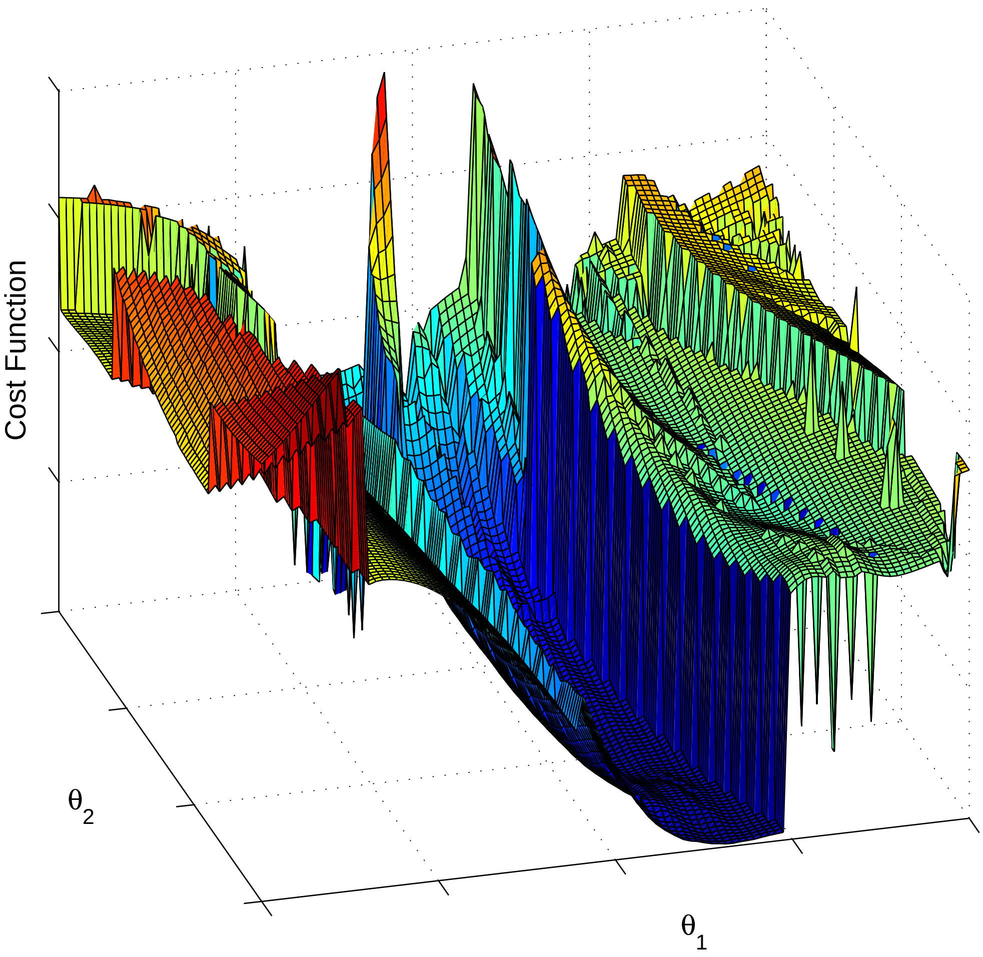


Figure 2.9: An example cost function, which should be minimised by simulated annealing. Clearly, the surface is not smooth and a local optimisation algorithm might get stuck in one of the local minima.

that has to be minimised[1]. For the plot all but two of the seven parameters were fixed, while sweeping the two remaining. The plot shows a surface which is barely continuous. The minimum of the plot is situated in the centre and is surrounded by several local minima. Consequently, the use of local optimisation algorithms would not lead to usable results. Known global optimisation algorithms are Simulated Annealing, the class of Evolutionary Algorithms, Taboo Search and others [Nelles, 2001]. While it is the aim of any of these algorithms to find the global optimum, and it has been proved at least for Simulated Annealing, that they find it in "infinite time", most often they return only a rather good local optimum. The quality of the result depends often more on how well the scientist understands the parameter he fits to the problem, than on the choice of the algorithm itself. For the identifications performed during the work of this thesis, the

---

[1] The minimum in the middle of the plot corresponds to the minimum found while identifying the Colpitts time series in the next chapter (Chapter 3).

Simulated Annealing algorithm was used.

Simulated Annealing owes its name to a physical analogy, namely the algorithm simulates the cooling process of a warm particle in a potential field. Once cooled down the particle will be positioned at the location with minimal potential, but, as long as it is warm, it moves around randomly trading in its kinetic energy against potential energy. This possibility to move upward, in opposite sense of the steepest descent, makes it possible for the particle to escape local minima, thus the algorithm is good for global optimisation: the cost function symbolising the potential field.

The algorithm can be summarised as follows:
First, the initial temperature T_0 has to be chosen. Its value depends on the smoothness of the cost function and no standard value can be given. Then, an initial parameter vector $\boldsymbol{\theta}_0$ has to be chosen, often the initial parameter vector is a random point in the search space. The parameters, themselves, define the nonlinearity of the Lur'e system $f_{\boldsymbol{\theta}}(x)$, in this context. Finally, evaluate the cost function and then iterate for $k = 1$ until the cost function has an acceptable value, that is specific to the problem:

1. Generate a new point $\boldsymbol{\theta}_k$ in the search space. The distance from the new point to the old one should follow a certain distribution, which in the original form of the algorithm was Gaussian, but the Cauchy distribution,

   $$g(|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}|) = \frac{T_k}{(|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}|^2 + T_k^2)^{\frac{D+1}{2}}}, \tag{2.11}$$

   (D being the dimension of $\boldsymbol{\theta}$) has shown to yield better results [Ingber, 1993].

2. Evaluate the cost function $I(\boldsymbol{\theta}_k)$.

3. Accept this new point with a probability

   $$h(I_k - I_{k-1}, T_k) = \frac{1}{1 + e^{\frac{I_k - I_{k-1}}{T_k}}}, \tag{2.12}$$

4. Decrease the temperature. For the Cauchy distribution the following annealing schedule finds a global minimum [Ingber, 1993]:

   $$T_k = \frac{T_0}{k}. \tag{2.13}$$

   When the final precision or temperature is met, the simulation should be stopped.

Cost Function The cost function gives a measure on how well the current parameters are suited to the problem. The most straightforward implementation is the one step prediction error, which, applied to the situation in Figure 2.8 (b), would be the difference between the output of the system at time $t + 1$ given the system at time t, $y'(t + 1|t)$, and the input at time $t + 1$, $y(t + 1)$,

$$\varepsilon = \frac{1}{r} \sum_{t=1}^{r} \big(y'(t + 1|t) - y(t + 1)\big)^2. \tag{2.14}$$

r is the number of samples that are considered for the calculation. Unfortunately, the prediction error is dependent on the initial state, which is usually put to zero and therefore would account for a systematic error in the cost function. In case the time series to identify are long enough, the transient can be discarded and the effect of the initial condition could be held neglectably small.

A cost function, not prone to the same error, can be constructed from the covariance matrix of the identified parameters of the linear model [De Feo, 2001]. In practise, however, such a cost function did not yield better results and was not used and will not be discussed.

In either case the cost function has to be completed with a second term, a second term, which penalises trivial or badly suited minima of the cost function. In fact, a very strong minimum is created by the trivial solution, where the static function is the identity function ($f(x) = x$). Therefore, a term has to be added, which penalises nonlinearity functions $f_{\boldsymbol{\theta}}(x)$, that add little or no total harmonic distortion to the time series $y(t)$. The method used to penalise this trivial solution was the mean squared error of the nonlinear function $f_{\boldsymbol{\theta}}(x)$, after the best straight line fit had been subtracted.

Furthermore, to avoid the degeneracy of the linear filter $G(z)$, some additional terms have to be added, namely

$$\varepsilon_{\text{lin},1} = \sum_i \frac{1}{|\mu_i|}, \tag{2.15}$$

which penalises filters, that contain poles in the origin (pure delays),

$$\varepsilon_{\text{lin},2} = \sum_{i \neq j} \frac{1}{|\mu_i - \nu_j|}, \tag{2.16}$$

which prohibits zero pole cancellation,

$$\varepsilon_{\text{lin},3} = \sum_i \frac{1}{|a_i|} \quad \text{and} \tag{2.17}$$

$$\varepsilon_{\text{lin},3} = \sum_i \frac{1}{|b_i|}, \tag{2.18}$$

which avoid that the order of the filter is reduced. a_i and b_i, are the filter coefficients and μ_i and ν_i the poles, respectively zeros.

The Mathematical Form of the Nonlinearity In general any basis for the Hilbert space L^2 can be used to parameterise the nonlinear function $f_{\boldsymbol{\theta}}(x)$ of the Lur'e system. Well known bases that can be used include Legendre and Tchebychev polynomials, piecewise linear functions, piecewise polynomials or splines.

Finally, only *splines in B-form* were considered for optimisation, which are, in turn, piecewise polynomials with additional continuity conditions that control the smoothness of the spline [de Boor, 1994]. In fact, they are constructed in such a way that their order indicates how many times they are continuously differentiable, a spline of order k is C^{k-2}. The fundamental element of a spline in B-form is the basis function the B-spline $B_{j,k}$. For a given knot sequence $t_1, \ldots, t_{n+k}$, there are n translated B-splines indexed by j and defined recursively by [de Boor, 2003]

$$B_{j,k}(x) = \frac{x - t_j}{t_{j+k-1} - t_j} B_{j,k-1}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}} B_{j+1,k-1}(x) \tag{2.19}$$

and

$$B_{j,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

The complete spline in B-form, finally, is the weighted sum

$$f(x) = \sum_{j=1}^{n} a_j B_{j,k}(x). \tag{2.21}$$

The free parameters, which form the vector in search space named θ above, are the knot sequence $t_1, \ldots, t_{n+k}$ and the weights a_j. Usually the first and the last knots, t_1 and $tn + k$, have a multiplicity higher than one. By adding knots at fixed values for t, *i.e.* t_1 and t_{end}, the smoothness can be increased without increasing the total number of parameters.

LINEAR IDENTIFICATION

The linear identification is a pure application task, the field has been deeply studied and results are available in the literature. For what is relevant here, the linear identification aims at finding the parameters a_i and b_i of the linear model,

$$y(t) = -\sum_{k=1}^{n_p} a_i\, y(t-k) + \sum_{l=0}^{n_z} b_i\, u(t-l), \tag{2.22}$$

where n_p and n_z give the number of poles and zeros, which has the transfer function,

$$G(z) = \frac{\sum_{l=0}^{n_z} b_i z^{n_p - l}}{z^{n_p} + \sum_{k=1}^{n_p} a_k z^{n_p - k}}. \tag{2.23}$$

In these equations any noise source has been omitted, in fact, a noise source can be located at different places, which causes the variety of models known in linear identification, as for example the ARMAX, Box-Jenkins or Output Error model structure [Ljung, 1999]. Ideally, all variability of the time series should be modelled by the nonlinear dynamics generating chaos, therefore no particularly coloured noise should be necessary to model the variability. This need is best observed by the Output Error model structure, which foresees a source of white noise $e(t)$ at the output of the model,

$$y(t) = -\sum_{k=1}^{n_p} a_i\, y(t-k) + \sum_{l=0}^{n_z} b_i\, u(t-l) + e(t). \tag{2.24}$$

The choice of the order depends heavily on the time series to be identified and its properties. Like mentioned above, a minimum is indicated by the the correlation dimension (Section 2.1.2), or also by the Hausdorff dimension [Ott, 1993]. Those dimensions indicate the minimal order necessary to construct the attractor in state space, in addition to that a higher order can be chosen to give more details to the attractor (see next chapter).

Finally, an algorithm has to be used to estimate the parameters. Because the final system operates in closed loop, the frequency based methods are not suitable; considering the Output Error model, the use of the prediction error based method is indicated [Ljung, 1999].

2.2.2 DRAWBACKS

Major drawbacks exist with this way of identifying chaotic systems. The first drawback is the need to run an optimisation algorithm on the parameters of the nonlinear, static part. Even though the different algorithms have been proved to converge, their use is, by their nature of not making use of any additional information than the cost function, very computational power consuming. Recent work has, however, addressed this problem in a different way than changing the model [Bagni, 2004].

The second major drawback is a little more subtle, but not less drastic. In fact, once the identification finished, the resulting system will in most cases not enter in qualitative resonance with the time series, like promised earlier, because the identified system usually exhibits subharmonic (Feigenbaum-like) chaos, while the phenomenon of qualitative resonance, on the other hand, requires homoclinic (Shil'nikov-like) chaos. In most cases the identified system can then be driven to homoclinic conditions [De Feo, 2001].

Another, third, drawback is the need for a time series with a large, almost white, spectrum. This is due to the linear identification involved [Ljung, 1999]. Time series with a rather narrow spectrum, which means with a small realisation variation $\delta(t)$, make the algorithm not converge, or yield a periodic system, not usable for classification. The next section presents this problem occurring with gait signals.

CHAPTER 3

DETERMINISTIC MODELLING OF GAIT SIGNALS

Brief — In this chapter the results of an attempt to model the diversity of gait signals with a Lur'e model are given. In the following these time series are analysed with standard tools and the change of model is motivated.

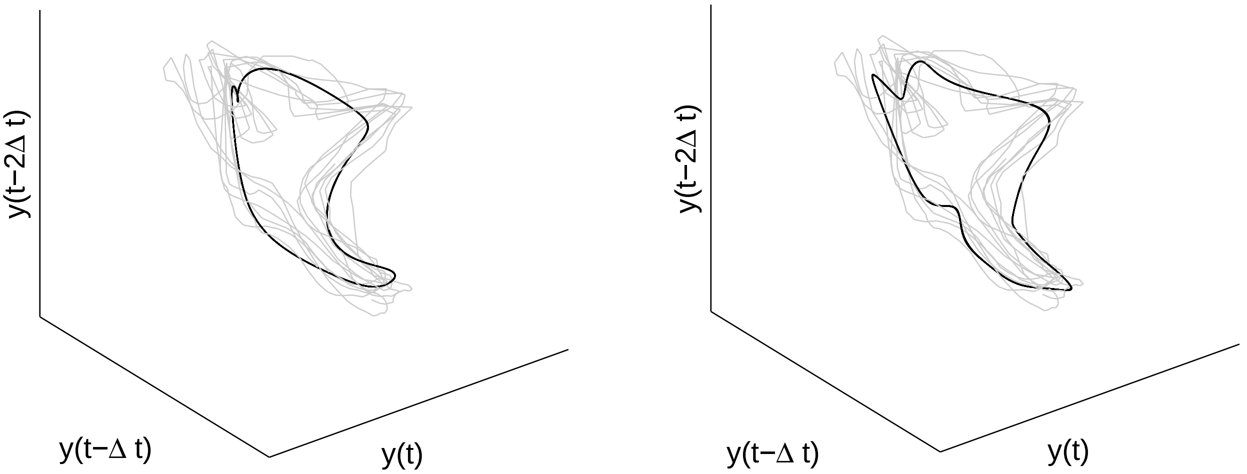
Personal Contribution — The work presented in this chapter is based on theory given in the chapter before. The analysis carried out and the conclusions are original, however.

The method of qualitative resonance seemed to be a good candidate as an algorithm for classifying gait signals, as they are intrinsically approximately periodic time series. The first step toward classification by means of qualitative resonance is the identification of the corresponding time series, which was done using the iterative approach discussed in Section 2.2.1.

Figure 3.1 shows the result of the identification of a sample gait signal. Represented are attractor reconstructions of the original signal in grey and the output of the identified system. The figure shows a good identification in general, the details of the original time series can be found in the identified time series as well. The higher the order of the linear system, the better is the resolution of the identified attractor. Due to better filtering, the trajectory can have sharper angles. On the other hand chaos does not emerge in any case during identification. This makes it impossible to use the identified systems for classification with qualitative resonance.

In contrast, Figure 3.2 shows the result of the same identification applied to the Colpitts time series. The identification terminated on a chaotic attractor, which shows, that the reason for the failure in the first case, lies in a structural difference between the time series. To proceed and propose an improvement to the existing algorithm, the structural difference between the time series has to be found. The first thing to do is to apply the tools presented in section 2.1.2.

The order of the linear system had been set between 4 and 7, orders that worked for the identification of Colpitts time series. In Section 2.1.2 the correlation dimension



(a) order 4

(b) order 7

FIGURE 3.1: Identification of gait signal. The black line shows the reconstructed attractor of the identified system, the order of the linear system is indicated below the subfigure. In grey the reconstruction of the gait signal is shown. The identified attractors are periodic.

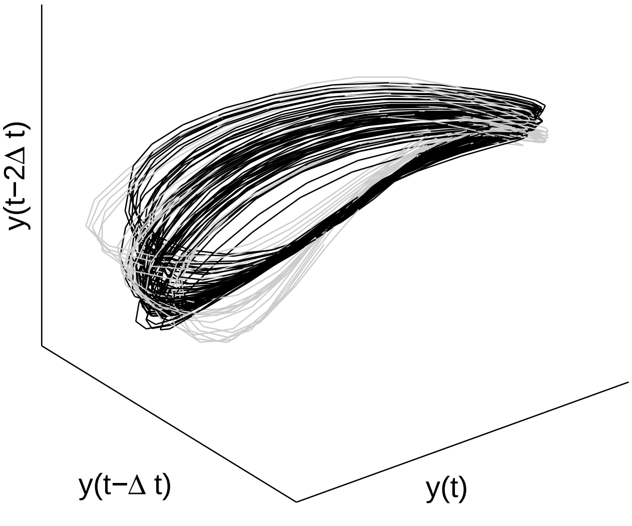


FIGURE 3.2: Identification of time series of the Colpitts oscillator with a model order 3. The same algorithm as above was used, the resulting attractor is chaotic.

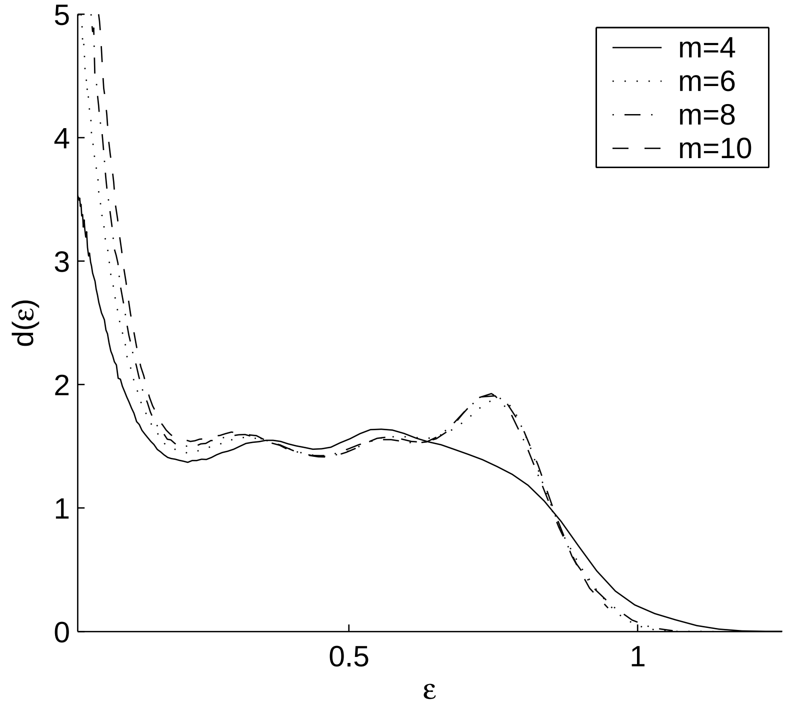


FIGURE 3.3: Estimation of the Correlation Dimension for gait signals. The flat part would indicate a fractal dimension of approximately 1.5, which is very low. The large slopes for small ε indicate the noise on the time series.

had been presented to estimate the minimal order to represent the time series, which constitutes also the minimal order of the linear system. Figure 3.3 shows a plot of $d(\varepsilon, N)$ (2.9), which is used to estimate the correlation dimension. N was the the total number of samples in the time series. The plot shows the same calculation for different embedding dimensions m.

For all embedding dimensions the curve is similar and indicates a fractal dimension around 1.6, but the curves have a large slope for small values of the radius ε. Noisy time series usually show this behaviour [Kantz and Schreiber, 1999]. Furthermore, a fractal dimension barely higher than 1.6 would indicate that the geometric object formed by the attractor is very similar to a line. Weak chaoticity or evolution on a small torus can not be excluded, but seem rather improbable for such small values. Consequently, the blurred appearance is due to noise and not to dynamic evolution. In practise this would mean that in the generation of gait signals nonlinear dynamics do not play a significant role and the variation does not bear information on the future evolution, but stems rather from exogenous noise. Noise sources include wind, unevenness of the ground or even the momentary perception of the walker. Accordingly, gait signals would be modelled more appropriately by conventional methods. Yet, whether or not nonlinear dynamics are present in the time series is better indicated by an estimation of the maximal Lyapunov Exponent than the correlation dimension. But even an estimation of the maximal Lyapunov Exponent does not reveal any nonlinear structures in the time series. Figure 3.4 shows the divergence sum of the time series, equation (2.10). Trajectories in a chaotic system diverge exponentially and in a logarithmic plot like in figure 3.4, straight

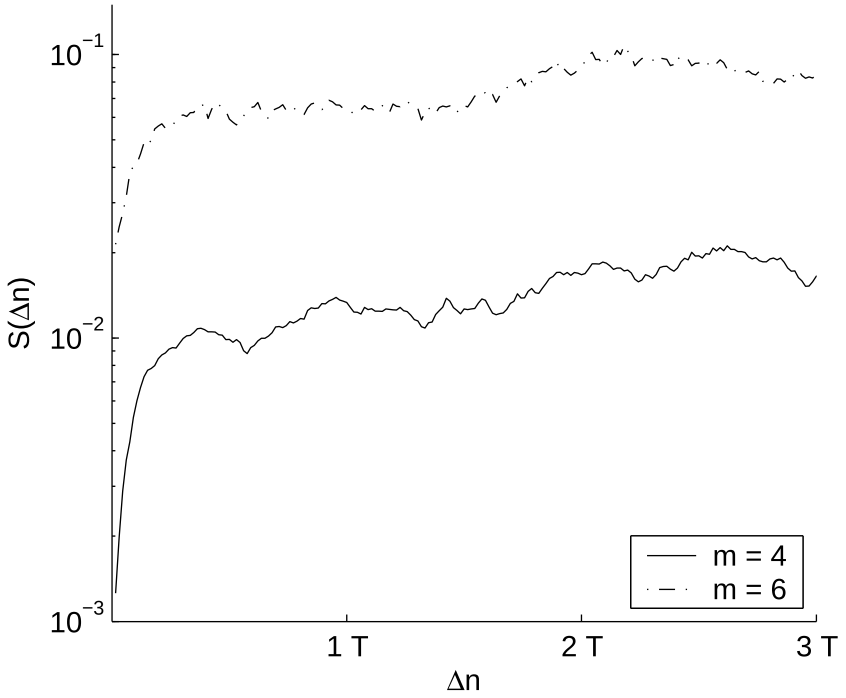


Figure 3.4: Estimation of the Maximal Lyapunov Exponent. $1T$ is one period. In no range of Δn the sum $S(\Delta n)$ shows robust linear increase.

lines should be visible. The plot, in contrary, shows a fast divergence for small differences in time ($\Delta n \ll 1T$), but divergence slows down afterwards. The sum shows in no interval a robust linear increase. Hence, gait signals have to be considered strongly periodic and are best modelled as a periodic oscillator with a exogenous noise source, or with a time variant model.

On the other hand, the way gait signals are best modelled does not necessarily correspond to the way they are best classified, it can only explain why identification fails. In the present case the strong periodicity of the signals involves an identified system with a periodic attractor. From this point of view the identification is successful, only it is not usable for the purpose it was done. A way out of this problem could be constituted:

1. by a rework of the linear identification;
2. a post processing of the linear system;
3. or a complete change of the model and identification algorithm.

The first possibility did not seem appropriate. The model and the identification algorithm were chosen because al theory behind was known. Having to rewrite the identification reduces substantially the motivation behind the choice. Post processing the linear system, like proposed as second possibility, is made by moving zeros or poles in the linear system. It was finally turned down in favour of the third possibility, as it did not promise for reliable progress.

The model finally used as replacement for the Lur'e model, and presented in the next chapter, was pointedly announced as solution for every nonlinear modelling purpose [Jaeger and Haas, 2004], but badly documented. This was a perfect opportunity to investigate the capabilities of the new model and continue the work of temporal pattern classification.

CHAPTER 4

BIOLOGICALLY INSPIRED MODELLING OF DIVERSITY

Brief — A biologically inspired approach is considered to model the diversity of approximately periodic time series. After the model a measure to determine its fitness is defined and subsequently the role of the model's parameters are analysed.

Personal Contribution — The model itself has been developed by others and cannot be counted as personal contribution.

On the other hand, the introduction of the measure (the entropy) in this context is original, and consequently the analysis carried out and the conclusions are original.

The model considered here is built on the theory presented by W. Maass [Maass et al., 2002] and H. Jaeger [Jaeger, 2001] who call it *Liquid State Machine* or *Echo State Network*, respectively. The models have been introduced in the Introduction (Section 1.1.2 on page 6), where also a brief literature review is given. Despite the particular naming of W. Maass and H. Jaeger, the model shares still all relevant properties with ordinary recurrent, artificial, neural networks, especially after identification. This is why in this chapter to this model is referred to as the biologically inspired model.

4.1 PRESENTATION OF THE MODEL

The general topology is depicted in Figure 4.1, which clearly shows a recurrent, artificial, neural network with on layer of leaky integrator neurons, indeed similar to the network proposed in [Jaeger, 2001].

The formalism presented here suggests a system operating in discrete time, in this case the internal state at time k is a vector denoted $\mathbf{x}_k$:

$$\mathbf{x}_k = \begin{bmatrix} x_{1,k} \\ x_{2,k} \\ \vdots \\ x_{n,k} \\ \vdots \\ x_{N,k} \end{bmatrix}. \tag{4.1}$$

Its dimension N is the number of nodes and therefore the dimension of the network, throughout this work a network dimension of 60 was chosen, but that choice will be discussed in Section 4.3.3. In general the value $x_{n,k}$ $(= x_n[k])$ denotes the state value at node number $n = [1 \dots N]$ and time $k = [1 \dots K]$, where K is the number of samples we have. The internal nodes behave like shown in Figure 4.2:

$$x_{n,k+1} = \alpha\, x_{n,k} + \beta_n \tanh(b + w_1\, s_1[k] + w_2\, s_2[k] + \ldots + w_N\, s_N[k]), \tag{4.2}$$

where $s_1[k], s_2[k], \ldots, s_N[k]$ are the time series of the inputs to the node. Like shown in Figure 4.1 the inputs can be of 3 possible types, they can be:

1. an external signal, fed directly into the network: dotted arrows,
2. a signal coming from another node (recurrent network): solid arrows,
3. the output signal, fed back: dashed arrows.

The connection weights $(w_1, w_2, \ldots, w_N)$ are the networks main parameters and are held in different matrices, one for each possible type. A connection weight of 0 means no connection. W_{in} $(N \times 1)$ holds the connection weight from the input node to the different internal nodes (dotted lines in Figure 4.1). W_{int} $(N \times N)$ is the weighted, directed adjacency matrix of the system and holds the connection weights from one node to another (solid lines), the last matrix, W_{back} $(N \times 1)$ holds the feedback weights (dashed lines). b_n is a bias that need not be present, it will be discussed in Section 4.3.1, the

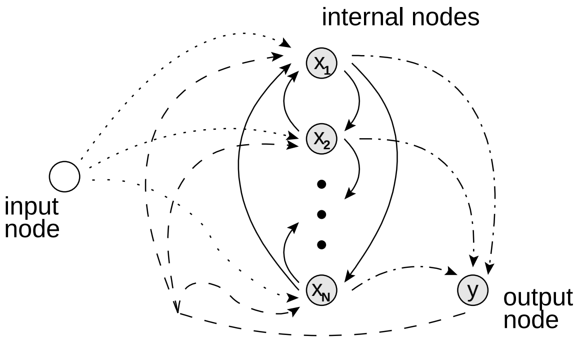


FIGURE 4.1: General topology of the biologically inspired model. Solid arrows indicate internal connections held in W_{int}, dashed arrows feedback connections held in W_{back}, dotted arrows input connections held in W_{in} and dash-dotted arrows output connections held in W_{out}.

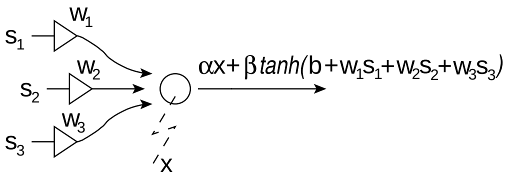


Figure 4.2: The nodes dynamics. The s_i are signals from other nodes, w_i the corresponding weights and x is the internal state of the node.

individual weights b_n form a $N \times 1$ matrix B. Finally, α and β are time constants, whose function will be explained in Section 4.1.2.

The output node differs from the internal nodes in the way that it has no memory, but it saturates as well. The weights with which it is connected to the internal nodes (dash-dotted lines) are held by the matrix W_{out} ($N \times 1$):

$$y_k = \tanh(w_{out,1}\, x_{1,k} + w_{out,2}\, x_{2,k} + \ldots + w_{out,N} x_{N,k}). \tag{4.3}$$

Combining everything the state equation for this circuit can be written:

$$\begin{aligned} \mathbf{x}_{k+1} &= \alpha \mathbf{x}_k + \beta \tanh(W_{int}\mathbf{x}_k + W_{in}u_k + W_{back}y_k + B), \qquad (4.4)\\ y_k &= \tanh(W_{out}\mathbf{x}_k), \qquad (4.5) \end{aligned}$$

where the hyperbolic tangent function is meant elementwise.

4.1.1 Design Parameters

The state equation has been given (4.4), and in this section the meaning of the different parameters, all present in the state equation, will be discussed further. In fact, any term in the state equation, except W_{out}, can be seen as a parameter with which the properties of the system can be controlled. This section lists the different parameters and shows how they are linked together.

The Connection Matrices

The connection matrices W_{int}, W_{back} and W_{in} are the network's most apparent parameters and they are truly design parameters, in contrast to W_{out}, which will be fixed by the identification process (see Section 4.1.4). W_{in} is not of major interest for system identification, the resulting system should exhibit self sustained oscillations, without need for external excitation.

The feedback connections are drawn from an uniform distribution between -1 and 1. This assures that distortion is different throughout the network. Too high values should be avoided because this leads to loss of information. This problem is the same as too high an amplitude of the time series to identify, and is discussed in the next section.

The properties of W_{int} and how they affect the network are discussed in great detail in Section 4.3.

The Saturation Function and the Time Series

The time series – which is to be identified – is of course not in the proper sense a design parameter of the system, nevertheless it has to be preprocessed in order to permit a good identification. If, for example, the amplitude of the time series is too high, too much information of the signal gets lost in the saturation of the nodes, the same applies if the feedback weights are too high, but not only the amplitude is important, the time series might carry the information only in a part of the amplitude band, in this case either the saturation function has to be adapted, or the time series further preprocessed [Baier, 2003]. This reflection shows the close relationship between time series and the saturation function.

Besides the amplitude and its distribution, another important property of the time series is its length. A very short time series leads to overfitting, this gives a lower bound for the length of the time series, a similar reason for an upper bound cannot be given, during simulations it showed however that too long time series do not lead to a successful identification. For instance, in the case of surrogate data where after a while the data becomes contradictory, it can be observed that as a consequence the identified system becomes periodic and no chaos emerges to justify diversity [Baier, 2003].

The saturation function, in general, only needs to be invertible, for reasons that will be shown in Section 4.1.4, the choice made throughout the analysis was however simply the $\tanh(\cdot)$ function, because it is widely used to model saturation effects.

4.1.2 The Memory of the Nodes

The memory of the nodes is controlled by α and β. β scales the strength of the network connections, whereas α controls the speed of the exponential decrease of a nodes state, in the absence of an input signal. $|\alpha| = 1$ describes nodes which show signs of their past for all time, they never forget, whereas $\alpha = 0$ describes nodes which show no signs of their past at all at the output. Values outside the interval $[0, 1)$ are possible choices, but make no sense in this context, indeed a value $\alpha \geq 1$ implies an unstable system, unstable systems are sensitive on initial conditions and do not have asymptotically unique behaviour. Systems with asymptotically unique behaviour have also been called having fading memory [Maass et al., 2002, Boyd and Chua, 1985] or having echo states [Jaeger, 2001].

If the network does not have an asymptotically unique behaviour, and the state at time $k > 0$ depends heavily on the state at time 0, the system cannot be brought into a known state just by exciting it long enough, because the transient will be infinitely long. For the identification process, which will be explained later, it is necessary that the initial state becomes insignificant after the transient.

If the network has asymptotically unique behaviour not only depends on the memory of the nodes, but also on the way the nodes are connected. A sufficient as well as a necessary condition for asymptotically unique behaviour of the network can be given. The necessary condition is

$$|\alpha + \beta\lambda_{max}| < 1, \tag{4.6}$$

where λ_{max} is the maximal eigenvalue of the adjacency matrix W_{int}. This is equivalent to say that the largest eigenvalue of the matrix A,

$$A = \alpha\mathbf{I} + \beta\mathbf{W}_{int}, \tag{4.7}$$

has to be smaller than 1. This is a necessary condition. This simplified version is slightly stronger than the condition given and proved in the appendix A.1. The sufficient condition, along with its proof, is given in appendix A.2.

4.1.3 Diversity Pool Interpretation

The general structure of the network has been presented, in the following an intuitive way of understanding how the circuit works can be given. With this explanation in mind the need for a measure for diversity – like presented in Section 4.2 – becomes evident.

The key word has already been given in the title of this section: *Diversity pool.* The one layer of nodes (or neurons) is seen as a black box containing modified versions of the input signals. The input signals can be some arbitrary input signal, fed in through the input node, or the output itself in a feedback situation. It is not important but can, nevertheless, be noted at this place that it is the feedback-only case in which we are particularly interested here.

Considering this black box with the input signal applied and the feedback signal loop closed, the output can be seen as N (the number of nodes) outputs, each time series lying on a node is an output, each of these outputs is a modified version of the input and bears certain properties of distortion and filtering. The fundamental idea of the approach is that any other time series related to the input time series can be constructed out of the time series lying on the nodes, the output function of the black box, which serves in this case as a base. Out of this base it would be possible to create any time series that belongs to the same class as the time series given as input, but which is not identical to it, like a vowel of different speakers. The nodes in the network would create the diversity associated with a certain class of time series and act therefore as a diversity pool, like mentioned in the beginning.

To adapt the network to a particular need, it would be necessary to adapt only the output weights: In the ideal case every node bears a distinctive temporal mark that could be present in the time series wanted at the output node. Then, to recompose the output the different nodes would be tapped proportional to the intensity of their corresponding distinctive mark as present in the reference signal (output).

In addition to the reflections above, [Maass et al., 2002] requires the separation property to hold that means two different input signals have to lead the system into different states. The larger the state space that can be explored by the system is, the better is the theoretical limit for the separation property. Then the question arises: “given our system, how have the parameters to be set in order that the whole state space can be explored?” This will be discussed in the rest of this report, just after the formal presentation of the identification process is given for completeness.

4.1.4 Identification Process

Once all the preliminary work, which includes generation of matrices W_{int} and W_{back} (in the case of an input signal present even W_{in}) and preprocessing of the time series to be identified y_{train}, has been done, the identification is made straightforward.

Consider a time series y_{train} should be identified. At first we note that if the system would have been identified correctly already, it would be possible to observe y_{train} on the output node, which would be fed back through the network. In the network, at the same time, the state sequence that would give y_{train} as an output would be observed. Now suppose the output node would be manually forced to y_{train} ($K \times 1$ vector), then call the resulting state sequence

$$\mathbf{X}_{\text{train}} = \begin{bmatrix} {\mathbf{x}_1}^T \\ {\mathbf{x}_2}^T \\ \vdots \\ {\mathbf{x}_K}^T \end{bmatrix}. \tag{4.8}$$

If y_{train} was a valid output sequence of the given system, the state equation (4.4) would need to be observed. This means:

$$\tanh(\mathbf{X}_{\text{train}} W_{out}) = y_{\text{train}}. \tag{4.9}$$

As the $\tanh(\cdot)$ function is invertible this can be written

$$\mathbf{X}_{\text{train}} W_{out} = \text{atanh}(y_{\text{train}}). \tag{4.10}$$

In the case the length of y_{train} is greater than the network dimension N, this is a linear, overdetermined system. It is solved in the least square sense by [Anton and Rorres, 1994]

$$W_{out} = ({\mathbf{X}_{\text{train}}}^T \mathbf{X}_{\text{train}})^{-1} \mathbf{X}_{\text{train}}\, \text{atanh}(y_{\text{train}}). \tag{4.11}$$

To calculate that expression numerically it is on the other hand not necessary to calculate the inverse explicitly.

The identification of the system is done in exactly that way. First the time series to identify is fed into the network through the feedback connections, while the output connections are disabled. The resulting state sequence is retained and used in the least squares equation above to calculate the output weights W_{out}.

It has to be said at this point, that even though the network had been constructed in a way that it shows asymptotically unique behaviour, nothing can be said on the system once the loop is closed.

4.2 Entropy as a Measure for Diversity

The equation to solve an overdetermined system in the least square sense has been given in (4.11), it is repeated here for convenience

$$W_{out} = ({\mathbf{X}_{\text{train}}}^T \mathbf{X}_{\text{train}})^{-1} \mathbf{X}_{\text{train}}\, \text{atanh}(y_{\text{train}}).$$

It follows, the problem is well posed if the matrix ${\mathbf{X}_{\text{train}}}^T \mathbf{X}_{\text{train}}$ is invertible. To measure the quality of the inversion the eigenvalues of that matrix are considered. When some of the eigenvalues are close to zero, the matrix is close to singular thus badly conditioned and the inversion is likely to give bad results. Therefore, a function that quantifies the condition of a matrix can be used to determine whether that same matrix can be used for identification in a biologically inspired network; such a function is the entropy, which is defined as [Tononi et al., 1998, Wackermann, 1999, Rieke et al., 1999]

$$H = -\frac{1}{\log_2(N)} \sum_{i=1}^{N} \lambda_i' \log_2(\lambda_i'), \quad \text{with } \lambda_i' = \frac{\lambda_i}{\text{trace}(C)}. \tag{4.12}$$

In this equation λ_i is the i^{th} eigenvalue of the cross-correlation matrix C of the time series, which, in occurrence, is the matrix to invert, $\mathbf{X}_{\text{train}}{}^T\mathbf{X}_{\text{train}}$, and is written

$$C = \begin{bmatrix} \bar{x}_1[1] & \bar{x}_1[2] & \cdots & \bar{x}_1[K] \\ \bar{x}_2[1] & \bar{x}_2[2] & \cdots & \bar{x}_2[K] \\ \vdots & \vdots & & \vdots \\ \bar{x}_N[1] & \bar{x}_N[2] & \cdots & \bar{x}_N[K] \end{bmatrix} \begin{bmatrix} \bar{x}_1[1] & \bar{x}_2[1] & \cdots & \bar{x}_N[1] \\ \bar{x}_1[2] & \bar{x}_2[2] & \cdots & \bar{x}_N[2] \\ \vdots & \vdots & & \vdots \\ \bar{x}_1[K] & \bar{x}_2[K] & \cdots & \bar{x}_N[K] \end{bmatrix}. \qquad (4.13)$$

$\left[\bar{x}_n[1],\ \bar{x}_n[2], \ldots, \bar{x}_n[K]\right]$ are line matrices giving the time series on the nodes, which have been detrended and normalised in energy.

Indeed, in the worst case, all nodes are identical and, hence, all columns in the matrix $\mathbf{X}_{\text{train}}$ equal. The cross-correlation matrix of that training matrix would have all entries equal to 1, the energy of the time series on the nodes. The corresponding eigenvalues would be all but one zero and the entropy would amount to zero. On the contrary, if the matrix $\mathbf{X}_{\text{train}}$ would hold N uncorrelated columns, the cross-correlation matrix would be equal to the identity matrix and it follows a maximal entropy. Consequently, the parameters of the network should be set to values, such that a high entropy of the cross-correlation matrix results.

Other measures to find a good identification could be considered. Notably, a maximal condition number of the matrix appears to paraphrase more precisely the need for a good inversion of the cross-correlation matrix. A reason which can be held against the condition number is its severity against networks that have two strongly correlated nodes, but that otherwise would be connected very well. Such networks could still perform very well, or at least better than networks that have all their nodes middlingly correlated. The entropy would favour the first one, the condition number the second. In general, the entropy manages to express formally very well the metaphor of the "diversity pool".

It could be argued, that no measure at all is needed and successive applications of the network to different problems will show what it is capable to do. This is, however, not a very scientific approach, as it could hide the fact that the network is trained to specific applications, thereby contradicting the hypothesis that the internal connections do not need to be trained. Furthermore, under the assumption that an "universal" every fitting network exists, a measure may help in finding it; while seeking for the best performances from case to case would leave it unveiled.

4.3 Optimising the System's Entropy

Since the entropy has been formally defined, it can be used to search parameter ranges which yield particularly high entropy and are therefore convenient choices for network generation. In the following the dependence of the entropy on several parameters is analysed. To this end a large number of networks was generated, with parameters as specified in the corresponding subsection. One or two parameters were swept over a certain range, the networks were excited and the entropy from the state matrix $\mathbf{X}$ were calculated.

To excite the network three types of time series were used. The first is a time series from the Colpitts oscillator operating in its chaotic range [Maggio et al., 1999]. This choice was made to assure that a 3-dimensional system should be enough to produce

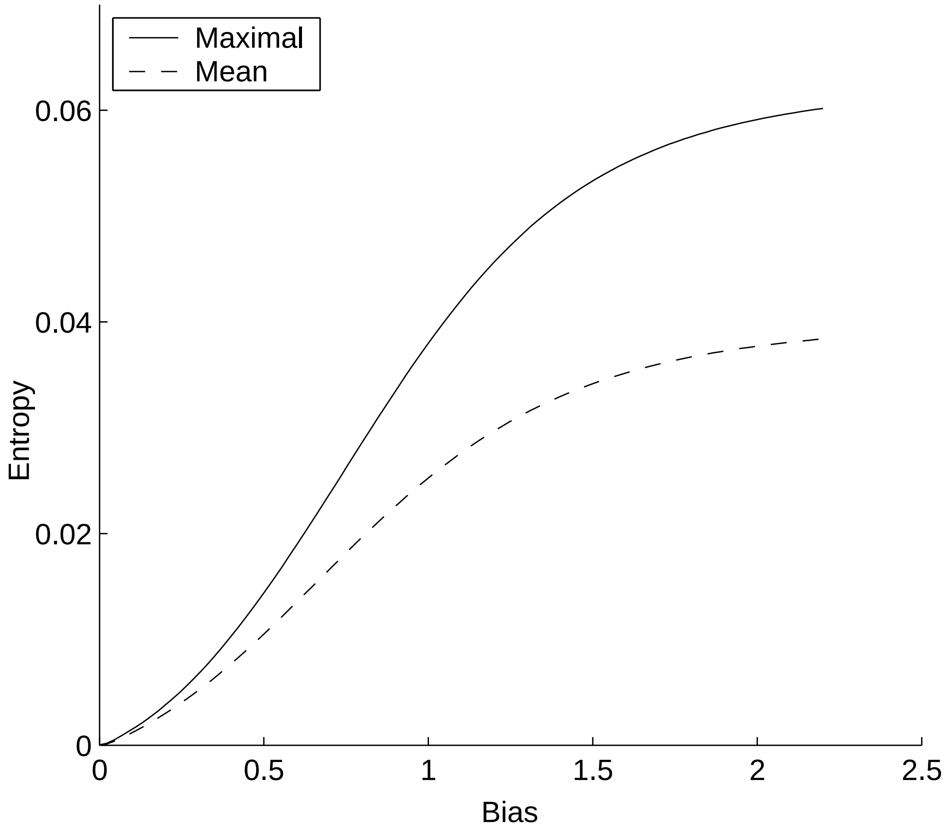


FIGURE 4.3: The effect of nonuniform bias on the entropy in a network without connections.

the time series. These time series had already been used to do identification with that network [Baier, 2003]. In addition to the Colpitts time series, its Surrogate was used as a second type of signal. It is made of numbers of a random process, which are rearranged to have the same spectral properties like the Colpitts, but with arbitrary phase noise. A signal crafted in this way can be used to test the system's generalisation ability. The third type is white Gaussian noise with a standard deviation $\sigma_x = 0.5$. It was brought to an interval $(-1, 1)$ by applying the $\tanh(\cdot)$ function on it. This type of signal has been mainly used in Section 4.3.4 for avoiding problems given by Colpitts time series in a linear identification framework.

4.3.1 NONUNIFORM BIAS

Jaeger [2001] proposes to put a nonuniform bias on the nodes. The bias is controlled by the parameter B in the state equation (4.4) and it is nonuniform, because for different nodes there are different weights. In particular 50% of the nodes have no bias, 25% a positive and 25% a negative, positive and negative bias have all the same value, as suggested by [Jaeger, 2001].

In the trivial case, where there are no connections between the nodes ($W_{int} \equiv 0$), the bias alone would make entropy apparent. Before proceeding, it is important to verify that the amount of entropy generated through simple distortion does not make up for

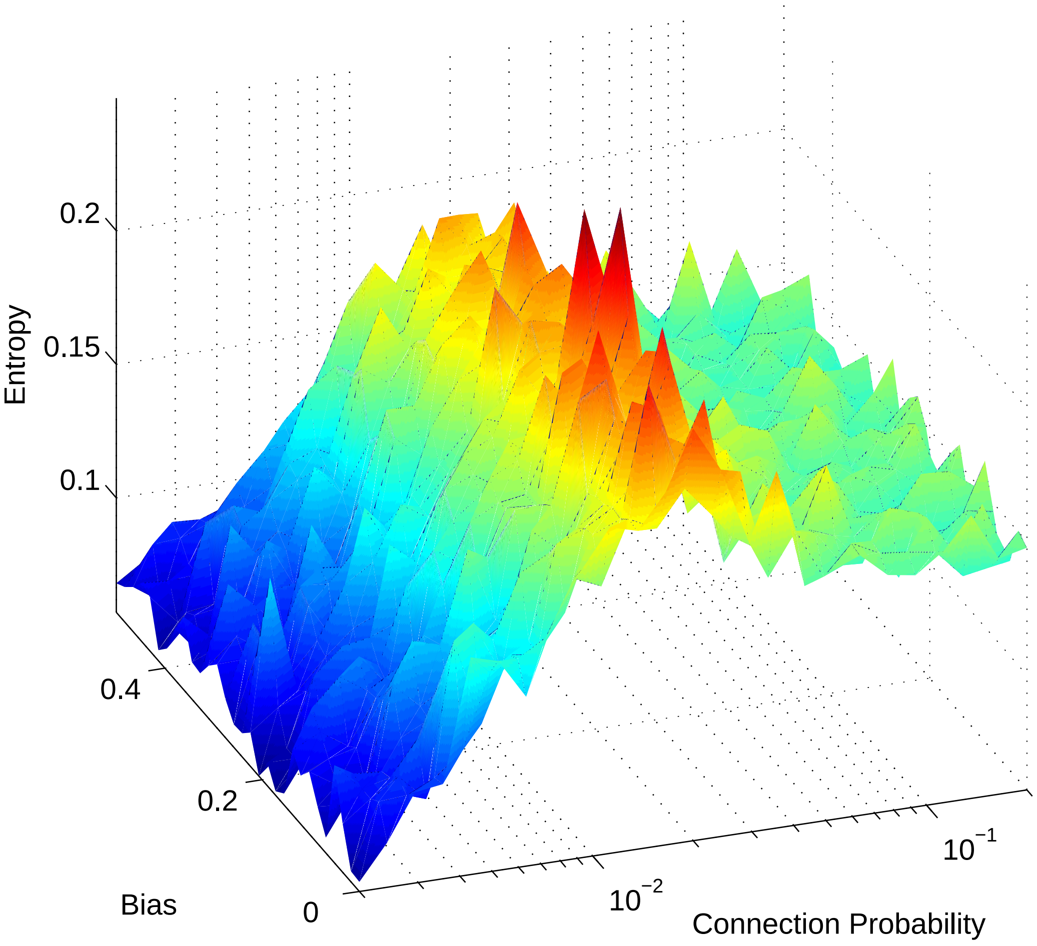


Figure 4.4: The entropy in a network in function of the nonuniform bias and the connection probability. From the plot can be seen that the nonuniform bias has little influence on the entropy.

the greater part of entropy that will be measured for the different setups later on.

Figure 4.3 shows the entropy in function of the applied bias, when the nodes are not connected with each other and the surrogate data of the Colpitts is applied to the nodes. The abscissa gives the value of the bias b. The vector B in the state equation (4.4) has 25% of its values at $+b$, 25% at $-b$ and 50% at zero. Depending on the constellation of feedback weights and bias, the entropy varies. On the plot the mean and the maximal value of 2'000 realisations of the corresponding matrices are reported. The entropy is monotonically increasing and reaches for high values of bias a maximum, which is negligible in comparison to the values obtained in a fully connected network.

The above assumption is confirmed by Figure 4.4, for which successively connections had been added to the network and different levels of bias applied, the excitation was made with the surrogate data. The surface visible shows the maximal entropy that could be obtained for 50 randomly drawn networks. The role of the number of connections, or of the connection probability is subject of Section 4.3.4 and is not discussed here. It

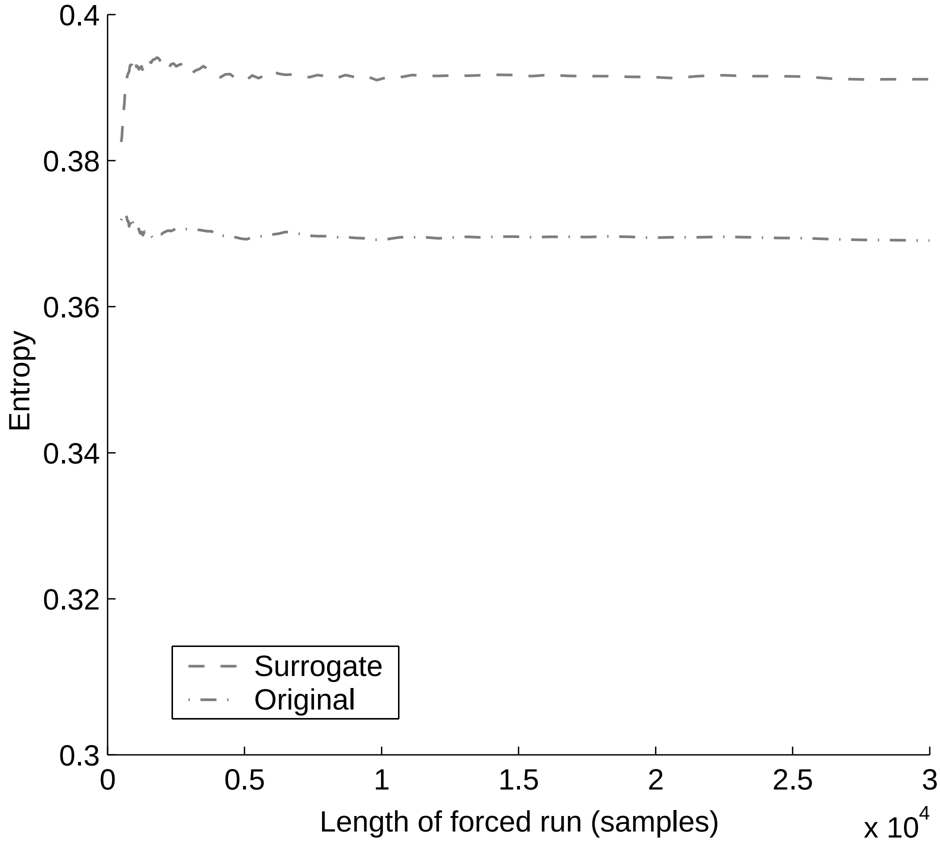


Figure 4.5: The entropy in function of the length of the training sequence. Training length has little or no effect on the entropy, except for absurdly small values.

can be seen on the figure that in a connected network the bias has a major influence on the entropy of the network. It seems that a bias in the order of 0.1 has a positive effect on the entropy, while a higher bias has a negative effect on the entropy; the bias is a parameter that has to be used with caution.

4.3.2 Training Length

In practise, the training length has shown to have a major influence on the identification process. Figure 4.5 shows how the entropy of the state matrix $\mathbf{X}$ changes in function of the training length. Both cases, the original and the surrogate, are reported in the graph. Clearly the surrogate data yields a higher entropy, this is not surprising, because of the arbitrary phase noise, which reduces the autocorrelation of the surrogate.

The overall variation of the entropy is very low, less than 5%, this is not enough to explain eventual difficulties during identification with particular lengths of the training sequence. On the other hand, the rapidly increasing entropy for small lengths of the surrogate time series, indicates that for small training lengths the memory of the network could take up more information than there actually was in the time series. This leads to overfitting and the training length should be at least as long as the length which yields maximal entropy.

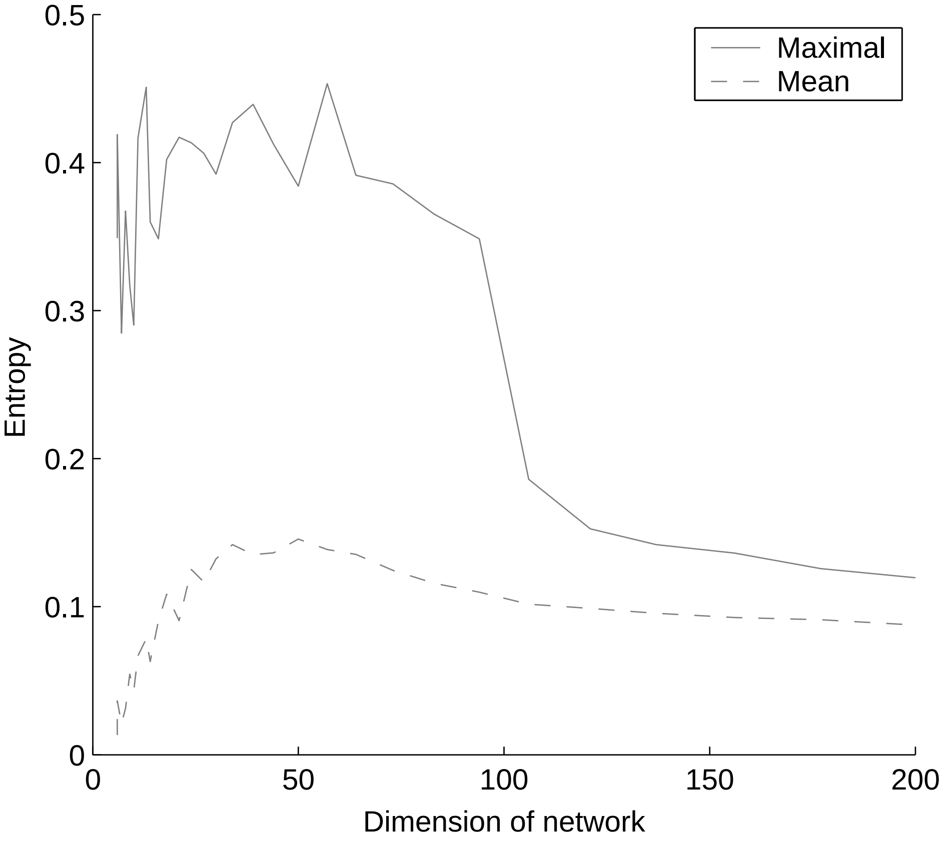


Figure 4.6: The entropy in function of the dimension of the network. Around dimension 60 mean and maximal entropy show a maximum.

4.3.3 Dimension

Biologically inspired models generally are sparsely connected, but have a higher dimension than their dense counterparts. In the case of the Colpitts oscillator, it is known that three dimensions would be enough to produce the time series. In this section dimension ranges with high entropy are searched. For this all the other parameters remained fixed. In particular the connection probability was 0.0135 and the maximal eigenvalue was 0.9, for each dimension 2'000 adjacency matrices were drawn. The results are shown in Figure 4.6, reported are the normalised maximal and mean values of the 2'000 entropy values obtained.

Surprisingly the entropy is not monotonically increasing, but reaches a maximum and decreases afterwards. The maximum is situated around dimension 60. One could conclude that dimension 60 is the ideal size for a the network topology considered here, however it should not be forgot that during the simulations the other parameters were fixed and that their tuning depends on the network dimension. In particular this might be the case for the overall connection probability, which was in this case 0.0135.

4.3.4 CONNECTION PROBABILITY AND MEMORY

In the previous section the comparison between high-dimensional sparse systems and low-dimensional dense systems was taken up and the influence of the dimension on the entropy was analysed. In this section the influence of the sparsity will be investigated, in the same time the influence of the memory inherent to the network is analysed. It has been mentioned in Section 4.1.2 that the memory comes to one part from the memory in the nodes and from another part from the connections held in W_{int}. As a measure for the memory simply the maximal eigenvalue of the matrix $A = \alpha\mathbf{I} + \beta\mathbf{W}_{int}$ (4.7) can be taken. To adjust the eigenvalue the adjacency matrix W_{int} is scaled.

Figure 4.7 shows the maximal value of the entropy that was obtained for 2'000 realisations of the network for each combination of maximal eigenvalue and connection probability. The network was excited with white Gaussian noise and both variants of the Colpitts time series, but represented on the figure is only the case of excitation with white noise, as the shape of the surface plot was in all three cases the same. The network dimension was 60 and no external bias was applied. The figure shows clearly a strong dependence of the network on the two parameters considered. Most interestingly the network shows different behaviour for small or high values of the connection probability.

Consider first a strongly connected network (high connection probability), which is the case on the right side of the plot. There the network shows a strong dependence on the maximal eigenvalue: Entropy grows linearly with the the maximal eigenvalue. It seems as if network behaves predominantly linear and that linear theory could be used to describe the circuit.

On the contrary for small connection probabilities, on the left side of the plot, the network does not show dependence on the spectral radius, but a very strong dependence on the connection probability. Entropy grows with increasing number of connections, right until the critical value, that separates the regions with qualitatively different behaviour. The distortion present on the different nodes interacts with the other nodes to generate diversity.

In brief, the network appears to behave highly linear on one side, and highly nonlinear on the other side. If this were true the time series on the nodes should in the linear case easily be identified by a linear black box model, in the other case, the same linear black box model would have great difficulties to identify the time series on the nodes. This will be tested in the next subsection.

1-STEP PREDICTION ERROR

To verify if the above assumptions were true a standard linear black box model with 5 poles and 4 zeros was taken and used to identify the time series on each node, when the network was excited with white Gaussian noise. Then on a validation time series the 1-step prediction error was calculated [Ljung, 1999]. The results confirm the assumptions, Figure 4.8 shows the 1-step prediction error in function of the maximal eigenvalue and the connection probability. Indeed the error is rather flat and low for high connection probability, showing a good identification of the different time series. It is interesting to note that for a large maximal eigenvalue and high connection probability the entropy (Figure 4.7) is high, but the 1-step prediction error is low anyway, indicating that the mechanisms generating the entropy are not the same in both cases. Changing the orders of the linear model changes naturally the 1-step prediction error, the general form stays

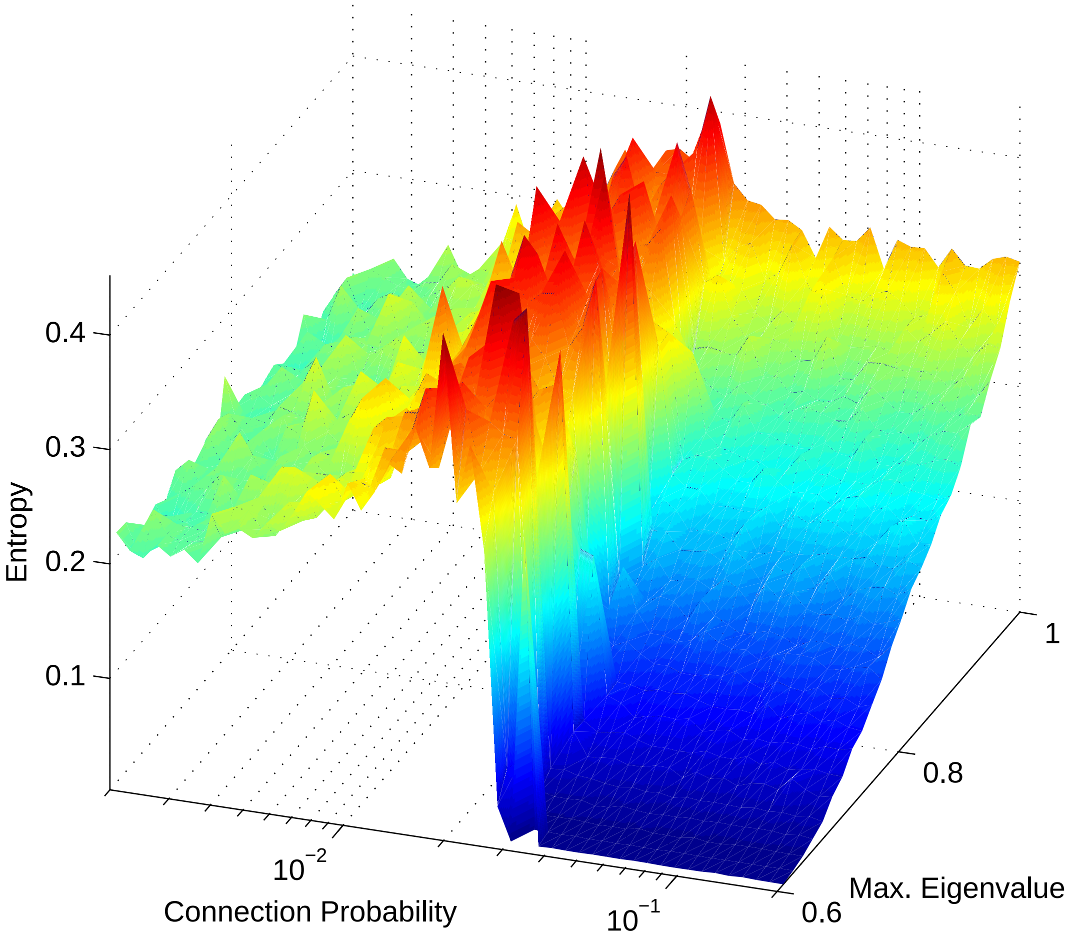


Figure 4.7: The entropy in function of the connection probability and the maximal eigenvalue of the connection matrix. On the right side plot indicates a linear behaviour of the network, the entropy increases almost linearly with the maximal eigenvalue. On the left side the network does not show a significant dependence on the maximal eigenvalue, but on the connection probability.

however the same.

4.3.5 Matrices with Components

The critical reader might already have calculated $60 \cdot 2 \cdot 10^{-2} = 1.2$ which is the expectation for the number of connections per node in the middle of the plots in Figures 4.7 and 4.8. This number close to one. This means that not every node has an outgoing (or incoming connection), therefore components are built up in the adjacency matrix. In fact, [Maass et al., 2002] proposes – in order to add "computational power" to an existing network – to add a new network in parallel to the existing one. The adjacency matrix of the newly created network is then built up of components. An interesting question is if the high entropy seen for low connection probability in Figure 4.7 is due to the particular form of the adjacency matrix. To verify this the maximal entropy that could be obtained

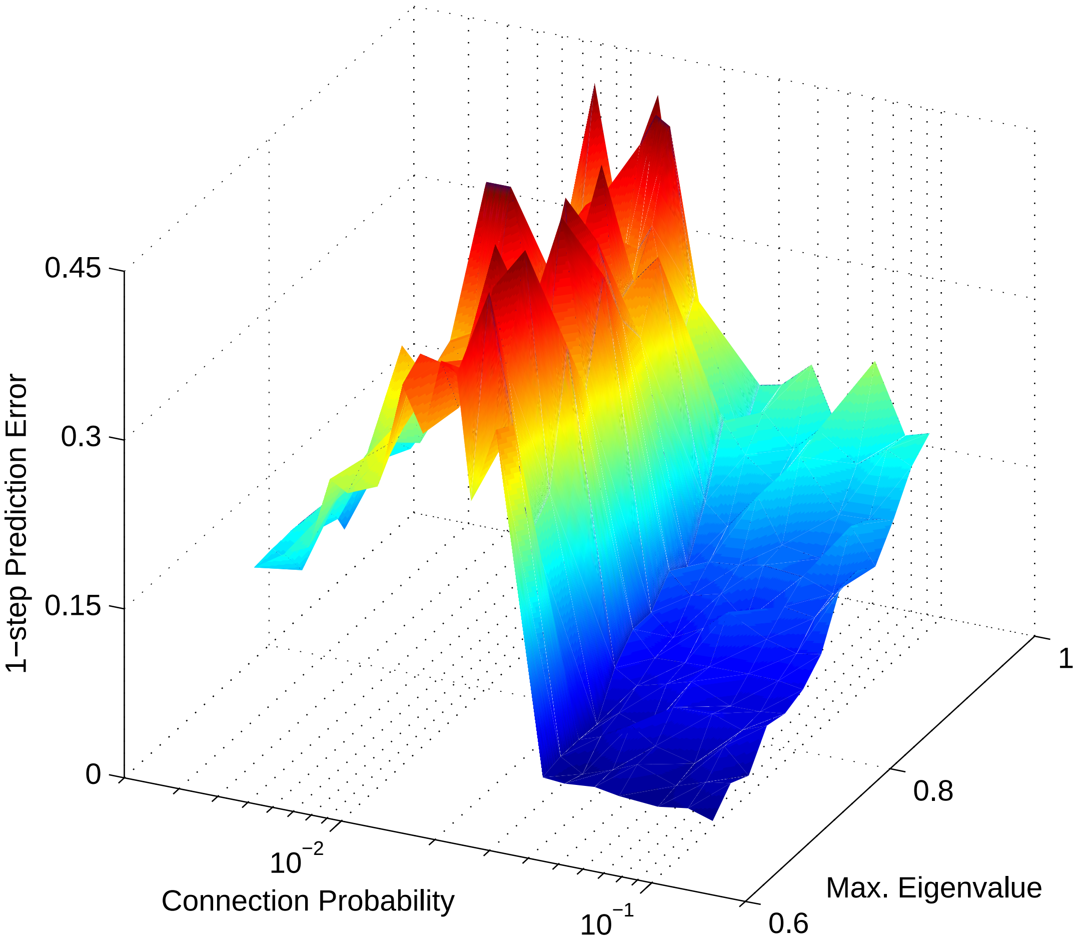
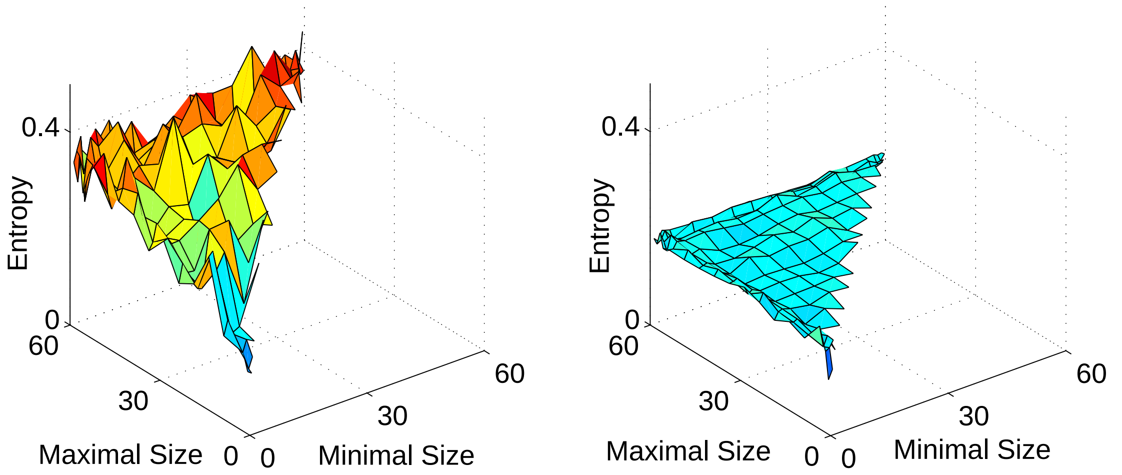


FIGURE 4.8: The mean 1-step prediction error when modelling the time series of the nodes with a standard linear model. The linear model has more difficulties to describe a network with few connections, than a densely connected network.

with a particular form of the adjacency matrix was calculated.

Both plots in Figure 4.9 show the entropy in function of the maximal and minimal component size in the network. To construct the networks components with sizes uniformly distributed between the maximal and the minimal sizes were chosen. Then within the components, connections were randomly assigned with a probability, such that the overall connection probability equals $9{\cdot}10^{-2}$ in the case of Figure 4.9 (a) and $2{\cdot}10^{-1}$ in the case of (b). Within the components, it was assured that each node had at least one incoming connection. If this was not the case already after randomly assigning the connections, supplemental connections were assigned.

In both cases, the entropy does not show a strong dependence on neither the maximal or the minimal component size, only for maximal sizes approaching 1 the entropy drops, which is in perfect correspondence with the earlier results. A maximal component size of 1 means no internal connections at all, like for the simulations analysing the role of the input bias (Section 4.3.1). In the case of such a flat entropy it would be possible to add components to an existing network and the resulting network would not be less



(a) Low connection probability

(b) High connection probability

FIGURE 4.9: Very low connection probability can lead to components, which are not connected to each-other, in the network. These plots show the entropy in function of the minimal and the maximal size of the components in the network. In general, a network splittered into components has less entropy than a completely connected network.

performant than one built up directly with the right size and the connection graph irreducible, *i.e.* when the adjacency matrix has only 1 component.

By comparing the two plots it is clearly visible that it is not the building up of components, which makes the difference in entropy for high and low connection probabilities, but rather the connection probability itself. For low connection probability (Figure 4.9 (a)) the entropy is considerably higher for all possible component sizes than for high connection probability (Figure 4.9 (b)).

A possible explanation for this could be that the less connections there are, the larger their weights will be, for a given maximal eigenvalue. The larger the weights are, the more distortion the time series will undergo. Another explanation would be that the sparsity itself really modifies the behaviour of the network. That in a sparse network due to the missing interconnections each node develops its own dynamic, which would not be the case in a more dense network, where the nodes somehow stay synchronised.

4.4 PERFORMANCE

So far the model was analysed operating in open loop configuration, identification was not done. This analysis permitted insight in how the network has to be designed to deliver optimal performance. The knowledge gained thereby, is now to be applied to perform the actual identification.

When finally doing the identification the hypothesis that the identification is made only by adapting the output weights has to be verified [Jaeger and Haas, 2004]. This means it should be checked, whether the diversity pool interpretation (*cf.* Section 4.1.3) makes sense. Continuous generation of random networks and always picking the best network for a particular task, could hide the fact that the internal connections have to be trained to the task. The only difference to conventional recurrent, artificial, neural networks would be the training algorithm, human supervision of random generation of networks, a rather inefficient algorithm.

To check the validity of the diversity pool interpretation, the network used in this section was the same for all tasks. It has been generated with parameters that proved to yield high entropy and it has been verified that the actual realisation was one, which yields high entropy.

The network had a dimension of 60. The matrix holding the internal connections W_{int} had a total number of 48 connections, resulting in a connection probability of $1.33 \cdot 10^{-2}$, which is, in fact, where entropy is maximal (*cf.* Figure 4.7). A visualisation at least to some extent of this important matrix is shown in Figure 4.10. In this plot every connection is symbolised by a dot. The x-axis gives the number of the node from which the connection is outbound and the y-axis to which it is inbound. The bias was +0.14 on a fourth of the nodes an −0.14 on another fourth, half of the nodes had no bias, this configuration, indeed, yielded highest entropy in Figure 4.4. This was the experimental setup throughout this section, without exception.

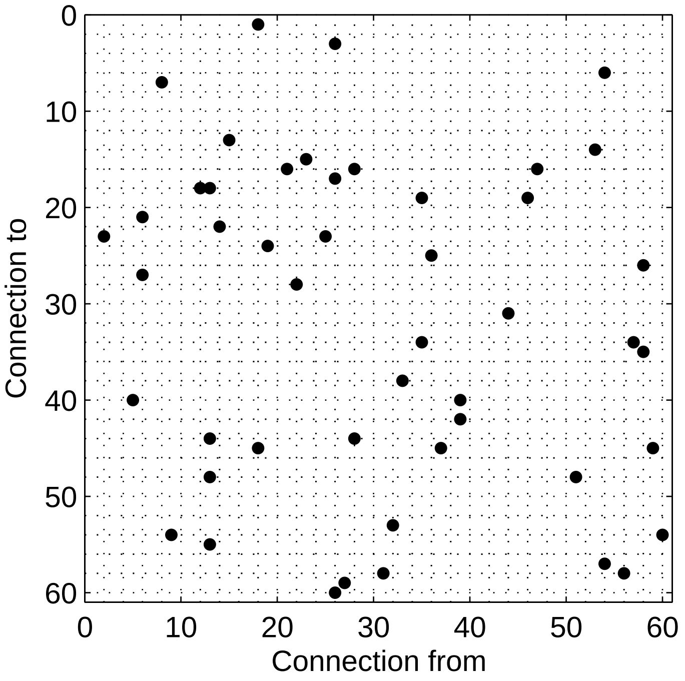


Figure 4.10: The internal connections of the network used for all identifications in Section 4.4. The dots on the grid indicate a connection.

4.4.1 Generalisation Ability

An important capability of an identification algorithm is what was named generalisation ability in Section 2.1.2; the ability to give a description of an observed pattern, even though this description is not strictly how the pattern was produced. In every day life generalisation ability is the ability to recognise a hand drawn circle as a circle, even though it is far from being round. Applied to temporal pattern recognition with synchronisation, generalisation ability focuses on how an identification algorithm reacts to data, which does not necessarily originate from a dynamical system, like surrogate data, or data from a stochastic process in general. However, the only way a dynamical system can produce an approximately periodic time series is by showing chaotic behaviour. Thus, during the identification chaos has to appear to model the realisation variation.

It has been seen, in Chapter 3, that during iterative identification chaos had difficulties to emerge, thus the iterative identification algorithm shows little generalisation ability. This was one of the reasons why this biologically inspired approach was chosen, and therefore a series of tests were performed on it to get an overview on its capabilities, *i.e.* exploring the conditions when it is capable to build a deterministic model even for non-deterministic data. More precisely, different surrogate data has been used for identification and the resulting attractors have been compared. The results will be shown below, first for the surrogate data of the Colpitts oscillator, then for synthetic time series based on the stereotype of the Colpitts time series and additional coloured noise.

SURROGATE DATA FROM COLPITTS OSCILLATOR

The identification of different data from dynamical systems had already been done with the kind of network considered [Jaeger, 2001], and it can be supposed it is possible to identify Colpitts time series as well with that kind of network and indeed it is possible to identify them rather well. The question remains whether it is possible to identify these time series because the underlying dynamics are low dimensional and well defined, or whether any data with similar spectral properties can be identified.

To verify, surrogate data of the Colpitts was created like formulated in Section 2.1.1. The result of the identification is shown in Figures 4.11 and 4.12. The first plot shows the attractor reconstruction of the identified system in black, along with some samples of the surrogate data in grey. The second plot shows the spectra of the identified system in black and the surrogate data in grey. Both plots give essentially the same information, and show a very good match, the higher harmonics have been identified well, and the subharmonic seems to be smeared out. The resulting attractor is a little bit flatter than the original. Whether in general the identification is well enough for classification will be seen below (4.4.2), when gait signals should be classified, but before the algorithm is used on clearly synthetic data, which is not inspired by nature.

SYNTHETIC TIME SERIES

The surrogate time series of the Colpitts oscillator have been identified well, still it is not sure whether there exists a omnipotent network that is identified solely by adapting the output weights. Furthermore, the spectral properties of the time series of the Colpitts oscillator are those of a simple chaotic system, it should be fairly easy for an overdimensioned network to identify them well.

To address these problems, a set of time series were used for identification, which in nature are very rarely observed. These were, more precisely, synthetic time series, which were constructed with the mean stereotype of the Colpitts time series and additional

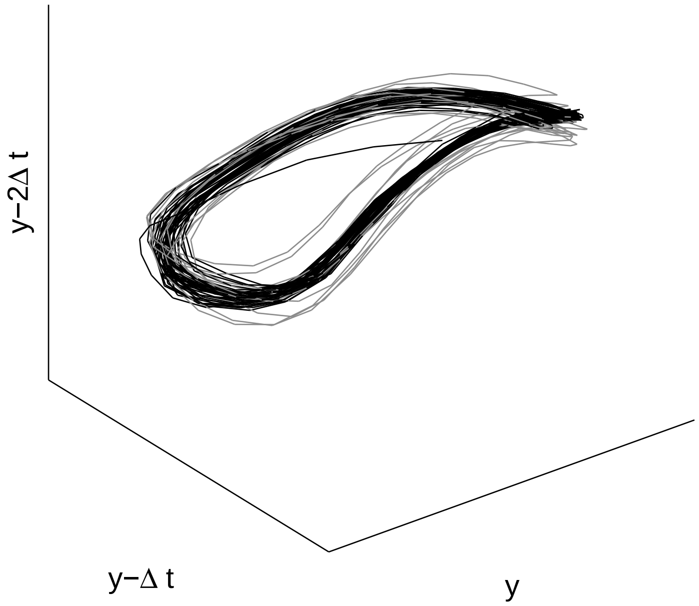


FIGURE 4.11: Identification of Colpitts surrogate data, attractor reconstruction.

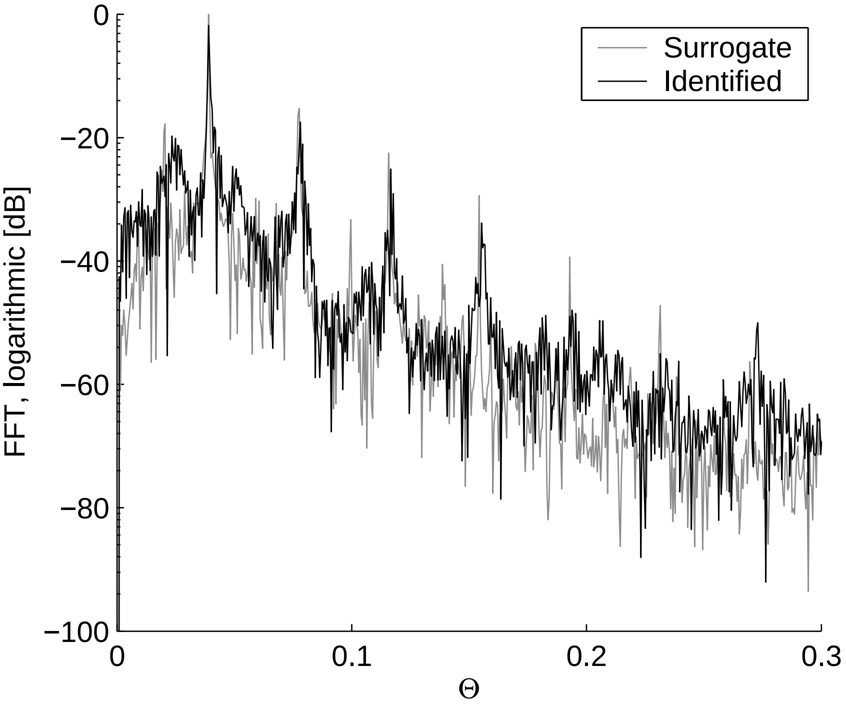


FIGURE 4.12: Identification of Colpitts surrogate data, FFT.

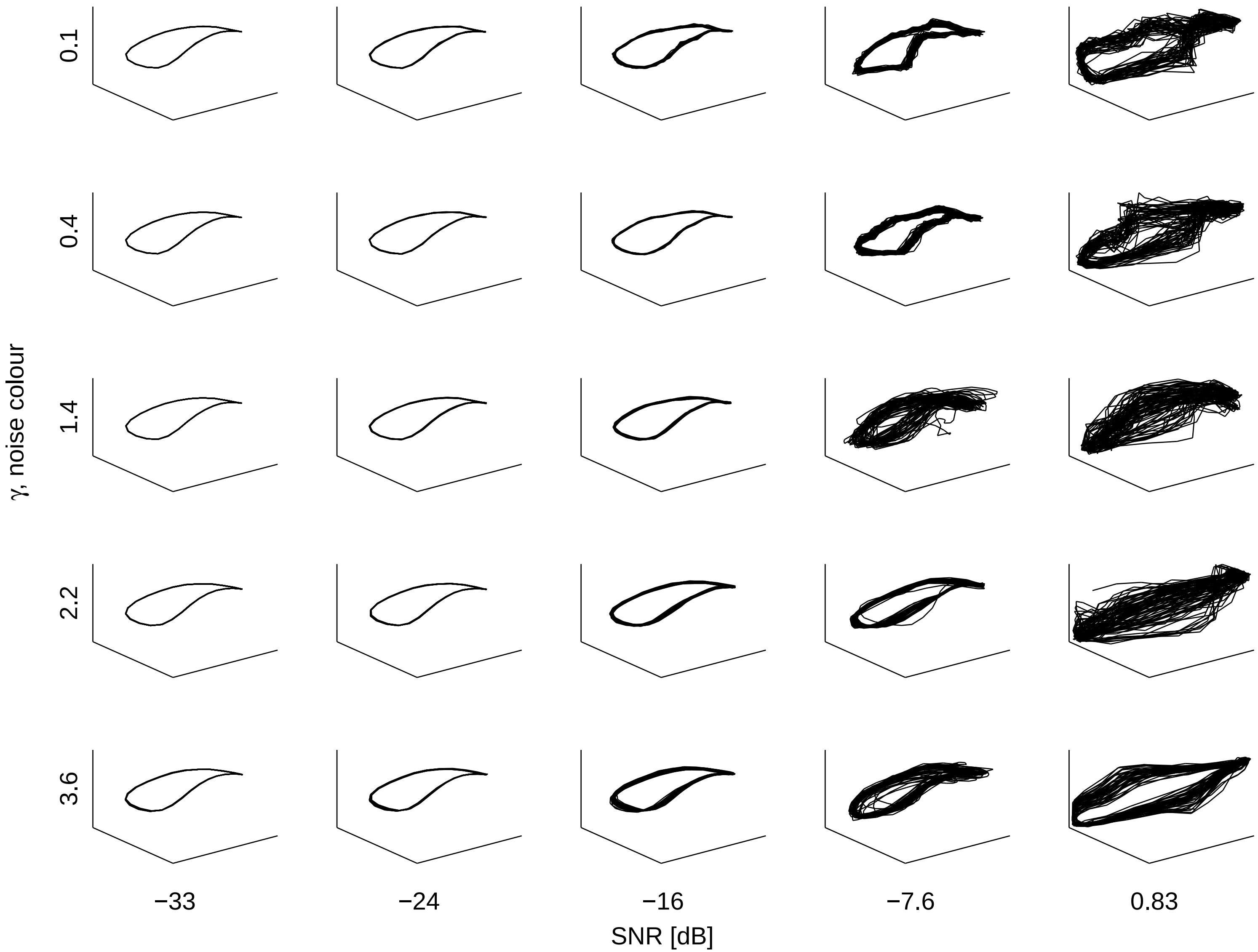


FIGURE 4.13: Identification of synthetic time series, reconstruction.

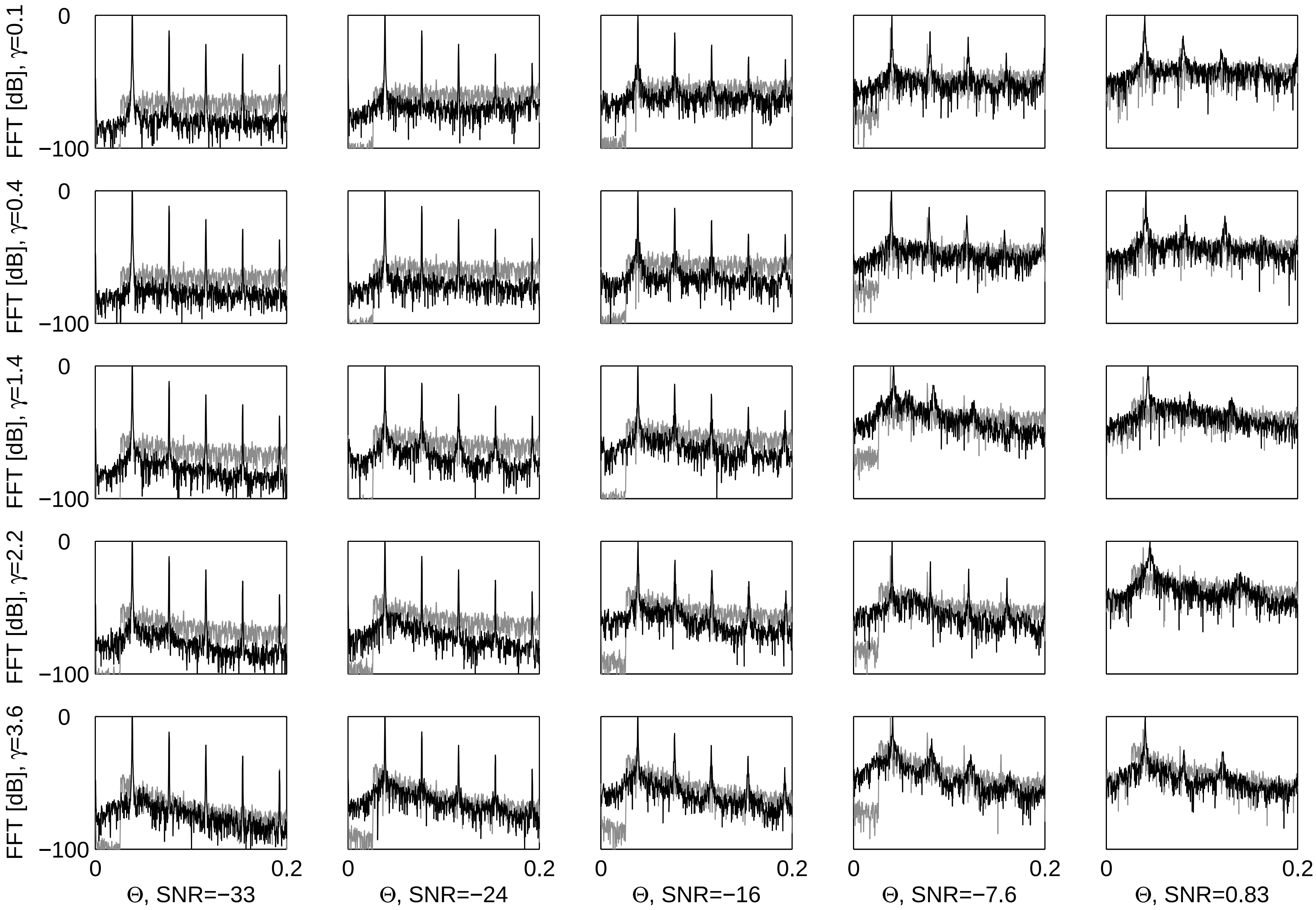


Figure 4.14: Identification of synthetic time series, FFT.

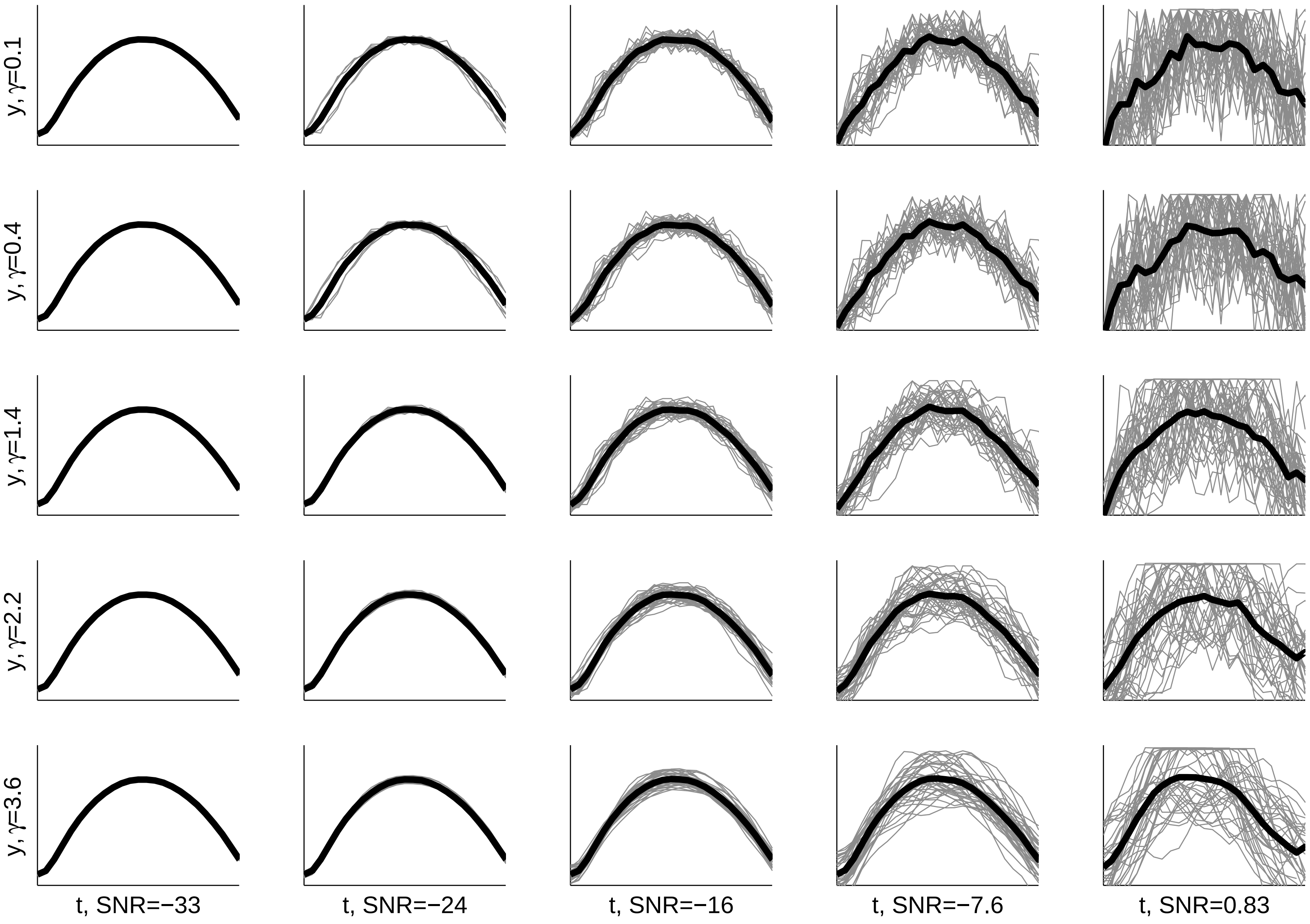


FIGURE 4.15: Identification of synthetic time series, stroboscopic plot of the training sequences.

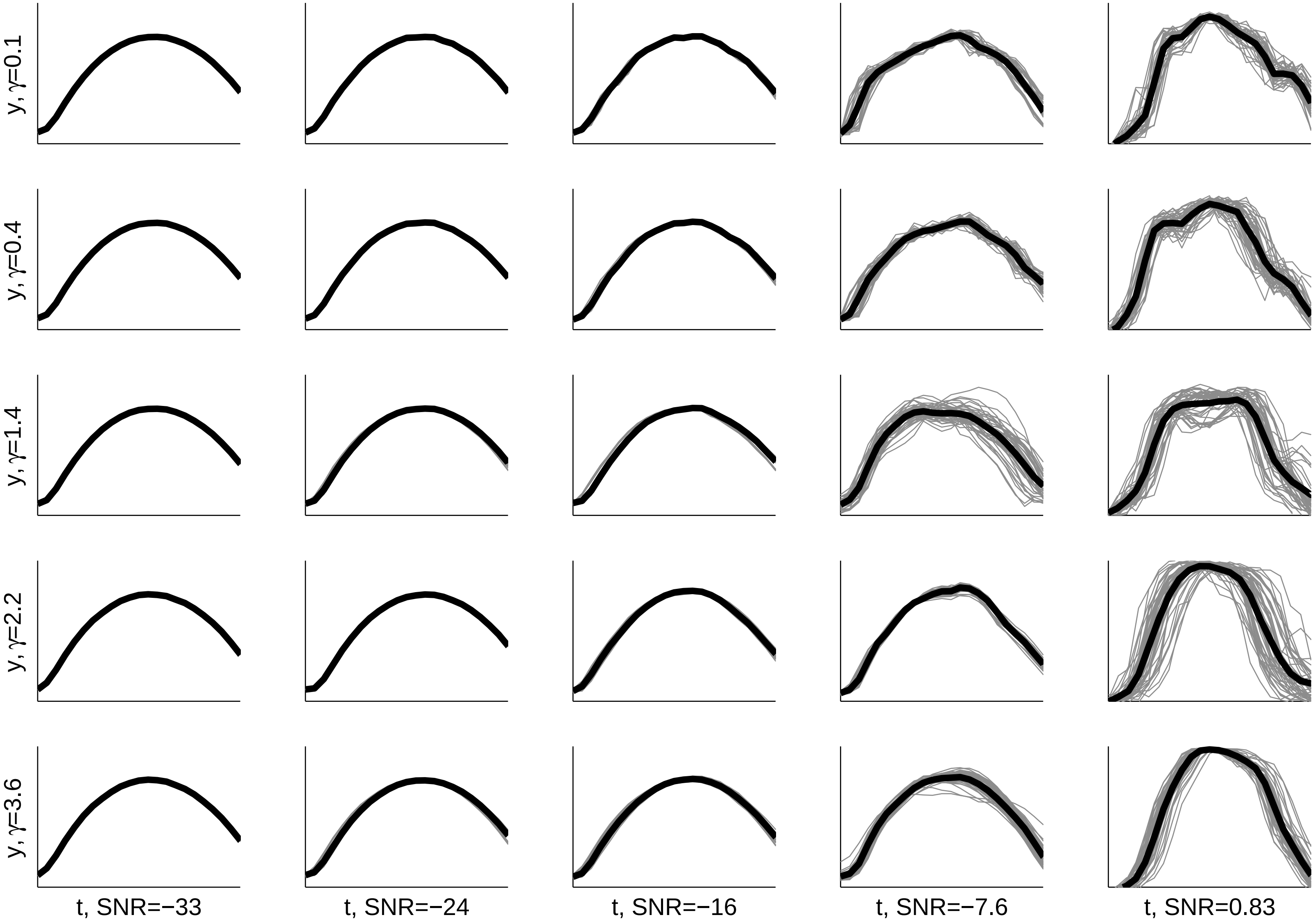


Figure 4.16: Identification of synthetic time series, stroboscopic plot of the identification.

coloured noise $\delta(t)$, with a spectrum like

$$\mathcal{F}(\delta(t)) = \begin{cases} 0 & \omega < \omega_0 \\ \frac{1}{(\omega-\omega_0+1)^{\frac{\gamma}{2}}} & \omega \geq \omega_0. \end{cases} \tag{4.14}$$

ω_0 was taken about two third of the fundamental frequency of the Colpitts time series, in such a way the noise does not appear modulated. Various values where taken for noise level and colour. The network, apart from the output connections assigned by the identification, was the same as the one used for identification of the surrogate Colpitts time series above. Figure 4.13 shows the attractor reconstruction of the identified systems, Figure 4.14 shows the spectra and Figures 4.15 and 4.16 show the stroboscopic plots of the training sequences and the identified system respectively. The spectrum of the time series to be identified is represented in grey and the spectrum of the identified system is shown in black. For better visualisation a 4096-Hanning window has been applied.

Qualitatively, even for this clearly non-deterministic, stochastic data the identification algorithm built a deterministic model. Quantitatively, the identified spectra match very good the original ones. In general, colour matches closely, whereas the identification underestimates the level constantly, however differences are up to the possibilities of what kind of spectrum a dynamic system can produce. Finally the time series were constructed specifically to be difficult to be identified. Even the discontinuity in the spectrum at w_0 is partially honoured.

When the signal to noise ratio gets in favour of noise, the characteristics of the attractor get lost. This can also be seen in spectrum. There the peaks of the higher harmonics disappear. The problem, which lies behind, is to know whether the continuous coloured spectrum is the noise, or whether it is the peaks. It appears the identification algorithm judges in favour of the contained energy.

4.4.2 CLASSIFICATION RESULTS

The most elementary classifier is the *dichotomizer.* It knows only one class and decides whether a given sample belongs to that class or not. It does this usually by evaluating a risk function and comparing it to a threshold, when the risk function is lower than the threshold, the dichotomizer decides that the sample belongs to the first class. The concept is easily expanded to the case with several categories. Such a classifier – a polychotomizer – would instead of comparing one risk function to a threshold compare several risk functions to each other. Each risk function would be associated to a particular category and its value indicates how high the risk of making a false decision is [Duda et al., 2001]. The aim of any classifier is of course to keep the false decisions as few as possible.

In a synchronisation framework for classifying approximately periodic time series the *synchronisation error* ε is used as risk function. The synchronisation error for a given time k is defined as the difference between the output of the forced model y_k and the time series to classify $y_{\mathrm{class},k}$ (*cf.* Figure 4.17). The root mean squared value over the length of the time series to classify is then the risk function itself. Finally, to do the classification, the synchronisation error would be compared to a threshold in the case of a single category, whereas in the multicategory case the different synchronisation errors would need to be normalised by their own threshold, before they can be compared.

To apply the algorithm to gait signals data from two healthy subjects was recorded. Both walked for 10 times one minute on flat ground, upstairs and downstairs. The data was then normalised in amplitude and mean frequency. The time series were then divided into deemed stationary pieces using a linear modelling technique and chopped into vectors with a length of 400 samples. Each vector holds approximately 15 periods and corresponds to 7 to 12 seconds of walking. This procedure yielded between 40 and 80 vectors for each category and subject. The data for walking on stairs were generally in a poorer condition than the data for walking on flat ground, so some pieces of stationary data were smaller than 400 samples and therefore were discarded. Furthermore the period length for walking on stairs was generally longer, leading to less periods recorded in the same time.

Three situations were considered for this data set. In a first time two different classifiers should classify the data for each subject separately; in a second step a third classifier should classify the data of both subjects. This implies the identification of 9 models, one per situation (first and second subject separately and both in one time) and category (downstairs, flat and upstairs walking).

For the identification and training of the models a few vectors were used from the data set. More precisely, one vector for each identification was considered. The models had to be identified with at least one vector for each category, furthermore, if by visual inspection the identified attractor was not satisfying, a second was taken. Then 3 vectors of the same category (positives) and 2 of any other category (negatives) were taken as training data for tuning the threshold. As an example, Figure 4.18 shows the attractors relevant for the flat model in the joint situation, two vectors were used, on from each subject. The plots show again that the realisation variation is slightly underestimated by the identification, like observed in the previous section (4.4.1).

After the identification the model has to be slightly adapted from its form shown in Figure 4.1 on page 38 to allow for a forcing with an external signal, y_{class}. The feedback loop is broken up and as entry to the model a sum of the external forcing signal and the output system is taken, this leads to the modified state equations,

$$\mathbf{x}_{k+1} = \alpha\mathbf{x}_k + \beta\tanh(W_{int}\mathbf{x}_k + W_{in}u_k + W_{back}((1-\rho)y_k + \rho y_{\text{class},k}) + B) \tag{4.15}$$

$$y_k = \tanh(W_{out}\mathbf{x}_k), \tag{4.16}$$

where, in fact, only the term in W_{back} has been changed; Figure 4.19 illustrates how the system is modified.

The parameter ρ gives the strength with which the system is forced. For a value $\rho = 0$, the system is not forced at all and $y_{\text{class},k}$ has no influence on y_k, whereas in the case of $\rho = 1$ no output at all is fed back and the model operates as filter. The stronger the system is forced (large ρ) the more likely the system synchronises with the input $y_{\text{class},k}$. The value of ρ is set such that for all time series belonging to the category the

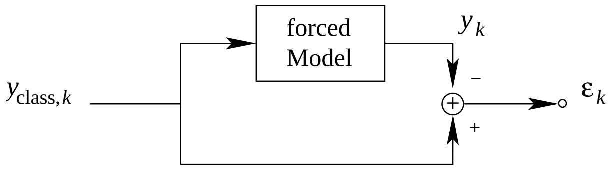
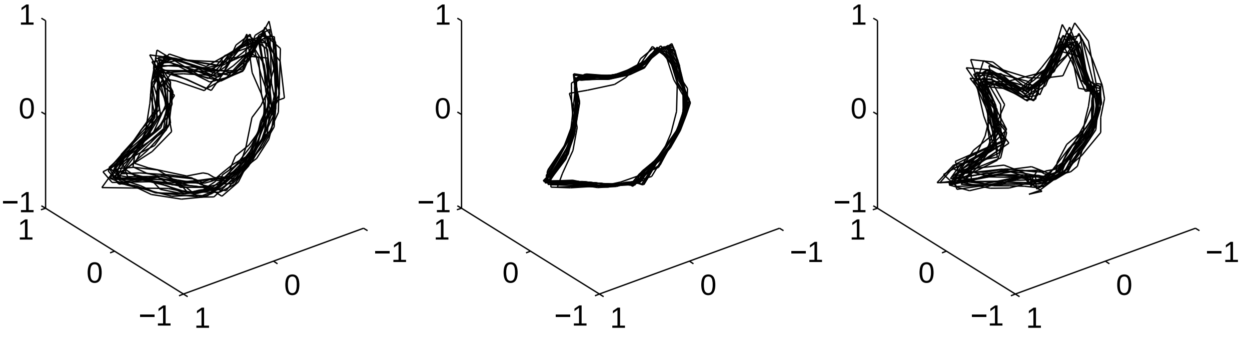


FIGURE 4.17: Estimation error for the classification.



(a) First Subject (b) Identified (c) Second Subject

FIGURE 4.18: Attractors for flat walking. In the middle the reconstructed attractor of the system used for classification, to left and right those of the time series used for identification of the former.

model is brought to synchronisation, but coupling is loose enough to permit the system to evolve freely on its own attractor in case of a foreign time series. The forcing strength has to be set in function of the particular model and data set. To determine it, the synchronisation error is plotted against the forcing strength for some time series belonging to the category and for some time series that should be rejected, *i.e.* not belonging to the same category. The difference between these two curves is the discrimination that is achieved by the classifier. Ideally, the synchronisation error behaves differently for members of the category and others. Because for members synchronisation phenomena are involved it is expected to be fractal in function of ρ, up to a certain critical value of ρ after which the synchronisation error monotonically decreases. Any other signal, not being a member of the category should not bring the system to synchronisation and therefore the synchronisation error should stay large. Consequently, the optimal value for the forcing strength is the value that maximises the discrimination, which is usually right above the critical value, for which the fractal behaviour of the discrimination error ends. Figure 4.20 shows the synchronisation errors and the average discrimination in the case of the model for both subjects and for flat walking. The plot suggests an optimal forcing strength of $\rho = 0.09$, like indicated by the dash-dotted line. The threshold or the weighting factor (in case of a dichotomizer or polychotomizer respectively) can be read out out of the same plot. It would be the mean between the synchronisation error for positives and negatives, which is 0.022 in this case.

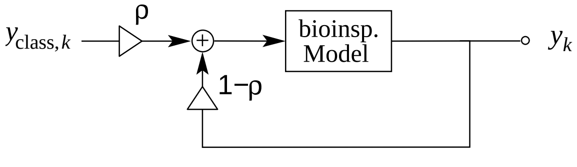


FIGURE 4.19: Excitation for classification.

as **in**	down	flat	up
down	92.5%	0%	7.5%
flat	0%	96.43%	3.57%
up	9.3%	2.7%	88%

Table 4.1: Results for classifying gait signals, first person.

as **in**	down	flat	up
down	87.76%	2%	10.24%
flat	0%	100%	0%
up	11.32%	1.89%	86.80%

Table 4.2: Results for classifying gait signals, second person.

as **in**	down	flat	up
down	79.1%	11%	9.9%
flat	0%	100%	0%
up	43.9%	16.8%	39.3%

Table 4.3: Results for classifying gait signals, one model for both persons. Both represented.

as **in**	down	flat	up
down	57.5%	20%	22.5%
flat	0%	100%	0%
up	5.6%	29.6%	64.8%

Table 4.4: Results for classifying gait signals, one model for both persons. Results for first person.

as **in**	down	flat	up
down	96%	4%	0%
flat	0	100%	0%
up	83%	3.8%	13.2%

Table 4.5: Results for classifying gait signals, one model for both persons. Represented only second person.

Similarly, all nine models were identified and their parameters estimated. In the end the remaining vectors, neither used for identification, nor during tuning of ρ, were classified. The results are shown in Tables 4.1 to 4.5. The first two tables show the results for classifying the data of the first and second subject separately. The third table shows the results of classifying the data of both subjects with one model for each category. For the results in each of the fourth and fifth table only the data of one subject have been used, but the classification was done with the model identified for classifying the data of both subjects.

Tables 4.1 and 4.2 show a remarkable success in telling flat data from other data. Indeed, the attractors of the different models synchronise only with difficulties to a time series not belonging to the given class; the attractors appear very selective in this respect. The attractors of the low dimensional modelling technique discussed in Chapter 3 were less selective, consequently a specific anti-synchronisation mechanism had to be built into the algorithm to avoid false classifications [De Feo, 2004a, De Feo, 2004b]. A possible explanation for this behaviour lies in the high dimensionality of the system, which permits higher filter capabilities. Unfortunately the same method has slightly more difficulties in making the distinction between up and down data. The cause of this lies in the dynamic properties of the data produced walking stairs, they seem to be "stiffer" than their flat counterparts. In Figure 4.21 this shows by a rather round, regular and differentiated shape of the waveform, whereas in the case of the stairs data the bear peaks seem to point out of a noisy ground. Arguably, the up time series appear already less stiff and effectively flat and up are more likely to be miss-detected than flat and down. This holds

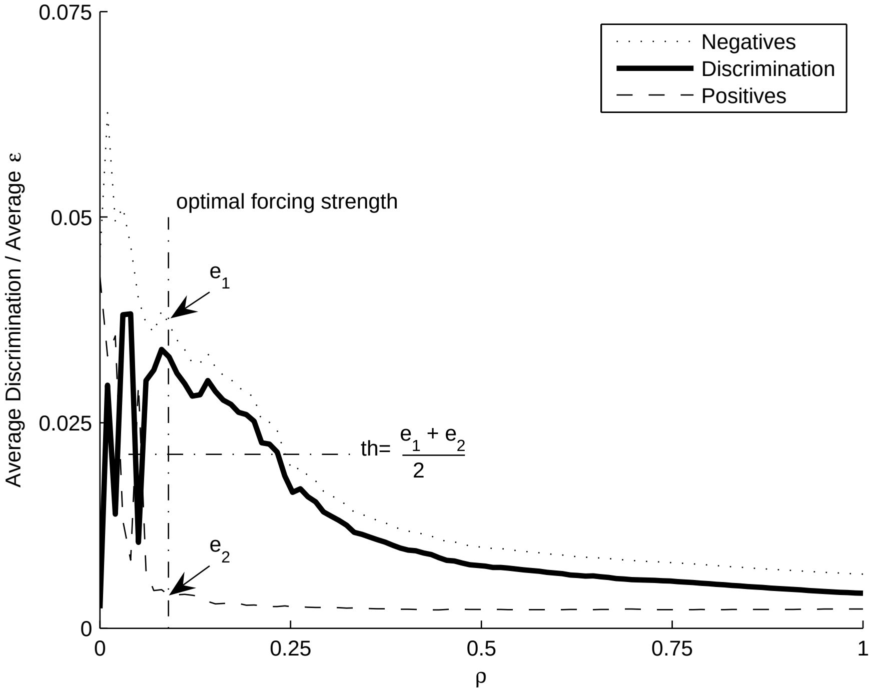
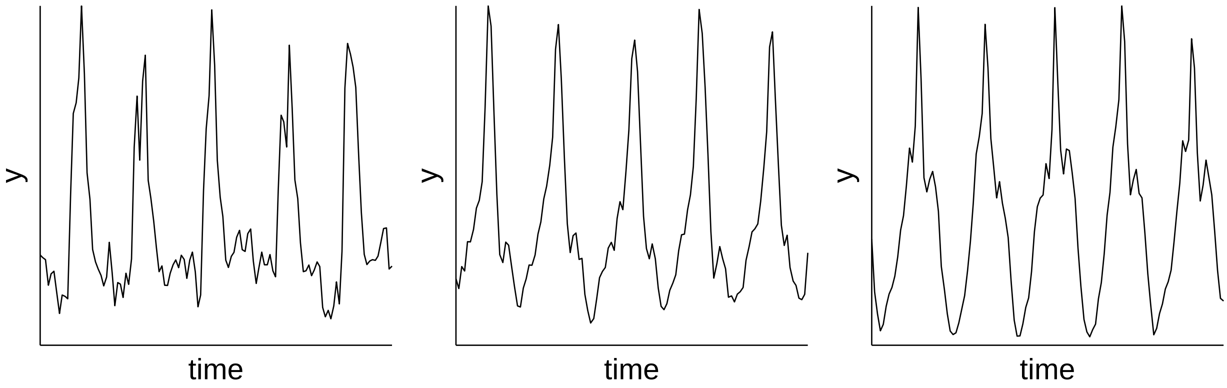


Figure 4.20: The average mean squared error between the model output and the time series to classify in function of the forcing strength ρ. The difference between the two curves is the average discrimination.



(a) Down (b) Up (c) Flat

Figure 4.21: Some pseudo periods of the different walking styles. Walking on stairs produces rather high and narrow peaks, whereas walking on flat ground gives smoother time series.

true at least for the first subject (Table 4.1), from which also stem the time series shown in Figure 4.21.

The implementation anti-synchronisation in the considered biologically inspired model could improve the classification between up and down data, though this would be the issue of future work. One way to do it would be, of course, to create a cost function being the sum of the synchronisation error of good examples and the reciprocal of the synchronisation error for bad examples; then to minimise this function with an optimisation algorithm. This approach would clearly have the disadvantage that the identification is no longer a simple regression, but again an optimisation. Another approach would be to continue using the regression for identification, but to add bad examples. The targeted output using regression would have to be crafted in such a way that the synchronisation error of the identified model is high for bad signals.

Table 4.3 shows a drastic degradation of the method, when it is applied to more than one subject. A priori, three possible reasons for this degradation can be given.

1. The diversity between subjects is greater than the diversity between classes, in which case the classifier would assign arbitrarily classes to the samples.
2. Dynamical aliasing makes samples generated by one subject and belonging to particular class appear like samples generated by another subject not belonging to the same class.
3. The identification procedure fails in this slightly more complex case (more complex compared to the single subject case).

Analysing further the results in Table 4.3, it can be seen that the degradation is caused mainly by a large amount of up samples being classified as down (high value in third row, first column). Still a large amount of down and flat samples have been classified correctly. This observation is in contradiction with the first and third reasons for the degradation, so reason two seems to be the most likely. Tables 4.4 to 4.5 support further this assumption. In the first of these tables, where only the samples of the first subject are classified with the joint model, shows a clear accentuation of the main diagonal. On the contrary Table 4.5, which shows the same classification for the samples of the second subject, has no longer a clearly accentuated main diagonal, but almost all up samples were classified as down (high value in third row, first column). This suggests that the samples of the second subject walking up are very similar to the samples of the first subject walking down. The rather high value in the first row, third column in Table 4.4 indicates the same. The presence of dynamical aliasing in the method does not permit its use in a anonymous setup. The classifier has to be trained to a specific subject.

Summarising it can be said that the proposed way of classifying gait signals performs very well in case the classifier can be trained to a particular subject. The classification rates obtained this way are contained in the band of 87% to 100%, which clearly outperforms the results given by Kern and Schiele [2003] (*cf.* Section 2.1.1 on page 2.1.1). The method performs in line or slightly worse for the results given by Coley et al. [2005] and Sekine et al. [2000], though it has to be stressed that the placement of the sensor considered by Coley et al. is less convenient for the subject and facilitates classification. Furthermore, in the classification considered by Coley et al. the subject is not supposed to walk downstairs.

CHAPTER 5

CONCLUSIONS & FINAL DISCUSSION

Brief — The results and insights of the previous chapters are summarised. Based on the results and remaining open questions, perspectives for future research are outlined.

During the work a previously developed framework for classification, or more precisely for feature extraction has been considered. The framework proposes a novel approach to model diversity deterministically, but suffered from drawbacks. The most important drawback was arguably that the identification of the reference model and the corresponding tuning procedure were very tedious. In fact, for synchronisation and anti-synchronisation to work within this framework the reference model had to be transformed, involving complicated computations. Furthermore, using the proposed identification algorithm, chaos did not emerge when identifying gait signals, but a chaotic reference model is a necessary condition in the proposed framework. This particularity was explained by the strong periodicity of the gait signals, which makes them differ from other signals considered.

To address this problems a change of reference model and its identification algorithm was considered. The same signals for which chaos did not emerge during identification proved to be easily identifiable with the new model and identification algorithm and chaos did emerge. Additionally, the whole procedure from identification to classification was considerably simplified. The identification does not involve an optimisation anymore, the tuning to synchronisation is done by adjusting a scalar and not a matrix, without the need for solving equations that stem from periodic control. Overall the computational power involved for classification could be decreased by orders of magnitude.

While introducing the new reference model into the classifier, two further questions have been addressed. First, the model, despite announced several time in literature for different tasks, still has never been analysed in depth. Here, a measure has been given, which expresses the intuitive interpretation how the model works. An interpretation also given by the authors of those reports, which introduced this kind of model [Jaeger, 2001, Maass et al., 2002]. Based on that measure favourable parameter settings could be given. Though it can be doubted that the model with maximal measure, *i.e.* maximal entropy, really coincides with the model that delivers best classification results, by construction, a high entropy is a conditio sine qua non for a successful identification of the reference model.

Finally, during the research, it could definitely be verified that deterministic chaos can be used even to model probabilistic patterns in view of their classification. Up to now, this was assumed, but never verified, although this constitutes a necessary condition for the classification method considered here (and also in its preceding version) to be universal. An important result also for the European project "APEREST", which aims at modelling information processing in the brain.

5.1 Future Work

The optimisation of the systems entropy was limited to case of varying the system's global parameters. A possible next step is to consider the system's microscopic parameters, *i.e.* the connections itself, as it was found that different networks randomly created with the same global parameter values, can have significantly different values for their entropy. For a start the entropy of matrices in companion form could be considered [Chen, 1999].

Still considering the microscopic structure of the network, an analogy to real neural networks could be searched. By analysing neural tissue it should be possible to record the synaptic connections within in a connection matrix. This connection matrix could then be used in the biologically inspired model and decided whether or not it outperforms randomly created connections. In case it performs well, from the same matrix the entropy could be determined and conclusions on the reliability of the entropy could be drawn and in the given case another measure could be defined. In a wider sense, these results could indicate, whether information processing in real neural networks (brains) exploits with synchronisation-like phenomena.

Alternatively, the periodic Lur'e systems abandoned in Chapter 3 could be further considered. Other phenomena that permit reliable classification (feature extraction), possibly synchronisation-like, could be searched and exploited.

APPENDIX A

ASYMPTOTICALLY UNIQUE BEHAVIOUR

Consider the state equations of the circuit (4.4) in open loop operation and with a general saturation function $f(\cdot)$ instead of the hyperbolic tangent $\tanh(\cdot)$

$$\mathbf{x}_{k+1} = \alpha \mathbf{x}_k + \beta f(W_{int}\mathbf{x}_k + W_{in}u_k + W_{back}y_k + \mathbf{B}). \tag{A.1}$$

A.1 NECESSARY CONDITIONS

Suppose u_k and y_k are constants and thus the system is

$$\mathbf{x}_{k+1} = \alpha \mathbf{x}_k + f(W_{int}\mathbf{x}_k + \tilde{\mathbf{B}}), \tag{A.2}$$

where

$$\tilde{\mathbf{B}} = \mathbf{B} + W_{in}u + W_{back}y \tag{A.3}$$

is constant. Suppose also $\boldsymbol{\xi}$ is a fixed point, *i.e.*

$$\boldsymbol{\xi} = \alpha\boldsymbol{\xi} + f(W_{int}\boldsymbol{\xi} + \tilde{\mathbf{B}}). \tag{A.4}$$

Now consider the linearisation around this fixed point:

$$\mathbf{x}_k = \boldsymbol{\xi} + \Delta\mathbf{x}_k, \tag{A.5}$$

$$\Rightarrow \mathbf{x}_{k+1} \approx \alpha\Delta\mathbf{x}_k + F(\boldsymbol{\xi})W_{int}\Delta\mathbf{x}_k. \tag{A.6}$$

Hence for the asymptotic stability of the fixed point, it is necessary that the matrix M,

$$M = \alpha I + F(\boldsymbol{\xi})W_{int}, \tag{A.7}$$

has all its eigenvalues with absolute value smaller than 1. This is a necessary condition for unique asymptotic behaviour. Here the Jacobian is

$$F(\boldsymbol{\xi}) = \begin{bmatrix} \frac{df}{dx}\big((W_{int}\boldsymbol{\xi})_1 + \tilde{B}_1\big) & 0 & 0 & \cdots & 0 \\ 0 & \frac{df}{dx}\big((W_{int}\boldsymbol{\xi})_2 + \tilde{B}_2\big) & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & & 0 \\ 0 & 0 & \cdots & 0 & \frac{df}{dx}\big((W_{int}\boldsymbol{\xi})_N + \tilde{B}_N\big) \end{bmatrix} \tag{A.8}$$

Note that if $f(0) = 0$, $\tilde{\mathbf{B}} = 0$, then $\boldsymbol{\xi} = 0$ is a fixed point and the corresponding matrix of the linearisation (A.7) is

$$M = \alpha I + \beta f'(0) W_{int} \tag{A.9}$$

and for $f(x) = \tanh(x)$ $f'(0) = 1$ and thus (A.9) becomes

$$M = \alpha I + \beta W_{int}. \tag{A.10}$$

If the function $f(\cdot)$, as is the case for $\tanh(\cdot)$, has

$$0 < \frac{df}{dx} \leq 1, \tag{A.11}$$

then $F(\boldsymbol{\xi})$ is a diagonal matrix with diagonal entries between 0 and 1. It may appear that taking 1 still would be necessary, *i.e.* that we need for asymptotically unique behaviour that the largest eigenvalue of

$$M = \alpha I + \beta W_{int}$$

needs to be smaller than 1. This is not the case, as can be seen for $N = 1$.

Let the system be

$$x_{k+1} = \alpha x_k + \beta f(w x_k + \tilde{b}), \tag{A.12}$$

and let

$$0 < f'(x) \leq 1, \qquad f'(0) = 1, \qquad w < 0,$$

and

$$g(x) = \alpha x + \beta f(w x + \tilde{b}).$$

Then

$$x_{k+1} = g(x_k). \tag{A.13}$$

Suppose the situation as depicted in Figure A.1. Clearly, $\boldsymbol{\xi}$ is a globally asymptotically stable fixed point. On the other hand:

$$g'(-\frac{\tilde{b}}{w}) = \alpha + \beta f'(0) w = \alpha + \beta w.$$

Here $\alpha + \beta w < -1$ is possible without destroying the asymptotically unique behaviour. A necessary condition is that the fixed point ξ is asymptotically stable, namely that

$$|\alpha + \beta f'(\xi) w| \leq 1. \tag{A.14}$$

A.2 SUFFICIENT CONDITIONS

Sufficient conditions are typically obtained by requiring that the Jacobians are contractive in some (common) metric. If the Euclidean norm is taken, then it could be required that all singular values of all matrices

$$\alpha I + \beta F(\boldsymbol{\xi}) W_{int} \tag{A.15}$$

are smaller than 1. A singular value of a matrix A is the square root of an eigenvalue of $A^T A$.

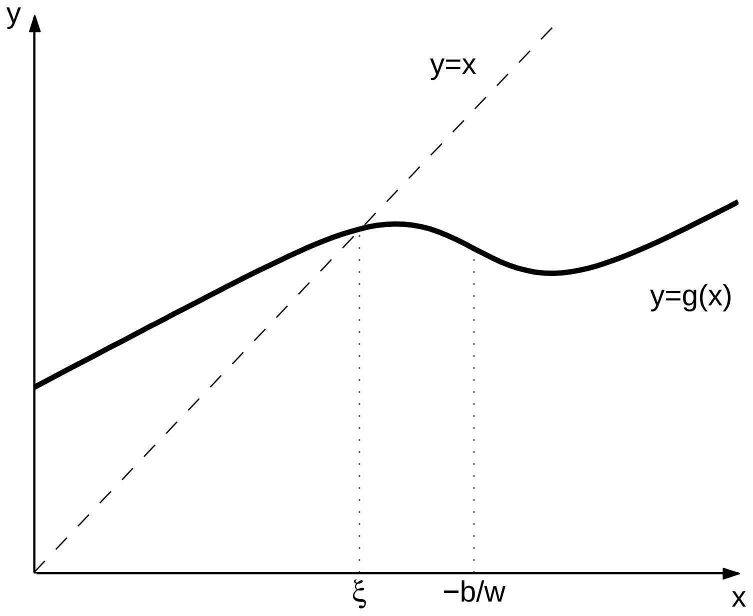


FIGURE A.1: Situation for the one dimensional case.

The proof goes as follows. We have to prove that any two solutions of the equations

$$\mathbf{x}_{k+1} = \alpha \mathbf{x}_k + \beta f(W_{int}\mathbf{x}_k + W_{in}u_k + W_{back}y_k + B) \tag{A.16}$$

converge to each other. Consider two such solutions

$$\mathbf{x}_k^{(0)} \quad \text{and} \quad \mathbf{x}_k^{(1)},$$

with initial conditions

$$\mathbf{x}_0^{(0)} \quad \text{and} \quad \mathbf{x}_0^{(1)},$$

respectively. By interpolating the initial conditions

$$\mathbf{x}_0^{(\mu)} = (1-\mu)\mathbf{x}_0^{(0)} + \mu \mathbf{x}_0^{(1)}. \tag{A.17}$$

define a continuous family of solutions: $\mathbf{x}_k^{(\mu)}$ is solution to $\mathbf{x}_0^{(\mu)}$.

Then,

$$\mathbf{x}_k^{(1)} - \mathbf{x}_k^{(0)} = \int_0^1 \frac{d}{d\mu}\mathbf{x}_k^{(\mu)} d\mu. \tag{A.18}$$

And,

$$\begin{aligned} \frac{d}{d\mu}\mathbf{x}_k^{(\mu)} &= J(\mathbf{x}_{k-1}^{(\mu)})\frac{d\mathbf{x}_{k-1}^{(\mu)}}{d\mu} \\ &= J(\mathbf{x}_{k-1}^{(\mu)})J(\mathbf{x}_{k-2}^{(\mu)})\cdots J(\mathbf{x}_{k-1}^{(\mu)})\frac{d\mathbf{x}_0^{(\mu)}}{d\mu} \\ &= J(\mathbf{x}_{k-1}^{(\mu)})\cdots J(\mathbf{x}_0^{(\mu)})(\mathbf{x}_0^{(1)} - \mathbf{x}_0^{(0)}), \end{aligned} \tag{A.19}$$

where

$$\begin{aligned} J(\mathbf{x}) =& \frac{\partial}{\partial \mathbf{x}} \left[\alpha x + \beta f(W_{int}\mathbf{x} + W_{in}u + W_{back}y + \tilde{\mathbf{B}}) \right] \\ =& \quad \alpha I + \beta F(W_{int}\mathbf{x} + W_{in}u + W_{back}y + \tilde{\mathbf{B}}) W_{int} \end{aligned} \tag{A.20}$$

For the largest eigenvalue λ_{max} of $J(x)$, it can be written

$$\begin{aligned} ||J(\mathbf{x})\mathbf{v}|| =& \sqrt{||J(\mathbf{x})\mathbf{v}||^2} \\ =& \sqrt{\mathbf{v}^T \mathbf{J}^T(\mathbf{x}) \mathbf{J}(\mathbf{x}) \mathbf{v}} \\ \leq& \sqrt{\mathbf{v}^T {\lambda_{max}}^2 \mathbf{v}} = \lambda_{max} ||\mathbf{v}|| \end{aligned} \tag{A.21}$$

where $||\cdot||$ denotes Euclidean norm. Hence,

$$\begin{aligned} ||\mathbf{x}_k^{(1)} - \mathbf{x}_k^{(0)}|| \leq& \int_0^1 ||\frac{d}{d\mu}\mathbf{x}_k^{(\mu)}|| d\mu \\ \leq& \int_0^1 ||J(\mathbf{x}_{k-1}^{(\mu)}) \cdots J(\mathbf{x}_0^{(\mu)})|| (\mathbf{x}_0^{(1)} - \mathbf{x}_0^{(0)})|| d\mu \\ \leq& \int_0^1 {\lambda_{max}}^k ||(\mathbf{x}_0^{(1)} - \mathbf{x}_0^{(0)})|| d\mu = {\lambda_{max}}^k ||(\mathbf{x}_0^{(1)} - \mathbf{x}_0^{(0)})|| \end{aligned} \tag{A.22}$$

If $\lambda_{max} < 0$, then $||(\mathbf{x}_0^{(1)} - \mathbf{x}_0^{(0)})|| \xrightarrow[k\to\infty]{} 0$. Thus, the condition is that the largest singular value of all matrices M,

$$M = \alpha I + \beta F(\mathbf{x} W_{int}),$$

is smaller than some number smaller than 1.

If $f(\cdot) = \tanh(\cdot)$ this becomes the largest singular value of all matrices $\alpha I + \beta D W_{int} \leq \lambda_{max} < 1$, where D is a diagonal matrix of entries between 0 and 1, where the 1 is included in the interval, but 0 not.

It may seem that it is sufficient to impose the condition on $\alpha I + \beta W_{int}$. This is, however, not true, already for the one dimensional case, $N = 1$, this is not true. In this case the condition is

$$(\alpha + \beta d w)^2 \leq \lambda < 1, \qquad \text{for } \ 0 < d. \leq 1 \tag{A.23}$$

Suppose that $w = -\frac{\alpha}{\beta}$. Then $(\alpha + \beta w)^2 = 1$, but for $0 < d < 1 \quad (\alpha + \beta w)^2 > 0$ and for large enough α and β it may be larger than 1.

Note that in both counterexamples in this section w was negative. For positive values of w it would be sufficient to consider $d = 1$. Similar properties hold in higher dimensions. However, in the considered networks both, excitatory and inhibitory, interactions can be found and therefore the analysis should not be limited to such cases.

LIST OF FIGURES

1.1 Classification block scheme . . . 2
1.2 General topology of a neural network . . . 6

2.1 Example of a vowel, time domain . . . 12
2.2 Example of a vowel, stroboscopic plot . . . 13
2.3 Example of a time series from the Colpitts oscillator . . . 15
2.4 The sensor of the Physilog. . . . 16
2.5 Example of a gait signal . . . 18
2.6 Example of an attractor reconstruction . . . 20
2.7 Observation function . . . 23
2.8 System in Lur'e form . . . 24
2.9 Not smooth cost function . . . 25

3.1 Identification of gait signal . . . 32
3.2 Identification of Colpitts . . . 32
3.3 Correlation dimension of gait signals . . . 33
3.4 Maximal Lyapunov exponent of gait signal . . . 34

4.1 General topology of biologically inspired model . . . 38
4.2 The nodes dynamics . . . 39
4.3 Nonuniform bias without connections . . . 44
4.4 Nonuniform bias and connection probability . . . 45
4.5 Length of the training sequences . . . 46
4.6 Dimension of the network . . . 47
4.7 Connection probability and maximal eigenvalue . . . 49
4.8 1-step prediction error of linear model . . . 50
4.9 Adjacency matrix with components . . . 51
4.10 Final adjacency matrix . . . 52
4.11 Identification of surrogate data, attractor reconstruction . . . 54
4.12 Identification of Colpitts surrogate data, FFT . . . 54
4.13 Identification of synthetic time series, reconstruction . . . 55
4.14 Identification of synthetic time series, FFT . . . 56
4.15 Identification of synthetic time series, stroboscopic plot of training sequences 57
4.16 Identification of synthetic time series, stroboscopic plot of identification . 58
4.17 Estimation error for the classification. . . . 60
4.18 Attractors relevant for classification of flat walking. . . . 61
4.19 Excitation for classification. . . . 61
4.20 Average discrimination . . . 63
4.21 Time domain representation of gait signals. . . . 64

A.1 Asymptotically unique behaviour 73

BIBLIOGRAPHY

[Aminian and Najafi, 2004] Aminian, K. and Najafi, B. (2004). Capturing human motion analysis using body-fixed sensors: Outdoor measurements and clinical applications. *Computer Animation & Virtual Worlds*, 15:79 – 94.

[Anton and Rorres, 1994] Anton, H. and Rorres, C. (1994). *Elementary Linear Algebra.* John Wiley & Sons, seventh edition.

[Arbib, 1995] Arbib, M. A., editor (1995). *The Handbook of Brain Theory and Neural Networks.* MIT Press.

[Bagni, 2004] Bagni, G. (2004). *Reduced Model Identification for Chaotic Signals.* PhD thesis, Università degli Studi di Firenze.

[Bai, 2003] Bai, E.-W. (2003). Frequency domain method of hammerstein models. *IEEE Transactions on Automatic Control*, 48(4):530 – 542.

[Baier, 2003] Baier, N. U. (2003). Attractor learning with recurrent, artificial, nonlinear, neural network. In *Proceedings of the European Conference on Circuit Theory and Design 2003*, pages III–417 – III–420.

[Baier et al., 2000] Baier, N. U., Schimming, T., and De Feo, O. (2000). Nonlinear structures in voiced speech signals. In *Proceedings of International Symposium on Nonlinear Theory and its Applications (NOLTA) 2000*, volume 2, pages 727 – 730.

[Beauchet et al., 2003] Beauchet, O., Kressig, R. W., Najafi, B., Aminian, K., Dubost, V., and Mourey, F. (2003). Age-related decline of gait control under a dual-task condition. *Journal of the American Geriatrics Society*, 51(8):1187.

[Bernardo and Smith, 1996] Bernardo, J. M. and Smith, A. F. M. (1996). *Bayesian Theory.* Wiley.

[Bittanti and Colaneri, 1999] Bittanti, S. and Colaneri, P. (1999). *Periodic Control*, pages 59–74. John Wiley & Sons, New York, NY.

[Bittanti and Picci, 1996] Bittanti, S. and Picci, G., editors (1996). *Identification, Adaptation, Learning: The Science of Learning Models from Data.* Springer-Verlag, New York, NY.

[Boyd and Chua, 1985] Boyd, S. and Chua, L. O. (1985). Fading memory and the problem of approximating non-linear operators with volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150 – 1161.

[Bülthoff et al., 2003] Bülthoff, H. H., Lee, S.-W., Poggio, T., and Wallraven, C., editors (2003). *Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 282 – 293. Springer.

[Caudill and Butler, 1990] Caudill, M. and Butler, C. (1990). *Naturally Intelligent Systems*. MIT Press.

[Chen, 1999] Chen, C.-T. (1999). *Linear System Theory and Design*. Oxford University Press.

[Cohen et al., 1999] Cohen, J. E., Newman, C. M., Cohen, A. E., L., P. O., and Gonzalez, A. (1999). Spectral mimicry: A method of synthesizing matching time series with different fourier spectra. *Circuits, Systems and Signal Processing*, 18(3):431 – 442.

[Coley et al., 2005] Coley, B., Najafi, B., Paraschiv-Ionescu, A., and Aminian, K. (2005). Stair climbing detection during daily physical activity using a miniature gyroscope. *Gait & Posture*. Accepted for Publication.

[Cover and Hart, 1967] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transaction on Information Theory*, IT-13(1):21 – 27.

[Cristianini and Shawe-Taylor, 2004] Cristianini, N. and Shawe-Taylor, J. (2004). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge.

[de Boor, 1994] de Boor, C. (1994). *A Practical Guide to Splines*. Number 27 in Applied Mathematical Sciences. Springer.

[de Boor, 2003] de Boor, C. (2003). *Spline Toolbox - For use with* MATLAB. The MathWorks, version 3 edition. Version 3.2.

[De Feo, 2001] De Feo, O. (2001). *Modeling Diversity by Strange Attractors with Application to Temporal Pattern Recognition*. PhD thesis, Ecole Polytechnique Federal de Lausanne. N. 2344.

[De Feo, 2003] De Feo, O. (2003). Self-emergence of chaos in the identification of irregular periodic behaviour. *Chaos*, 13(4):1205 – 1215.

[De Feo, 2004a] De Feo, O. (2004a). Qualitative resonance of Shil'nikov-like strange attractors, part I: Experimental evidence. *International Journal of Bifurcation & Chaos*, 14:873–891.

[De Feo, 2004b] De Feo, O. (2004b). Qualitative resonance of Shil'nikov-like strange attractors, part II: Mathematical analysis. *International Journal of Bifurcation & Chaos*, 14:893–912.

[De Feo et al., 2000] De Feo, O., Kennedy, M. P., and Maggio, G. M. (2000). The colpitts oscillator: Families of periodic solutions and their bifurcations. *International Journal of Bifurcation and Chaos*.

[Deller et al., 1993] Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (1993). *Discrete-Time Processing of Speech Signals*. Prentice Hall.

[Diks, 1999] Diks, C. (1999). *Nonlinear Time Series Analysis*, volume 4 of *Nonlinear Time Series and Chaos*. World Scientific.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, second edition.

[Fernando and Sojakka, 2003] Fernando, C. and Sojakka, S. (2003). Pattern recognition in a bucket. In *Lecture Notes in Artificial Intelligence*, volume 2801, pages 588 – 597. Springer.

[Gardner, 1991] Gardner, W. A. (1991). Exploitation of spectral redundancy in cyclostationary signals. *IEEE Signal Processing Magazine*, 8:14 – 36.

[Gardner, 1994] Gardner, W. A., editor (1994). *Cyclostationarity*. IEEE Press.

[Grewal and Andrews, 1993] Grewal, M. S. and Andrews, A. P. (1993). *Kalman Filtering: Theory and Practice*. Prentice Hall.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

[Haykin, 1998] Haykin, S. (1998). *Neural Networks. A Comprehensive Foundation*. Prentice Hall, second edition.

[Herzel, 1993] Herzel, H. (1993). Bifurcations and chaos in voice signals. *Applied Mechanics Reviews*, 46:399 – 413.

[Highleyman, 1962] Highleyman, W. H. (1962). Linear decision functions, with application to pattern recognition. *Proceedings of the IRE*, 50(6):1501 – 1514.

[Holden et al., 1991] Holden, A., Tucker, J., and B.C., T. (1991). Can excitable media be considered as computational systems? *Physica D*, 49:240 – 246.

[Ijspeert, 2003] Ijspeert, A. (2003). Vertebrate locomotion. In Arbib, M., editor, *The handbook of brain theory and neural networks*, pages 649–654. MIT Press.

[Ingber, 1993] Ingber, L. (1993). Simulated annealing: Practice vs. theory. *Mathematical and Computational Modelling*, 18(11):29 – 57.

[Ishizaka and Flanagan, 1972] Ishizaka, K. and Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal*, 51(6):1233 – 1268.

[Jaeger, 2001] Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks. Technical Report 148, Fraunhofer Institute for Autonomous Intelligent Systems.

[Jaeger, 2002a] Jaeger, H. (2002a). Short term memory in echo state networks. Technical Report 152, Fraunhofer Institute for Autonomous Intelligent Systems.

[Jaeger, 2002b] Jaeger, H. (2002b). Tutorial on training recurrent neural networks. Technical Report 159, Fraunhofer Institute for Autonomous Intelligent Systems.

[Jaeger and Haas, 2004] Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304:78 – 80.

[Kantz, 1994] Kantz, H. (1994). A robust method to estimate the maximal lyapunov exponent of a time series. *Physics Letters A*, 185(1):77 – 87.

[Kantz and Schreiber, 1999] Kantz, H. and Schreiber, T. (1999). *Nonlinear Time Series Analysis.* Number 7 in Cambridge Nonlinear Sciences Series. Cambridge University Press, paperback edition.

[Kern and Schiele, 2003] Kern, N. and Schiele, B. (2003). Multi-sensor activity context detection for wearable computing. In *European Symposium on Ambient Intelligence*, Eindhoven, The Netherlands.

[Krogh et al., 1994] Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). Hidden markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5).

[Kuznetsov, 1998] Kuznetsov, Y. A. (1998). *Applied Bifurcation Theory.* Springer.

[Liebert, 1991] Liebert, W. (1991). *Chaos und Herzdynamik.* Number 4 in Reihe Physik. Harri Deutsch.

[Ljung, 1999] Ljung, L. (1999). *System Identification – Theory for the User.* Prentice Hall, second edition.

[Maass et al., 2002] Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14:2351 – 2560.

[Maass et al., 2003] Maass, W., Natschläger, T., and Markram, H. (2003). A model for real-time computation in generic neural microcircuits. In Becker, S., Thrun, S., and Obermayer, K., editors, *Proc. of NIPS 2002, Advances in Neural Information Processing Systems*, volume 15, pages 229–236. MIT Press.

[Mackey and Glass, 1977] Mackey, M. C. and Glass, L. (1977). Oscillations and chaos in a physiological control system. *Science*, 197(287).

[Maggio et al., 1999] Maggio, G. M., De Feo, O., and Kennedy, M. P. (1999). Nonlinear analysis of the Colpitts oscillator and applications to design. *IEEE Transactions on Circuit and Systems—I*, 46:1118 – 1130.

[Makhoul et al., 1985] Makhoul, J., S., R., and H., G. (1985). Vector quantization in speech coding. *Proceedings of the IEEE*, 73(11):1551 – 1588.

[Mao, 1997] Mao, X. (1997). *Stochastic Differential Equations & Applications.* Mathematics & Applications. Horwood Publishing.

[Minsky and Papert, 1969] Minsky, M. L. and Papert, S. A. (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press.

[Najafi et al., 2002] Najafi, B., Aminian, K., Loew, F. c., Blanc, Y., and Robert, P. A. (2002). Measurement of stand-sit and sit-stand transitions using a miniature gyroscope and its application in fall risk evaluation in the elderly. *IEEE Transactions on Biomedical Engineering*, 49(8):843 – 851.

[Najafi et al., 2003] Najafi, B., Aminian, K., Paraschiv-Ionescu, A., Loew, F. c., Büla, C. J., and Robert, P. A. (2003). Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. *IEEE Transactions on Biomedical Engineering*, 50(6):711 – 723.

[Natschläger et al., 2002] Natschläger, T., Maass, W., and Markram, H. (2002). The "liquid computer": A novel strategy for real-time computing on time series. *Telematik*, 8(1).

[Nelles, 2001] Nelles, O. (2001). *Nonlinear System Identification*. Springer.

[Ott, 1993] Ott, E. (1993). *Chaos in Dynamical Systems*. Cambridge University Press.

[Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *Annal of Mathematical Statistics*, 33(3):1065 – 1076.

[Puskorius and Feldkamp, 1994] Puskorius, G. V. and Feldkamp, L. A. (1994). Neurocontrol of nonlinear dynamical systems with kalman filter trained recurrent networks. *IEEE Transactions onf Neural Networks*, 5(2):279 – 297.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286.

[Rabiner and Juang, 1993] Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall.

[Rieke et al., 1999] Rieke, F., Warland, D., de Ruyter, v. S., and Bialek, W. (1999). *Spikes: Exploring the Neural Code*. MIT Press.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386 – 408.

[Rosenstein et al., 1993] Rosenstein, M. T., Collins, J. J., and De Luca, C. J. (1993). A practical method for calculating largest lyapunov exponents. *Physica D*, 65(1-2):117 – 134.

[Sauer and Yorke, 1993] Sauer, T. and Yorke, J. A. (1993). How many delay coordinates do you need? *International Journal on Bifurcation and Chaos*, 3(3):737 – 744.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, Alexander, J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press.

[Schreiber and Schmitz, 2000] Schreiber, T. and Schmitz, A. (2000). Surrogate time series. *Physica D*, 142(3 – 4):197 – 385.

[Sekine et al., 2002] Sekine, M., Tamura, T., Akay, M., Fujimoto, T., Togawa, T., and Fukui, Y. (2002). Discrimination of walking patterns using wavelet-based fractal analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(3):188 – 196.

[Sekine et al., 2000] Sekine, M., Tamura, T., Togawa, T., and Fukui, Y. (2000). Classification of waist-acceleration signals in a continous walking record. *Medical Engineering & Physics*, 22(4):285 – 291.

[Story and Titze, 1995] Story, B. H. and Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97(2).

[Strogatz, 1994] Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos.* Perseus Books.

[Tononi et al., 1998] Tononi, G., Edelman, G. M., and Sporns, O. (1998). Complexity and coherency: Integrating information in the brain. *Trends in Cognitive Sciences*, 2(12):474 – 484.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer.

[Wackermann, 1999] Wackermann, J. (1999). Towards a quantitative characterisation of functional states of the brain: From the non-linear methodology to the global linear description. *International Journal of Psychophysiology*, 34:65 – 80.

Curriculum Vitae

Norman Urs Baier
rue du Petit-Beaulieu 7
CH-1004 Lausanne

normanurs.baier@gmail.com
Tel. +41 79 710 00 32

Marital Status: Single
Born in Biel (CH), march 24, 1976
Swiss and German citizenship

Work Experience

2001 – 2004	PhD Student and Assistant Swiss Federal Institute of Technology Lausanne Identification of approximately periodic signals with a biologically inspired algorithm Part of Work Package 2 of European Project “Approximately Periodic Representation of Stimuli” (APEREST)
1999	Technician at Nokia, part time (5 months) Measures of a GSM net in Switzerland
1998	Practical training at ETA ltd, part of Swatch Group (3 months) Development of a 930 MHz Oscillator for use in a pager

Studies

2000 – 2001	Doctoral School Department of Communication Systems, Swiss Federal Institute of Technology Lausanne (SFIT)
1995 – 2000	Diploma as electrical engineer Swiss Federal Institute of Technology Lausanne
1997/98	Exchange year at Linköpings Tekniska Högskola, Linköping, Sweden
1991 – 1995	Grammar School, Collège St-Michel, Fribourg, Switzerland

Skills

Informatics	Operative systems:	Linux, Windows
	Programming:	C/C++, Matlab, PHP, HTML
	Applications:	LaTeX, Office
Languages	Good knowledge in German (Mother’s tongue), English and French	
	Basic knowledge in Swedish and Italian	

Spare Time

Sports	Yachting, swimming, windsurfing, skiing
Other	Black & white photography and darkroom, languages