Biomechanically Informed Nonlinear Speech Signal Processing

Max A. Little

Exeter College

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Michaelmas Term 2006

Abstract

Biomechanically Informed Nonlinear Speech Signal Processing

Max A. Little, Exeter College, University of Oxford

Linear digital signal processing based around linear, time-invariant systems theory finds substantial application in speech processing. The linear acoustic source-filter theory of speech production provides ready biomechanical justification for using linear techniques. Nonetheless, biomechanical studies surveyed in this thesis display significant nonlinearity and non-Gaussianity, casting doubt on the linear model of speech production. In order therefore to test the appropriateness of linear systems assumptions for speech production, surrogate data techniques can be used. This study uncovers systematic flaws in the design and use of existing surrogate data techniques, and, by making novel improvements, develops a more reliable technique.

Collating the largest set of speech signals to-date compatible with this new technique, this study next demonstrates that the linear assumptions are not appropriate for all speech signals. Detailed analysis shows that while vowel production from healthy subjects cannot be explained within the linear assumptions, consonants can. Linear assumptions also fail for most vowel production by pathological subjects with voice disorders. Combining this new empirical evidence with information from biomechanical studies concludes that the most parsimonious model for speech production, explaining all these findings in one unified set of mathematical assumptions, is a stochastic nonlinear, non-Gaussian model, which subsumes both Gaussian linear and deterministic nonlinear models.

As a case study, to demonstrate the engineering value of nonlinear signal processing techniques based upon the proposed biomechanically-informed, unified model, the study investigates the biomedical engineering application of disordered voice measurement. A new state space recurrence measure is devised and combined with an existing measure of the fractal scaling properties of stochastic signals. Using a simple pattern classifier these two measures outperform all combinations of linear methods for the detection of voice disorders on a large database of pathological and healthy vowels, making explicit the effectiveness of such biomechanically-informed, nonlinear signal processing techniques.

Acknowledgements

This thesis is first and foremost dedicated to my long-suffering wife Maya, without her emotional support, this thesis would have been impossible. I owe her an impossibly large debt.

It has been a privilege to have been supervised by some great scholars in Oxford. I direct unreserved gratitude to Patrick McSharry, who joined in later but whose patient and steadfast advice and guidance has been substantial and critical. Irene Moroz, particularly in the early stages, helped with my induction to research life at Oxford and later helped to keep me on track and on time. Steve Roberts at the engineering science department has been a major source of knowledge and enthusiasm, our wide-ranging conversations on topics in engineering mathematics were thrilling and I am particularly grateful for his eternally positive attitude which paid off during difficult periods.

The willingness of several people in Oxford to act as critical reviewers of this work has been invaluable. For this, David Allwright, Nick Hughes and Gesine Reinert all deserve special thanks. Outside Oxford, Liam Clarke at the London School of Economics and Gernot Kubin at the Technical University of Graz in Austria have been of great help. Dan Sinder's numerical simulations of aeroacoustic noise were also very helpful. I am indebted to Martin Burton and Declan Costello at the Radcliffe Infirmary in Oxford and Adrian Fourcin at University College London for invaluable clinical advice. Conversations about mathematical and engineering topics with other research students were an endless source of inspiration: Christina Orphanidou, Reason Machete and Oscar Martinez-Alvarado all helped to make the subject enjoyable. I am lucky to know a small army of professional proof-readers: Jacky Barrett, Julia Sadler and Sara Jansson all deserve special thanks in helping to uphold the quality of the text and keep the typos at bay.

I have been extraordinarily privileged to receive financial support during this work from the EPSRC through the mathematics department, for this I am grateful to Sam Howison and John Ockenden for persuading the department to fund me over many other talented students. Finally, I wish to thank my brother Crispin at Leeds University for persuading me to follow him down this academic path.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This thesis is an investigation of how best to use mathematics to analyse certain signals utilising software, in this case *speech signals*. Motivating this investigation are the possibilities opened up by new mathematics and new technology. Existing approaches have a long history but the conceptual foundations were laid down long before the mathematical and technological advances, and novel evidence of the kind produced and described in this thesis, were available. These advances suggest close scrutiny of the mathematical foundations of current models and techniques. As a result of this critical examination, the specific information about the mathematical limitations of current techniques can be uncovered. Armed with this information, it is then possible to create new techniques, based upon more appropriate mathematical models, that do not suffer from these limitations.

1.1 Speech Models

> "Since all models are wrong, the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Ockham he should seek an economical description of natural phenomena." – George Box [1].

Mathematical models of reality are just that – models [2]. They are not reality any more than a map is the territory it represents. Nonetheless, mathematical models can be successful at representing physical situations, since they can produce outputs that are consistent to a degree with observational data from those situations. These models are enormously useful in *engineering* [1] – in which the application of these models facilitates many practical ends. Of relevance here are mathematical models that form the basis of certain engineering data processing methods, as is the case in digital speech processing, the subject of this thesis.

[1] Engineering: "The profession of designing and constructing works of public utility, such as bridges, roads, canals, railways, harbours, drainage works, gas and water works, etc." [3]. Engineering as conceived in this thesis is much broader than this and includes, for example, telecommunications and information engineering: building efficient machines for transmitting, coding, processing and manipulating data.

Two different approaches to mathematical modelling can be distinguished: *first principles* and *data-driven*. The former often proceeds by organising known, fundamental processes (the first principles) that the modeller believes to be relevant into hypothetical mechanisms [4]. These mechanisms are assumed to be responsible for generating the observations of the physical situation. Verification of these mechanisms can be obtained by comparing the output of the model with the observations. Refinements to the model may then be necessary if the observations do not match the model output, but in general, two different models may be indistinguishable by their resulting outputs alone. This choice of models requires the (implicit) invocation of the *principle of parsimony*, otherwise known as Ockham's razor [3]: preference should be given to the simpler of two competing explanations. In this way, the skilful choice of mathematics can be used to create simple models that behave in a manner consistent with observations from the physical situation. It is also possible to make predictions of future behaviour of the physical situation if enough confidence can be obtained in the match between model output and observations [4]. Such mathematical models apply to a vast range of physical circumstances of practical importance, and have the power to explain this large range of phenomena with just a few assumptions [5]. For more information about first principles modelling, see [6, 7].

Data-driven modelling involves very little information over and above the observational data itself. A simple "black-box" model is written down that is assumed to be general enough to be capable of representing the data [8, 6]. It will usually have a number of *free parameters* that are not known in advance. This model is encoded into a program that can be run on a computer. The observational data is transformed into a format suitable for storage in computer memory. Then a 'matching' process takes place whereby the parameters of the model are altered by some *fitting algorithm* so that the model output matches the observational data as closely as possible. If the model is not a good fit to the data, then more free parameters may be added to the model or the assumptions in the model may be changed in an attempt to improve the match. As above, two different models may fit the data equally well, and in this case the simpler model should be preferred. If enough confidence can be gained in the match of model output to observations, predictions may also be possible. Data-driven modelling is described in more detail in [1, 8].

As discussed above, an essential part of the process in both approaches is the application of the principle of parsimony. Any one set of observations cannot be expected to represent the full range of behaviour that the physical phenomena of interest may potentially exhibit. A trade off has to be achieved between selecting highly complex models

that can match a given set of observations extremely well, and selecting simpler models that can *generalise* well to unavailable observations.

There are no set rules to constructing mathematical models for physical situations [4], and all models of these situations will to a greater or lesser extent remain *imperfect* when compared to actual observational data [9]. Furthermore, how to characterise precisely the simplicity or complexity of a model in a rigorous mathematical sense is an active and open area of research, and general results are few. It is important also to decide in advance what information one wishes to obtain from the model in the first place [4]. Thus, in choosing a model to represent a physical situation, the purpose of this model must also be a guiding factor in the appropriateness of the mathematics.

Although this thesis is partly concerned with the selection of appropriate data-driven models for speech, much is known from first principles about the *biomechanics* of speech production, and common sense suggests that this is additional, valuable information that can be used to facilitate the construction of data-driven models. If such a model can be constructed that is also as consistent as possible with the known biomechanics, this lends additional authority to the modelling choices. The biomechanics then informs the choice of processing methods for digital speech data.

1.2 Speech Technology

Humans have evolved a variety of different methods for communication. Principal amongst these are written text, diagrams, figures and other illustrations, gestures made using the body, singing and the spoken word, as transmitted in the sounds of speech. People make physical speech sounds for communicating ideas using their vocal organs, comprising the lungs, *larynx* (the voice box), the *vocal tract* (the throat), the mouth including the tongue and lips, and the muscles that move these organs, and finally the nerves that orchestrate that movement. Speech sounds are communicated from the speaker's mouth to the listener's ear through fluctuating sound pressure variations in air [10].

Even ignoring the conversational interaction between speaker and listener, speech, as a human behaviour, is a highly intricate activity, involving precise synchronisation between vocal organ muscles and, since a speaker hears their own speech sounds, feedback from sensory organs including the ear. One well-established theory of the organisation of speech is that the sounds are considered to comprise a basic catalogue of units called *phonemes* that form the lowest level of a hierarchy that groups phonemes together into *syllables*,

which are themselves grouped together to make words, which are then grouped together to make clauses and sentences [10].

Speech is a focus of scientific study in the speech sciences of *phonetics* and *linguistics*. Some specialist subdisciplines of psychology study speech, and there exist many other subdisciplines and inter-disciplines in the speech sciences. Phonetics is defined as the study and classification of speech sounds, especially with regard to the physical aspects of their production [3]. This includes some *biomechanics* [10]. Linguistics can be broadly defined as the study of human language: how it is structured and used to represent and communicate ideas and meaning, and how it is formed and decoded [3]. Since the ability to communicate by speech is critically important to normal human functioning, there are medical specialisms that deal with the various types of voice disorder that can arise due to disease, accident or the result of medical intervention. For example, *otolaryngology* includes the study and practice of the diagnosis and treatment of voice disorders which arise due to biological problems including larynx tissue disease or damage to the nerves that control the muscles of the larynx, which can have a profound effect on the ability of the patient to produce speech sounds [11].

Speech has received considerable attention in *telecommunication engineering*, and speech technology has become a ubiquitous part of modern life. The ability to transmit, store, reproduce, analyse and synthesise speech using machines has enormous practical value. The most visible example of speech technology in everyday usage is perhaps the oldest: the telephone (or, more recently, the wireless mobile telephone network). Originally making use of analogue electronics (namely, conductive wires, transformers, resistors, capacitors, the vacuum valve and later the transistor), the basis of the telephone system was the amplification and transmission of speech sounds encoded into fluctuating electronic currents (called a *signal* [12]) via a microphone at the transmitting end. The receiver contains a loudspeaker, which converts the transmitted signal into a reconstruction of the speech sounds at the transmitter [13].

Speech technology also has uses as tools for speech scientists and medical practitioners. In phonetics, for example, the *spectrogram* is fundamental to the analysis of speech sounds [14, 10]. The spectrogram is a visual representation of the speech signal allowing the user to see the breakdown of the speech signal into *frequency components* [12] that occur within each phoneme, and how these frequency components change in time. The particular arrangement of the frequency components in a phoneme is a strong indicator of the associated phonetic category [14]. Under certain restrictions, similar and related

analysis of speech sounds produced by patients can be a valuable aid to the diagnosis and progress monitoring in the course of medical treatment for voice disorders [11].

An important change occurred in speech technology, towards the end of the previous century, in the widespread introduction of *analogue-to-digital* (ADC) and *digital-to-analogue convertors* (DAC), allowing the storage, transmission and processing of purely *digital* signals which have significant engineering advantages over analogue signals [12]. This move to digital encoding of speech signals [15] has coincided with ever-increasing computer speed, computer memory and storage size, the rapid increase in the volume of digital data that may be transmitted through telecommunication systems and the global interconnectivity of the internet. Some of these developments in hardware have enabled software techniques to be applied directly to the processing of such digital speech signals, including digital *speech compression*, which is fundamental to mobile and internet speech telephony, and advanced technologies for human-machine interaction such as *automatic speech recognition*.

1.3 Mathematics and Speech Signal Processing

Underlying the kind of software utilised by most digital speech technology are appropriate mathematical models and methods. New mathematics may well have an important role to play in more advanced software and technologies for digital speech processing, such as compression, storage, transmission, analysis and manipulation both by the mathematical formalisation of new developments in speech science, and by theoretical advances in mathematics itself. Such formalisations may then be programmed in software and applied to digital speech signals, to exploit the power implicit to these scientific advances. It can be expected that, within reason, the capacity of the computer hardware required to implement such new methods will generally become widespread.

The mathematical models and methods underlying most current digital speech technology in common usage are the set of techniques typically encountered in the engineering subdiscipline of *digital signal processing* [12], although there is some overlap with techniques from other areas such as *statistical time series analysis* [16] and *information theory* [17]. Many mathematical models and theorems comprise such techniques, and new ones are being included all the time. The core set of classical ideas of *linear, time-invariant (LTI) systems theory* [12] is thoroughly investigated and understood. Nonetheless, simply because they are well understood does not automatically imply that they are appropriate

for all digital signals. This is because although they can be applied to process signals, transform them, or extract information, if the signals are fundamentally incompatible with the mathematical assumptions underlying the technique, then such application is flawed. For example, if the signal does not obey the assumptions of LTI systems theory, then information extracted from these signals using techniques based around such theory is suspect. Therefore, important questions must be settled about the validity of any mathematical technique before it is used with a signal, if the resulting information is to be meaningful and reliable.

In the context of digital speech technology, a more recent innovation is the introduction of methods from the emerging discipline of *nonlinear time series analysis*: theory and mathematical techniques for the analysis and processing of signals that are assumed to have derived from some mathematical model that cannot be completely described within the framework of LTI systems theory [8]. Because the discipline deals with signals, and the assumptions underlying the techniques are not linear, this area could also be described as a form of *nonlinear signal processing*. [2] As a relatively new discipline, there are many outstanding open problems, and by contrast to LTI systems theory and associated signal processing algorithms, little is known about the reliability, robustness, performance and appropriateness of these new techniques in general. Similarly, there are many open problems concerning how these techniques can be usefully applied to digital speech processing and analysis.

Nonetheless, some initial research work conducted in speech science and engineering communities have shown that nonlinear signal processing methods could offer important advantages over and above the classical LTI techniques [18, 19, 20, 21, 22, 23, 24]. From this it appears that nonlinear signal processing approaches are promising, in that they may well offer explanatory power in speech science. Such new scientific understanding could also have engineering applications to speech technology, and hence improve current speech processing software. The overall aim of this study is to investigate the fundamental appropriateness of new mathematical models and methods for analysing and processing speech signals, and explore their application in the context of a typical speech technology application.

[2] In this thesis, by "nonlinear" methods we mean methods not conforming to LTI systems assumptions. This includes non-Gaussian methods, therefore.

1.4 Research Statement and Summary of Results

The research questions stem from the following argument put forward in this thesis:

"Nonlinear signal processing methods are valuable for digital speech analysis, barring important limitations."

This leads to the following set of hypotheses:

- Based upon knowledge in speech science and evidence from speech signals themselves, the mathematical assumptions of LTI systems theory cannot represent all the dynamics of all speech,
- LTI systems theory is only appropriate for some limited cases of speech phonemes,
- Nonlinear, non-Gaussian stochastic assumptions are particularly important to some speech phonemes, and some disordered speech,
- Appropriate nonlinear signal processing methods are, in some aspects, better than LTI systems methods in voice disorder detection,
- Nonlinear, non-Gaussian assumptions for speech signals offer a simplified, mathematical framework that explains more phenomena with fewer assumptions than classical LTI assumptions, and as such can offer improvements in engineering reliability, robustness and performance,
- Not all the standard, nonlinear time series analysis algorithms are robust enough to be of practical value to speech processing, so that new, nonlinear algorithms are required.

The first three are "foundational" scientific statements of the validity, appropriateness and reliability of nonlinear time series analysis methods applied to digital speech signals, and are one focus of this thesis. Another focus of the thesis is the development of the last three statements, with particular reference to an application case study from biomedical engineering in otolaryngology.

It will thus be demonstrated, through a succession of theoretical arguments and experimental results, that certain nonlinear signal processing methods can indeed be valuable, and make a practical contribution to speech signal processing, under the right conditions. This study supports this argument by combining:

- Information from speech science,

- Evidence gained from rigorous statistical tests,
- The principle of parsimony, and,
- Performance comparisons against classical LTI signal processing methods in an example engineering application.

1.5 Scope of the Thesis

This thesis is essentially a systematic investigation of the merits of nonlinear, non-Gaussian signal processing approaches to digital speech signal analysis, and signal processing is usually considered to be an engineering discipline. In order to do justice to the many issues raised by the use of nonlinear digital signal processing methods on speech signals, the thesis is essentially limited in scope. It does not address issues of the purer foundations of the mathematical concepts it uses. Similarly, whilst making use of certain results from speech science (such as biomechanical modelling and phonetics), it does not address issues of the validity of the first-principles mathematical modelling choices in these domains in depth. Also, although it presents an application example from the medical science of otolaryngology, it is not directly concerned with confronting the many clinical issues involved.

Thus, the thesis is limited to the choice and application of certain mathematical concepts and algorithms to processing real data in the form of digital signals. In order to tackle the problems raised, there is a significant mathematical component. This requires some mathematical concepts from LTI systems theory, probability, stochastic processes, nonlinear dynamics, information theory, and statistics. It also makes use of some previously developed biomechanical models.

1.6 Summary of Contributions

The thesis reports several contributions to the state of the art of knowledge in the discipline of nonlinear digital signal processing, of which it forms a part:

- The systematisation and improvement of a statistical surrogate data test for nonlinearity/non-Gaussianity in digital signals,

- Application of this test to the largest database assembled to date, assessing the evidence for and against nonlinearity/non-Gaussianity in the predominant classes of speech phonemes and in disordered voices,
- The introduction and justification for a new, parsimonious, nonlinear/non-Gaussian model for speech signals, and,
- The development of a novel method for characterising the nonlinear/non-Gaussian dynamics represented in a signal, and the case study application of this method to the automated detection of voice disorders.

1.7 Structure of the Thesis

The thesis begins, in Chapter 2, with a review of the relevant biomechanics of speech production and phonetics. This review discusses existing models of vocal tract and vocal folds and how they interact. It then examines models of turbulent airflow phenomena in speech. Next, the mathematics and practice of classical, linear, digital speech signal processing is reviewed in Chapter 3. This chapter is a detailed exposition of the well-known theory of LTI systems, with the focus on making the underlying mathematical assumptions explicit, since these assumptions will be the subject of subsequent critical examination.

The thesis then moves on to an overview of the mathematical foundations of nonlinear time series analysis in Chapter 4. The first part of this chapter is a review that explains the core set of mathematical assumptions of nonlinear time series analysis which lie outside those of LTI systems, and are thus a generalisation of LTI systems theory. The relevance of these assumptions and their consequences to speech production is an important aspect of this thesis, and in order to assess this relevance rigorously, the last part of the chapter develops a new surrogate data test against the appropriateness of LTI systems assumptions for real speech signals. This new test overcomes some of the flaws of existing surrogate techniques.

All the latest evidence, and new evidence presented for the first time in this thesis against the appropriateness of LTI systems for speech is gathered together in Chapter 5. This evidence is obtained from real speech signals analysed using the new surrogate data test developed in the previous chapter. In conjunction with the principle of parsimony discussed in the introduction, and information from first-principles speech models reviewed

in earlier chapters, this evidence justifies a new model for nonlinear speech signal processing applications, introduced in the final part of this chapter.

Subsequently, Chapter 6 introduces a novel practical algorithm for detecting and characterising the existence of the nonlinear structure of speech identified in earlier chapters, and demonstrates the effectiveness of this algorithm in the context of a clinical application. This chapter demonstrates that the nonlinear speech signal processing methods, developed upon the basis of the new nonlinear model of speech signals, can outperform traditional LTI systems methods of classical linear digital signal processing, thus further justifying the new speech signal model.

Chapter 7 is a discussion of the overall thesis, drawing conclusions and making tentative generalisations to other nonlinear signal processing applications. It ends with suggestions for future work in the field of nonlinear signal processing based upon the methods and techniques introduced in this study. The appendices contain additional details including mathematical proofs of results referenced in the body of the thesis.

CHAPTER 2

Brief Overview of Biomechanics and Phonetics

As discussed in the introduction, first-principles models of speech production, although they necessarily entail simplifying assumptions, contain valuable information that can be used to inform our choice of data-driven models. This chapter will therefore discuss and explore the behaviour of some of the most well-established models of speech sound production that have been developed in the speech science communities. This will help to shed light on the basic biomechanics at work in speech production, this information acting as a guiding principle in later chapters. The focus of this thesis is the development of novel nonlinear signal processing algorithms which are guided by biomechanical knowledge, rather than new or improved first-principles models. Therefore, this chapter presents a necessarily brief account of the relevant biomechanics, with pointers to more in-depth treatments in the published literature. It will also introduce some basic concepts from phonetics which will provide similarly useful information and a context for the more detailed investigations of particular speech sounds covered in this thesis.

2.1 Anatomy

The human vocal apparatus is comprised of three main organs: the lungs, the vocal folds and the vocal tract [10]. The lungs can be considered as a flexible bag with a tube (the windpipe or trachea) attached that can be expanded by muscles to suck air in or contracted to push air out. The vocal folds, situated in the larynx, are a pair of band-like soft membranous tissues that can be positioned by muscles in the larynx. During *voiced* sounds such as vowels (see §2.3), they are stretched across the larynx and act as a slit-like constriction to the airflow from the lungs that vibrates when air is blown over them. Finally, the vocal tract consists of three coupled cavities (pharyngeal, oral and nasal). These cavities resonate at particular frequencies which are affected by the position of the jaw, tongue, lips and the cartilaginous *velum* (or soft palate) which can be raised to shut off the nasal tract, stopping air from flowing out of the nose [10]. Figure 2.1 is a diagram

showing how these organs are arranged inside the head and neck.

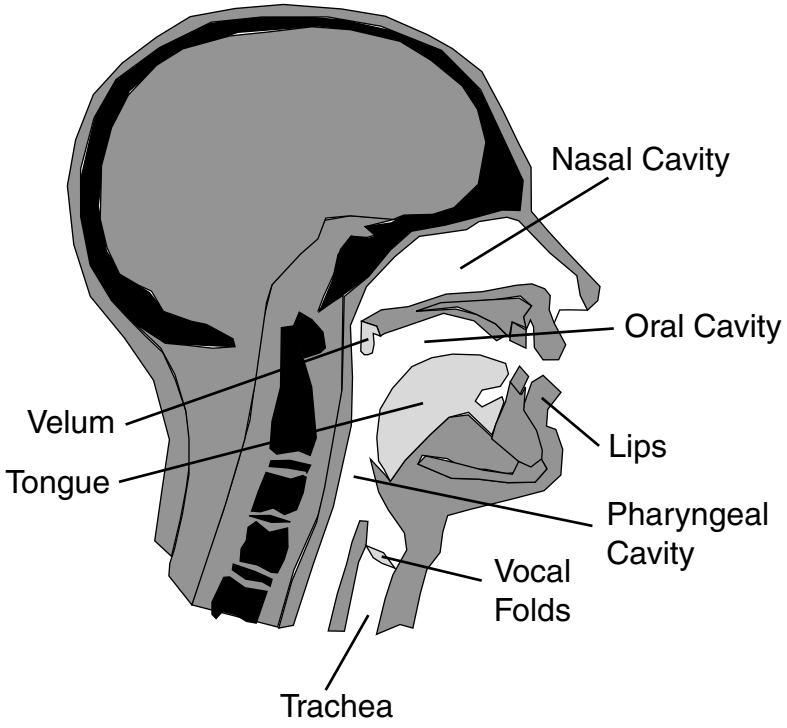

Figure 2.1: Arrangement of the vocal organs inside the head and neck.

2.2 Review of Biomechanical Models of Speech Production

Focusing on the physical mechanism responsible for the generation of audible speech sounds, there are several dynamical variables of interest. These are the air pressure, air flow rate and expansion and contraction of the various components of the larynx. Most speech production models divide the system into two major subcomponents, the vocal tract and the vocal folds [25]. As we will show, this appears to account successfully for the mechanisms of audible speech in vowels. This, however, ignores the noise-like sound produced during speech due to "turbulence" in the airflow: more recent models incorporate such effects as well, and are thus able also to model consonants and breath noise.

2.2.1 The Vocal Tract – Lossless Acoustic Tube Model

Vocal tract modelling has a long history. At least as far back as the 1700's, with the pioneering work of von Kempelen in his mechanical speaking machine [26], it was realised

that the essential role of the vocal tract is that of a (mainly passive) *acoustic resonator*, although detailed mathematical models were only developed later.

One successful model of the vocal tract is the *lossless acoustic tube* model [13]. We will follow the development of this model here. The basic structure of the model for the vocal tract is an acoustic tube of slowly-varying cross-section with rigid walls. The vocal folds are attached at one end of the tube, and the lips are the opening at the other end. The tube is assumed to have cross-sectional area that varies smoothly along the length of the tube. All energy loss mechanisms inside the tube are ignored [13]. It is assumed that only *planar* acoustic wave motion is important, and all nonlinearities are small enough to be neglected.

Under these assumptions, the governing equation in the tube is the quasi one-dimensional, linear *acoustic wave equation* [27]. The relationship between pressure $p(x,t)$ and flow rate $u(x,t)$ is:

$$\begin{aligned} -\frac{\partial}{\partial x}p(x,t) &= \frac{\rho}{A(x)}\frac{\partial}{\partial t}u(x,t), \\ -\frac{\partial}{\partial x}u(x,t) &= \frac{A(x)}{\rho c^2}\frac{\partial}{\partial t}p(x,t). \end{aligned} \tag{2.1}$$

with $A(x)$ representing cross-sectional area, c the speed of sound in air, x the spatial co-ordinate running along the axis of symmetry of the tube, t time and ρ the constant equilibrium density of the air. The boundary conditions will be determined later.

Our goal in solving this model will be to relate flow rate fluctuations at the vocal folds to corresponding changes in flow rate at the lips, determined by the *acoustic transfer function* of the tube model. For *linear systems* such as (2.1), the *superposition principle* holds: any linear combination of solutions of the equation is also a solution. Therefore the special approach of *Fourier transforms* may be used, representing the solution in terms of a sum of weighted *complex exponentials* of a given radian frequency $\omega = 2\pi f$ (in units of radians per second where f is in Hertz). Such *frequency analysis* can be carried out by inserting exponential functions into the equations (2.1). Subsequent calculations determine the required transfer function in terms of these complex exponentials. [1]

The pressure and flow rate are expressed in terms of complex exponentials:

$$p(x,t) = P(x,\omega)e^{i\omega t}, \quad u(x,t) = U(x,\omega)e^{i\omega t}, \tag{2.2}$$

[1] Such complex exponentials are actually *eigenfunctions* of linear systems such as (2.1) [5].

so that equations (2.1) become the pair:

$$\begin{aligned} -\frac{d}{dx}P(x,\omega) &= \frac{i\omega\rho}{A(x)}U(x,\omega), \\ -\frac{d}{dx}U(x,\omega) &= \frac{i\omega A(x)}{\rho c^2}P(x,\omega). \end{aligned} \tag{2.3}$$

Eliminating the pressure variable from equations (2.3) obtains the second-order *Webster's horn equation*:

$$\frac{d^2}{dx^2}U(x,\omega) - \frac{1}{A(x)}\frac{d}{dx}A(x)\frac{d}{dx}U(x,\omega) + \frac{\omega^2}{c^2}U(x,\omega) = 0. \tag{2.4}$$

It remains to discuss the boundary conditions at both ends of the tube to complete the model. It is assumed that the tube is forced at one end by the vocal fold oscillation. We let the function $U_{\mathrm{f}}(\omega)$ denote the amplitude of the driving flow rate of the vocal folds at a given radian frequency ω. This forms the first boundary condition for the tube end at $x = 0$.

An expression for the *acoustic impedance* $Z(\omega)$ (the ratio of air pressure to air flow rate) of the radiative opening derived in [28] provides a second boundary condition at the lip end. The (frequency-dependent) real part of $Z(\omega)$, called the *radiation resistance*, is proportional to the amount of energy in the acoustic tube absorbed by the surrounding medium, while the imaginary part is the amount of mass loading of the surrounding air on the acoustic tube. The boundary conditions for equation (2.4) are then:

$$\begin{aligned} U(0,\omega) &= U_{\mathrm{f}}(\omega), \\ P(L,\omega) &= Z(\omega)U(L,\omega), \end{aligned} \tag{2.5}$$

where L is the length of the tube. Using equation (2.1) above

$$P(x,\omega) = \frac{i\rho c^2}{\omega A(x)}\frac{d}{dx}U(x,\omega), \tag{2.6}$$

the lip end boundary condition becomes:

$$\left.\frac{d}{dx}U(x,\omega)\right|_{x=L} = \frac{\omega A(L)}{i\rho c^2}Z(\omega)U(L,\omega). \tag{2.7}$$

The vocal tract at the lip end is modelled as a simple piston in an *infinite baffle* [27], that is, the equivalent of a tube opening at one end on to the surface of an infinite flat plane. All the effects of interaction between the radiated sound and facial features are ignored. This approximation also treats the normal flow rate as uniform over the tube area. Then the acoustic impedance function $Z(\omega)$ has the following form [28]:

$$Z(\omega) = \frac{\rho c}{\pi r^2}\left[R_1(2kr) + iL_1(2kr)\right], \tag{2.8}$$

where

$$R_1(x) = 1 - \frac{2J_1(x)}{x}, \quad L_1(x) = \frac{2\mathbf{H}_1(x)}{x}, \tag{2.9}$$

and r is the radius of the (circular) lip opening, with $k = \omega/c$ the *wavenumber*. The function $J_1(x)$ is the Bessel function of the first kind, and $\mathbf{H}_1(x)$ is the first *Struve function* [29].

Finally, the required *transfer function* $H(\omega)$ of the tube evaluated at an arbitrary frequency ω is:

$$H(\omega) = \frac{U(L,\omega)}{U(0,\omega)} \tag{2.10}$$

In order to solve the boundary value problem to find the transfer function, we need to specify the area function $A(x)$. The cross-sectional area of the vocal tract can be obtained from X-ray or MRI (magnetic resonance imaging) [30]. Typically these measurements result in a series of point area measurements along the length of the tract which must somehow be interpolated to create the smooth area function $A(x)$. In this study, a 9th order polynomial was fitted to published area measurement data obtained by X-ray measurement, [2] after [13]. Figures 2.2 and 2.3 show the measured and interpolated area functions for two different vowels. [3]

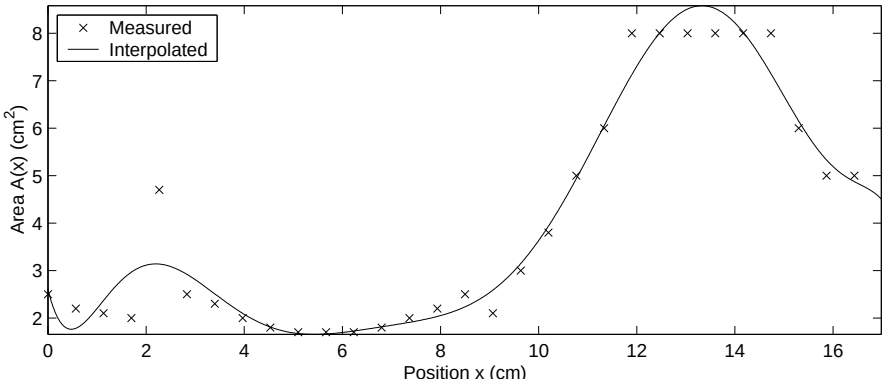

Figure 2.2: Measured and interpolated vocal tract area functions for vowel /aa/.

Since the system is linear and obeys the superposition principle, we can replace the vocal fold boundary condition with a delta function, or *impulse* in time, and solve the system to find the "impulse response" solution. Now, any arbitrary boundary condition function may be expressed as a linear superposition of delta functions weighted by this boundary function evaluated at each instant in time. Subsequently, the solution to the

[2] This interpolation method and order were chosen to provide the best compromise between satisfying the smoothness assumptions of the model and capturing the details of the 30 area measurement points. The polynomial was fitted using least-squares regression.

[3] A description of the vowel codes used in this study is given in §2.3.

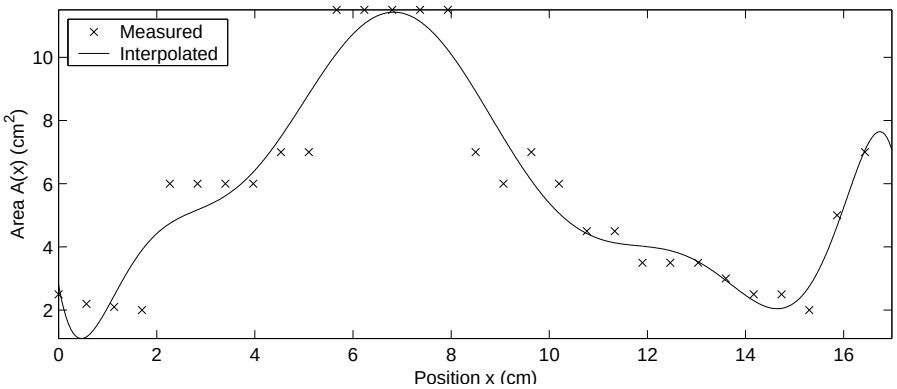

Figure 2.3: Measured and interpolated vocal tract area functions for vowel /eh/.

system with this arbitrary boundary condition may be obtained by convolving the impulse response with the boundary condition. Essentially, the impulse response contains all the information about the structure of the solution to the differential equation (2.4) with the given lip end boundary condition, so that we can solve for the transfer function $H(\omega)$. This account is a simplification of a somewhat delicate theory, for more detailed information see, for example [5].

The equivalent to this delta function in the Fourier representation is the constant function $U(0, \omega) = 1$, so that the transfer function at a given radian frequency is:

$$H(\omega) = U(L, \omega). \tag{2.11}$$

Unfortunately, this problem as posed is not solvable analytically: here an approximate solution may be obtained using a numerical method. Thus the equation was discretised spatially by replacing the derivatives with finite differences, and forming a system of linear equations to be solved for flow rate at each discretised point in space. The numerical calculations are detailed in Appendix §A.1.

Figure 2.4 shows the *power spectrum* of the resulting *frequency response* $|H(\omega)|^2$ where $\omega = 2\pi f$ of the model evaluated over a range of frequencies in which the model can be considered accurate, for the vocal tract configurations $A(x)$ of two different vowels. The vocal tract length was $L = 17$cm, typical of an adult, and a mouth opening radius of $r = 1$cm. The other parameters were $c = 343\,\mathrm{m\,s^{-1}}$ and $\rho = 1.13\,\mathrm{kg\,m^{-3}}$.

Note that for a uniform tube (with constant area function $A(x) = \text{const}$) of length 17cm closed at one end and open at the other, the natural resonance frequencies are at $f_n = (2n-1)c/(4L), n = 1, 2, 3, \ldots$ which evaluates to approximately $f_n = 500, 1500, 2500 \ldots$Hz. This accounts for the general pattern of resonant peaks seen in figure 2.4 – the modifications

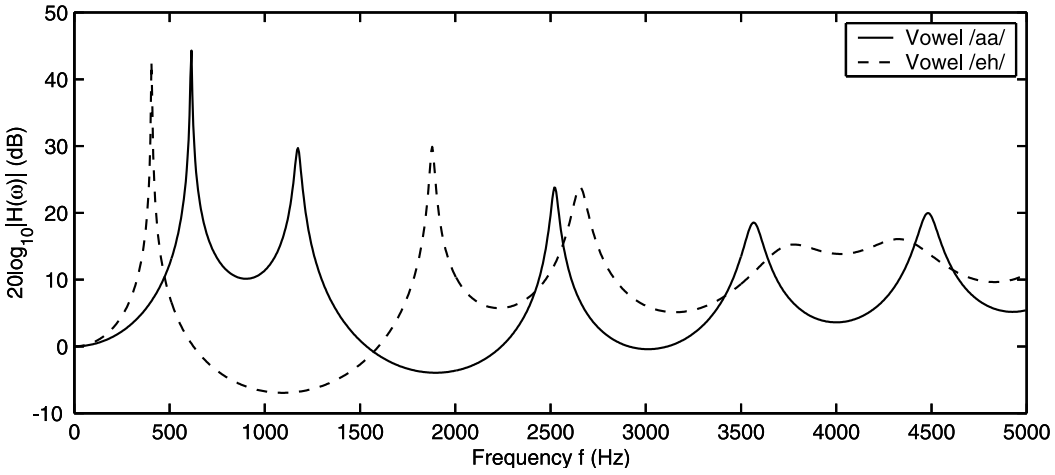

Figure 2.4: Frequency responses in decibels with $\omega = 2\pi f$ for frequency f in Hertz of a varying area acoustic tube model of the vocal tract with infinite plane baffle acoustic open termination at the lips, for two different vowel configurations, /aa/ and /eh/.

in the frequency location of these resonances are due to the area variation $A(x)$, and the decreasing sharpness of the peaks with increasing frequency is mainly an effect of the radiative lip opening. In phonetics these resonant peaks in the transfer functions are called *formants*. When the tongue and other *articulators* such as the lips change position they alter the geometry of the vocal tract, hence changing the frequency and sharpness of these resonances. It is mostly by these changing patterns of resonances that we are able to distinguish one vowel from another, discussed in more detail in §2.3. As can be shown using *digital formant analysis*, discussed in Chapter 3, this changing patterns of resonances is very similar to that observed in real speech signals.

We now discuss the limitations of this model due to the choice of modelling assumptions. This tube model has no loss mechanisms at all, which is clearly not very realistic. There are many forms of losses that can occur in acoustic systems such as this. For example, the walls of the vocal tract are not perfectly rigid and so can vibrate in sympathy with the air in the tube and dissipate energy, or the air itself will lose energy due to viscous friction, but extensive investigations have shown that the most important effect is that of wave energy loss that occurs due to the lips being opened [13].

Only planar wave motion has been considered in this model [13]. Non-planar waves in a (constant $A(x)$) cylindrical tube model such as this are *evanescent* at frequencies below the first non-planar mode *cut-on frequency* of $f_{\text{cut}} = 1.84c/(2\pi r)$, where r is the radius of the tube [27]. That is, non-planar wave modes decay in amplitude rapidly with distance along the tube and so their propagation can be neglected [27]. Given an average vocal tract area

of approximately 5cm^2, this leads to an approximate value of $f_{\text{cut}} \approx 8\text{kHz}$. Therefore, to investigate higher frequency behaviour of this model we would need to include non-planar wave motion as well.

The lip end boundary condition of a simple piston in a tube opening out onto an infinite plane baffle is reasonable when the size of the lip opening is small compared to the size of the head [27], but it ignores all the diffraction effects of a tube opening out on to the surface of a sphere, which is, of course, a much more realistic representation of the shape of the head. At around 1300Hz the acoustic wavelength is approximately 26cm, which is roughly the size of the head. Therefore, above this frequency the infinite baffle is a good approximation; below this the approximation is worse. However, in [31], three different models were compared: the simple piston in an infinite baffle, a simple piston in a spherical baffle and a pulsating sphere, concluding that the piston in an infinite plane baffle model is reasonable for the physical dimensions and frequencies considered in this thesis.

The straight tube model is a notable simplification of real vocal tracts that are actually curved. This means that there will be reflection and refraction at the bend for planar waves. Secondly, the assumption about cylindrical tubes is not that realistic: 3D MRI studies show that the vocal tract departs significantly from this geometry in some places and for some particular tongue configurations. Whether these affect the resonances significantly would require a different geometrical model. Thirdly, the trachea and lungs are connected to the vocal tract when the vocal folds are open and the vocal folds are of course coupled to the tract: these effects have been explored [15] and have been shown to alter the frequency of the lowest resonance.

However, attempting to capture all these detailed effects would complicate the model and analysis unnecessarily, whilst leading to differences that would be very hard to verify from real acoustic speech signal measurements.

It must also be explained that this model does not include the effect of resonances in the nasal tract. For the production of most phonemes the velum is raised so that the nasal tract is not coupled to the rest of the vocal tract. Therefore this omission from the model does not significantly affect the accuracy. However, for certain phonemes (see §2.3) the nasal tract plays a critical role in generating the appropriate resonance patterns. In this thesis we will only be concerned with those phonemes for which the nasal tract is not coupled to the rest of the vocal tract.

2.2.2 The Vocal Folds – Two-Mass Model

There are two basic, relevant components to consider in a model of the vocal folds. The first is the vocal fold tissue (consisting of the *mucosal membrane* over a cartilaginous structure), and the second is the air flowing through that structure. A simplified picture of the vibratory mechanism of the folds in terms of dynamical forces in balance is that of air pressure exerted by the lungs on the closed vocal folds which, although under tension, are forced apart from the bottom. Air then flows freely through the vocal folds. Since the air flow rate is now large, the pressure is lowered in the larynx and this drop in pressure, combined with the elastic restoring force of the vocal fold tissue, wins out over the inertia of the tissue mass. The folds therefore snap back together sharply, cutting off the airflow abruptly. On closing, an impact restoration force acts in the opposite direction due to the vocal fold tissue now being in compression. The viscous damping of the vocal folds would keep them closed, but the air pressure from the lungs builds up and forces them apart again, and the cycle repeats. Thus the vocal folds act as a vibrating valve, disrupting the constant airflow coming from the lungs and forming it into regular puffs of air.

In general the governing equations are those of fluid dynamics coupled with the elastodynamics of a deformable solid. In one approach to solving the problem, the airflow is modelled as a modified quasi-one-dimensional Euler system which is coupled to the vocal fold flow rate, and the vocal folds are modelled by a *lumped* two mass system [32]. Such an approach requires significant computational resources. A somewhat simpler, *semi-continuum* approach models the vocal fold tissue as two lumped masses. Since the region near the vocal folds is much shorter than the acoustic wavelength of the vocal tract, the air in this region can be considered incompressible such that the incompressible Navier-Stokes equations can be used [33]. An even simpler model, requiring many fewer degrees of freedom than the continuum models, is the lumped sixteen mass model of [34]. However, all these models are complex and obscure the basic mechanisms that account for the vibration of the vocal folds. Furthermore, it has been shown (using PCA decomposition into eigenmodes [4]) that only the first two or three vibrational modes of the vocal folds dominate [35]. Three eigenmodes always account for 90% of the variance of the vibration, justifying simpler models.

Simple mathematical modelling of the vocal folds has focused on capturing some im-

[4] If the dynamical variables are taken together to represent vectors, then PCA (Principal Components Analysis) can be used to find a smaller, linear subspace of the original vector space onto which to project these dynamical vectors. This subspace is spanned by a set of new orthogonal basis vectors.

portant observed effects. Primarily, these are:

- Self-sustained vibration,
- The relationship between the frequency of vibration and the tension in the vocal folds,
- The overall "waveshape" of the air flow rate against time which falls very quickly but rises more slowly in each cycle,
- The percentage of the duration of each cycle of vibration in which the vocal folds are open,
- The smallest lung pressure needed to maintain self-sustained vibration, and
- The *mucosal wave*: synchronised wave-like motion running up the vertical inside faces of the vocal folds [36].

A popular model that addresses these effects is the two-mass model in [36], further simplified in the asymmetric [37, 38] and symmetric model of [39], which we will describe here. Figure 2.5 shows the simplified model configuration. For a comprehensive overview of the modelling assumptions and derivation of the equations of motion of this model see [39].

In this model on each half the vocal folds are divided into two separate masses connected by a viscoelastic spring k_{c}, giving four masses in total. However, due to the symmetry only one half of the system is modelled. The other half is assumed to behave identically but with motion in the opposite direction reflected about the vertical line of symmetry. The first, larger mass is driven by static air pressure from the lungs, Bernoulli forces inside the larynx and partly by the second mass through the connecting spring k_c. Dissipation due to the viscous damping r_1, r_2 implies that energy is provided to m_1 by Bernoulli forcing F. It also implies that any energy imparted to the second mass comes only from the motion of the first transmitted through the connecting spring k_{c}.

On impact (when the fold position exceeds the resting position, i.e. when $x_1 < -x_{01}$ or $x_2 < -x_{02}$) the elasticity of the folds is considered to be far higher than when open, hence the elasticity constant changes discontinuously. Due to the spring coupling, the smaller mass will in general oscillate at the same frequency as the larger mass, but with a time lag. Thus although the model does not replicate the mucosal wave motion itself,

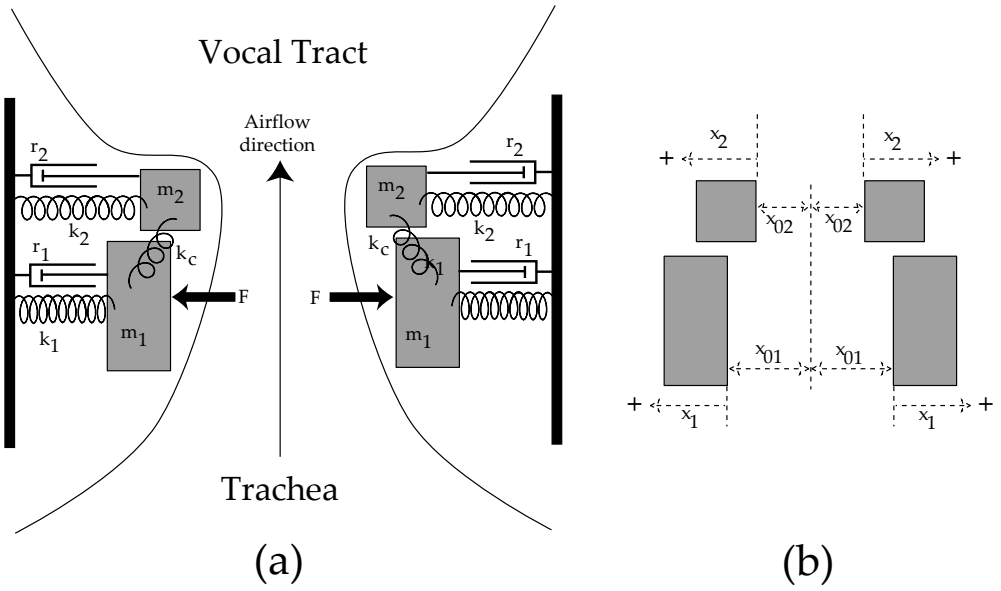

Figure 2.5: Two-mass vocal fold model. The system is symmetric, with left and right halves identical. (a) Mass (m_1, m_2), stiffness (k_1, k_2, k_c), damping (r_1, r_2) and internal forcing (F) components and configuration, (b) Coordinate configuration. The positions x_1, x_2 are measured positive when the vocal folds are open larger than the resting position x_{01}, x_{02} and negative when the vocal folds are closer together than the resting position. F represents the Bernoulli and static lung pressure forces acting only on the larger mass. See text for a more detailed description.

it captures the two important vibrational modes and the duration that the mucosal wave takes to propagate up the membrane.

The resting positions x_{01}, x_{02} and summation of both left and right halves give rise to the two *phonation neutral area* constants:

$$\begin{aligned} a_{01} &= 2lx_{01}, \\ a_{02} &= 2lx_{02}. \end{aligned} \tag{2.12}$$

where l is the length of the vocal folds perpendicular to the plane drawn in figure 2.5. The modelling configuration and assumptions lead to the following equations [39]:

$$\begin{aligned} m_1\ddot{x}_1 + r_1\dot{x}_1 + k_1x_1 + \Phi(-a_1)c_1\frac{a_1}{2l} + k_c(x_1 - x_2) &= p(t)ld_1, \\ m_2\ddot{x}_2 + r_2\dot{x}_2 + k_2x_2 + \Phi(-a_2)c_2\frac{a_2}{2l} + k_c(x_2 - x_1) &= 0. \end{aligned} \tag{2.13}$$

where the dot indicates differentiation with respect to time. The Heaviside step function Φ is used here to differentiate the collision from non-collision conditions, i.e. $\Phi(x) = 1$ for $x > 0$ and $\Phi(x) = 0$ for $x \leq 0$. Here, m_1, m_2 are the masses, and $a_1 = a_{01} + 2lx_1$,

$a_2 = a_{02} + 2lx_2$ are the lower and upper areas of the vocal folds, k_1, k_2, k_c are the elasticity constants, and r_1, r_2 the viscoelastic damping constants, d_1 is the height of mass one, l is the length of the larynx and c_1, c_2 are the additional collision elasticity constants.

The driving force $p(t)$ is:

$$p(t) = p_s \left[1 - \Phi\left(a_{\min}\right) \left(\frac{a_{\min}}{a_1} \right)^2 \right] \Phi\left(a_1\right) \tag{2.14}$$

and the vocal fold flow rate at the top of the vocal folds $u_f(t)$ is:

$$u_f(t) = \sqrt{\frac{2p_s}{\rho}} a_{\min} \Phi(a_{\min}) \tag{2.15}$$

where $a_{\min} = \min(a_1, a_2)$, p_s is the static lung pressure and ρ is the constant equilibrium density of air.

Although this model captures the phenomena listed above, this list ignores some considerably more complex, *nonlinear dynamical behaviour* [40] which has been observed in the motion of real vocal folds, particularly in cases of vocal fold disorders [41, 42]. Equation (2.13) is however an example of a *piecewise smooth*, nonlinear coupled oscillator [43], and it has been demonstrated numerically that it exhibits a rich variety of nonlinear dynamical behaviours [39]. Furthermore, for systems with discontinuous equations of motion such as this, there exists the possibility of *border-collision* and *grazing bifurcations* (changes in dynamical behaviour as a parameter is altered), which are phenomena not appearing in systems with smooth equations of motion [43]. We will next demonstrate, using numerical integration, that this model exhibits behaviour that ranges from simple and regular (periodic) to irregular (apparently chaotic).

Numerical simulations [5] of normal and irregular oscillation (using parameters from [39]) are shown in figures 2.6 and 2.7 respectively. Three-dimensional *state space plots* for the same parameters are shown in figure 2.8.

While some limited forms of *bifurcation analysis* [40] are possible on the two-mass model above, the large number of parameters makes this a difficult task. In the previous section where a model for the vocal tract was presented, the transfer function was determined using frequency analysis. It will therefore be useful to use frequency analysis for this section to determine a representation of the vocal fold flow rate $u_f(t)$ in terms of complex exponentials. This is obtained by finding the *power spectrum* [12] of the vocal fold flow rate, denoted $|U_f(\omega)|^2$. Figure 2.9 shows the numerically estimated power spectrum

[5] First-order Euler finite differences with $\Delta t = 0.02s$. The theoretical difficulty of the existence of discontinuous functions which do not have derivatives defined everywhere was not taken into account – this did not pose any stability problems however.

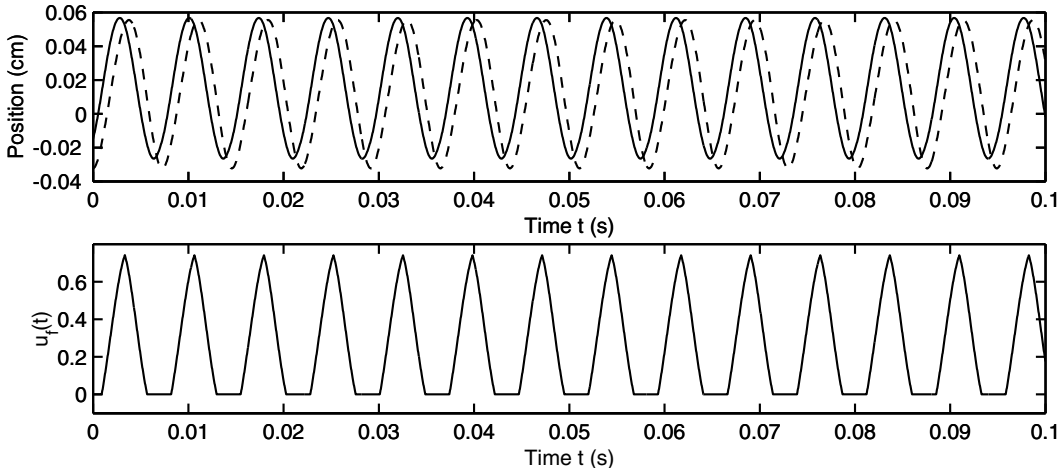

Figure 2.6: Numerical simulation of regular vibration of the vocal folds, parameters $m_1 = 0.125, m_2 = 0.025, k_1 = 0.08, k_2 = 0.008, l = 1.4, d_1 = 0.25, \rho = 0.00113, r_1 = r_2 = 0.05, a_{01} = a_{02} = 0.02, k_c = 0.025, P_s = 0.008$. Top panel shows $x_1(t)$ (solid line), $x_2(t)$ (dashed line), bottom panel the output flow rate $u_{\rm f}(t)$. Note that $u_{\rm f}(t)$ is never less than zero, i.e. the airflow is always from the lungs to the lips.

[6] for the two example vocal fold model outputs obtained earlier. The fact that the power spectrum declines in magnitude gently with frequency is often called *spectral roll-off*.

It is also clear from figure 2.9 that the flow rate output $u_{\rm f}(t)$ produces energy at frequencies that lie in the ranges of all the formants of the vocal tract shown in figure 2.4. Hence the decreasing sequence of Fourier harmonics excites the vocal tract into resonance at all the formant frequencies. However, the sequence of harmonics decreases in amplitude with increasing frequency, and the rate of decrease in amplitude will be partly responsible for the tonal quality of the spoken speech, i.e. whether the voice sounds "harsh" or "soft".

Although the time series shown in figure 2.7 is not long enough to identify visually the irregular behaviour as chaotic, in [39] one positive *Lyapunov exponent* was numerically estimated from the output given these parameters which is a good indicator of (but not conclusive evidence for) chaos [8]. Furthermore, we also note that the regular vibration has energy at several regularly spaced peaks, indicating the periodic behaviour. By contrast, the irregular vibration appears to contain energy at most frequencies, and there is only one obvious peak at the dominant frequency of oscillation. From the state space plots of figure 2.8 it can be seen that while the regular vibration leads to a simple closed loop, the irregular vibration is a more complicated object, and this complexity is born out in the

[6] Estimates were obtained without windowing or transient removal using the `fft` command in Matlab over 10,950 time steps of the model, which for the normal oscillation parameters was exactly 30 fundamental cycle periods at a fundamental frequency of almost 137Hz. The 0Hz component was removed by subtracting the mean from the output signal $u_{\rm f}(t)$.

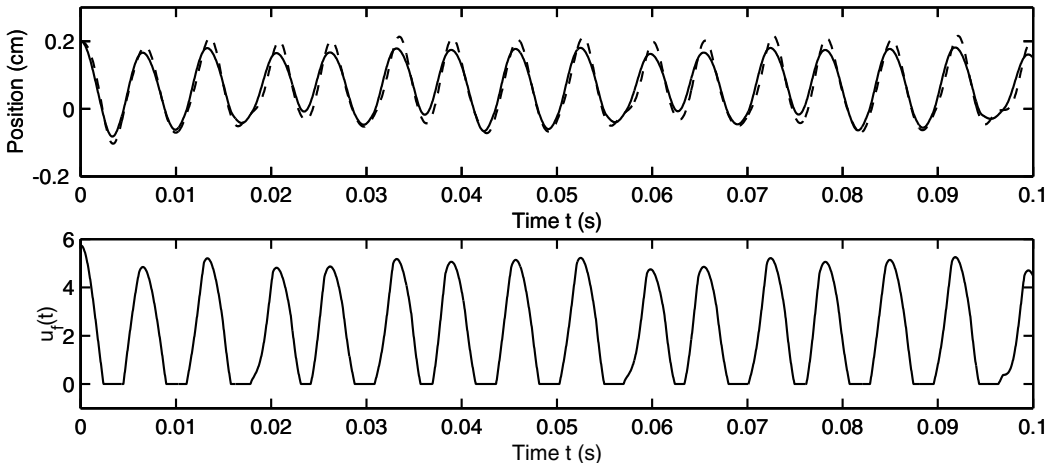

Figure 2.7: Numerical simulation of irregular, apparently chaotic behaviour of the vocal folds. All parameters are the same as figure 2.6 except $a_{01} = 0.02, a_{02} = 0.01, k_c = 0.09, P_s = 0.05$. Top panel shows $x_1(t)$ (solid line), $x_2(t)$ (dashed line), bottom panel the output flow rate $u_{\mathrm{f}}(t)$.

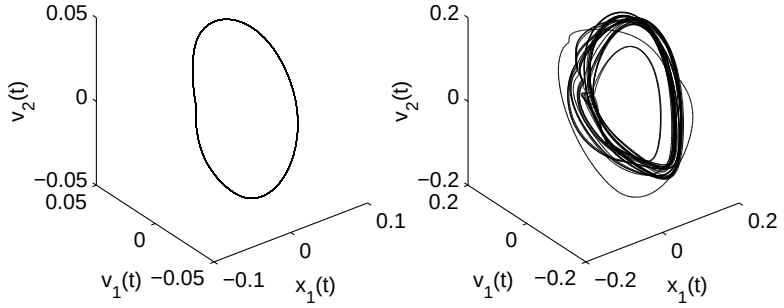

Figure 2.8: Numerical simulation of typical behaviours of the vocal fold model in state space, where $v_1(t) = \dot{x}_1(t)$ and $v_2(t) = \dot{x}_2(t)$. Left panel regular motion, right panel irregular motion. Parameters as in the figures 2.6 and 2.7 respectively.

lack of clear harmonic structure in the power spectrum of figure 2.9.

There are many assumptions made in deriving the two-mass model. Just one of these is the linearisation of the vocal fold tissue; actual elastic tissue springs are nonlinear [44, 36]. Therefore for large deflections, the linearised model will be inaccurate, as may happen when the static lung pressure and tension are great, i.e. when the amount of energy in the system is large. Also, the assumption about steady flow upon which the use of the Bernoulli equation is founded is probably inaccurate. The air flow in the larynx is likely to be complex and so we might expect some *vorticity* – a rotational component to the air flow. The validity of these and many other assumptions have been studied in detail by several researchers – for a comprehensive overview see [36, 37, 38].

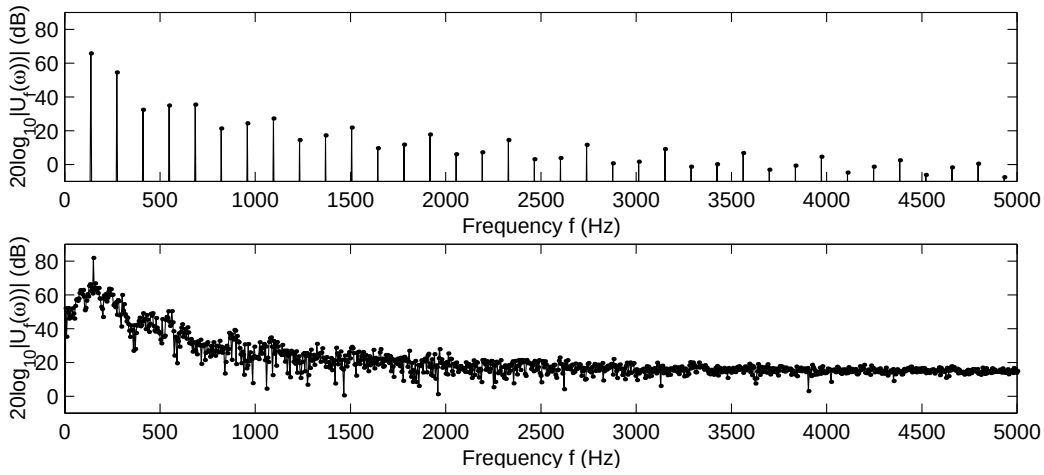

Figure 2.9: Numerical power spectra in decibels $|U_{\mathrm{f}}(\omega)|^2$ with $\omega = 2\pi f$ of two example vocal fold model outputs. Top is the regular vibration, bottom the irregular vibration. Model parameters as in figure 2.6 and 2.7 respectively.

2.2.3 Vocal Tract and Fold Models Combined

As shown in figure 2.1 the vocal folds are situated at the base of the vocal tract just below the pharyngeal and oral cavities. Assuming no feedback from the vocal tract to the vocal folds, the flow rate output of the vocal folds $u_{\mathrm{f}}(t)$ forces the vocal tract cavities downstream into resonance. If we have the Fourier transform $U_{\mathrm{f}}(\omega)$ of the vocal fold flow rate signal, given a specific lip opening area and tongue configuration, we can model the resulting flow rate at the lips $U(L,\omega)$. However, in practice, we obtain measurements of the speech pressure signal at the lips, using a microphone. Therefore of interest is the ratio of the pressure at the lips to the flow rate at the vocal folds:

$$Z_{\mathrm{p}}(\omega) = \frac{P(L,\omega)}{U_{\mathrm{f}}(\omega)} \tag{2.16}$$

which is:

$$\begin{aligned} Z_{\mathrm{p}}(\omega) &= \frac{P(L,\omega)}{U(L,\omega)} \times \frac{U(L,\omega)}{U_{\mathrm{f}}(\omega)} \\ &= Z(\omega) \times \frac{U(L,\omega)}{U_{\mathrm{f}}(\omega)} \\ &= Z(\omega)H(\omega) \end{aligned} \tag{2.17}$$

From now on we will refer to $P(L,\omega)$ as $P_{\mathrm{L}}(\omega)$. Then:

$$P_{\mathrm{L}}(\omega) = Z(\omega)U_{\mathrm{f}}(\omega)H(\omega) \tag{2.18}$$

As we can see, in the frequency domain, the acoustic pressure at the lips is the product of the radiation impedance, the vocal fold output and the acoustic transfer function of

the vocal tract. It is equation (2.18) that motivates the ubiquitous *source-filter theory* of voice production [13, 15], with the vocal fold flow $U_{\mathrm{f}}(\omega)$ acting as the "source" of vibration energy, and the product $Z(\omega)H(\omega)$ acting as a "filter" that enhances or suppresses various frequency components present in the source spectrum.

Such a combined model assumes that the vocal tract is passively driven by the vocal folds, so that there is no influence of the vocal folds on the vocal tract. The extent of the validity of this assumption has been studied extensively. For a review of the approaches see for example [15]. What has been discovered is that the lowest resonances of the vocal tract are affected somewhat by any interaction, and that the vocal fold flow rate output develops a slight "ripple" at the frequency of the lowest resonance due to the loading of the air mass of the vocal tract. Therefore the independent models we have described here are not entirely accurate in this respect. However, they are reasonable approximations that are often used in practice.

2.2.4 Aeroacoustic Noise

The models presented above appear to account successfully for audible speech, but only for voiced sounds such as vowels (see §2.3). However, a significant component is missing: that of *frication* and *aspiration* noise. Such noise is produced when the air is forced through a narrow constriction at sufficiently high speeds that "turbulent" airflow is generated, which in turn produces noise-like pressure fluctuations. Frication noise is deliberately employed when speaking to produce consonants (see §2.3) whereas aspiration noise is an unavoidable, involuntary consequence of airflow from the lungs being forced through the vocal organs, and can be heard in vowels and, to a lesser extent, in consonants as well. Also, certain voice pathologies are accompanied by a significant increase in such aspiration noise, which is perceived as increased "breathiness" in speech. This noise is therefore an important part of sound generation in speech. One significant deficiency in the above models is due to the assumptions about fluid flow upon which their construction is based [45].

These models have made very many simplifying assumptions about the airflow in the vocal organs, for example, that the *acoustic limit* [5] holds in which the fluid is nearly in a state of uniform motion. Similarly, the simple *Bernoulli's equation* applies if the fluid is assumed inviscid and irrotational. For more detailed information about these common assumptions in fluid dynamics, please see [46, 27]. The important point for this thesis is that these assumptions forbid the development of complicated, "turbulent" fluid flow motion, in which the flow follows convoluted paths of rapidly varying velocity, with eddies

and other irregularities at all spatial scales [47]. This breakdown of regularity occurs at high *Reynolds number*, the dimensionless quantity:

$$Re = \frac{\rho u l}{\eta}, \tag{2.19}$$

where $\eta = 1.76 \times 10^{-5}\,\mathrm{kg\,m^{-1}\,s^{-1}}$ is the typical shear coefficient of viscosity for air, and $\rho = 1.13\,\mathrm{kg\,m^{-3}}$ the typical density of air [48]. For the length scales l of a few centimetres in the vocal tract and for subsonic air flow speeds u typical of speech [49], this number is very large (of order 10^5), indicating that airflow in the vocal tract can be expected to be turbulent. Under certain assumptions, turbulent structures, and *vortices* in particular (fluid particles that have rotational motion), can be shown to be a source of *aeroacoustic sound* [48].

Turbulence is a very complex phenomenon, itself an open and active area of research, let alone turbulence in the vocal organs. As such, a detailed mathematical treatment is beyond the scope of this thesis. Instead, we will give a qualitative account of some of the most pertinent results and discoveries. Over and above phenomenological approaches that make use of simple electrical or digital noise sources and empirical observations of noise in speech [31], there are two broad classes of mathematical models that have been formulated to attempt to incorporate the effects of aeroacoustic noise generation in speech:

- Solving numerically the full partial differential equations of gas dynamics (e.g. the Navier-Stokes equations), and,
- Using the theory of *vortex sound* [48].

Numerical solutions to the Navier-Stokes equations require significant computational resources, but have the advantage that very detailed simulations of the vorticity patterns due to particular vocal organ configurations can be obtained [50, 51]. For example, the study of [51] focused on the production of aspiration noise generated by vortex shedding at the top of the vocal folds, simulated over a full vocal fold cycle. It was shown that when the vocal folds are closed, a stable jet of air forms. As the vocal folds begin to open, the jet Reynolds number increases such that vortices are generated downstream. As the folds close, the jet Reynolds number reaches a maximum such that the vortex generation is maximum. Finally, on closing, the jet reverts to a stable configuration, and the vorticity is minimal. This study demonstrates that the computed sound radiation due to vorticity contains significant high frequency fluctuations when the folds are fully open and beginning to close. On the basis of these results, it can be expected that if the folds do

not close completely during a cycle (which is observed in cases of more "breathy" speech), the amplitude of high frequency noise will increase.

The second class of models, which makes use of *Lighthill's acoustic analogy*, promises at least a partial analytical approach to finding the acoustic pressure due to turbulence [48]. These models are based around the theory of vortex sound generated in a cylindrical duct [48], where, essentially, each vortex shed at an upstream constriction acts as a source term for the acoustic wave equation (2.4) in the duct, as the vortex is convected along with the steady part of the airflow. The resulting source term depends upon not only the attributes of the vortex itself, such as size and circulation, but also upon the motion of the vortex through the streamlines of the flow [52, 48]. This modelling approach has only recently been used, so that there exist few complete models of vortex sound generation mechanisms in the vocal organs [53]. The most complete model that uses this approach involves the numerical simulation of two overall components: the mean steady flow field and the acoustic wave propagation in the vocal tract [49]. Vortices are assumed to be shed at random intervals at constrictions at particular locations in the vocal tract, for example, at the vocal folds or between the roof of the mouth and the tongue. Each vortex is tracked as it is convected along the mean flow field, following the shape of the vocal tract created by the particular configuration of articulators such as the tongue. Each vortex contributes to the acoustic source term at each spatial grid point. Numerical acoustic pressure simulations [7] at the lips for the consonants "sh" and "s" (see §2.3) are shown in figure 2.10, along with the corresponding power spectra. [8]

An important observation is that these simulated pressure signals appear as *stochastic processes* [54], i.e. a sequence of random variables. It is also noticeable from the spectra that although the signals are stochastic, they exhibit significant non-zero *autocorrelation* (see Chapter 3), since the spectral magnitudes are not entirely constant. Similarly, although beyond the scope of this thesis, one explanation for turbulent fluid flow is in terms of vortex motion on all length scales transferring energy from the largest scales to the smallest, where the energy is dissipated in viscosity [47]. Thus we may expect that the resulting pressure signals will have particular *self-similarity* properties [47, 45], an observation that will play a role in later chapters. We note also that the particular shape of the spectra is one important factor that a listener uses to determine the difference between consonants, analogous to the way in which a listener separates different vowels by their

[7] Numerical simulations calculated by Dr Daniel Sinder, used here with permission.

[8] Estimates were obtained using a Hanning window [12] and the `fft` command in Matlab over 28,000 time steps of the model output.

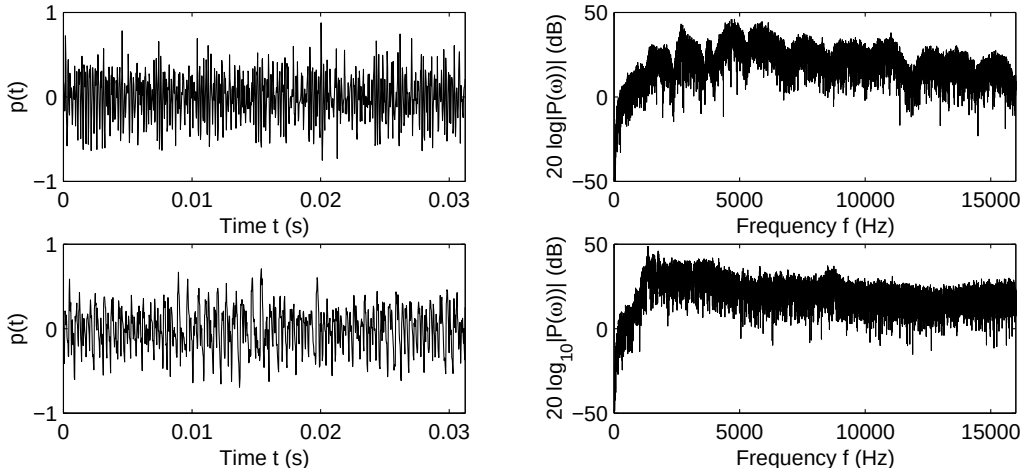

Figure 2.10: Simulated pressure signals and numerical power spectrum $|P(\omega)|^2$ in decibels (with $\omega = 2\pi f$) of two example aeroacoustic simulations of acoustic frication noise, from the model of [49]. The top row is the consonant "s", the bottom row the consonant "sh". The left column shows the time series over a short interval in time for clarity; the vertical scale is dimensionless signal amplitude. The time discretisation interval was $\Delta t = 31.25\,\mu$s. The right column shows the power spectra of these pressure signals.

patterns of resonances.

2.3 Basic Phonetics

In this section we will review some basic phonetics of interest to this thesis. This will cover only a fraction of what is naturally a vast discipline, given the large number of human spoken languages that exist and their many dialects and individual and regional variations. Of interest will be the two major categories of sounds that make up all languages: *vowels* and *consonants*. Where two such sounds can be used to differentiate one word from another, they are classed as different *phonemes* [10].

Vowels are always *voiced* phonemes, in that the source of sound energy is the steady oscillation of the vocal folds that excite the vocal tract into a specific pattern of resonances: the formants introduced in §2.2.1. Examples of vowels are the sonorous, tonal sounds in the middle of each of the following words: "hard", "bird", "beat", "bit", "bat", "bet". They can be short or long, held constant (*monophthongs*) or slowly varied from one to another (*diphthongs*). They vary significantly from one language to another, but the vowels shown in table 2.1 can be found in quite a large number of languages, including British English. Since this thesis is not concerned with detailed aspects of different languages, only those vowels mentioned in that table will be studied. We note, however, that due

Table 2.1: Vowels, consonants and codenames used in this study.

Type	Example word	Codename
Vowels	f**a**rther	/aa/
	b**i**rd	/er/
	b**ea**t	/iy/
	b**i**t	/ih/
	b**a**t	/ae/
	b**e**t	/eh/
	b**oo**t	/uw/
	p**u**t	/uh/
	p**o**t	/ao/
	b**u**t	/ah/
Consonants	**s**igh	/ss/
	shy	/sh/
	fee	/ff/
	thigh	/th/

to significant variability in the pronunciations of the given examples [10], any results in this thesis will not be explicitly predicated upon any idiosyncratic formant variations of speakers with differing accents.

Consonants, by contrast to vowels, have a noise-like "hissing" or "explosive" character, as exemplified at the start of words such as "spy" or "try". The source of acoustic energy is mostly turbulent vortices generated at constrictions created by articulators such as the tongue, teeth, lips and vocal folds. These vortices impact upon later obstructions in the airstream, creating sound, as described above in section §2.2.4. Consonants can be classified into different phonemes according to the following configurations of the vocal organs [10]:

- Varying the position in the vocal tract of the vortex-generating constriction (for example, by placing the tongue tip at varying locations along the roof of the mouth),
- Causing the vocal folds to vibrate simultaneously (voiced) or remain fully open and static (unvoiced),
- Expelling air through the mouth or just the nose,
- Holding the sound constant (*fricatives*) or either abruptly stopping the flow of air, or generating single puffs of air after stopping the airflow (*stops*).

These configurations combine to produce a very large number of possible phonemes. In this study we will only be concerned with the fricatives shown in table 2.1.

Phonemes combine in particular temporal sequences to form *syllables*, which then combine to form different words. We note that there is a significant tendency for articulations during one sound to anticipate those in the following sound, a phenomena known as *anticipatory co-articulation*. Thus phonemes pronounced in isolation *citation form* will vary considerably from those in continuous, natural or *running* speech. Since this study is concerned largely with the basic acoustic properties of speech signals and not their linguistic content, we will only study phonemes that are unaffected by such co-articulation.

2.4 Chapter Summary

In this chapter we have reviewed several biomechanical models of the vocal organs. We have shown that a good approximation to the vocal tract is a varying cross-sectional area acoustic tube with an infinite plane baffle opening at the lips, and that this model can be understood by the associated patterns of resonance frequencies.

For the vocal folds, we have shown that a simple model with four degrees of freedom is able to capture most of the observed dynamics of vocal fold oscillation. We have shown that this model, a nonlinear dynamical system, is capable of both simple, regular vibration and more complex, apparently chaotic motion. We have shown that the model output has spectral components that are responsible for exciting the resonances of the vocal tract. Combining the models motivated the source-filter theory of voice production.

The importance of the source-filter theory in speech science cannot be underestimated. It forms the basis of most speech analysis technologies. This theory underpins the ubiquitous technique of *digital formant analysis* presented in Chapter 3 and many other technologies such as digital speech compression and speech recognition, mentioned in the introduction. However, as we will see in later chapters, this theory does not account for the full dynamics encountered in real speech signals.

Also discussed in this chapter was the source of ubiquitous noise-like aeroacoustic sound that forms an important part of sound production in the vocal organs, and it was demonstrated how a considerably simplified model of turbulent phenomena generated autocorrelated stochastic pressure signals. This observation will inform tests in later chapters that will be performed on real speech signals, and will also inform the use of a particular signal processing method for analysing this noise component for changes indicative of certain speech pathologies.

Finally, we have introduced some basic aspects of phonetics which set a context for

the further analysis of speech signals in later chapters.

CHAPTER 3

Classical Linear Digital Speech Analysis

Linear digital signal processing is currently the mainstay of scientific and commercial telecommunications and speech processing. It is a focus of this thesis to identify the limitations of these techniques by analysing the appropriateness of the mathematical foundations of these methods for speech. This chapter therefore introduces and discusses the relevant basic concepts of these techniques. These foundational concepts will be the subject of scrutiny in this chapter and later in the thesis.

3.1 Signals, Sampling and Quantisation

In the context of this thesis, the term *signal* is defined as a scalar, real valued measurement of a physical quantity that can change with time [12], and will be denoted by $s(t)$ for all $t \in \mathbb{R}$. Such *continuous time signals* arise in the context of speech as measurements of the change in pressure in air near the mouth of a speaker, obtained from a microphone. A *system* is defined as a physical device that operates on signals, and the operation of passing this signal through a system is called *signal processing* [12]. Note that this definition includes systems that are implemented as algorithms in computer software, however, continuous time signals are not directly suitable for processing in software on a computer.

Conversion of a continuous time signal into a *discrete time signal* is carried out by an ADC. This electronic device performs two actions [12]:

- *Time discretisation*, also known as *sampling*. The ADC produces a *discrete time signal* denoted by the sequence $s_n = s(n\Delta t)$. This is just the values of the continuous time signal at the instances in time $n\Delta t$, for the time index $n \in \mathbb{Z}$, and the *sampling interval* $\Delta t \in \mathbb{R}$ is a (small, positive) number that has the units of time in seconds. The *sampling frequency* or *sample rate* is the inverse of the sampling interval, Δt^{-1} and has the units of frequency in Hertz. Typically, speech signals are sampled at a rate of between 8000 to 44100 Hz.

- *Quantisation* maps the real valued discrete time signal s_n, which can assume an infinity of possible values, to a signal s_n^q that can assume only a finite number of values, or *quantisation levels*, for subsequent processing. Typically this finite number will be between $Q = 2^8 = 256$ and $Q = 2^{16} = 65536$ possible values. One common type of quantisation mapping is truncation:

$$s_n^q = \lfloor q s_n \rfloor , \tag{3.1}$$

  where q is a real valued *amplification factor* and $\lfloor\rfloor$ is the floor operation. For a bounded signal $-1 \leq s_n \leq 1$ and $q = 2^{15} - 1 = 32767$, this mapping takes the interval $[-1, 1]$ to the range of integers $-32767, -32766, \ldots 32767$, so that $Q = 65535$.

It should be noted that the many-one operation of sampling maps some distinct signals $s_1(t), s_2(t)$ to the same signal s_n. This *sampling error* places limitations upon the signals $s(t)$ that may be unambiguously represented by the discretised signal s_n. Quantisation is a many-one mapping that introduces *quantisation error*. These errors may have an effect upon further processing and are considered nuisances that can be mitigated by a variety of tricks. For example, the *sampling theorem* [12] states that for a sinusoidal, continuous time signal $s(t) = \sin(2\pi F t)$ where F is frequency in Hertz, if the signal is sampled at a rate of $F_s > 2F$ then $s(t)$ can be exactly reconstructed from the sampled signal s_n, using appropriate interpolation functions. Therefore, by restricting the range of frequencies of any sinusoidal components in a signal $s(t)$, such ambiguity may be avoided. Similarly, quantisation error may be decreased by increasing Q, and there are quantisation functions other than truncation that introduce less error. For further details of the issues of sampling and remedies for sampling error, see [12]. An in-depth discussion of quantisation and quantisation error is presented in [55].

We generally consider in this thesis that these issues of sampling and quantisation error have been resolved sufficiently for our purposes at the ADC stage. Therefore the quantisation error is considered to be very small so that $s_n^q \approx s_n$, and it is considered that no signal ambiguity has been introduced by sampling. Therefore, the signal s_n^q will not generally be mentioned further, reference to s_n will be made instead, but it should be understood that the signals are actually quantised and stored in computer memory as binary representations. Such binary representations are generally referred to as *digital signals*.

There exists a number of useful special signals. This chapter will make use of the *unit*

sample sequence, defined using the Kronecker delta:

$$\delta_n = \begin{cases} 1, & \text{if } n = 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

This is also referred to as the *unit impulse*. Similarly, the *unit step function* is also very useful:

$$\theta_n = \begin{cases} 1, & \text{if } n \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

For the complex-valued discrete time signal case, another important class are the *complex exponential signals* [12]:

$$x_n = Ae^{i\omega n} = A\left(\cos \omega n + i \sin \omega n\right), \tag{3.4}$$

where $A > 0$ is called the *amplitude* of the exponential, and $0 \leq \omega \leq \pi$ the *frequency* in radians per sample. Such signals are important in speech processing, as will be described later. If instead we use frequency $0 \leq F \leq F_s/2$ in Hertz, then $\omega = 2\pi F/F_s$.

We will also discuss *stochastic signals* that are sequences of random variables we denote by w_n. Such signals are also known as (examples of) discrete time *stochastic processes* [54] and, independent, identically distributed (i.i.d.) processes have the important property that their joint distributions are time-invariant, known as *strong stationarity*. See [54] for a more in-depth discussion. An example of particular importance is the zero mean *Gaussian i.i.d. process* $w_n \sim \mathcal{N}(0, \sigma^2)$, where the variance σ^2 is finite.

Of the various mathematical operations that can be applied to such discrete time signals, of importance to this chapter is the *time delay operator*:

$$z^k[s_n] = s_{n-k}, \tag{3.5}$$

where $k \in \mathbb{Z}$ is called the time delay.

It is sometimes convenient to describe discrete time signals that are zero for $n < 0$ and non-zero for $n \geq 0$ as *causal*, and we will use this terminology later.

3.2 Linear Time-Invariant Discrete Time Systems Theory

A (quantised) signal s_n stored in computer memory is subsequently processed by a signal processing system. We are interested in this chapter in a class of *discrete time systems* that can be described as *linear* and *time-invariant*. Discrete time refers to the fact that

these systems act only on sampled signals. Such systems can be described mathematically as functions $F : \mathbb{R}^M \to \mathbb{R}$:

$$y_n = F(\mathbf{s}_n), \tag{3.6}$$

where $\mathbf{s}_n$ is an M-dimensional *vector* of discrete time signals. The discrete time signal output y_n of such a system is therefore only defined at the time instants $n\Delta t$. Linearity and time-invariance are two *mathematical* properties with important implications for speech processing; these properties will be the subject of critical examination later in the thesis.

3.2.1 Time-Invariance

A *time-invariant* system is one whose function F does not change with time. This is embodied in the following property [12]:

$$F(z^k[\mathbf{s}_n]) = z^k[F(\mathbf{s}_n)]. \tag{3.7}$$

In other words, applying F to the input first and then delaying it will produce exactly the same output as first delaying the input and then applying F.

3.2.2 Linearity

A *linear system* F is one that has the following property:

$$F(a_1\mathbf{u}_n + a_2\mathbf{v}_n) = a_1F(\mathbf{u}_n) + a_2F(\mathbf{v}_n), \tag{3.8}$$

where a_1, a_2 are arbitrary real constants, and u_n, v_n are arbitrary, discrete time signals. This property can be extended by induction to any weighted linear combination of signals. This property is also known as the *superposition principle* [12].

3.2.3 Recursive Linear Filters

All the LTI systems studied in this thesis belong to a class of functions described as *constant-coefficient difference equations* [12]:

$$y_n = \sum_{k=1}^{P} a_k y_{n-k} + x_n, \tag{3.9}$$

where the a_k are P, real-valued constants, the coefficients of the system. For time indices $n \geq n_0$ and some initial time n_0, the P *initial conditions* $y_{n_0-1}, y_{n_0-2}, \ldots y_{n_0-P}$ together with the input signal x_n are required to calculate all subsequent values of y_n. Therefore, this is an example of a *recursive* system. The fact that the system requires past outputs

y_{n-k} in order to calculate the current output means that this system has internal *memory*, and this memory is described as the *system state*.

These ubiquitous systems are otherwise known as *digital filters* in the signal processing literature. They are also known as order-P *autoregressive, AR(P)* systems. These systems satisfy the linearity property described above, the proof of this can be obtained by induction [12]. They are also time-invariant since the coefficients do not change with the time index n.

Such recursive systems are also *causal*: the output of the system does not depend upon future values of the input or output. In other words, at some time instant, say, $n = n_0$, the output of the system depends only upon values of x_n and y_n for $n \leq n_0$.

3.2.4 Convolution

Since the linear recursive system defined above satisfies the properties of linearity and time-invariance, we can use *convolution* and the *impulse response* to predict the behaviour of the system (with zero initial conditions, i.e. zero initial system state) to any arbitrary input sequence. This will be valuable for understanding the appropriateness of such linear filters in speech processing. Convolution is an associative, commutative and distributive binary operator $*$ that acts on two signals to produce a third signal [12]:

$$u_n * v_n = \sum_{i=-\infty}^{\infty} u_i v_{n-i} = \sum_{i=-\infty}^{\infty} v_i u_{n-i}. \tag{3.10}$$

3.2.5 Impulse Response

The *impulse response* h_n of a linear system is the solution of the system with zero initial conditions, when the input is the unit impulse function δ_n. [1] This special solution is useful in the following way. Since the superposition principle applies to all LTI systems, any linear combination of solutions of the system is another solution to the system. The impulse response of the system is the solution of the system when the input is a unit impulse, given zero initial conditions. Furthermore, as we will show next, any signal x_n can be written as a linear combination of unit impulses. It follows that we can determine the solution of the system to an arbitrary input signal by forming a linear combination of solutions to unit impulses, weighted by the input signal at each time instant.

[1] This is similar to the machinery of *Green's functions* used in the theory of partial differential equations [5].

Any signal x_n can be decomposed into a weighted sum of unit impulses:

$$x_n = \sum_{k=-\infty}^{\infty} x_k \delta_{n-k}. \tag{3.11}$$

In the special case of the recursive filter system of equation (3.9), the impulse response of the system is calculated as:

$$h_n = \sum_{k=1}^{P} a_k h_{n-k} + \delta_n, \tag{3.12}$$

for $n \geq 0$ and $h_j = 0$ for $j < 0$. It is thus a causal signal.

Therefore, the solution y_n for any system (3.9) with zero initial conditions given any arbitrary input x_n is formed as the linear combination of impulse responses h_n weighted by the corresponding input signal x_n at time instant n. This is therefore the *convolution* of the impulse response with x_n [12]:

$$y_n = h_n * x_n = \sum_{k=-\infty}^{\infty} h_k x_{n-k}. \tag{3.13}$$

For the case of equation (3.9), in general h_n is non-zero for all values of $n \geq 0$. Such systems are therefore known as *infinite impulse response* (IIR) filters. Closed form expressions do exist for h_n in this case, using the *direct* solution to the difference equation (3.9) [12], or indirectly through the *z-transform* representation, which will be described later. The z-transform representation is powerful in that it provides additional, useful information about the behaviour of the system. In practice, the (truncated) impulse response of any desired finite duration $n = 0, 1 \ldots N$ can be determined using (3.12) computationally.

3.2.6 Stability

We will, in general, only treat *stable* linear systems in this thesis, i.e. those systems that, given a bounded input signal produce a bounded output signal (*BIBO stability* [12]). For recursive linear systems (3.9), this condition can be shown to be equivalent to the requirement that the impulse response is *absolutely summable* [12]:

$$\sum_{k=-\infty}^{\infty} |h_k| < \infty. \tag{3.14}$$

In turn, via the closed form for h_n, BIBO stability translates into the condition that all the roots λ_k of the *associated homogeneous difference equation* to (3.9) have a magnitude of less than unity. The homogeneous difference equation is the difference equation with zero input term. For a derivation of these results, please see [12].

3.2.7 z-Transforms and Transfer Functions

A useful tool in the analysis of LTI systems is the *z-transform.* [2] This transform will allow us to derive the *transfer function* of the system (3.9), which in turns allows the calculation of the *frequency* and *phase* responses of this system.

The (one-sided) z-transform applied to an arbitrary signal x_n is defined by the following equation:

$$\mathcal{Z}[x_n] = \sum_{n=0}^{\infty} x_n z^{-n} = X(z), \tag{3.15}$$

where $z \in \mathbb{C}$. This infinite power series converges only for certain values of the variable z. The *region of convergence* (ROC) is the set of all values of z for which $X(z)$ is finite. For causal signals this z-transform is unique, and the ROC is the exterior of some circle in the complex plane [12]. This transform has a number of useful and important mathematical properties, see [12] for a list of these. For the purposes of this thesis we will make explicit the linearity, time delay, and convolution properties.

The z-transform is linear, in that obeys the superposition principle:

$$\mathcal{Z}[a_1 v_n + a_2 u_n] = a_1 V(z) + a_2 U(z). \tag{3.16}$$

For a signal x_n under time delay of k time indices, $z^{-k}[x_n] = x_{n-k}$, the (one-sided) transform has the following behaviour [12]:

$$\mathcal{Z}[x_{n-k}] = z^{-k}\left[X(z) + \sum_{n=1}^{k} x_{-n} z^{n}\right], \tag{3.17}$$

and this collapses down to:

$$\mathcal{Z}[x_{n-k}] = z^{-k}[X(z)], \tag{3.18}$$

for purely causal signals.

Convolution in time n is equivalent to multiplication in z [12]:

$$\mathcal{Z}[u_n * v_n] = U(z)V(z). \tag{3.19}$$

This, combined with the impulse response of an LTI system leads to a powerful method for describing the behaviour of the system.

Assuming that the initial conditions of the system (3.9) are zero, i.e. $y_{-j} = 0$ for $1 \leq j \leq P$, then the system solution given an arbitrary input signal can be obtained by convolution, equation (3.13). Using the convolution property of the z-transform:

$$\mathcal{Z}[y_n] = Y(z) = \mathcal{Z}[h_n * x_n] = H(z)X(z). \tag{3.20}$$

[2] This plays a similar role to the Laplace transform for continuous-time linear systems [5].

The function $H(z)$ is known as the *transfer function* of the system. The transfer function for the system (3.9) with non-zero initial conditions may be obtained explicitly as follows [12]:

$$Y(z) = \sum_{k=1}^{P} a_k z^{-k} \left[Y(z) + \sum_{n=1}^{k} y_{-n} z^n \right] + X(z), \tag{3.21}$$

giving

$$Y(z) = \frac{X(z) + \sum_{k=1}^{P} a_k z^{-k} \sum_{n=1}^{k} y_{-n} z^n}{1 - \sum_{k=1}^{P} a_k z^{-k}} = H(z)X(z) + H(z)N_0(z), \tag{3.22}$$

with $N_0(z) = \sum_{k=1}^{P} a_k z^{-k} \sum_{n=1}^{k} y_{-n} z^n$. The transfer function $H(z)$ is:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}. \tag{3.23}$$

The complete response (3.22) of the system (3.9) to an arbitrary input with non-zero initial conditions can therefore be seen as a sum of two terms, the first due to the input entirely, and the second due to the initial conditions (the initial state of the system). It can be shown that if the P *system poles*, which are the roots λ^k, $1 \leq k \leq P$ of the denominator $A(z)$ of the transfer function $H(z)$ satisfy $\left|\lambda^k\right| < 1$, then the term due to the initial conditions decays to zero as n tends to infinity [12]. This decaying term is referred to as the *transient response* of the system. The rate of decay depends upon the magnitude of the roots: the smaller the magnitude, the faster the decay.

We mention an important point about stability. For a causal, recursive system such as (3.9), described by a transfer function $H(z)$ that is the ratio of two polynomials, BIBO stability is equivalent to the condition that the poles of the transfer function lie inside those set of points $z \in \mathbb{C}$ for which $|z| = 1$ (the *unit circle*) [12]. Also, although it will not be required in this thesis, it should be mentioned that there exists an inverse z-transform that allows the *indirect*, explicit calculation of the output in time of a recursive system such as (3.9) [12].

3.2.8 Stochastic Processes and Recursive Linear Filters

Of importance to this thesis is the case of stochastic signals x_n as input to recursive filters such as those described above, and in particular the Gaussian i.i.d. process $w_n \sim \mathcal{N}(0, \sigma^2)$. These input processes are special in that under action of the linear recursive system, the output signal y_n is also a Gaussian process (although no longer i.i.d.) This is because the

linear combination of any number of Gaussian random variables is also a Gaussian random variable, see Appendix §A.2.1. Thus the output of a recursive filter driven by such a signal defines a *Gaussian process*, in that the joint density of any finite collection of members of the process is a multivariate Gaussian [54].

We remark here that Gaussian probability densities are fully parameterised by first and second order statistical moments only, i.e. mean and variance [54].

3.2.9 Cross-correlation and Autocorrelation

Closely resembling convolution is the operation of *cross-correlation*, which can be interpreted as a measure of the similarity between two signals u_n, v_n at *time lag* l:

$$r_{uv}(l) = \sum_{n=-\infty}^{\infty} u_n \overline{v}_{n-l} = u_l * \overline{v}_{-l}. \tag{3.24}$$

where the overbar denotes complex conjugation.

In the special case when $u_n = v_n$, we have the *autocorrelation*, which is then the similarity of the signal u_n to itself:

$$r_{uu}(l) = \sum_{n=-\infty}^{\infty} u_n \overline{u}_{n-l} = u_l * \overline{u}_{-l}. \tag{3.25}$$

We note that for signals that are not absolutely summable, i.e. for which $\sum_{n=-\infty}^{\infty} |x_n|^2$ is infinite, we take the limit over normalised finite sums in these definitions of cross- and autocorrelation, so that, for example, the cross-correlation becomes:

$$r_{uv}(l) = \lim_{M\to\infty} 1/(2M+1) \sum_{n=-M}^{M} u_n \overline{v}_{n-l}. \tag{3.26}$$

An important example is the sampled autocorrelation for finite length signals of length N over the range of lags $l = 0, \pm 1, \pm 2 \ldots \pm (N-1)$ which is:

$$r_{uu}(l) = 1/N \sum_{n=0}^{N-|l|-1} u_n \overline{u}_{n+|l|}. \tag{3.27}$$

Then for $l \geq N$, $r_{uu}(l) = 0$. Cross-correlation has the following property:

$$r_{uv}(l) = r_{vu}(-l), \tag{3.28}$$

so that autocorrelation is an even function of l:

$$r_{uu}(l) = r_{uu}(-l). \tag{3.29}$$

We will make use of the autocorrelation of certain special signals. For w_n a zero mean, Gaussian i.i.d. signal of variance σ^2:

$$r_{ww}(l) = \sigma^2 \delta_l \tag{3.30}$$

For a proof of this, see Appendix §A.2.2. This result can be used to predict the autocorrelation of the output y_n of an LTI system with impulse response h_n when given w_n as input [12]:

$$r_{yy}(l) = \sigma^2 \sum_{k=-\infty}^{\infty} h_k h_{k+l}. \tag{3.31}$$

3.2.10 Discrete Fourier Transform and Frequency Response

Frequency analysis is an important tool for speech signal processing: *Fourier analysis* allows the representation of a signal in terms of a weighted linear combination of complex exponential signals, called a *spectrum* [12]. For discrete time signals, the *discrete time Fourier transform* is:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x_n e^{i\omega n}. \tag{3.32}$$

A sufficient condition for uniform convergence of this sequence is that the signal x_n is absolutely summable [12]. This function $X(\omega)$ is periodic with period 2π, a consequence of the fact that, due to sampling, the frequency range for a discrete time signal is limited to $0 \leq \omega < 2\pi$, with frequencies outside this interval mapped onto frequencies inside it [12].

For computation in software, a convenient approach is to evaluate this spectrum at N regularly-spaced frequency points $\omega_k = 2\pi k/N$. In addition, all practical, causal signals are of finite length L so that $x_j = 0$ for $j < 0$ and $j \geq L$. We define the *Discrete Fourier Transform* (DFT):

$$\mathcal{F}[x_n] = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} = X(k), \tag{3.33}$$

for $k = 0, 1, \ldots, N-1$. If $L \leq N$, then $X(k)$ is a unique representation of the finite length signal x_n for the range $n = 0, 1 \ldots N-1$. In order to reconstruct x_n in this range we can make use of the associated *inverse Discrete Fourier Transform* (IDFT):

$$\mathcal{F}^{-1}[X(k)] = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{i2\pi kn/N} = x_n, \tag{3.34}$$

for $n = 0, 1, \ldots, N-1$. There exists a very efficient algorithm for calculating the DFT of signal, the Fast Fourier Transform (FFT) [12].

There are two very important remarks that must be made at this point. Firstly, the DFT $X(k)$ is unique for a finite time segment $0 \leq n \leq N-1$ of a signal x_n. Outside this finite range the representation is only unique if the signal is infinitely periodic with period N. Formally, if $X(k) = \mathcal{F}[x_n]$ then:

$$x_n = x_{n+N} \tag{3.35}$$

$$X(k) = X(k+N), \tag{3.36}$$

for all n and k.

Secondly, if $N < L$ so that the DFT operates on a truncated version of the finite length signal, artefactual "ripples" will be introduced into the spectrum $X(k)$. These *Gibb's phenomena* are generally unwanted [12].

Other properties of the DFT closely resemble those of the z-transform. It obeys the superposition principle:

$$\mathcal{F}[a_1 u_n + a_2 v_n] = a_1 U(k) + a_2 V(k). \tag{3.37}$$

One important difference between the z-transform and the DFT is that multiplication of the DFT of two signals is equivalent to the *circular convolution of length* N $\circledast$ of the two signals in time:

$$\mathcal{F}[u_n \circledast v_n] = U(z)V(z). \tag{3.38}$$

where circular convolution is defined as:

$$u_n \circledast v_n = \sum_{i=-\infty}^{\infty} u_i v_{n-i(\mathrm{mod}N)} = \sum_{i=-\infty}^{\infty} v_i u_{n-i(\mathrm{mod}N)}. \tag{3.39}$$

The *circular cross-correlation* of the signals u_n and v_n:

$$\tilde{r}_{uv}(l) = \sum_{n=-\infty}^{\infty} u_n v_{n-l(\mathrm{mod}N)} \tag{3.40}$$

has the following DFT [12]:

$$\mathcal{F}[\tilde{r}_{uv}(l)] = U(k)\overline{V}(k) \tag{3.41}$$

The transfer function $H(z)$ described earlier is valuable for determining the response of the system to arbitrary sums of complex exponential signals [12]. Given an exponential signal of frequency ω as input, the output of the linear system (3.9) is:

$$y_n = \sum_{k=-\infty}^{\infty} h_k \left[A e^{i\omega(n-k)} \right] = A \left[\sum_{k=-\infty}^{\infty} h_k e^{-i\omega k} \right] e^{i\omega n}. \tag{3.42}$$

However, the term in square brackets is just the discrete time Fourier transform of the impulse response of the system, which we write as $H(\omega)$. This is called the *frequency response* of the system. The output of the system (3.9) is then:

$$y_n = AH(\omega)e^{i\omega n}. \tag{3.43}$$

Therefore, when complex exponential signals act as inputs to LTI systems described by difference equations, the output will have the same frequency as the input exponential but the amplitude and *phase* (the complex argument) will be altered according to $H(\omega)$. It is useful to express this magnitude and phase change induced by $H(\omega)$ as separate functions of the radian frequency ω, i.e. the *magnitude response*:

$$M(\omega) = \left|H(e^{i\omega})\right|, \tag{3.44}$$

and the *phase response*:

$$\Phi(\omega) = \arg H(e^{i\omega}). \tag{3.45}$$

3.2.11 Power Spectrum and the Wiener-Khintchine Theorem

One important characteristic of a signal is the *energy*, defined as [12]:

$$E_x = \sum_{n=-\infty}^{\infty} |x_n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega)|^2 \, d\omega. \tag{3.46}$$

This is *Parseval's relation* for discrete time signals, see [12] for the proof of this. The quantity:

$$E_{xx}(\omega) = |X(\omega)|^2, \tag{3.47}$$

is the distribution of energy as a function of frequency ω, called the *energy spectral density* [12]. In practice, we have finite duration signals and perform spectral analysis using the DFT. We then define the (discrete) *power spectrum* as the square magnitude of $X(k)$:

$$P_{xx}(k) = |X(k)|^2. \tag{3.48}$$

The Wiener-Khintchine theorem makes the connection between the *circular autocorrelation* and the power spectrum:

$$P_{xx}(k) = \mathcal{F}\left[\tilde{r}_{xx}(l)\right] \tag{3.49}$$

where $\tilde{r}_{xx}(l)$ is defined as:

$$\tilde{r}_{xx}(l) = \sum_{k=0}^{N-1} x_n \overline{x}_{n-l(\mathrm{mod}N)}. \tag{3.50}$$

For a proof of the Wiener-Khintchine theorem see Appendix §A.2.3. In other words, the power spectrum is the DFT of the circular autocorrelation. Thus, the circular autocorrelation and the power spectrum contain the same information, but all information about the *phase* of complex exponentials that make up x_n is lost so that x_n cannot be reconstructed from either the circular autocorrelation or the power spectrum.

3.2.12 Linear Prediction Analysis

Given a particular signal x_n, the question often arises whether there exists some system that can reproduce this signal. This question is equivalent to finding an appropriate data-driven model for the signal. Assuming that an AR(P) model such as equation (3.9) is appropriate, the problem of estimating the parameters $a_k, k = 1, 2, \ldots P$ for this model is called *optimum filter design* in the signal processing literature. These parameters fully characterise the transfer function of the model, so that the assumed spectrum of the model can be analysed using the magnitude and phase response of this transfer function. Analysis such as this involving linear system parameter estimation is also called *linear prediction analysis* (LPA) in the context of speech processing [56].

Optimum parameter estimation can be approached from several different directions. We will discuss three distinct approaches that are often cited in the speech analysis literature. Each of these leads essentially to the same mathematical formalisation. All three approaches must solve the problem of obtaining the parameters that produce the best model for the signal x_n using equation (3.9).

Error Minimisation by Least-Squares Optimisation

Assume the existence of an *error signal* e_n that represents the error entirely due to the parameters a_k of the current value x_n of the signal and the estimate produced by the linear system:

$$x_n - \sum_{k=1}^{P} a_k x_{n-k} = e_n. \tag{3.51}$$

The best model has parameters a_k that minimise the error signal e_n for all n. One such measure of the overall error is the sum of squares of e_n:

$$E^2 = \sum_{n=-\infty}^{\infty} e_n^2 = \left[x_n - \sum_{k=1}^{P} a_k x_{n-k} \right]^2. \tag{3.52}$$

Note that E^2 has one global minimum with respect to the parameters a_k, which can be found by setting the partial differentiation with respect to these parameters to zero:

$$\frac{\partial E^2}{\partial a_k} = \frac{\partial}{\partial a_k} \sum_{n=-\infty}^{\infty} \left[x_n - \sum_{j=1}^{P} a_j x_{n-j} \right]^2 = 0, \tag{3.53}$$

for $k = 1, 2, \ldots P$ which leads to the following matrix problem to be solved for the a_k:

$$\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1P} \\ R_{21} & R_{22} & \cdots & R_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ R_{P1} & R_{P2} & \cdots & R_{PP} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = - \begin{bmatrix} R_{10} \\ R_{20} \\ \vdots \\ R_{P0} \end{bmatrix}, \tag{3.54}$$

where $R_{jk} = \sum_{n=-\infty}^{\infty} x_{n-j} x_{n-k}$. An important special case arises when the signal x_n has finite length L, i.e. when $x_j = 0$ for $j < 0$ or $j \geq L$. Then:

$$R_{jk} = \sum_{n=|j-k|}^{L-1} x_n x_{n-|j-k|} = r_{xx}(|j-k|), \tag{3.55}$$

which, since x_n is real-valued, is just the autocorrelation of x_n at time delay $|j-k|$ over $L-1-|j-k|$ samples. Hence all the entries along a given diagonal of the matrix in equation (3.54) are equal to the autocorrelation at time delay $|j-k|$:

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(P-1) \\ r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(P-2) \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ r_{xx}(P-1) & r_{xx}(P-2) & \cdots & r_{xx}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = - \begin{bmatrix} r_{xx}(1) \\ r_{xx}(2) \\ \vdots \\ r_{xx}(P) \end{bmatrix}. \tag{3.56}$$

For the proofs of these results, see [57]. The system of equations (3.56), called the *Yule-Walker* equations, can be solved very efficiently [12]. The resulting system is always stable [12].

Likelihood Maximisation with Gaussian System Input

This approach requires that the input to the linear system whose parameters are to be estimated is a zero mean, Gaussian, i.i.d. stochastic process of variance σ^2, w_n:

$$x_n - \sum_{k=1}^{P} a_k x_{n-k} = w_n. \tag{3.57}$$

Denote the density function of each random variable w_n by $p(w)$. The probability of obtaining a certain realisation w_n, $n = 0, 1, \ldots N$ of the stochastic process given a certain

set of parameters a_k is:

$$P\left(\mathbf{w}|\,\mathbf{a}\right) = \prod_{n=0}^{N-1} p\left(x_n - \sum_{k=1}^{P} a_k x_{n-k}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{n=0}^{N-1} \exp\left(-\frac{1}{2\sigma^2}\left[x_n - \sum_{k=1}^{P} a_k x_{n-k}\right]^2\right), \tag{3.58}$$

where $\mathbf{w}$ is the length N vector of samples w_n, and $\mathbf{a}$ is the length P vector of parameters a_k. In the maximum likelihood approach, the specific parameter vector $\mathbf{a}_{\mathrm{ML}}$ that maximises this probability leads to the best model:

$$\mathbf{a}_{\mathrm{ML}} = \arg\max_{\mathbf{a}} P\left(\mathbf{w}|\,\mathbf{a}\right). \tag{3.59}$$

Since this probability is always positive, we can minimise the negative of the natural logarithm instead:

$$-\ln P\left(\mathbf{w}|\,\mathbf{a}\right) = N\ln\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[x_n - \sum_{k=1}^{P} a_k x_{n-k}\right]^2 \tag{3.60}$$

At the minimum of this quantity, the variation with respect to the parameters a_k is zero. Equating the partial derivative with respect to a_k to zero gives:

$$\frac{\partial}{\partial a_k}\left[-\ln P\left(\mathbf{w}|\,\mathbf{a}\right)\right] = \frac{\partial}{\partial a_k}\left[N\ln\left(\sqrt{2\pi\sigma^2}\right)\right] + \frac{\partial}{\partial a_k}\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[x_n - \sum_{k=1}^{P} a_k x_{n-k}\right]^2 = 0, \tag{3.61}$$

for $k = 1, 2 \ldots P$. However, the first term in the middle expression does not depend upon the parameters, leaving the second term, for which the constant scaling factor $1/2\sigma^2$ cancels. This leads to the following set of equations:

$$\frac{\partial}{\partial a_k}\sum_{n=0}^{N-1}\left[x_n - \sum_{k=1}^{P} a_k x_{n-k}\right]^2 = 0, \tag{3.62}$$

for $k = 1, 2 \ldots P$. But this is exactly the same as the least-squares error formulation of the previous section. Therefore, the least-squares approach and the Gaussian system input, maximum likelihood approach are mathematically equivalent.

System Input Energy Minimisation

As in the least-squares approach above, consider that the system input of finite length N is an unknown, real-valued signal e_n that has finite energy:

$$E_e = \sum_{n=0}^{N-1} |e_n|^2 = \sum_{n=0}^{N-1} e_n^2. \tag{3.63}$$

If there is good cause to believe that the energy in the system output $E_x = \sum_{n=0}^{N-1} x_n^2$ is nearly all due to the response of the system rather than the input, then it is reasonable

to state that the best model is one whose parameters minimise the input signal energy. Since this expression is just the sum of squares of the input signal, this approach again leads to the least-squares approach.

3.3 Applications and Limitations for Speech Processing

As mentioned earlier, the techniques of signal processing based around LTI, discrete time systems theory presented earlier have found their way into a large number of practical applications in speech processing. This section focuses on two fundamentally important techniques: *LPA formant analysis* and *power spectral density estimation*, due to their ubiquity in technological applications. It will then discuss the limitations of these techniques due to their origins in LTI systems theory.

The application area of *speech compression* is an ideal case study. Digital speech signals are transmitted over telecommunications networks or stored in computer memory as *binary signals*, using only the binary digits 0 and 1. The *bit rate* (in bits per second) required to transmit the digital speech signal determines the *bandwidth* of the network. Similarly, the bit rate determines the amount of computer memory required to store the speech signal [58]. Typically, good quality digital speech signals are sampled using 16 bits per sample (giving $2^{16} = 65536$ different quantisation levels) at a sampling rate of 8kHz, leading to a bit rate of $128,000$ bits per second. The cost of a network is largely determined by the required bandwidth, so that there is an economic imperative to reduce the bit rate of speech signals to build more cost-effective networks. Speech *codecs* (a contraction of encoder/decoder) are digital devices that perform bit rate reduction (compression) of speech signals. There exist a very large number of standard codecs in current use, but the most ubiquitous of these are those that can be grouped under the umbrella term *Code-Excited Linear Prediction* (CELP) . Examples of such codecs and variants are integrated into the current mobile telephone networks of Europe, Japan and North America [59]. See figure 3.1 which shows, in block diagram form, the basic overall structure of the typical CELP codec.

The second application area is the calculation of the *spectrogram*. *Spectrographic analysis* is a fundamental technique in acoustic phonetics [10, 14]. It is based essentially upon the power spectrum of a speech signal, updated at regular intervals in time. A spectrogram is a graphical display of the changing magnitude of frequency components in a (discrete time) signal, with time on the horizontal and frequency on the vertical axes. The

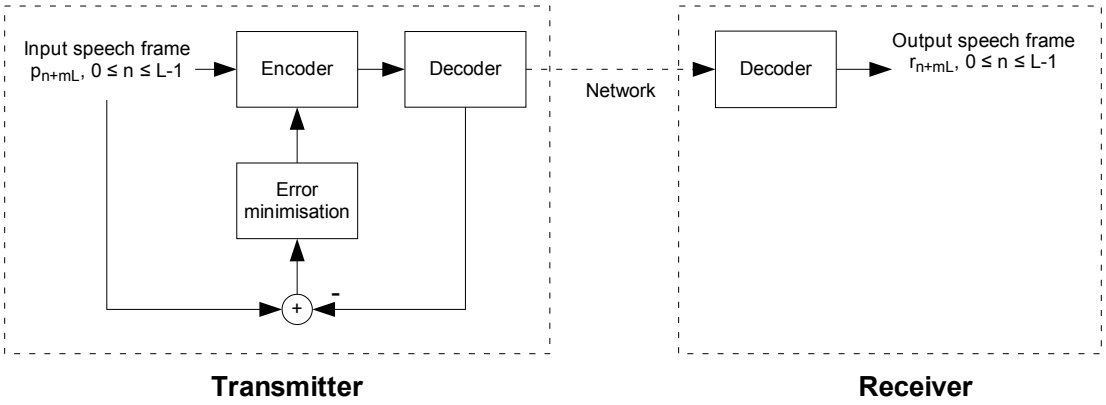

Figure 3.1: Block diagram of the structure of the typical CELP codec. The input speech signal p_n is processed in frames of length L samples, for the frame number $m = 0, 1, \ldots$. The frame is processed in the encoder at the transmitting end with LPA to extract the linear system parameters and calculate the residual e_n. The coded frame data is then passed on to a local copy of the decoder, which reconstructs the speech frame. The difference between this reconstruction and the input speech frame is passed to an error minimisation step, which informs the encoder to produce a better encoding for the speech frame. This process of encoding, decoding and error minimisation proceeds iteratively until an acceptable quality encoding for the frame is produced. This best encoding is transmitted over the network to the receiver, where an identical decoder reconstructs the speech frame. This overall process is repeated frame by frame to create the reconstruction r_n of the speech signal.

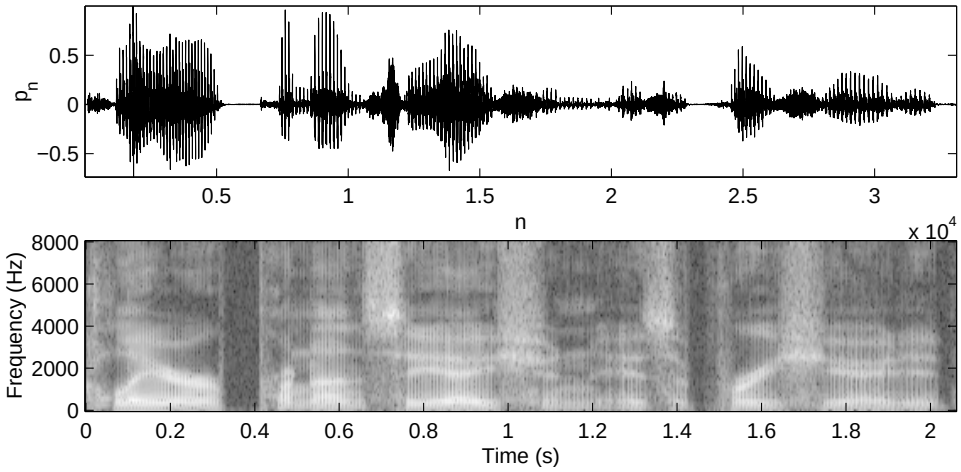

Figure 3.2: Spectrogram of the spoken phrase "Clear pronunciation is appreciated" from a male speaker, recorded in 16 bits, this recording at a sample rate of 16kHz (which is faster than typical telecommunications systems). The top panel shows the speech pressure signal p_n, the bottom the spectrogram, using 128 point DFTs, with 50% overlap and the Hanning window. The bright bands on the spectrogram show the changing formants. The speech data is taken from the TIMIT database [60].

brightness/darkness or colour of the plot at each time/frequency location on the graph is proportional to the square magnitude of the frequency component at that time and frequency. The spectrogram is useful for visually identifying the changing formants in the phonemes that make up spoken words [14]. Figure 3.2 shows a typical spectrogram of a spoken phrase from the TIMIT database [60].

3.3.1 Digital Formant LPA

This application of LPA is motivated by a discrete time version of the source-filter equation (2.18), in the following way. The continuously-varying cross-sectional area acoustic tube vocal tract model of Chapter 2 is instead approximated by a series of concatenated, rigid acoustic tubes each with constant cross-sectional area, and any losses due to viscosity and heat conduction are ignored. [3] It can be shown [13] that this concatenated tube system, as a whole, has a rational acoustic transfer function $H(\omega)$ with only denominator terms, for which the discrete time counterpart of this tube is simply the LTI system of equation (3.9) described above, with transfer function $H(z)$, equation (3.23). Thus, the discrete time speech pressure signal p_n is taken to be the output of an LTI system driven by an

[3] Similarly, the bend in the vocal tract is ignored as discussed in Chapter 2.

input signal e_n:

$$p_n = \sum_{k=1}^{P} a_k p_{n-k} + e_n, \tag{3.64}$$

with initial conditions p_{-j} for $1 \leq j \leq P$ determined from the actual speech signal p_n. Taking the z-transform then gives:

$$P(z) = H(z)E(z). \tag{3.65}$$

Given the system coefficients a_k and the speech pressure signal p_n, the equation (3.64) may be solved for e_n, and thus we can, in effect, calculate the input driving signal to the model of equation (3.9). Assuming that the simplified, piecewise constant cross-sectional area biomechanical model is correct, LPA can therefore be used to identify the coefficients a_k of the linear system with transfer function $H(z)$ that represents the combined effect of the resonances of the vocal tract and the radiation impedance at the lips. The term $E(z)$ then represents the input to this system, i.e. the flow rate at the top of the vocal folds for voiced sounds or the vortex sound generation sources in the vocal tract for unvoiced sounds. For a more in-depth exposition of these concepts, see [15]. The resonances of the vocal tract of the speaker, entirely represented in the system coefficients a_k, contain important information about the phonemic content of the spoken words. Figures 3.3 and 3.4 show the results of LPA applied to one example each of a voiced and unvoiced speech signal.

The basic process of CELP speech compression uses LPA at the transmitting end of the network to identify formants of the speaker's phonemes. LPA analysis is performed on a small time interval of the speech signal, called a *frame*. [4] The resonances of the vocal tract of the speaker at the transmitting end are represented in the system coefficients a_k. These coefficients are digitally encoded and transmitted, along with a coded representation of the *residual*, which is the error signal e_n of equation (3.51), over the network to the receiver. At the receiving end, the coefficients and the residual are decoded. Given the initial conditions p_{-j}, $1 \leq j \leq P$ and the residual together with the system coefficients, equation (3.64) is used at the receiver to reconstruct the original speech signal p_n for this frame. This process of LPA analysis, coding, transmission, decoding and reconstruction is repeated for the next time frame, and so on. Refer to figure 3.1 for a diagram of this process.

[4] Typically each frame is 20 to 30 milliseconds in length, which is between 160 and 240 samples at a sample rate of 8kHz.

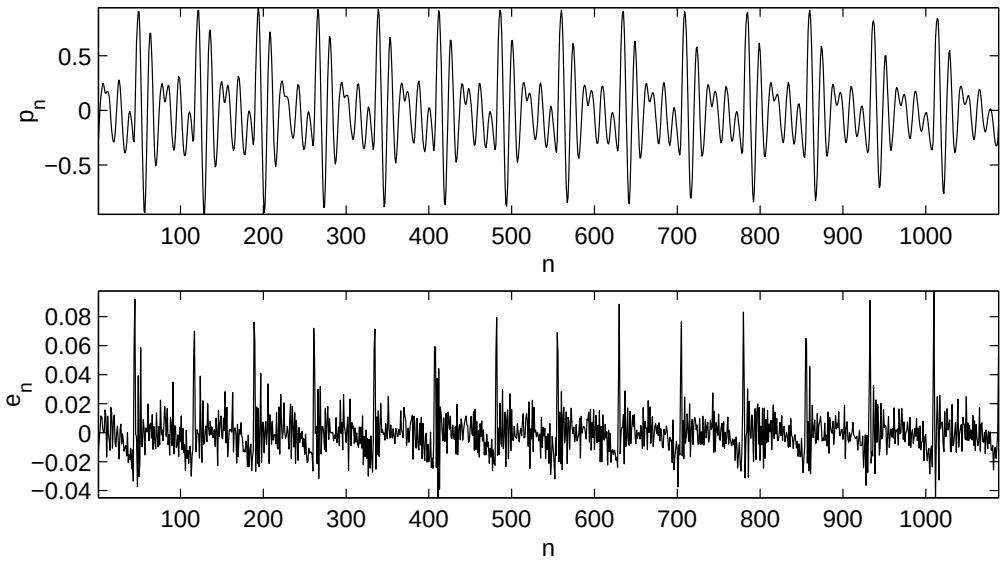

Figure 3.3: Linear prediction analysis applied to a voiced speech signal, of system order $P = 20$. Top panel is the original speech pressure signal p_n, bottom panel is the term e_n obtained by solving for this term in equation (3.64) with zero initial conditions $p_{-j} = 0$ for $1 \leq j \leq P$. The initial transient response of the system (100 samples) has been discarded.

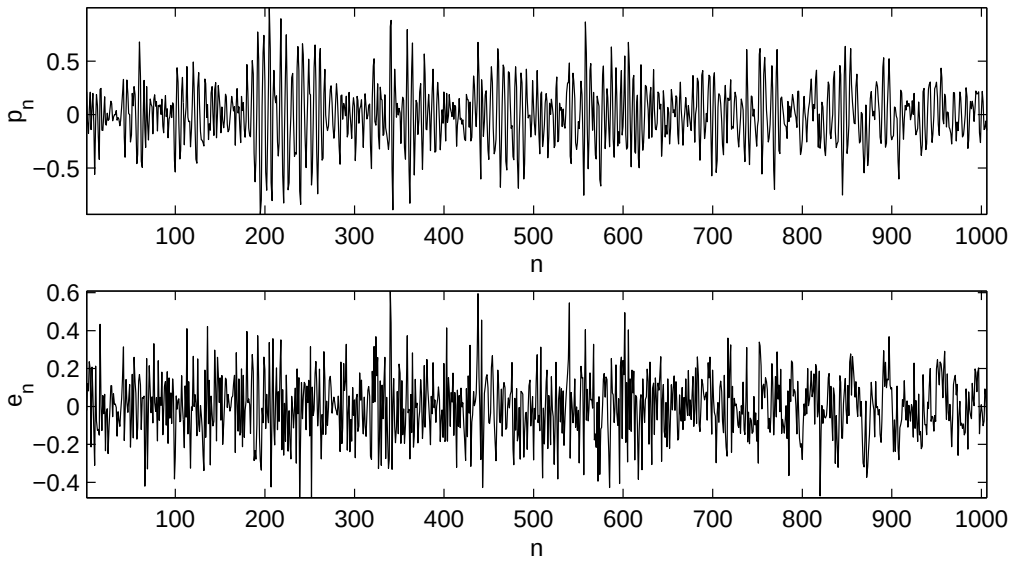

Figure 3.4: Linear prediction analysis applied to an unvoiced speech signal. Top panel is the original speech pressure signal p_n, bottom panel is the term e_n obtained by solving for this term in equation (3.64) with zero initial conditions $p_{-j} = 0$ for $1 \leq j \leq P$. The initial transient response of the system (100 samples) has been discarded.

Advanced CELP codecs, can, at best, reduce bit rates [5] from 128,000 down to 800 bits per second [59]. This compression arises mainly due to the application of an efficient, parameterised representation of the residual signal e_n. For voiced sounds, it is found experimentally that the residual consists of a series of regular impulses superimposed onto a small amplitude, noise-like signal. For unvoiced sounds the residual is generally noise-like (see the residuals in figures 3.3 and 3.4). CELP codecs calculate a coded representation of these impulses for voiced sounds, along with a coded representation of the noise-like signal. It is this coded representation that is sent over the network to the receiver, rather than the residual signal. At the receiving end, the representation is decoded to create an approximate reconstruction of the residual. There will certainly be significant error in this reconstruction, but the reconstruction is of sufficient quality that the final, reconstructed speech pressure signal is intelligible. However, the bit rate of the coded representation of the residual is far smaller than that of the original residual, thus achieving significant bit rate reduction [58].

This is a highly simplified account of CELP codecs, which have been iteratively improved with many refinements over several decades. Nonetheless, the basic application of LPA remains unchanged, even if the coding schemes and algorithms have become exceedingly elaborate and sophisticated. Formant LPA is an essential component of these systems [56]. However, discussion on the validity of LTI systems theory in formant LPA is required, to which we now turn.

LPA Error Minimisation by Least-Squares Optimisation

Assuming that the signal was generated by a recursive LTI system such as (3.64), the ideal goal of least-squares optimisation is to find the system coefficients that make the system input "error" signal e_n zero. When the system input is zero, the system (3.64) has no effective input, and only oscillates if the initial conditions are non-zero [12]. Such a zero-input model can be considered as a discrete version of a continuous-time acoustic resonator with no input. However, we know from the relevant biomechanics introduced in Chapter 2 that voiced speech production requires excitation of the vocal tract by acoustic coupling to the vocal folds, or to vortex sound generation sources. Therefore, minimisation of the input "error" signal is not well justified biomechanically. It has been suggested that this justification for formant LPA is valid when the vocal folds are completely closed [61]. However, for some individuals and for some cases of voice disorder, the vocal folds do

[5] Usually at the expense of a considerable loss in perceived quality of the reconstructed speech.

not close completely during normal oscillation [11], and identifying the time intervals in which the vocal folds are closed from the acoustic speech signal alone is a difficult problem. Similarly, there are no *apriori* reasons to conceptualise the vocal fold flow rate signal as an "error" that must be minimised to obtain accurate system coefficients, which can be used to calculate the formants of the vocal tract.

LPA Gaussian Maximum Likelihood

It is clear from Chapter 2 that the vocal folds do not produce a stochastic excitation signal driving the vocal tract, let alone being an i.i.d. Gaussian stochastic process. Only for vortex sound generation sources is the stochastic excitation signal assumption plausible, but in that case it is not i.i.d. Using LPA on a known LTI system driven by an unknown signal will generally lead to significant errors in the estimation of the system coefficients, casting considerable doubt that the Gaussian maximum likelihood justification for formant LPA can be adequately interpreted as recovering the actual resonances of the vocal tract.

The following simple experiment demonstrates the problem. Consider a simple linear system $P = 1$ of the form of equation (3.9) with $a_1 = -0.9$. Starting with zero initial conditions ($y_{-1} = 0$), apply a Gaussian, stochastic i.i.d. time series of unit variance and zero mean w_n to the system input (i.e. set $x_n = w_n$). Then, for 30 realisations of an input signal of length $N = 1024$ samples, the subsequent application of LPA to the system output y_n obtains a mean estimate of $a_1 = -0.900$ to three decimal places. The standard deviation is 0.014. Assuming that this estimate has a Gaussian distribution, the 95% confidence interval is $[-0.928, -0.873]$ to three decimal places.

Now we replace the stochastic input term with the sampled x co-ordinate of a system of nonlinear ordinary differential (ODE) equations, the Rössler system [40]. The parameters of the nonlinear ODE system were $a = 0.2, b = 0.4, c = 8.0$ and the initial conditions were $x(0) = 5, y(0) = 1, z(0) = 0.1$, solved using 4th-order Runge-Kutta integration. The input signal x_n is then samples of the x-co-ordinate of the ODE system obtained at each integration time step n. For 30 successive time intervals of the integration of this system, each of the same length $N = 1024$ samples, LPA obtained a mean estimate of $a_1 = -0.999$ to three decimal places. The standard deviation is 0.001, and the 95% Gaussian confidence interval is $[-0.998, -1]$ to three decimal places. Thus it can be seen that LPA applied to the output of an LTI system with a nonlinear dynamical system input signal introduces significant error in the estimation of the true system coefficients.

We have seen, in Chapter 2, that the biomechanical models of the vocal folds are a

nonlinear dynamical system acting as an input to the linear vocal tract system. Thus, from biomechanical considerations, the Gaussian maximum likelihood interpretation of LPA used for formant analysis is inconsistent with the physical situation. This inconsistency, coupled with the above demonstration, casts doubt over the effectiveness of LPA in the estimation of the vocal tract system coefficients.

LPA by System Input Energy Minimisation

Minimising the energy in the input signal implies an assumption that the vocal tract resonator system accounts for nearly all the energy in the speech signal. However, there exists no obvious reason to believe in advance that this is true, indeed, for voiced speech the major source of energy is the exhalation of air from the lungs, and this energy is transferred partly into the vocal folds to sustain vibrations. The vocal tract is modelled as a passive resonator that merely vibrates in sympathy with this source of oscillation energy.

Time-Invariance

As described in Chapter 2 the production of spoken words and phrases involves the complex, co-ordinated articulation of the vocal muscles to shape the formants of the speech pressure signal. This leads to an inherently non-stationary process whereby the phonemes merge into one another – it is never entirely clear where the boundaries between phonemes are located in general [10]. Thus the time-invariance requirement of LTI systems theory is fundamentally inappropriate for ordinary, running speech.

3.3.2 Power Spectral Density Estimation

Characterisation of signals in terms of constituent components is a useful tool in signal processing. In speech processing, the power spectrum carries important information about the phonemic content of the speech signal. For finite length discrete time signals, the DFT can be used as a basic technique in *nonparametric Power Spectral Density estimation* (PSD) [12]. For a signal x_n the discrete power spectrum $P_{xx}(k)$ can form the basis of nonparametric power spectral density estimates obtained, for example, by calculating several overlapping DFTs and averaging the $P_{xx}(k)$ values [12]. For a more in-depth discussion of PSD estimation, see [12].

One approach to calculate the spectrogram of a speech signal is to use PSD estimation

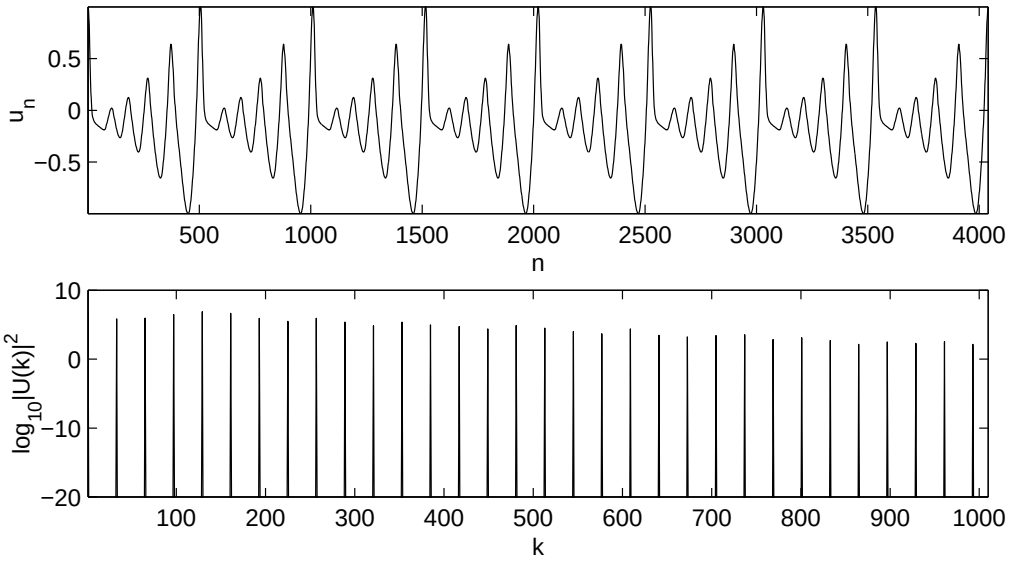

Figure 3.5: Power spectrum of a periodic signal. The top panel shows the signal u_n, the bottom panel the power spectrum calculated using the DFT with rectangular windowing. For clarity only part of the signal and the spectrum are shown. The signal length is 16160 samples.

which is often based on the DFT. Estimation is carried out on successive finite intervals of the speech pressure signal p_n. The averaged values of $P_{xx}(k)$ for each interval are the data from which the spectrogram plot is constructed.

It can be shown that for (discrete time) periodic signals, the discrete power spectrum consists of a series of regularly-spaced unit impulses. The spacing between these impulses is inversely proportional to the period of the signal [12]. Similarly, for zero mean, i.i.d. Gaussian stochastic processes, the power spectrum is non-zero and constant for all values of k. This is because, as we have observed, the autocorrelation of the i.i.d. Gaussian stochastic signal is the variance multiplied by the unit impulse, and the DFT of this is just the variance, which is constant. Therefore, by the Wiener-Khintchine theorem, the discrete power spectrum is also constant. However, in Chapter 2 it was demonstrated that some forms of vocal fold oscillation are highly complex and irregular. Thus the resulting speech pressure signals will also be irregular. Experimentally, the DFT of such irregular signals, even if they are not stochastic, is indistinguishable from a stochastic process that has non-zero autocorrelation for time lags l greater than zero. This is demonstrated in figures 3.5, 3.6 and 3.7 which show the power spectrum of a periodic signal, a stochastic signal and a chaotic signal (one of the co-ordinates of the Rössler system in a chaotic regime [40]).

Thus, complex, irregular and chaotic signals are generally difficult to distinguish from

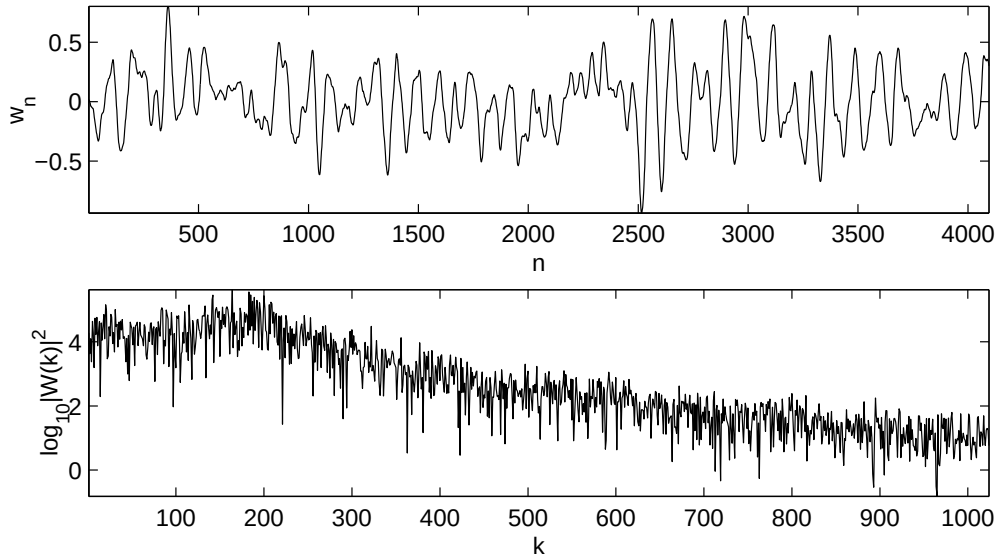

Figure 3.6: Power spectrum of a stochastic process. The top panel shows the signal w_n, the bottom panel the power spectrum estimated using the DFT with rectangular windowing. For clarity only part of the signal and the spectrum are shown. The signal length is 16384 samples.

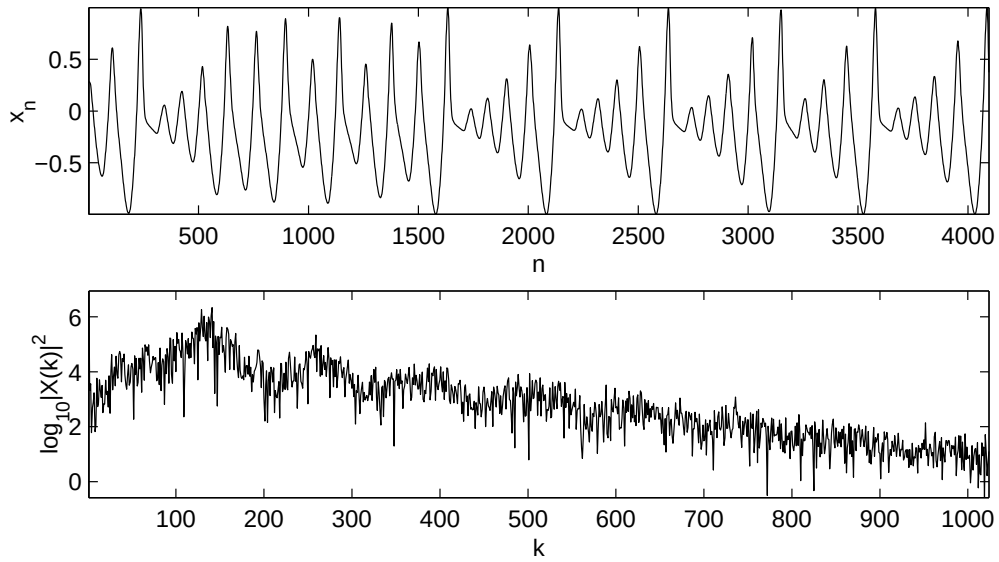

Figure 3.7: Power spectrum of a chaotic signal, the x-co-ordinate from 4th-order Runge-Kutta numerical integration of the Rössler system of ODEs [40]. The ODE system parameters were $a = 0.2, b = 0.4, c = 8.0$ and the initial conditions were $x(0) = 5, y(0) = 1, z(0) = 0.1$. The top panel shows the signal x_n, the x-co-ordinate of the ODE, normalised to the range $[-1, 1]$, obtained at each integration step n. The bottom panel the power spectrum estimated using the DFT with rectangular windowing. For clarity only part of the signal and the spectrum are shown. The signal length is 16384 samples.

stochastic signals from the power spectrum alone. It is quite possible to misinterpret a spectrogram based upon the DFT as showing unvoiced phonemes when actually the vocal folds are in vibration. This is a fundamental limitation stemming from the assumptions of Fourier analysis.

3.4 Chapter Summary

In this chapter we have introduced the mathematical foundations of LTI systems theory, and the techniques of linear digital signal processing based upon those foundations. This was followed by an overview of the widespread use of two of these techniques, as exemplified in two ubiquitous speech processing technologies. A critical examination of the validity of the assumptions underlying these techniques concluded that, with reference to current understanding of the biomechanics of speech production, there are certain inconsistencies that undermine the use of these methods for speech signal processing purposes. For example, there is no interpretation of LPA that can be said to correctly identify the vocal tract resonances from speech signals, and, as we shall see in later chapters, there is, in addition, clear evidence that real speech signals do not conform to the assumptions of LTI systems theory.

CHAPTER 4

Nonlinear Time Series Analysis

The previous chapters have established theoretical and initial empirical evidence that supports the claim that LTI systems theory is not adequate for representing all the dynamics of speech production. One approach to tackling this inadequacy is to relax some of the underlying mathematical assumptions, and, in particular, the fairly stringent requirement of linear superposition. Moreover, we can actually consider the LTI systems introduced in Chapter 3 as special cases of discrete time, *nonlinear dynamical systems*. As such, the latter are a natural generalisation of LTI systems, a generalisation we make in an attempt to produce new, discrete time models of speech production that are capable of capturing more of the dynamics of speech than linear techniques can. This chapter introduces the foundational mathematics required (which will be used in later chapters) to produce more extensive and rigorous evidence for nonlinearity in speech, and subsequently demonstrates how that nonlinearity might be exploited in new signal processing methods with practical applications.

4.1 Discrete-Time, Nonlinear, Random Dynamical Systems

In §3.1 we described how in practice, continuous time signals are sampled and quantised to create a digital version stored in computer memory for subsequent processing. Much as practical digital formant and spectral analysis uses digital representations of speech production, here we will also restrict our attention to such digitised signals. We assume that such a representation is accurate enough to create useful, parsimonious models. [1] We introduce in this section useful discrete time models belonging to the class of nonlinear dynamical systems with P-element *state space* vectors $\mathbf{y}_n \in R$ where R is a compact subset

[1] As a preliminary note, we state that, for practical reasons, since all the physical signals x_n we will encounter are bounded ($|x_n| < \infty$), and to make use of some powerful theorems, in general, unless stated otherwise, all the subsets we describe are compact, and all the functions C^1 diffeomorphisms, that is, they are differentiable and have a differentiable inverse.

of $\mathbb{R}^P$. The system is also *forced* by a P-element input vector signal $\mathbf{x}_n$:

$$\mathbf{y}_n = \mathbf{F}\left(\mathbf{y}_{n-1}, \mathbf{a}\right) + \mathbf{x}_n, \tag{4.1}$$

where the vector *system function* $\mathbf{F} : \mathbb{R}^P \times \mathbb{R}^Q \to \mathbb{R}$ maps the past system state $\mathbf{y}_{n-1}$ onto the current state $\mathbf{y}_n$. The parameter vector $\mathbf{a} = [a_1, a_2 \ldots a_Q]^T$ contains real elements. Such systems do not generally obey the superposition principle, although we can represent the system of (3.9) in the form of (4.1) and this model *is* LTI. See Appendix §A.2.4 for a proof of this. Exactly as with the LTI system, for time indices $n \geq n_0$ and some initial time n_0, the initial condition vector $\mathbf{y}_0$ together with the input signal $\mathbf{x}_n$ are required to calculate all subsequent values of $\mathbf{y}_n$.

Under mild restrictions (namely, for reasons described later in this chapter, we require x_n to be drawn from a compact probability space), the forcing vector $\mathbf{x}_n$ can be any vector stochastic process, and is not required to be i.i.d. or Gaussian. The resulting signals $\mathbf{y}_n$ have a natural *discrete time Markov chain* description [62, 54], since $\mathbf{y}_n$ can be known from information contained only in $\mathbf{y}_{n-1}$, formally:

$$P\left(\mathbf{y}_n \,|\, \mathbf{y}_j, j = 0, 1 \ldots n-1\right) = P\left(\mathbf{y}_n \,|\, \mathbf{y}_{n-1}\right), \tag{4.2}$$

so that the Markov property holds for the stochastic process $\mathbf{y}_n$. This property allows us to view the system of equation (4.1) as a source of discrete time stochastic processes which are generally non-Gaussian, opening up the possibility of analysis with tools from probability theory. On the other hand, we can view such systems as deterministic nonlinear systems forced by stochastic input. Both of these different viewpoints have value in bringing out distinct aspects of the behaviour of the system and in this thesis we will switch between them as appropriate. This interesting intersection between nonlinear dynamical systems and probability theory is a large and open area of research beyond the scope of this thesis, for more details see for example [62, 63].

4.2 Deterministic Maps

The system (4.1) without the forcing vector $\mathbf{x}_n$:

$$\mathbf{y}_n = \mathbf{F}\left(\mathbf{y}_{n-1}, \mathbf{a}\right), \tag{4.3}$$

is completely determined by the system function $\mathbf{F}$, parameter vector $\mathbf{a}$, and the initial condition vector. Such systems are called *maps* in the nonlinear dynamical systems literature [64]. There are many special mathematical tools that have been developed to study

such nonlinear systems (for example, bifurcation theory) [40, 64] but these tools usually require an explicit expression for the system function $\mathbf{F}$. In this thesis we will not assume that we have this expression and therefore we cannot use these tools. However, the class of systems where $\mathbf{F}$ is a diffeomorphism have certain special properties which we now describe.

4.2.1 Orbits

A (half) *orbit* is a sequence of points $\{\mathbf{y}_n, n \geq 0\}$ defined by the system (4.3). Any initial point $\mathbf{y}_0$ generates a unique orbit [64].

4.2.2 Invariant Sets

Orbits that diverge away to infinity are of little interest in this thesis; we only study here systems that produce orbits bounded within R. For such systems, *invariant sets* can arise, here defined simply as those sets $A \subset R$ that are solutions to the equation:

$$A = \lim_{k \to \infty} \mathbf{F}^k(A, \mathbf{a}), \tag{4.4}$$

for $k \in \mathbb{N}$ and do not contain any subsets that are themselves solutions to this equation. Here $\mathbf{F}^k$ denotes the k-th composition of $\mathbf{F}$ such that $\mathbf{F}^0$ is the identity and $\mathbf{F}^{k-1}(\mathbf{F}^1) = \mathbf{F}^k$. When a system function $\mathbf{F}$ admits such an invariant set, that set may be characterised into one of several distinct types. We will now discuss two of these types of importance to this thesis.

Periodic Orbits

Invariant sets A composed of K distinct points are called *periodic orbits*, and the special case of $K = 1$ are called *fixed points*. A sufficient condition for the existence of a unique fixed point contained in some subset $D \subset R$, where D is closed and bounded, is that the function $\mathbf{F}$ is a *contraction* on D: the (Euclidean) distance between any two distinct points in D gets smaller under $\mathbf{F}$. This is essentially an application of the contraction mapping theorem [47, 64].

Chaotic Orbits

Systems such as (4.3) admit much more complicated invariant sets than periodic orbits. Informally, there exist system functions $\mathbf{F}$ that will eventually, under repeated iteration,

separate two arbitrarily close points until they are any given distance apart. There exist systems that can separate points exponentially fast. The average rate at which two nearby points in state space are separated, called the (global) *Lyapunov exponent*, is a measure of the overall, exponential expansion rate of the system [64]. Systems which separate points at a positive exponential rate in at least one direction are said to have *sensitive dependence upon initial conditions*, since any small perturbation of the initial conditions gets amplified until it affects the orbit on the scale of the size of the set R in which it is contained. This is the commonly accepted definition of *chaos* [64].

To remain bounded within R, a chaotic system must simultaneously expand distances in some part of state space and contract them in others. This combination of expansion and contraction can lead to very complicated invariant sets that sometimes display *self similarity* in state space. Such sets are composed of parts that are geometrically similar to the whole set, but scaled by some constant factor. Such sets are commonly called *fractals* [47]. Due to this geometric self similarity these sets also do not have integer *dimension* [47]. It is important to note that there are several different commonly-used dimension measures (for example box-counting dimension, Hausdorff dimension, correlation dimension) which can have quite different values for the same set [47]. We will discuss fractal sets in a later section of this chapter.

We make the informal remark that signals x_n produced by such chaotic systems can display considerable irregularity, which is apparently at odds with their entirely deterministic and often quite simple origins in equations such as (4.3) or the Rössler system used in the demonstrations of §3.3.2 [64].

4.3 Recurrence

Of importance to random dynamical systems (4.1) and deterministic systems (4.3) is the concept of *recurrence* in state space [64, 65, 66]. Whilst there are many definitions of recurrence in the literature used for specific, technical purposes, [2] in this thesis we will define *recurrent orbits* $\{\mathbf{y}_n, n \geq 0\}$ as those that return to a given subset of state space after a time delay $\Delta n > 0$ [65]:

$$\mathbf{y}_n \subset B(\mathbf{y}_{n+\Delta n}, r), \tag{4.5}$$

[2] For example, *nonwandering* and *chain recurrent* sets embody a much weaker concept of recurrence than we use in this thesis [64].

where $B(\mathbf{y}_n, r)$ is a closed ball of radius $r > 0$ around the point $\mathbf{y}_n$ in state space, and $\mathbf{y}_n \not\subset B(\mathbf{y}_{n+m}, r)$ for $0 < m < \Delta n$. Each different n may be generally associated with a different Δn, called the *recurrence time*. An important remark to make here is that a periodic orbit is a special kind of recurrent orbit in which $r = 0$ and $\Delta n = K$, the period of the orbit, is the same for all n, so that:

$$\mathbf{y}_n = \mathbf{y}_{n+\Delta n}. \tag{4.6}$$

Lastly, we will, for the purposes of this thesis, define an *aperiodic orbit* as recurrent but not periodic. [3] These concepts of periodic and aperiodic are therefore mutually exclusive, but are both special cases of the more general concept of recurrence.

We will see later in this thesis that *recurrence time statistics* [66] provide valuable information about the properties of nonlinear, random and deterministic dynamical systems [67] that will find practical usage. [4]

4.4 Time-Delay Reconstruction

Although we assume that the model (4.1) is responsible for generating the system state $\mathbf{y}_n$, in practice we usually do not have access to the precise values of the system state at any one time n. By contrast, we usually only have a measurement of a single element of the system state vector available through a smooth *measurement function* $h : \mathbb{R}^P \to \mathbb{R}$ that maps the system state $\mathbf{y}_n$ on to a univariate digital signal s_n:

$$s_n = h\left(\mathbf{y}_n\right). \tag{4.7}$$

It is not immediately obvious but despite the fact that the system state $\mathbf{y}_n$ lies in a subset of the P-dimensional space and the measured signal is one dimensional, s_n actually contains much useful information about the original system function $\mathbf{F}$. In this thesis we will make use of two *embedding theorems* that, informally, allow the *reconstruction* of the system function $\mathbf{F}$ from the measurements of s_n alone. They are both based around the construction of a *time-delay reconstruction map* $\Theta : R \to \mathbb{R}^d$ which is defined as:

$$\Theta(\mathbf{y}_n) = \left[h(\mathbf{y}_n), h(\mathbf{y}_{n-\tau}), h(\mathbf{y}_{n-2\tau}) \ldots h(\mathbf{y}_{n-(d-1)\tau})\right]^T, \tag{4.8}$$

[3] This usage departs somewhat from the literature where aperiodic has a technical meaning in studies of nonlinear dynamical systems – here we are simply concerned with expressing what we mean by recurrence which is not strictly periodic.

[4] For example, recurrence analysis forms the basis of the method of recurrence plots in nonlinear time series analysis [68].

where $d \in \mathbb{N}$ is called the *reconstruction dimension*, and $\tau \in \mathbb{N}$ is the *reconstruction delay*.

The first theorem, commonly referred to as *Taken's Embedding Theorem* [69], which applies exclusively to deterministic systems such as (4.3), states that for typical $\mathbf{F}$ and h, and for the compact manifold R of dimension m, if $d \geq 2m+1$, then the time-delay map Θ is an *embedding* (that is, a diffeomorphic map) of R on to a compact subset S of the *embedding state space* $\mathbb{R}^d$. See [69, 70] for a rigorous proof of this. [5]

This theorem implies the existence of a *dynamical conjugacy*: for typical $\mathbf{F}$ and h, $S = \Theta(R)$ is equivalent to R, up to the coordinate change Θ. We can define a new system on S with the system function $\mathbf{G} = \Theta \circ \mathbf{F} \circ \Theta^{-1}$, which shares all the coordinate independent attributes of $\mathbf{F}$ such as Lyapunov exponents, existence of invariant sets A and other topological properties [69]. It is in this sense that time-delay reconstruction allows the recovery of $\mathbf{F}$ from the observations s_n alone.

There are some practical difficulties with the use of this theorem: for example, we usually do not know the dimension m of any invariant set A in advance. If d is too small then the reconstruction fails, and setting d too large introduces redundant coordinates which may lead to computational problems when handling an excessively large amount of data. There are a variety of practical algorithms that have been devised to find an appropriate value of d, including the method of *false-nearest neighbours* and *PCA embedding* [8]. Furthermore, we need to choose a particular reconstruction time delay τ. If τ is too small then points in the reconstructed space tend to cluster around the diagonal; at the other extreme when τ is too large the coordinates become increasingly dynamically unrelated, particularly if the orbit is chaotic. Many approaches exist for selecting an appropriate time delay: choosing the first time delay at which the autocorrelation crosses zero, or choosing the first minimum of the *time-delayed mutual information* [8].

The second theorem is a more recent extension of Taken's embedding theorem, and it applies to the more general, forced systems such as (4.1). Since in this thesis we will be concerned with *stochastic* forcing, of relevance here is the so-called *Stochastic Taken's Embedding Theorem* [70]. This states that, as in the deterministic case, the time-delay map Θ is also an embedding for $d \geq 2m+1$, where the state space of the system is confined to the set R of dimension m. However, the nature of the reconstruction differs from the deterministic version; this difference is made explicit in [70].

[5] To be more precise, the theorem states that there is an open and dense subset in the product of the space of all C^1 system functions $\mathbf{F}$ and C^1 measurement functions h for which the delay map Θ is an embedding. Also, the use of the term "typical" is technical and refers to specific set-theoretic properties, the detail of which is beyond the scope of this thesis.

In the stochastic forcing case, although a dynamical conjugacy $\mathbf{G}$ exists, it depends upon the particular realisation of the forcing terms $\mathbf{x}_n$ which are unknown in general. Nonetheless, the existence of an embedding Θ implies that the embedding space is still a faithful representation of the original system, and in some cases the forcing term may be small enough to be negligible in practical applications.

Finally, we point out here that quantising measurement functions such as those discussed in §3.1 are not C^1, and as a result the conditions of the embedding theorems are *technically* never satisfied in reality; nonetheless, it is common practice to assume that the quantisation resolution is sufficiently high that this issue can be ignored. We follow this practice in this thesis.

4.5 Information Theory and Time Series Analysis

We will have a variety of reasons to measure the *information* contained in a probability density. For example, dynamical systems such as (4.1) can be characterised in terms of the (instantaneous) probability densities of the stochastic processes that they generate. This will be used to produce a practical test for distinguishing linear from nonlinear or non-Gaussian systems.

4.5.1 Information and Entropy

For a probability density over the discrete random variable X, $P(X = i), i = 1, 2 \ldots N$, the *entropy*, or *average information content* is [17]:

$$H[X] = -\sum_{i=1}^{N} P(X = i) \ln P(X = i) = E\left[-\ln P(X)\right], \tag{4.9}$$

measured in units of *nats*, [6] using the convention $0 \ln 0 = 0$. Entropy satisfies the following properties:

- $H[X] \geq 0$,
- $H[X] = 0$ if and only if $P(X = i) = 1$ for one i only,
- $H[X] \leq \ln N$, and,
- $H[X] = \ln N$ if and only if $P(X = i) = 1/N$ for $i = 1, 2 \ldots N$.

[6] If the logarithm to base two is used instead of the natural logarithm, then entropy has the units of *bits*, coinciding with the usual meaning in computer science. The term *nat* suggests itself therefore when the natural logarithm is used instead.

In other words, entropy is non-negative and takes on the maximum value $\ln N$ for the uniform density. For these reasons, entropy is often called *uncertainty*, since a uniform density has the largest entropy and is the density for which we have the largest uncertainty about which outcome to expect in any particular trial.

For discrete random variables X and Y with joint density function $P(X = i, Y = j), i, j = 1, 2 \dots N$, the entropy extends naturally [17]:

$$H[X,Y] = -\sum_{i,j=1}^{N} P(X=i,Y=j) \ln P(X=i,Y=j) = E\left[-\ln P(X,Y)\right]. \tag{4.10}$$

As a consequence of this, if X and Y are independent, then $H[X,Y] = H[X] + H[Y]$.

Similarly, for conditional probability density functions $P(X = i|Y = j), i, j = 1, 2 \dots N$ the entropy satisfies:

$$H[X|Y] = -\sum_{i,j=1}^{N} P(X=i,Y=j) \ln P(X=i|Y=j) = H[X,Y] - H[Y], \tag{4.11}$$

which can be shown to follow from the definition $P(X|Y) = P(X,Y)/P(Y)$.

For a continuous probability density $p(x)$ over the random variable $x \in \mathbb{R}$ the *differential entropy* can be assigned similarly:

$$H[x] = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx, \tag{4.12}$$

whenever the integral exists. We note that differential entropy does not satisfy all the properties of the discrete case. In particular, it can be negative. Useful special cases include the continuous uniform density $p(x) = 1/(b-a)$ for $x \in [a,b]$ and $p(x) = 0$ otherwise, for which the (differential) entropy is $\ln(b-a)$. Also, we will make extensive use of the *multivariate Gaussian*:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M |\mathbf{C}|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)\mathbf{C}^{-1}(\mathbf{x}-\mu)^T\right], \tag{4.13}$$

for the M-element real-valued vector random variable $\mathbf{x}$ with mean vector μ and covariance matrix $\mathbf{C}$, where $|\mathbf{C}|$ is the determinant of $\mathbf{C}$. The entropy for this variable is [17]:

$$H[\mathbf{x}] = \frac{1}{2}\ln\left[(2\pi e)^M |\mathbf{C}|\right] = \frac{1}{2} M \ln[2\pi e] + \frac{1}{2}\ln|\mathbf{C}|. \tag{4.14}$$

4.5.2 Mutual Information

We will make use of measures of independence for random variables and stochastic processes. As seen in §3.2.8, Gaussian random variables are special in that they remain

Gaussian under linear transformations. We can therefore use methods based around second order statistical moments such as covariance and autocorrelation to characterise the stochastic time series generated by linear systems driven by Gaussian forcing terms. However, for more general nonlinear or non-Gaussian systems such as (4.1), we will need more general measures than this. [7]

In this thesis we will make use of the *mutual information* between pairs of random variables x, y:

$$I[x,y] = H[x] - H[x|y] = H[x] + H[y] - H[x,y] = E\left[-\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\right]. \qquad (4.15)$$

From this expression it can be seen that if x and y are independent, then the joint density factorises leading to $I[x,y] = 0$. Mutual information has the following useful properties:

- $I[x,y] \geq 0$,
- $I[x,y] = 0$ if and only if $p(x,y) = p(x)p(y)$, and,
- $I[x,y] = I[y,x]$.

For these reasons, mutual information is often described as a measure of independence between general non-Gaussian random variables, much as covariance is a measure of independence for Gaussian random variables.

In the context of dynamical systems such as (4.1), we consider the state at any instant n as a random (vector) variable. Then an estimate of the information shared between the states of the system at different instances in time separated by a time lag τ can be quantified using the *time-delayed mutual information* (TDMI) of the measured signal s_n from the system [8]:

$$I[s](\tau) = H[s_n] + H[s_{n-\tau}] - H[s_n, s_{n-\tau}], \qquad (4.16)$$

where $I[s](\tau)$ denotes $I[s_n, s_{n-\tau}]$, making use of estimates of the probability densities $p(s_n)$, $p(s_{n-\tau})$ and $p(s_n, s_{n-\tau})$. This measure can also be understood as a form of *nonlinear/non-Gaussian* autocorrelation.

It will be of use later in this chapter to find the TDMI for a known autocorrelation sequence $r_{ss}(\tau)$. Assuming that we have a signal s_n, which is the measured output y_n of a linear system such as (3.9) forced by a Gaussian, zero mean, strongly stationary, i.i.d.

[7] Higher-order statistical techniques (using higher order moments and cumulants) can also be used for this purpose; however such methods are unreliable for the very short signal lengths we consider in this thesis [24].

signal x_n, then this will always be possible. To see this, note that if these assumptions hold, then the signal s_n will itself be linear, zero-mean, Gaussian and strongly stationary. Therefore the joint density $p(s_n, s_{n-\tau})$ will depend upon the (absolute) time lag τ only, and we will denote this density as $p_\tau(u, v)$. Similarly, the marginal densities $p(s_n)$ and $p(s_{n-\tau})$ will be equal – we denote these densities as $p_0(u)$. The covariance matrix which describes the joint density $p_\tau(u, v)$ therefore has the following entries: [8]

$$\mathbf{C} = \begin{bmatrix} \sigma_{0,0} & \sigma_{\tau,0} \\ \sigma_{0,\tau} & \sigma_{\tau,\tau} \end{bmatrix} = \begin{bmatrix} r_{ss}(0) & r_{ss}(\tau) \\ r_{ss}(\tau) & r_{ss}(0) \end{bmatrix}, \tag{4.17}$$

where $\sigma_{i,j}$ is the covariance of the signal at time s_{n-i} and s_{n-j}. Using the expression (4.14) above we obtain:

$$I[s](\tau) = \frac{1}{2} \ln \left(\frac{r_{ss}^2(0)}{r_{ss}^2(0) - r_{ss}^2(\tau)} \right). \tag{4.18}$$

See Appendix §A.2.5 for a proof of this result. We will also need to know the TDMI for a zero mean, Gaussian i.i.d. stochastic signal w_n, which, making use of the expression (3.30) is:

$$I[w](\tau) = \frac{1}{2} \ln \left(2\pi e \sigma^2 \right) \delta_\tau. \tag{4.19}$$

where σ^2 is the variance of the Gaussian signal w_n and δ_τ is the Kronecker delta (as defined in Chapter 3).

4.5.3 Measuring Time-Delayed Mutual Information – A New Method

Later in this chapter we will wish to estimate the TDMI from a measured signal and compare this to cases where the autocorrelation is already known (such as equation (4.18)). Calculating the TDMI for a given, arbitrary signal s_n, assuming that it is a strongly stationary stochastic process, requires first estimating the probability densities $p_\tau(u, v)$ and $p_0(u)$. Subsequently, the entropy integral must be evaluated. Since the densities are not known in advance, and the integral is often analytically intractable, the entropies must be estimated numerically. This estimation introduces errors, which we now address.

It can be shown that estimating the densities by forming appropriate frequency histograms (counting the number of signal samples that fall into equal-width bins) and approximating the differential entropy integral using summation approaches the differential entropy asymptotically as the bin width tends to zero [17]. For finite bin width sizes, the entropy calculated using this summation is overestimated with an additive term.

[8] Note that in practice, the finite nature of real signals requires the use of circular autocorrelation estimates.

Similarly, for finite length signals, the smaller the bin width, the fewer points fall into each bin. Such a density representation leads to underestimates in the subsequent entropy value. At the other extreme, bins that are too large lead to almost uniform density representations and entropy overestimation. There will therefore be a best, compromise value of the bin width for each signal where the estimate is most accurate.

At the same time, error analysis due to finite length data from discrete probability densities shows bias that is also additive with the first order terms proportional to the number of bins (or the square of the number of bins in the case of joint random variables) and inversely proportional to the finite number of samples N [71, 72].

Finally, the differential entropy integral must be estimated using any one of a number of numerical integration methods, all of which have finite accuracy [73].

In order to mitigate these sources of error, we propose to use the simplest numerical integration method with accuracy better than Euler's method: the two-point *trapezoidal* method [73], which is accurate to order Δu^3 (where Δu is the bin width used in the density estimation.) [9] We propose further to cancel out any additive over- or under-estimates that depend upon the length of the signal N, the bin width Δu and the time lag τ. This correction is an (additive) calibration of the TDMI estimator using the known special case of the zero mean, i.i.d., Gaussian signal whose TDMI expression is known analytically (expression (4.19)), following [74]. The resulting TDMI estimator is denoted $I_{\mathrm{N}}[s](\tau)$ – see Appendix §A.3 for more details of the calculations involved.

4.6 Fractals

Whilst there is no universal agreement on the essential mathematical properties that constitute a fractal set [47], statistical or geometric *self-similarity* is often considered as definitive, and we adopt that convention in this thesis.

As mentioned above, many deterministic, nonlinear chaotic systems have fractal invariant sets. Also, when considered as *graphs* of n against s_n, the measured signals s_n produced by systems such (4.1) can exhibit statistical self-similarity. Furthermore, as described in §2.2.4, vortex sound generation mechanisms in speech may lead to signals which are statistically self-similar. As such, it will be useful to be able to characterise the statistical self-similarity properties of speech signals.

[9] This method is one member of a hierarchy of *Newton-Cotes* integration methods; see [73] for more details. Extensive investigation found that this is the most accurate Newton-Cotes method for the TDMI estimation performed in this study.

Although there exist methods for estimating the dimension of an orbit in the reconstructed state space, these methods place excessive requirements on data quality and quantity [8]. The speech signals we use in this thesis are too short and noisy to make this a tractable approach. In this study we will therefore only be concerned with measuring the statistical self-similarity properties of the graph of speech signals.

4.6.1 Statistical Scaling Exponents

Consider the real function f defined on a compact subset of the real line $[a, b]$ and define the *graph* of the function as the set:

$$\text{graph}f = \{(t, f(t)) : a \leq t \leq b\}. \tag{4.20}$$

Some continuous time stochastic processes have sets graph$f$ that are statistically self-similar, so that $f(t)$ has the same probability density as the rescaled process $g(t) = \gamma^{\alpha} f(t/\gamma)$ for some real $\alpha > 0$ and all real $\gamma > 0$. We call α the *scaling exponent* for the graph. In practice, we do not have access to the continuous function $f(t)$; instead we have a sampled and quantised version s_n, and assume this digital signal is a measured output of a system such as (4.1). Therefore, we can consider the measured signal s_n as a discrete time stochastic process which approximates a continuous-time stochastic process with a particular scaling exponent.

The digitisation process will have destroyed the smallest temporal and amplitude scales due to sampling and quantisation error. Also, the signal s_n is finite in length. The best we can hope for is a practical algorithm that can estimate, from the digitised signal s_n, the scaling exponent α. Practical algorithms that we will use in this thesis will be based upon fitting a straight line to an appropriate $\log - \log$ graph of some measured quantity against the inverse of the length scale.

4.7 Testing Against Gaussian Linearity

Given a signal s_n, it is of value to know, in advance of producing some model equation (4.1), what choices of function $\mathbf{F}$ might be most appropriate. Whilst data-driven model construction, as discussed in the introduction, requires fewer assumptions than first-principles modelling, it is still important to strive to make *appropriate* assumptions. As discussed earlier, one goal of this thesis is to test the assumptions of classical linear signal processing introduced in Chapter 3, due to their ubiquity in practical speech technologies, and the

evidence for nonlinearity from speech biomechanics discussed in Chapter 2. *Surrogate data tests* offer a practical way of testing precisely these kinds of modelling assumptions given speech signals alone [75, 74].

4.7.1 Hypothesis Test Design

Surrogate data tests are computational approaches to *statistical hypothesis testing*. An hypothesis test comprises the following components:

- A *null hypothesis* (call this H_0),
- An *alternative hypothesis* (call this H_1),
- A *test statistic*, and,
- The *level of significance* for this test statistic.

The hypotheses represent some supposed, mutually exclusive states of nature. Then the null hypothesis H_0 is rejected if the test statistic lies inside a *critical region*, which we can determine knowing the density of the test statistic given that H_0 is true, and the level of significance, S (a probability). We fail to reject H_0 otherwise. To decide upon the critical region we choose some level of statistical significance S which is the probability of rejecting H_0 when it is in fact true (the probability of making a Type I error, Type II being the failure to reject H_0 when it is in fact false). In practice, however, for general null hypotheses, the density of the test statistic given H_0 is often unknown. The goal of surrogate data testing is to compute surrogate data or *realisations* that conform to the null hypothesis. This will allow us to estimate the required density, and hence perform the test.

Since we specified a level of significance, we do not need to estimate the density and thus explicitly calculate the critical value; the test may instead be conducted using rank-order statistics. For a given S, we compute $M = 2/S - 1$ surrogate realisations [10] and the M test statistic values on these realisations. Then if the test statistic calculated on the original time series is the largest or smallest of all these $M + 1$ values, it has a probability S of producing a Type I error, as required.

[10] This is true for the *two-sided tests* we perform in this thesis.

4.7.2 Choice of Null Hypothesis

In Chapter 3, digital formant analysis and CELP coding were introduced as exemplary applications of LPA for speech technology. Therefore, we will address the underlying assumptions of LPA in the surrogate data test. This will augment the theoretical arguments already put forward with additional empirical evidence. As already mentioned, one interpretation of LPA is that the stochastic driving signal x_n is a zero mean, Gaussian, i.i.d., strongly stationary stochastic process. This leads to a zero mean, Gaussian, strongly stationary output stochastic process s_n with joint probability densities at different time steps which are jointly zero mean and Gaussian. It will therefore be valuable to posit the following null hypothesis and mutually exclusive alternative:

- H_0: The speech signal s_n was generated by a linear system such as (3.9) driven by a zero mean, strongly stationary, i.i.d., Gaussian stochastic process,

- H_1: The speech signal was not generated by the above system with the listed properties.

Clearly, rejection of H_0 entails the negation of any one of the listed properties (zero mean, strong stationarity etc.). A few interesting special cases that could lead to a rejection of H_0 include completely deterministic maps such as (4.3) and systems such as (4.1) driven by non-Gaussian stochastic processes, but also includes trivial alternatives such as an i.i.d., strongly stationary uniform stochastic process. Thus rejection of H_0 must be *taken in the context* of the theoretical arguments put forward in Chapter 2; without this context the test is interesting but not particularly informative.

4.7.3 Choice of Test Statistic

Having set up H_0, it is necessary to choose a particular test statistic [75]. The statistic must be capable of distinguishing between H_0 and H_1; however, consideration must also be given to other important factors.

Each additional *free parameter* used in the calculation of the statistic that affects the value of the statistic, that is, a variable in the statistic algorithm that must be chosen on the basis of experimentation alone, makes the test less reliable. This is because it is possible to "tune" this parameter to produce certain results on a particular data set, but changing this parameter can produce a different result on the same data set. There is

therefore no ultimately correct value for the parameter. For this reason we should prefer statistics that have as few free parameters as possible [76, 74].

Furthermore, we should prefer statistics for which *analytic* results are known, that is, for which the values of the test statistic can be computed explicitly for particular signals that either do or do not conform to H_0. This allows us to compare the results of the statistic against known special cases to ensure that the test is functioning correctly [74].

Similarly, statistics that are sensitive to other aspects of the time series independent of the status of H_0 or H_1 should be avoided. This is to guard against the problem that, for example, a statistic is sensitive to the variance of a signal, and the surrogate signals all have a larger variance than the original. These kinds of spurious sensitivities can lead to incorrect rejection of the null hypothesis [74, 77].

Finally, when testing large data sets it is important for practical reasons to choose a statistic that requires as little computational effort and time as possible.

There exist a very large variety of statistics that have been proposed in the literature on surrogate data testing, including correlation dimension [78], nonlinear prediction errors [79] and higher-order statistics – for more details see [75]. However, these statistics require setting several free parameters, there are few known analytical results about these statistics, they can be sensitive to incidental aspects of the time series such as variance or mean, and they require significant computational effort and resources. For our purposes this makes them less attractive than the time-delayed mutual information of §4.5.2, used by [80] for surrogate data testing, which is known analytically for our H_0, and, as we will demonstrate later, involves a minimum of free parameters. With the analytical results, we can introduce checks to screen for certain systematic errors, checks that we could not achieve with any of these other, less thoroughly understood statistics [74].

4.7.4 Generating Surrogates

As described earlier, in order to estimate the density of the test statistic given H_0, surrogate data tests involve the generation of M realisations u_n of the original signal s_n that are specifically designed to conform to H_0, in our particular case, surrogates that are jointly Gaussian, linear, zero mean, stochastic processes. There are several methods that have been devised to generate relevant surrogates: these include constrained-realisation using simulated annealing [75], temporal-shifting [81] and *amplitude-adjusted Fourier transform* (AAFT), and an *iterative* (IAAFT) version of the same.

In this thesis we choose the most computationally efficient method that requires the

fewest arbitrary parameters. Simulated annealing is generally computationally inefficient [75], and despite their computational simplicity, temporal-shift surrogates require the choice of three parameters per surrogate; any hypothesis test based upon the use of this method will therefore be sensitive to the choice of these parameters [81]. Therefore the IAAFT method, which requires only simple computational operations (FFT and sorting) and only one parameter (the number of iterations), is the most appropriate choice for this study. The IAAFT method involves the following steps [75]:

1. The power spectrum $P_{ss}(k)$ of the signal s_n is calculated using the FFT,

2. A shuffled version r_n of the original signal s_n is produced, that is, the samples at each time step n of s_n are randomly permuted,

3. The new signal u_n is generated from the FFT phase information of r_n and the square magnitude information of the original signal $P_{ss}(k)$, and,

4. The individual samples of u_n and s_n are rank ordered, and the samples of u_n are replaced by those of s_n in the corresponding rank order.

The second step destroys the original temporal ordering of the measurements, which removes any detectable dynamical origins of the signal such as those generated by a deterministic map. The third step imposes the spectral magnitude information onto the new surrogate signal u_n. Therefore, by the Wiener-Khintchine theorem, the surrogate and the original share the same *circular* autocorrelation information. The final step, the amplitude adjustment step, constrains the surrogate to have the same amplitude (probability density) as the original signal. The second to final steps are iteratively performed on the candidate surrogate signal, aiming at a better compromise between any spectral errors introduced by amplitude adjustment, and any amplitude (probability density) errors introduced by spectral magnitude changes. The iteration is guaranteed to converge – for more details please see [75]. Surrogates produced by the IAAFT method will have almost exactly the same circular autocorrelation and probability density as the original signal, yet have randomised phase (temporal) information.

Note that if we stop the IAAFT iteration at step three, then the candidate surrogate will contain precisely the same circular autocorrelation information as the original signal s_n. Conversely, stopping the iteration at step four will ensure that the candidate surrogate signal has exactly the same probability density as the original. In this thesis, we prefer to match the power spectrum to the original exactly, and therefore stop the iteration at

step three after a suitable number of iterations. This inevitably implies that the candidate surrogate will not have precisely the probability density we require. Please see [82] for more detailed investigations of the accuracy issues involved with the use of AAFT and IAAFT surrogates. In the next section, in order to mitigate this problem, we will develop a test for the severity of this probability density mismatch.

As discussed in §3.2.11, the power spectrum contains the same information as the circular autocorrelation. Calculating the power spectrum and using this to constrain the autocorrelation properties of the surrogates implicitly assumes that the original signal has periodic continuation outside the range of the DFT $0 \leq n \leq N-1$ (see §3.2.10). However, most signals we encounter do not conform to this periodicity requirement precisely, nor do they naturally contain any significant discontinuities (since the original continuous-time signal $s(t)$ can often be considered as continuous). The process of digitally sampling and then truncating such a signal to a finite time range N will often induce large, artificial discontinuities, that is, jumps in value across the beginning and end of the time range. These artificial discontinuities, which were not a feature of the original signal, contribute significant power into the power spectrum at all frequencies. This additional power is not a feature of the original, continuous-time signal, but will be a significant feature of the surrogates. This difference between the power spectrum of the original signal and the sampled signal with discontinuities can sometimes lead to spurious rejections of the null hypothesis – see [75] for further discussion. To guard against this possibility, it is important to ensure that the values s_0 and s_{N-1} and the derivatives at these end points are as close as possible. In practice, minimising the difference between these values and the difference between these first derivatives is usually a sufficient precaution [75], which is adopted in this study.

Note that if a signal conforms to the H_0 of this study, then it has a Gaussian probability density, so that here, in the final step of the algorithm, we modify the process slightly to constrain the amplitudes to have the same probability density as a *Gaussian, i.i.d., strongly stationary signal* of the same variance as the original signal s_n. This differs somewhat from the unmodified algorithm described above, where often the null hypothesis is taken to be that the original signal is a linear Gaussian stochastic process but transformed with some monotonic, time independent function [75].

4.7.5 A New Approach – Surrogate Data Integrity Testing

Generating surrogates that conform perfectly to H_0 is impossible – there will always be some sources of error [75, 82]. Nonetheless, we must always check that the surrogates are accurate enough; however there do not exist any systematic methods for performing these checks [82]. In this section we will therefore introduce a new solution to this problem.

Discussed in §4.5.2 was the TDMI which can be computed analytically for signals that conform to the null hypothesis H_0. As a shorthand we shall denote $I_{\mathrm{L}}[s]$ the "linear statistic" for the signal s_n calculated using circular autocorrelation estimates for the covariance matrix entries. Conversely, $I_{\mathrm{N}}[s]$ is the "nonlinear statistic" for the signal s_n calculated using (estimated) probability densities, numerical integration and calibration (suppressing the time lag τ for clarity). Assuming that these test statistics are reliable, it is possible to perform a test prior to applying the main test against H_0, to check that the generated surrogates conform to H_0.

Using circular autocorrelation estimates ensures that the linear statistic is reliable (it does not introduce any additional bias) since it is calculated using the same power spectrum information used to generate the surrogates. Similarly, we can assume, using the integration and calibration procedure described above, that the nonlinear statistic is reliable. Stopping the IAAFT iteration at the third step in the algorithm ensures that the circular autocorrelation of the surrogates matches precisely that of the original. Hence, by comparing $I_{\mathrm{N}}[u]$ against $I_{\mathrm{L}}[s]$, we can probe whether the surrogates conform to the null hypothesis H_0. If the surrogates are in some way flawed this will invalidate the surrogate data test against H_0.

Due to the inevitable probability density error introduced when generating surrogates, this test can only be approximate. There will be systematic differences, but it should be possible to assess whether the deviation between these two statistics is large enough to warrant uncertainty about the appropriateness of the surrogates for the null hypothesis.

4.7.6 Synthetic Examples

In this section we will demonstrate the practical application of the surrogate data test described above using synthetic signals, where we know the truth or falsehood of the null hypothesis. Two different signals, one which conforms to the null hypothesis and another which does not, will be tested by generating surrogates, testing whether these surrogates conform to the null hypothesis, and, assuming this preliminary test is passed, testing the

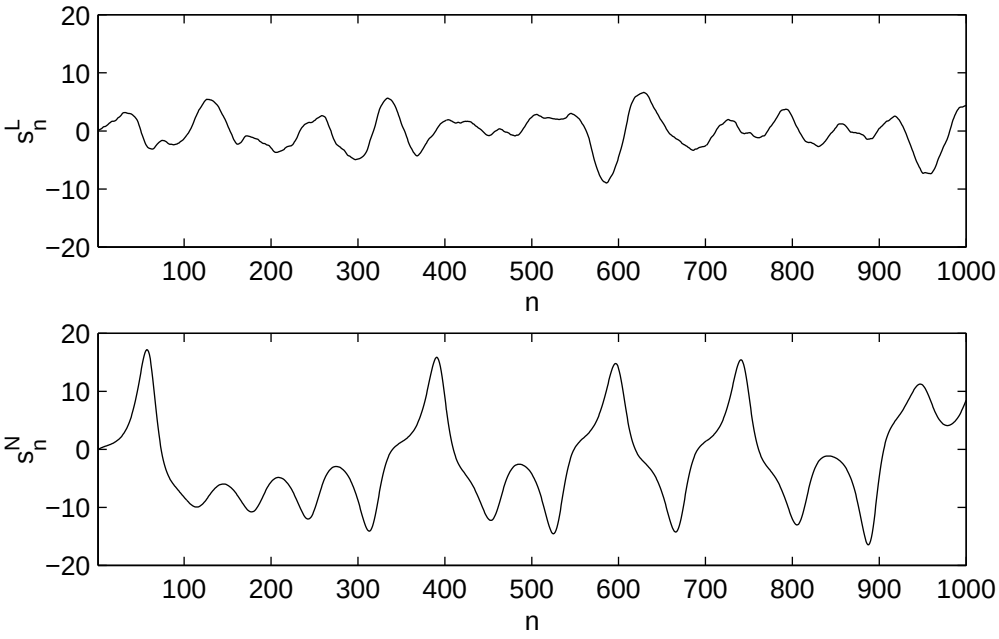

Figure 4.1: Linear s_n^{L} and nonlinear s_n^{N} synthetic signals for demonstrating surrogate data test. The top panel is the output of an AR(2) process with Gaussian, i.i.d., stochastic input signal, and the bottom panel is one coordinate from the output of the Lorenz equations, a deterministic nonlinear system. The horizontal axis is time index n. The top signal conforms to the null hypothesis H_0, and the bottom one does not (H_1 is true). Both signals are of length $N = 6358$.

linearity of the original.

Figure 4.1 shows two signals, the first, s_n^{L} obtained as the output from an AR(2) process such as equation (3.9) driven by a Gaussian, zero mean, i.i.d. stochastic signal, and the second s_n^{N} one of the coordinates of the *Lorenz system* [8] in the chaotic parameter regime. [11] The nonlinear signal s_n^{N} has been end-point value and derivative matched, in accordance with earlier observations in this chapter. Clearly the first system conforms to H_0, and the second conforms to H_1, since, for example, it is non-zero mean and has non-Gaussian joint densities at different time lags.

The next figure 4.2 shows the linear I_{L} and calibrated nonlinear I_{N} TDMI statistics calculated for both signals s_n^{L} and s_n^{N}. As can be seen, the linear and nonlinear statistics track each other closely, up to a certain time lag τ, for the linear signal, but they diverge significantly for the nonlinear signal. This experiment instills confidence that the statistics are capable of distinguishing H_0 from H_1. This figure shows, for the linear signal, that the accumulated sources of error in the corrected calculation of I_{N} amount to a small discrepancy at all time lags [74, 80].

[11] This is a nonlinear, deterministic set of ordinary differential equations which has been integrated using the finite difference method, which leads to a deterministic map such as (4.3).

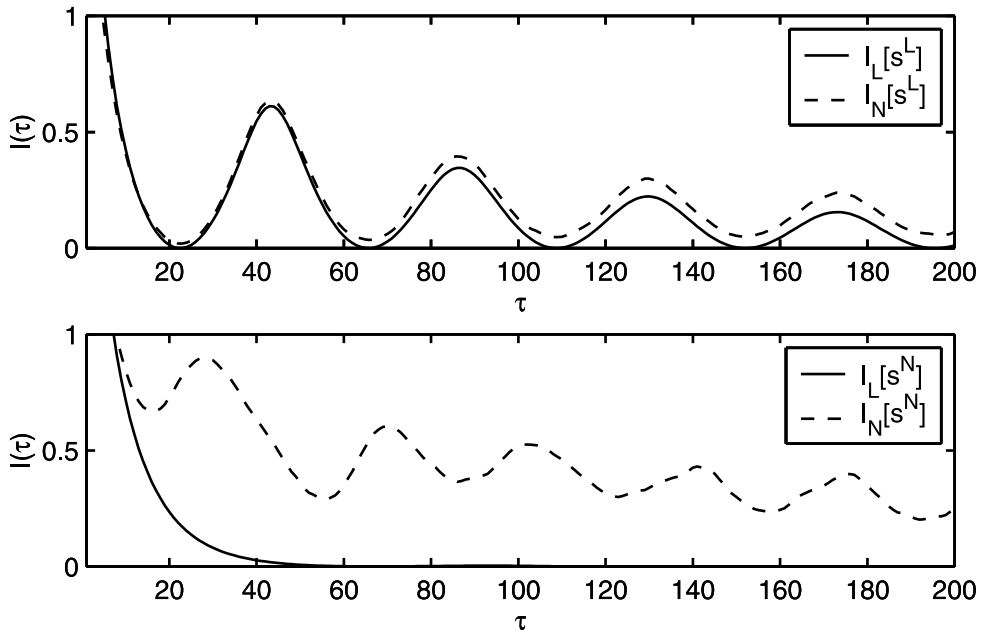

Figure 4.2: Linear and nonlinear TDMI statistics applied to synthetic linear s_n^{L} and nonlinear s_n^{N} signals. The top panel is the output of an AR(2) process with Gaussian, i.i.d., stochastic input signal, and the bottom panel is one coordinate from the output of the Lorenz equations, a deterministic nonlinear system. The horizontal axis is time lag τ, the vertical axis mutual information in nats. In both panels, the linear statistic I_{L} and the calibrated nonlinear statistic I_{N} have been calculated on that signal. The number of bins used in the nonlinear TDMI calculation was $Q = 20$ – see Appendix §A.3 for further details of this calculation. The signals were both of length $N = 6358$.

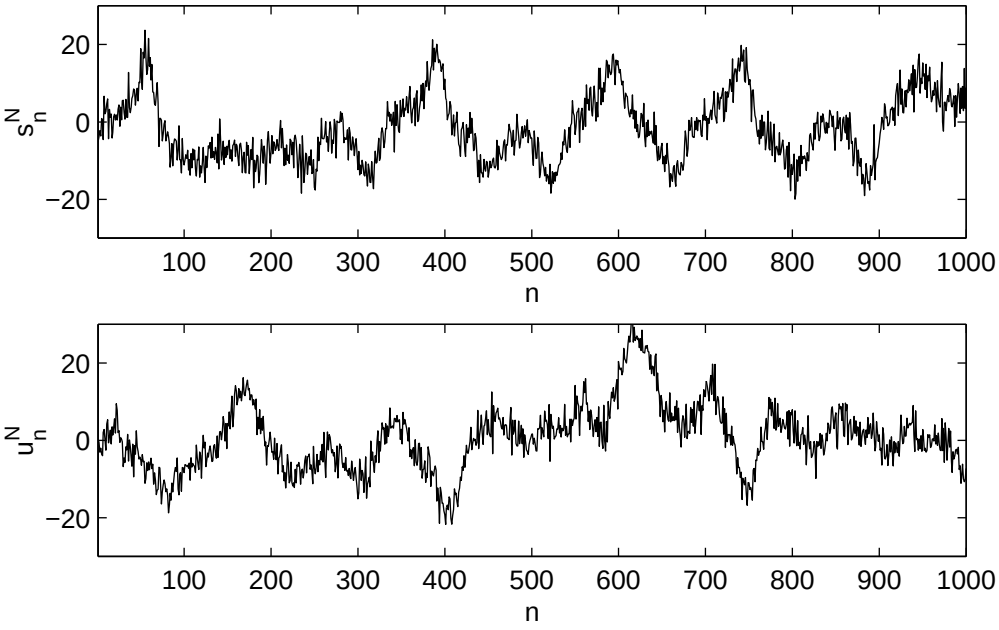

Figure 4.3: Synthetic nonlinear signal corrupted by additive Gaussian noise s_n^{N} (top panel), and one surrogate for this signal generated using the IAAFT method u_n^{N} (bottom panel) with 50 iterations. The horizontal axis is the time index n. Both signals are of length $N = 6358$.

For all real world signals, we should expect some *observational noise* contaminating the measurements. The source of such randomness can be measurement error or other confounding factors which we cannot control. We can simulate this by adding Gaussian, i.i.d., strongly stationary noise to the synthetic nonlinear signal to obtain the noisy signal s_n^{N} (here the observational noise has maximum amplitude range 30% of the maximum amplitude range of the original signal). This noisy signal is depicted in the top panel of figure 4.3. In the bottom panel of this figure is shown one surrogate generated using the IAAFT method described in the previous section. Although familiarity with the Lorenz system might allow detection, by eye, of the fact that it does not conform to H_0 where the surrogate does, these two signals share precisely the same power spectrum and so are indistinguishable by linear techniques alone.

Next, we perform the integrity check on the surrogates by comparing the linear statistic on the original to the nonlinear statistic calculated on the $M = 19$ surrogates, this number chosen to test H_0 to significance level $S = 0.1$. The results are shown in the top panel of figure 4.4. As can be seen, the nonlinear statistic on the surrogates closely follows the linear statistic on the original, to within the small, systematic errors introduced in the calculation of the nonlinear statistic. Thus confidence is instilled that the surrogates do indeed conform to H_0.

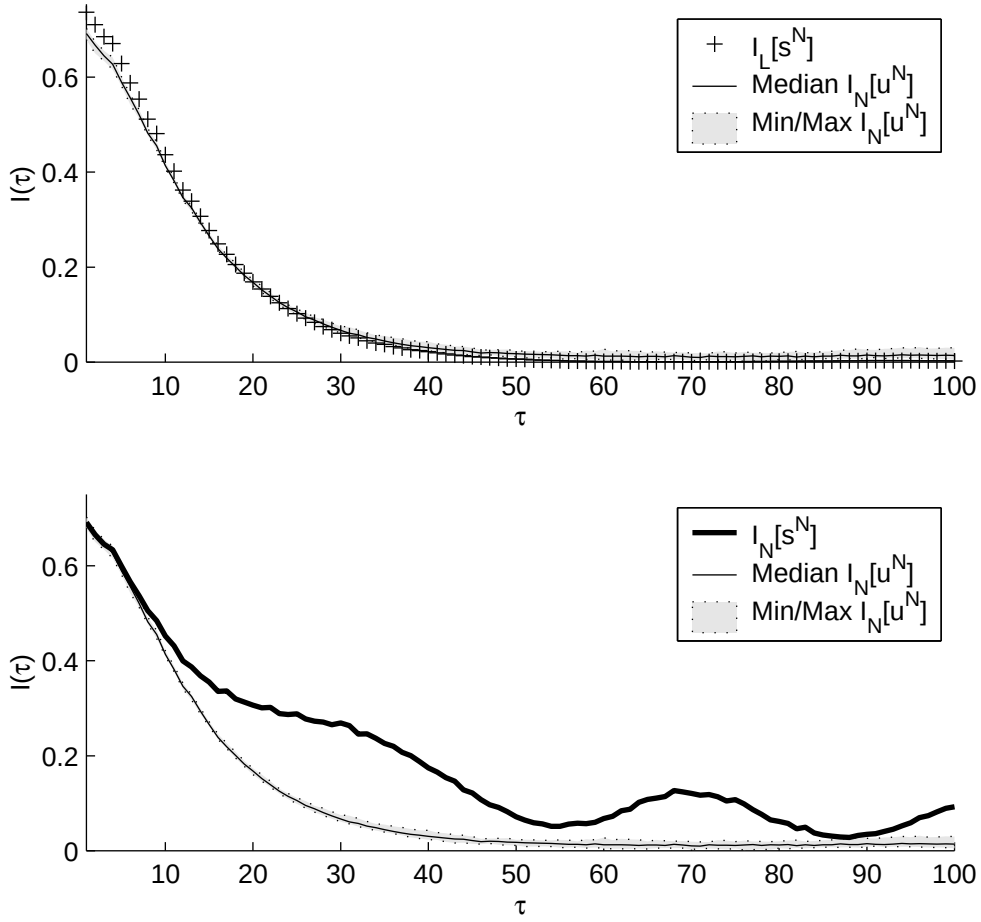

Figure 4.4: Surrogate data integrity check that IAAFT generated surrogates, using 50 iterations, conform to the null hypothesis H_0 of a Gaussian, stochastic, zero mean, strongly stationary linear process (top panel). The crosses are the linear statistic calculated on the original s^{N}, and the grey box encloses all the nonlinear statistic values calculated on the surrogates. The unbroken black line is the median value of the nonlinear statistic on the surrogates. The bottom panel shows the results of the hypothesis test – the thick black line is the nonlinear statistic calculated on the original. The grey box encloses, as in the top panel, the maximum extent of the nonlinear statistic values calculated on the surrogates. The thin black line is the median of the nonlinear statistic calculated on the surrogates. The horizontal axis is time lag τ, and the vertical axis is information in nats.

Finally, the null hypothesis test is carried out by calculating the nonlinear statistic on both the surrogates and the original, and comparing the results. As can be seen in the bottom panel of figure 4.4, for most time lags τ the nonlinear statistic on the original is larger than all the nonlinear statistic values calculated on the surrogates. Thus, for most time lags, given that we can be confident that the surrogates do conform to H_0, we can reject H_0 and conclude that the original signal conforms instead to H_1, as expected. Thus the surrogate data test functions correctly. [12]

As mentioned earlier, rejection of H_0 only tells us that one or more of the stated properties is violated: this is the content of H_1. Which of these properties is violated cannot be inferred from this result. Considerable caution should be exercised before reaching any conclusions based solely upon these results, but the existence of other, independent evidence can be used to inform the choice of model functions $\mathbf{F}$. In this synthetic case, for example, given that we know that the dynamical origin of the signal s_n^{N} is a nonlinear deterministic map, the interpretation of the results of the surrogate test are unambiguous: s_n^{N} is a deterministic, nonlinear signal. We can rule out the possibility that it is a stochastic system such as (4.1), and we also know that the dynamical function $\mathbf{F}$ does not change with time (so that this system is stationary).

4.8 Chapter Summary

This chapter has reviewed a selection of mathematical topics relevant to the nonlinear aspects of speech. Since they may act as parsimonious models for speech production, discrete time nonlinear stochastic and deterministic systems in state space and their properties were introduced, followed by an overview of methods used to analyse the measured outputs from these systems. These methods are drawn from a diverse set of mathematical disciplines, such as nonlinear time series analysis, fractal geometry and information theory. Finally, surrogate data hypothesis tests against the assumptions of LTI systems theory and their current limitations were discussed. Some novel solutions to overcome these limitations were then proposed, including a new calibration method that allows more rigorous testing of the null hypothesis by ensuring the suitability of the surrogate data signals.

These concepts will form a key part of this thesis. The nonlinear model frameworks

[12] Some interesting details are evident here. H_0 cannot be rejected for the first few time lags $1 \leq \tau \leq 6$ nor for time lag $\tau = 87$. Due to the smoothness of the Lorenz system, *local linearisation* [8] is appropriate, and this may go some way towards explaining the first observation. Similarly, we expect that the mutual information between time-delayed versions of the output of the nonlinear system decreases with increasing time delay, helping to explain the latter observation.

and analysis methods will be put to use with real speech signals, to produce new evidence against the applicability of LTI systems theory for speech, and to extract useful information with practical applications.

CHAPTER 5

Nonlinearity in Speech Signals

The theoretical models introduced in Chapter 2 provide good reason to conclude that nonlinearity and/or non-Gaussianity might be an important feature in speech production. Nonetheless, as discussed in the introduction, theoretical models alone are not sufficient evidence and must be verified against empirical data. Whilst ideally a direct, quantitative comparison between theoretical model outputs and physical speech signals could be performed, in practice such comparisons require estimates of the theoretical model parameters from signals, and parameter estimation for nonlinear models such as those introduced in this thesis is generally very difficult (and thus a broad and open topic of research) due to inherent model imperfections and numerous other confounding experimental factors. Quantitative matching procedures such as the least-squares approach often introduce significant errors in the estimated model parameters, as demonstrated in §3.3.1. In this thesis therefore we instead attempt to verify whether the underlying assumptions (of LTI systems theory) are valid for the data, using the hypothesis test developed in the previous chapter. The results of this test, in combination with the models put forward earlier in the thesis, will allow us to make more informed choices about which models might be more parsimonious than linear models, for subsequent applications.

5.1 Review of Previous Empirical Investigations

The issue of whether nonlinear signal processing approaches might offer improvements over classical approaches has attracted some attention in the speech analysis literature. Of importance to this thesis are other surrogate data tests that have been conducted; we now review two examples from the literature that are typical of the studies that address this topic.

Miyano [21] produced a surrogate data test using the *Wayland translation error statistic*, which is a nonlinear, geometric measure designed to test the extent to which orbits in the embedded state space from some continuous, nonlinear model such as (4.3) are paral-

lel when they come close (due to the continuity of the invariant sets in which the orbits are contained). Close trajectories that deviate from parallelism are therefore considered, under this statistic, to be indicative of discontinuity due to a stochastic forcing term such as that found in the system (4.1). Conversely, close trajectories that are nearly always parallel are indicative of equation (4.3). Using the AAFT method, the null hypothesis H_0 is that the speech signals are generated by a zero mean, Gaussian, linear, strongly stationary stochastic process transformed by a monotonic, nonlinear function. Surrogates are generated for two short excerpts of vowel signals from one male and one female subject. The test statistic is applied to both the surrogates and the original signals, finding that H_0 can be rejected at the 5% significance level. The authors conclude that the H_0 model for the speech signals can be rejected, and that a deterministic nonlinear system would be more appropriate. However, in a somewhat contradictory conclusion, using a *nonlinear predictor* as test statistic instead, [1] the null hypothesis could not be rejected.

There are a number of systematic errors in this result. The first is a misinterpretation of the alternative hypothesis H_1: as stated in §4.7, H_0 and H_1 must be completely mutually exclusive states of nature. In this case, H_1 true simply implies that one or more of the listed properties (linearity, Gaussianity transformed through a monotonic function, randomness or strong stationarity) of H_0 does not hold (at the stated level of statistical significance) for the speech signals. In this case, rejection of H_0 does not *necessarily* imply that a deterministic nonlinear system would be a better model for the signals, since a *non-stationary* (time variant) Gaussian, stochastic linear model, for example, might also be indicated. The authors state that numerical investigations of the test statistic reveal particular empirical values for Gaussian linear time series, and this evidence is used to support the conclusion of deterministic nonlinearity, but this is not *formally* a part of the hypothesis test. This is because the null hypothesis is determined by the structure of the surrogates, rather than the nature of the test statistic [8].

Secondly, there are eight parameters that must be chosen by hand in order to calculate the test statistic. Although a systematic search with one of these parameters is performed, the rest are chosen on a trial-and-error basis. Changing these parameters may affect the result of the hypothesis test. Thirdly, since the two speech signals and surrogates are not shown, we cannot be sure that cyclic autocorrelation artifacts due to end point discontinuities discussed in §4.7 have not crept into the surrogates. This may mean that

[1] It is possible to assume a particular parametric form for the system function **F** in system (4.3) and use a variety of methods to estimate the parameters. Once the parameters have been estimated, they are said to define a *nonlinear predictor* for the signal [8].

the surrogates are flawed systematically. The fourth issue, albeit a minor one, is that the hypothesis test is formulated as a *two-sided t-test* which assumes that the test statistic values are normally distributed. This is by comparison to non-parametric rank tests which make fewer restrictive assumptions about the density of the test statistic [83]. Finally, no analytical results are known for the test statistic, so that it is not possible, for example, to test the integrity of the surrogates before performing the hypothesis test.

The study of Tokuda [20] is designed to test the null hypothesis H_0 that for the mainly periodic vowel sounds, the individual cycles follow each other in a random sequence, as opposed to a deterministic sequence (H_1). Appropriate surrogate data signals for this null hypothesis are generated by the method of *spike-and-wave surrogates* [84], in which the original signal is split into separate cycles and then reconstructed by joining these cycles together end-to-end in a new, randomised order. By this process any deterministic dynamical structure at the joins (which depends upon the temporal ordering of the individual samples) is destroyed. Using the same Wayland translation error as the study discussed above, it was found that the null hypothesis could be rejected at the 5% significance level using a rank order test, for five different vowel samples. The study concludes that the individual cycles follow each other in a deterministic sequence in these vowels.

Again we find several systematic errors in this study. Apart from the reliability issues of the Wayland translation error statistic, spike-and-wave surrogates can introduce spurious discontinuities at the joins between cycles [78]. These discontinuities imply that certain other properties of the surrogates, such as stationarity and continuity, may well differ from those of the original, in addition to the property of determinism at the cycle joins. The test statistic may be sensitive to these other properties. Since the values of the test statistic obtained on the original and the surrogates are qualitatively the same (they actually "track" the values on the surrogates) and differ quantitatively only by a very small amount, it is quite plausible that the rejection of the null hypothesis is due to systematic problems with the generation of the surrogates, rather than the existence of deterministic structure joining the cycles.

Also, with spike-and-wave surrogates, assuming that the signal is generated by a nonlinear deterministic system, most of the signal will still retain deterministic nonlinearity, and only at the joins will there be any significant departure from this model. Therefore the chosen statistic must be highly sensitive in order to detect this subtle difference. The sensitivity of the Wayland translation error to such small differences is unknown in general, and, in fact, the quantitative differences displayed in the study are extremely slight.

We would prefer a statistic that can show a much larger difference [80], as is displayed in §4.7 in the difference between Gaussian linear and deterministic nonlinear systems with the TDMI statistic.

Systematic problems such as these cast doubt on the reliability of the results, particularly since only a handful of (Japanese) vowels are tested. Unfortunately, these systematic errors are typical of the surrogate tests in the literature [75], and the claims of deterministic nonlinearity, supporting models such as (4.3) for vowel signals, are therefore somewhat dubious. In order to address these deficiencies, we will, in this chapter, apply the more reliable test developed in this thesis to a large database of speech examples, paying careful attention to avoid the systematic problems discussed earlier. We shall then seek a (necessarily cautious) interpretation of the results and their significance for speech technology. The aim is to obtain more reliable conclusions than existing studies about the extent of the suitability of LTI systems theory in speech processing.

5.2 Applying the New Surrogate Data Test

One main focus of this thesis is to test whether the LTI systems assumptions hold for speech signals, despite changes in formants (differing vowels) or acoustic energy source (aeroacoustic noise in consonants versus vocal fold vibration in vowels). Furthermore, it has been suggested through simulation (see §2.2.2 and references [42], [37]) and empirical investigations [42] that nonlinear dynamics may be present in voice disorders. We will thus wish to test whether this is confirmed by empirical evidence. In this section therefore we will apply the new surrogate data test, described in §4.7, to three different classes of speech signals: stable vowels, consonants and stable vowels from subjects with various voice disorders.

5.2.1 Data

The data used in this study derives from two widely used sources of test speech signals: the DARPA TIMIT Acoustic-Phonetic Speech Corpus [60], and the Kay Elemetrics Disordered Voice Database [85].

The TIMIT database, primarily designed for automated speech recognition system research and construction, consists of speech samples from 630 male and female healthy adult speakers from the eight major regional dialects of US English. The subjects come from a variety of ethnic backgrounds. All the speech samples were recorded under quiet

acoustic conditions with minimal background noise. The speech samples consist of a variety of phrases of running speech (not in isolated phonemes). Every phoneme in the speech samples is labelled. The samples were quantised using 16 bit resolution and sampled at a rate of 16kHz.

Since the speech samples in this database contain running speech, we need to avoid anticipatory co-articulation to satisfy the stationarity assumption of the null hypothesis. Similarly, diphthongs are also avoided since they are considered to be non-stationary in the sense that the vocal tract resonances are changing with time. Thus the phoneme speech data for this study was selected carefully in order to avoid any formant and amplitude changes. This involved finding labelled, long monopthong and fricative phonemes in the database, and selecting a central part of each phoneme. The selected data thus contains speech samples from 26 different, randomly chosen subjects, 13 male and 13 female, with two representatives from each phoneme in table 2.1. Unfortunately, the selected data consists of only a few consonants, since it is extremely rare to find stationary consonants of sufficient duration from running speech [10].

The Kay database contains speech samples from 707 adult US subjects, including deliberately sustained /aa/ vowels and running speech phrases. Of these 707 subjects, 654 are patients with a wide variety of organic, neurological, traumatic and psychogenic voice disorders (which we will discuss in more detail in the next chapter). Diagnoses were performed by professional voice clinicians after extensive vocal function testing. The samples were recorded under quiet acoustic conditions and quantised at 16 bit resolution and at two different samples rates, 25kHz and 50kHz.

From this database, 26 disordered subjects were selected at random. Of these, 22 have diagnoses and therefore the data selected for the surrogate tests represents 22 different voice disorders. A small segment of speech data was extracted from the central part of each deliberately sustained /aa/ vowel pronounced by the subject.

Finally, all the data for the surrogate test has been standardised in the following way. Firstly, the signal amplitudes have been normalised to the range $[-1, 1]$. All the signals which were not originally recorded at 16kHz sample rate have been downsampled to 16kHz using high-order, anti-alias pre-filtering followed by decimation [12]. Furthermore, in order to avoid cyclic autocorrelation discontinuity problems (as discussed in §4.7), the start and end samples and gradients of the selected speech samples were matched by hand.

Thus the final data set for this chapter consists of 50 different speech signals; tables 5.1 and 5.2 list the source database file information, subject information, diagnoses and

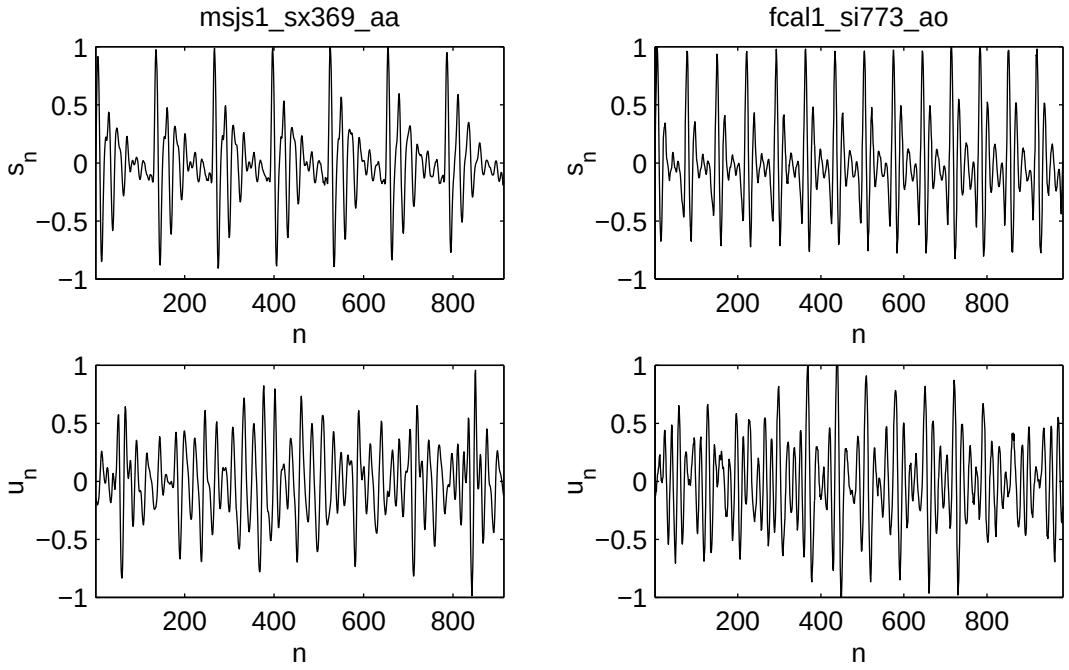

Figure 5.1: Two selected normal vowel speech signals s_n (top row) with one associated surrogate u_n for each signal (bottom row). The vertical axis is amplitude, and the horizontal axis is time index n in samples. For clarity only the first 1000 samples are shown.

sound signal lengths. The average length of these samples is 0.2 seconds. [2]

Six of these selected speech signals s_n are displayed in figures 5.1, 5.2 and 5.3, showing two vowels, two fricative consonants and two disordered vowel examples.

5.2.2 Results

The results of applying the surrogate data test to the selected speech data signals s_n are summarised in tables 5.3 and 5.4. For each selected sound signal, $M = 19$ surrogates u_n were generated using 50 IAAFT iterations each, for a (two-sided test) confidence level of 90%. The table reports the number of time lags at which the nonlinear statistic was either the largest or the smallest of the values calculated on the surrogates and the original. Six example surrogate signals are shown in figures 5.1, 5.2 and 5.3. The nonlinear TDMI statistic used $Q = 20$ bins, calibrated using 19 i.i.d. Gaussian signals of the same variance as the original signal s_n. For the selected data in figures 5.1, 5.2 and 5.3, the surrogate data integrity checks and null hypothesis test results are shown in figures 5.4, 5.5 and 5.6 respectively.

[2] For the purposes of independent verification of the results reported in this thesis, Microsoft WAV files of these signals and Matlab/C software to carry out the calibrated surrogate data tests are available from the URL http://www.maths.ox.ac.uk/~littlem/thesis/.

Table 5.1: Summary of information about TIMIT healthy speech data selected for the surrogate data test. All sounds are from healthy subjects. For phoneme codes, see table 2.1.

File name	Age	Sex (M/F)	Phoneme code	Sound length (seconds)
faks0_sx223_aa	29	F	/aa/	0.074
fcft0_sa1_er	23	F	/er/	0.069
fdac1_si844_iy	68	F	/iy/	0.071
fmaf0_si2089_ih	42	F	/ih/	0.063
fjwb1_sa2_ae	30	F	/ae/	0.080
fdkn0_sx271_eh	32	F	/eh/	0.078
fmjb0_si547_uw	23	F	/uw/	0.064
futb0_si1330_uh	26	F	/uh/	0.065
fcal1_si773_ao	30	F	/ao/	0.061
fmgd0_sx214_ah	55	F	/ah/	0.060
msjs1_sx369_aa	26	M	/aa/	0.057
mrws0_si1732_er	27	M	/er/	0.059
mreb0_si2005_iy	28	M	/iy/	0.071
mbwm0_sa1_ih	32	M	/ih/	0.071
mstf0_sa1_ae	27	M	/ae/	0.065
mbml0_si1799_eh	27	M	/eh/	0.075
mdbp0_sx186_uw	24	M	/uw/	0.059
mcsh0_sx199_uh	54	M	/uh/	0.065
mbjk0_si2128_ao	25	M	/ao/	0.058
mdld0_si913_ah	25	M	/ah/	0.062
faks0_sa1_ss	29	F	/ss/	0.063
fjem0_sa1_sh	25	F	/sh/	0.099
fjmb0_si547_ff	23	F	/ff/	0.065
faem0_sx312_th	26	F	/th/	0.077
mwbt0_sa1_ss	52	M	/ss/	0.078
mjsw0_sa1_sh	33	M	/sh/	0.064
mdwd0_sx450_ff	24	M	/ff/	0.114
mdwd0_sx90_th	24	M	/th/	0.059

Table 5.2: Summary of information about Kay Elemetrics disordered speech data selected for the surrogate data test. Where the age and sex are left blank they are unknown. All sounds represent phoneme /aa/ (for phoneme codes, see table 2.1).

File name	Age	Sex (M/F)	Sound length (seconds)	Diagnosis
EGT03AN_kay_aa	75	F	0.398	Parkinson's disease
CAC10AN_kay_aa	49	F	0.415	Inflammatory disease
CAR10AN_kay_aa	66	F	0.298	Contact granuloma
AXL04AN_kay_aa	53	F	0.448	Hyperfunction
SEC02AN_kay_aa	21	F	0.368	Asymmetric arytenoid movement
SWS04AN_kay_aa	26	F	0.428	Cyst
NMB28AN_kay_aa	42	F	0.276	Erythema
RMB07AN_kay_aa	48	F	0.365	Reinke's polypoid degeneration
GMM09AN_kay_aa	45	F	0.158	Laryngeal web
JXS01AN_kay_aa	70	M	0.354	Ventricular compression
JAF15AN_kay_aa	80	M	0.251	Gastric reflux
MWD28AN_kay_aa	38	M	0.381	Adductor spasmodic dysphonia
RPC14AN_kay_aa	76	M	0.393	Bowing
WFC07AN_kay_aa	56	M	0.368	A-P squeezing
WXE04AN_kay_aa	36	M	0.398	Atrophic laryngitis
BSA26AN_kay_aa	69	M	0.407	Paralysis
CBD19AN_kay_aa	71	M	0.375	Cordectomy
CMA06AN_kay_aa	56	M	0.451	Keratosis/leukoplakia
CTB30AN_kay_aa	36	M	0.404	Cricoarytenoid arthritis
DMG24AN_kay_aa	23	M	0.382	Haemorrhagic polyp
EFC08AN_kay_aa	66	M	0.405	Post microflap surgery
HWR04AN_kay_aa	76	M	0.472	Hyperfunction
NAK16AN_kay_aa			0.327	Undiagnosed disorder
CCM15AN_kay_aa			0.177	Undiagnosed disorder
CCP29AN_kay_aa			0.188	Undiagnosed disorder
CCP21AN_kay_aa			0.249	Undiagnosed disorder

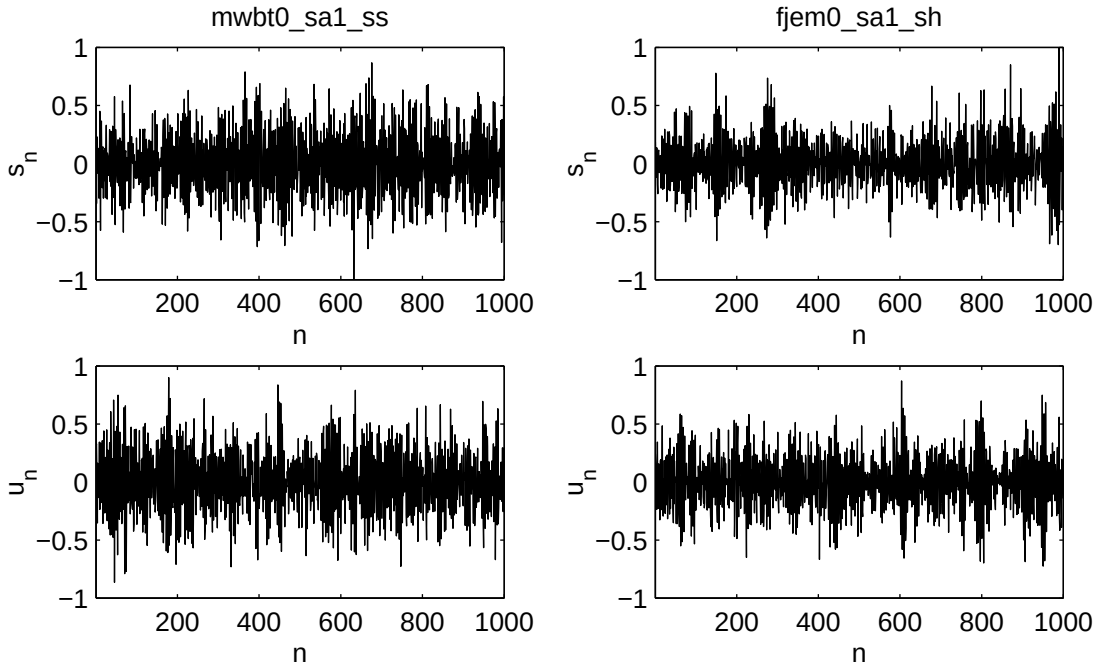

Figure 5.2: Two selected fricative consonant speech signals s_n (top row) with one associated surrogate u_n for each signal (bottom row). The vertical axis is amplitude, and the horizontal axis is time index n in samples. For clarity only the first 1000 samples are shown.

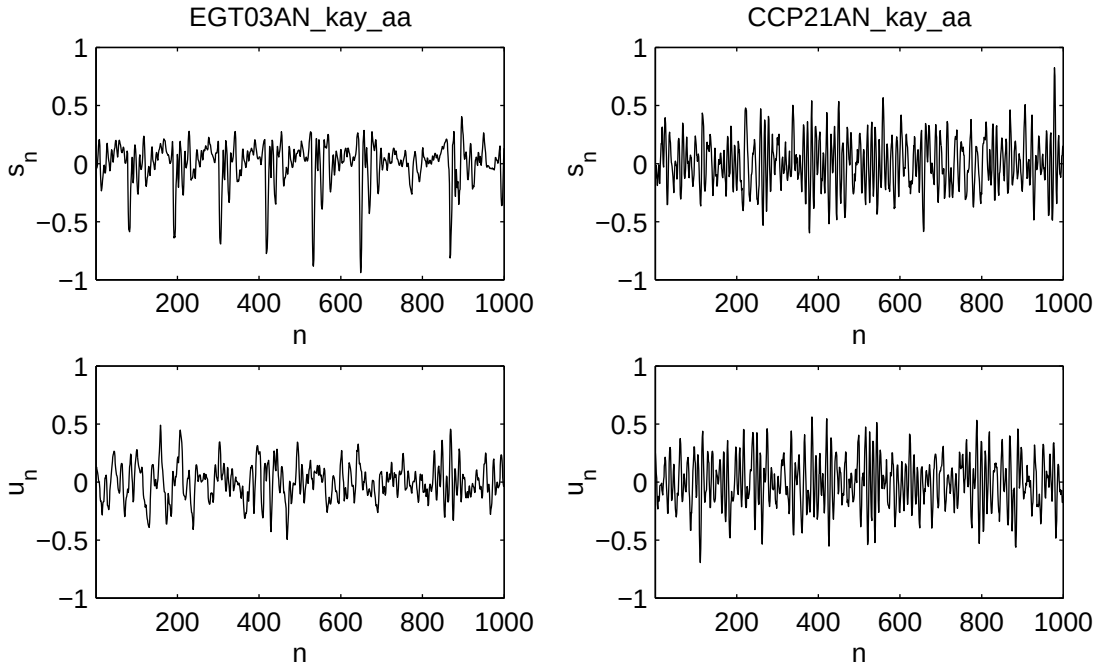

Figure 5.3: Two selected disordered speech signals s_n (top row) with one associated surrogate u_n for each signal (bottom row). The vertical axis is amplitude, and the horizontal axis is time index n in samples. For clarity only the first 1000 samples are shown.

Table 5.3: Results of the surrogate data null hypothesis test for every time lag $\tau \leq 1 \leq 200$ for selected speech data from the TIMIT database. The null hypothesis H_0 is that of a Gaussian, linear, zero mean, strongly stationary random process. Details of the speech data can be found in table 5.1.

File name	H_0 rejections	File name	H_0 rejections
faks0_sx223_aa	200	mstf0_sa1_ae	198
fcft0_sa1_er	194	mbml0_si1799_eh	200
fdac1_si844_iy	200	mdbp0_sx186_uw	111
fmaf0_si2089_ih	172	mcsh0_sx199_uh	190
fjwb1_sa2_ae	199	mbjk0_si2128_ao	198
fdkn0_sx271_eh	199	mdld0_si913_ah	200
fmjb0_si547_uw	190	faks0_sa1_ss	22
futb0_si1330_uh	200	fjem0_sa1_sh	14
fcal1_si773_ao	199	fjmb0_si547_ff	25
fmgd0_sx214_ah	181	faem0_sx312_th	14
msjs1_sx369_aa	199	mwbt0_sa1_ss	30
mrws0_si1732_er	195	mjsw0_sa1_sh	2
mreb0_si2005_iy	199	mdwd0_sx450_ff	5
mbwm0_sa1_ih	199	mdwd0_sx90_th	18

Table 5.4: Results of the surrogate data null hypothesis test for every time lag $\tau \leq 1 \leq 200$ for selected speech data from the Kay database. The null hypothesis H_0 is that of a Gaussian, linear, zero mean, strongly stationary random process. Details of the speech data can be found in table 5.2.

File name	H_0 rejections	File name	H_0 rejections
EGT03AN_kay_aa	200	WFC07AN_kay_aa	200
CAC10AN_kay_aa	190	WXE04AN_kay_aa	199
CAR10AN_kay_aa	100	BSA26AN_kay_aa	21
AXL04AN_kay_aa	198	CBD19AN_kay_aa	183
SEC02AN_kay_aa	200	CMA06AN_kay_aa	200
SWS04AN_kay_aa	192	CTB30AN_kay_aa	198
NMB28AN_kay_aa	194	DMG24AN_kay_aa	197
RMB07AN_kay_aa	199	EFC08AN_kay_aa	200
GMM09AN_kay_aa	186	HWR04AN_kay_aa	173
JXS01AN_kay_aa	199	NAK16AN_kay_aa	173
JAF15AN_kay_aa	197	CCM15AN_kay_aa	26
MWD28AN_kay_aa	198	CCP29AN_kay_aa	28
RPC14AN_kay_aa	199	CCP21AN_kay_aa	27

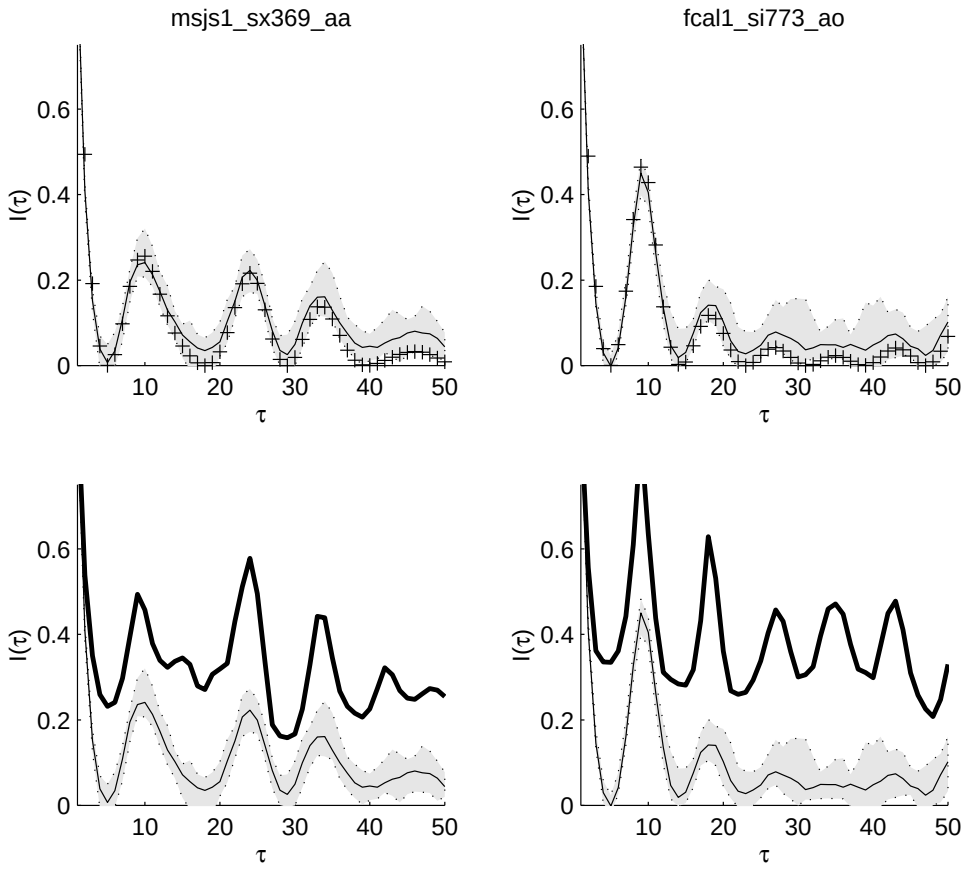

Figure 5.4: Surrogate data integrity check and hypothesis test results for two selected TIMIT vowels. (Top row) surrogate data integrity checks that the IAAFT generated surrogates, using 50 iterations, conform to the null hypothesis H_0 of a Gaussian process. The crosses are the linear statistic calculated on the originals s_n, and the grey box encloses all the nonlinear statistic values calculated on the $M = 19$ surrogates. The unbroken black line is the median value of the nonlinear statistic on the surrogates. (Bottom row) results of null hypothesis test, the thick black line is the nonlinear statistic calculated on the originals. The grey box encloses, as in the top row, the maximum extent of the nonlinear statistic on the surrogates. The horizontal axes are time lag τ in samples, shown for the limited range $1 \leq \tau \leq 50$ for clarity, and the vertical axes are mutual information $I(\tau)$ in nats. The nonlinear statistics were all calculated using $Q = 20$ bins.

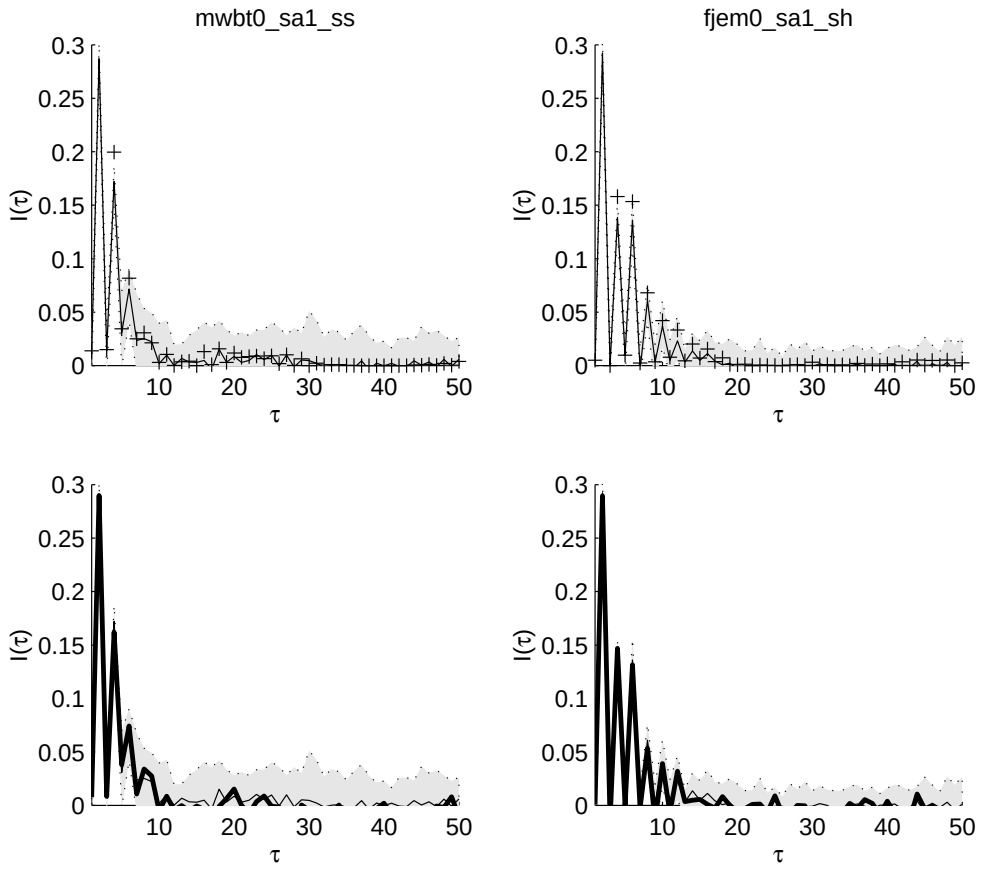

Figure 5.5: Surrogate data integrity check and hypothesis test results for two selected TIMIT fricative consonants. (Top row) surrogate data integrity checks that the IAAFT generated surrogates, using 50 iterations, conform to the null hypothesis H_0 of a Gaussian process. The crosses are the linear statistic calculated on the originals s_n, and the grey box encloses all the nonlinear statistic values calculated on the $M = 19$ surrogates. The unbroken black line is the median value of the nonlinear statistic on the surrogates. (Bottom row) results of null hypothesis test, the thick black line is the nonlinear statistic calculated on the originals. The grey box encloses, as in the top row, the maximum extent of the nonlinear statistic on the surrogates. The horizontal axes are time lag τ in samples, shown for the limited range $1 \leq \tau \leq 50$ for clarity, and the vertical axes are mutual information $I(\tau)$ in nats. The nonlinear statistics were all calculated using $Q = 20$ bins.

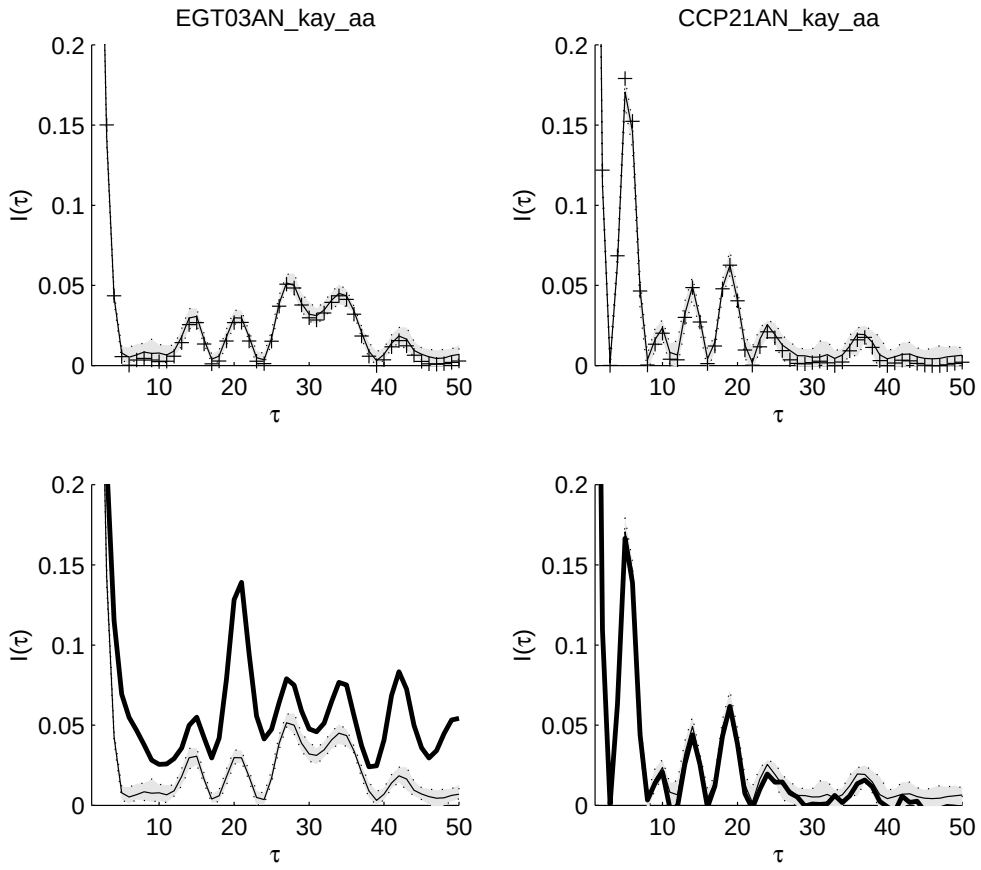

Figure 5.6: Surrogate data integrity check and hypothesis test results for two selected Kay disordered vowels. (Top row) surrogate data integrity checks that the IAAFT generated surrogates, using 50 iterations, conform to the null hypothesis H_0 of a Gaussian process. The crosses are the linear statistic calculated on the originals s_n, and the grey box encloses all the nonlinear statistic values calculated on the $M = 19$ surrogates. The unbroken black line is the median value of the nonlinear statistic on the surrogates. (Bottom row) results of null hypothesis test, the thick black line is the nonlinear statistic calculated on the originals. The grey box encloses, as in the top row, the maximum extent of the nonlinear statistic on the surrogates. The horizontal axes are time lag τ in samples, shown for the limited range $1 \leq \tau \leq 50$ for clarity, and the vertical axes are mutual information $I(\tau)$ in nats. The nonlinear statistics were all calculated using $Q = 20$ bins.

5.3 Interpretation and Discussion of Results

As can be seen in figures 5.4, 5.5 and 5.6, the surrogate data integrity check is satisfied, since the linear statistic on the original is very close in value to the nonlinear statistic on the surrogates. Thus we can have confidence that the surrogates all conform to H_0. This result is typical of all the other speech signals.

Referring to table 5.3, for the healthy TIMIT data set, we can see that for vowel sounds, the overwhelming majority of time lags in the range $1 \leq \tau \leq 200$ reject H_0 at the 90% confidence level. An exception is the vowel *mdbp0_sx186_uw* where this holds for only slightly more than half of the tested time lags. Conversely, for the fricative consonants, the results are almost completely the opposite: for nearly all the time lags we cannot reject H_0 as an explanation for the dynamics.

For the disordered speech examples from the Kay data set, table 5.4 shows that again, for most of the speech signals, over the vast majority of time lags tested, we can reject H_0. There are a few exceptions where the converse is true.

We can conclude, overall then, that for most healthy and disordered vowel sounds, linear systems such as (3.9) with zero mean, strongly stationary, Gaussian i.i.d. input signals (forcing signals) can be rejected as models for these signals, and that nonlinear models, either stochastic or deterministic, may be more suitable. For fricative consonants and some disordered speech on the other hand, we cannot rule out the linear model.

Assuming that the test statistics have sufficient discriminatory power, in performing these hypothesis tests we have, in most cases, found a statistically significant *effect* – the departure from the linear Gaussian model. Remembering that this test does not pin down the exact explanation of the dynamical origins of the effect, we will instead turn to knowledge of the biomechanics of speech production introduced in Chapter 2 to inform our interpretation of these results.

5.3.1 Aeroacoustic Noise and Gaussian Linearity

Looking at the speech signals and their associated surrogates, the signals which are most visually similar to their surrogates are the fricative consonants and the "breathy" disordered sounds (for example signal *CCP21AN_kay_aa*). For these signals, H_0 cannot generally be rejected. For these sounds, vocal fold oscillation ceases altogether, and the airflow through the vocal tract is not regularly interrupted. Such fricative consonants and aspiration noise are therefore produced solely by aeroacoustic sound mechanisms and can be modelled as

a random impulse train, one impulse for each vortex shed at the constriction, convolved with an impulse response that depends upon the shape of the vocal tract, the path of that vortex through the tract, and properties of the vortex itself. This mechanism could find a very parsimonious representation in the linear model (3.9), but there is no requirement for the forcing signal (the vortex impulse train) to be Gaussian, i.i.d., zero mean, or strongly stationary. It would appear, however, from the results of the surrogate data test, that a linear system driven by a Gaussian, i.i.d., zero mean and strongly stationary stochastic process is still the best candidate model here.

5.3.2 Periodic and Aperiodic Vocal Fold Dynamics

For the healthy and disordered vowels signals which exhibit the most regularity and periodicity (for example signals *msjs1_sx369_aa* and *fcal1_si773_a0*), the surrogates have qualitatively similar small fluctuations, but the regularity has been destroyed. These signals lead to the largest number of rejections of H_0. Such vowel signals are modelled from first principles as nonlinear dynamical systems that force the passive, linear system of the vocal tract into resonance at specific frequencies. The significant differences between the surrogates and the original signals leave us with little choice but to accept this first principles model as the best candidate. Digital models such as (4.3) are therefore still indicated.

However, the situation is somewhat less clear-cut with aperiodic disordered signals such as *EGT03AN_kay_aa*. The original signal s_n exhibits near periodicity and some aperiodicity, such that the surrogate, though lacking any obvious repetition, is harder to separate from the original visually. Nonetheless, figure 5.6 shows the clear rejection of H_0. The nonlinear statistic for the original is very close to that for the surrogates; therefore the size of the departure from the assumptions of H_0 is smaller than with the more periodic examples. It is harder in cases such as this to suggest an appropriate digital model, but certainly (4.1) would be capable of generating such signals.

5.3.3 Implications for Speech Technology

From the results of the surrogate data test and knowledge of the biomechanics of speech production, we conclude that over a short time interval in which the signals can be considered to have time invariant dynamics (stationarity), consonants and breathy disordered speech is best modelled with a classical, Gaussian linear model such as (3.9). For highly periodic healthy and disordered vowel sounds, a deterministic nonlinear model such as

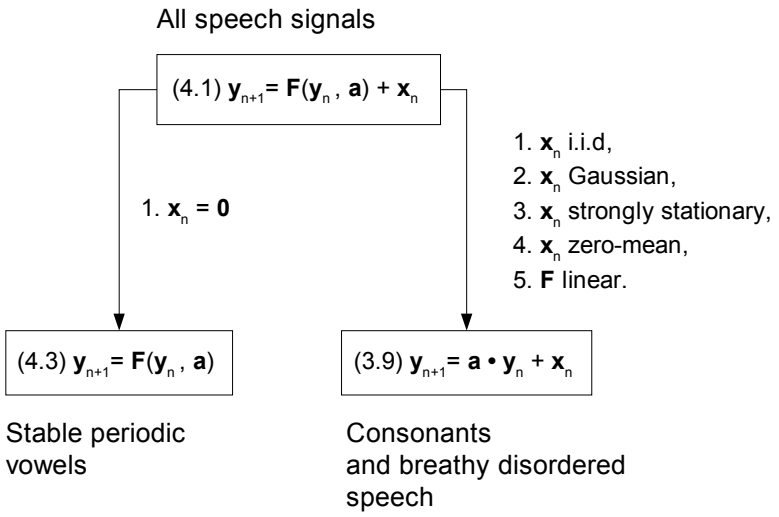

Figure 5.7: Graphical illustration of the hierarchical relationship between the candidate non-Gaussian, nonlinear model (4.1), the deterministic map (4.3) and the classical linear, Gaussian model (3.9). The left arrow lists the additional mathematical assumptions required to specialise the model at the top to the deterministic map case, and the right arrow shows the assumptions required to specialise to the classical Gaussian linear model. Alongside each model are the range of speech signals that each can reasonably explain, given the evidence from the surrogate data tests and the biomechanical, first principles knowledge described in this thesis. The inner product (dot) notation has been used as a shorthand for the summation in the linear model (3.9).

(4.3) is indicated, where for highly aperiodic (irregular) disordered vowel sounds, (4.1) is more appropriate.

Linear systems (3.9) are, however special cases of the more general, nonlinear, non-Gaussian models (4.1) (as shown in Appendix §A.2.4). Similarly, by setting the forcing term $\mathbf{x}_n = \mathbf{0}$, (4.3) becomes another special case of (4.1). Overall, therefore, we can model all the different speech signals we have encountered with just this one system. From the principle of parsimony, out of these several alternatives, we should prefer the model that can explain the dynamics of the most signals with the fewest restrictive assumptions. Since the linear Gaussian system and the deterministic map system are specialisations of equation (4.1), we should prefer this unified model. Figure 5.7 illustrates graphically the hierarchical relationship between these three different models, and how the model at the top is therefore the simplest, by virtue of needing the fewest mathematical assumptions.

The results of the surrogate data test are an empirical confirmation of the theoretical inconsistencies with the use of LPA for digital formant analysis identified in §3.3.1, for stationary vowel sounds. As we can see in figure 5.4, the departure from the H_0 required by the stochastic input interpretation of LPA is large. Therefore, there will be inaccuracies

in the estimates of the linear model parameters. This in turn implies that for CELP speech codecs, there will still be some additional structure contained in the residual that is not captured in the model parameters. Given that the compression efficiency of CELP codecs depends partly upon a parsimonious encoding of the residual, and that often the residual is assumed to have a Gaussian, i.i.d. component [56], compression could be improved by using, for example, non-Gaussian residual models. Whilst only small improvements may be possible in any one frame, the overall bandwidth savings could mount up substantially, especially considering that the inaccuracies arise for vowel sounds which represent the majority of the phonemes encountered in normal speech.

These inaccuracies will also permeate other digital speech technology that makes use of LPA. This includes speech recognition systems, for example. Nonetheless, these technologies appear to function fairly well despite these problems. One explanation for the apparent robustness of LPA is that, although the linear Gaussian hypothesis is ruled out significantly for a large proportion of phonemes, the nonlinear statistic on the original often appears to "track" the linear statistic on the original (see figure 5.4). The nonlinear statistic follows the increases and decreases of the linear statistic, to a certain extent. Therefore, there is qualitative agreement between the linear and nonlinear dependency structure at different time lags. This might be indicative of why, despite the failure of the linear model to account for all the dynamics in stable vowels, LPA still functions to extract a general picture of the time lag dependency structure [74].

Another explanation for the apparent success of LPA techniques is the observation that LPA residuals are often very close to Gaussian, i.i.d. stochastic processes, an observation that has been exploited in CELP codecs (discussed in Chapter 3). These codecs therefore originally used samples from a Gaussian stochastic process as a representation of the residual [56]. The near-Gaussianity and near-independence of these residuals is often cited as evidence that Gaussian AR models are a completely appropriate description for speech signals. However, an observation using information theoretic principles is that whatever the statistics of the original signal, least-squares AR model fitting always *increases the Gaussianity and independence of the residuals* [86]. Thus the appropriateness of Gaussian AR models for any signal cannot be judged by examining the properties of the residual, since the parametric fitting process will introduce a bias in favour of, in the least-squares case, a Gaussian linear model for the original signal.

A limitation of the surrogate data tests conducted in this chapter is the time invariance, or stationarity assumption. As discussed in §3.3.1, running speech is fundamentally dif-

ferent to isolated phonemes in that there is always substantial co-articulation, and where one phoneme ends and the other begins is often ambiguous. Thus the short time intervals selected for the test data in order to ensure stationarity are somewhat artificial. We must always remember that the alternative to the null hypothesis H_1 includes the possibility of linear systems driven by non-stationary Gaussian processes. Nonetheless, short time stationarity is a common assumption in current speech technologies making the results of the surrogate test more relevant to existing systems.

It is appropriate here to mention that there exists a possible contention between the complexity of the model selected above which is measured by the minimum number of special, restrictive mathematical assumptions needed to specify the function **F** completely, and the *effective complexity* [87] that counts more highly curved, nonlinear functions as more complex than smoother, linear functions. This contention is an open problem beyond the scope of this thesis which brings up deep issues of what is meant in general by mathematically "simple" and "complex", and involves theories of Bayesian complexity, minimum description length (MDL), minimum message length (MML), and Kolmogorov complexity. The interested reader is referred to, for example, Vitanyi [88] for more detailed discussions.

5.4 Chapter Summary

In this chapter, in order to overcome some of the deficiencies of earlier surrogate data studies, we have applied the improved surrogate data test developed in the previous chapter to try to refute the assumptions of LTI systems theory in a wide variety of speech signals. This led to the discovery that LTI systems theory cannot reasonably account for all the dynamics of the larger majority of speech signals, but for a small minority of signals it can. On the basis of this new empirical evidence and the theoretical modelling studies of earlier chapters, we concluded that a new digital model of speech production might better account for the wide range of phenomena encountered in real speech signals. This represents the most comprehensive and rigorous surrogate data test of LTI systems assumptions in speech signals carried out to date.

In the next chapter, we will develop new, practical methods for exploiting the modelling approach introduced here. This will be tested in a biomedical speech technology application.

CHAPTER 6

Clinical Applications of Nonlinearity in Speech

The previous chapters have motivated, introduced and tested empirically a new mathematical approach to modelling digital speech signals. This takes explicit account of current biomechanical knowledge and evidence from real signals. Such a development may be scientifically interesting, but the practical value must also be made explicit. It is the purpose of this chapter therefore to complement the theoretical developments of earlier chapters with the results of applying them to a practical engineering problem. This problem acts as a specific case study from which more general conclusions will be drawn later in the thesis.

It is clear from the surrogate data study of Chapter 5 that there is significant departure from LTI systems theory for normal speech. However, as argued in that chapter, significant practical or economic benefits will likely accrue only over long periods of time or over whole technological infrastructures (such as the entire mobile telephone network). For disordered speech however, there have been some pioneering studies claiming immediately obvious evidence of complex phenomena such as chaos and bifurcations [11, 42]. The analysis of disordered speech with methods from nonlinear time series analysis has thus received particular attention, and the practical advantages of such techniques may be more immediate and testable on a much smaller scale than is possible for normal speech. For these reasons, this chapter will investigate the use of the algorithms based upon concepts introduced in Chapter 4 to *disordered* voice analysis.

6.1 Nonlinear Clinical Measurement of Speech

Voice disorders arise due to physiological disease or psychological disorder, accident, misuse of the voice, or surgery affecting the vocal folds, and have a profound impact on the lives of patients. [1] This effect is even more extreme when the patients are professional voice users, such as singers, actors, radio and television presenters, for example. Commonly used

[1] This thesis only studies *voice* disorders which are but one of the many kinds of *speech* disorder.

by speech clinicians, such as surgeons and speech therapists, are acoustic tools, recording changes in acoustic pressure at the lips or inside the vocal tract. These tools [11], amongst others, can provide potentially objective measures of voice function. Although acoustic examination is only one tool in the complete assessment of voice function, such objective measurement has many practical uses in clinical settings, augmenting the subjective judgement of voice function by clinicians. These measures find uses, for example, in the evaluation of surgical procedures, therapy, differential diagnosis and screening [11, 89]. These measures can be used to portray a "hoarseness" diagram illustrating voice quality graphically [90], and there also exists a variety of techniques for automatically screening for voice disorders using these measures [91, 92, 93].

Phenomenologically, normal and disordered sustained vowel speech sounds exhibit a large range of phenomena. This includes *nearly periodic* or regular vibration, *aperiodic* or irregular vibration and sounds with no apparent vibration at all: some examples were shown in Chapter 5. All can be accompanied by varying degrees of noise which can be described as "breathiness". Titze [94] introduced a typology for these sounds, and this has been extended with subtypes [11]. Type I sounds are those that are nearly periodic: coming close to perfect periodicity. Type II are those that are aperiodic. They have no single, obvious or dominant period. The third class, Type III are those sounds that appear to have no pattern at all, and may even be noise-like, or random. Normal voices can usually be classed as Type I and sometimes Type II, whereas voice disorders commonly lead to all three types of sounds.

There exists a very large number of approaches to the acoustic measurement of voice function. The most popular of these are the *perturbation* measures *jitter* and *shimmer* and variants, and *noise-to-harmonics ratios* (NHR) [11, 90]. However, these measurement methods have limitations for the analysis of disordered speech. One reason is that they are only properly applicable when near periodicity holds: in Titze's typology only Type I sounds satisfy this property [11]. The behaviour of the algorithms for other sound types is not known theoretically and limited only to experimental results [90]. The source of this limitation is that they make extensive use of extraction of the *pitch period*, or *fundamental frequency* (defined as the inverse of the pitch period) from the acoustic signal [11]. Popular pitch period extraction techniques include *zero-crossing detection*, *peak-picking* and *waveform matching* [11]. The concept of pitch period is only valid for Type I sounds and therefore application of these methods based upon periodicity analysis, to any other type of sound is problematic [92]. Type II and III have therefore received much

less attention in the literature [94], such that there exist few methods for characterising these types, despite the fact that they exist in great abundance in clinical settings. This precludes the proper use of these tools on a large number of disordered speech cases, limiting the reliability of the analysis, since in fact some algorithms will not produce any results at all for Type II and III sounds [89].

Another reason for the limitations of these methods is that they are based upon classical linear signal processing methods (such as autocorrelation, the discrete Fourier transform, linear prediction analysis and cepstral processing) that are insensitive to the biomechanical nonlinearity and non-Gaussianity in speech [11].

Since standardised, reliable and reproducible results from acoustic measures of voice function are required for clinical applications, these limitations of perturbation methods are problematic in clinical practice [89]. It is clear that there is a clinical need for reliable tools that can characterise all types of disordered voice sounds for a variety of clinical applications, regardless of whether they satisfy the requirements of near periodicity, or contain significant nonlinearity, randomness or non-Gaussianity [94].

Furthermore, current disordered voice analysis techniques are complicated by the use of any arbitrary algorithmic parameters whose choice affects the analysis method – changing these parameters can change the analysis results. Such arbitrary parameters are not justified on the basis of theoretical principles; they are chosen by experimental and empirical evaluation alone. There exists the danger that these parameters are highly "tuned" to the particular data set used in any one study, limiting the reproducibility of the analysis on different data sets. It is necessary therefore to reduce the number of arbitrary parameters to improve the reproducibility of these measurement methods.

To address these limitations of classical linear techniques, there has been growing interest in applying tools from nonlinear time series analysis to disordered speech signals in order to attempt to characterise and exploit these nonlinear phenomena [11, 42]. Algorithms for calculating the *correlation dimension* [8] have been applied, which were successful in separating normal from disordered subjects [95]. Correlation dimension and *second-order dynamical entropy* [8] measures showed statistically significant changes before and after surgical intervention for vocal fold polyps [96], and Lyapunov exponents for disordered voices were found to be consistently higher than those for healthy voices [97]. It was also found that jitter and shimmer measurements were less reliable than correlation dimension analysis on Type I and unable to characterise Type II and (non-random) Type III sounds [98]. However, correlation dimension analysis was found to be less reliable

for analysis of *electroglottographic* [2] data from disordered voice sounds in another study [99], and inconclusive results were found for fractal dimension analysis of sounds from patients with neurological disorders, for both acoustic and electroglottographic signals [100]. Instantaneous nonlinear *amplitude* (AM) and *frequency* (FM) formant modulations were shown effective at detecting muscle tension dysphonias [101]. For the automated acoustic screening of voice disorders, *higher-order statistics* lead to improved normal/disordered classification performance when combined with several standard perturbation measures [93].

These studies show that nonlinear time series methods can be valuable tools for the analysis of voice disorders, in that they can analyse a much broader range of speech sounds than perturbation measures, and in some cases are found to be more reliable under conditions of high noise. Despite these successes of nonlinear time series analysis methods, common approaches such as time-delay reconstruction, correlation dimension and Lyapunov exponent calculation discussed in Chapter 4 require that the dynamics of speech be purely deterministic (so that the model of equation (4.3) holds), such that random Type III sounds have so far received little attention from nonlinear approaches. There are also numerical, theoretical and algorithmic problems associated with the calculation of nonlinear measures such as Lyapunov exponents or correlation dimensions for real speech signals, casting doubt over the reliability of such tools [8, 99, 100, 102]. For example, correlation dimension analysis shows high sensitivity to the variance of signals in general, and it is therefore necessary to check that changes in correlation dimension are not due simply to changes in variance [103]. Similarly, algorithms for the estimation of Lyapunov exponents or correlation dimensions require a very large amount of data with a low level of noise and the absence of other confounding factors, which is difficult to obtain in practice.

As we have shown in this thesis, the deterministic nonlinear dynamical model alone, whilst promising, is inadequate since randomness due to turbulence is an inherent part of speech production. The new, stochastic, nonlinear signal model introduced earlier can also account for Type III random speech sounds. The output of this model can then be analysed using methods that are able to characterise both nonlinearity and randomness. The deterministic component of the model can exhibit both periodic and aperiodic dynamics. It is proposed to characterise this component using *recurrence analysis* (see §4.3). The stochastic components can exhibit statistical self-similarity, which can be analysed

[2] Electroglottography measures the changes in electrical resistance through the larynx as it opens and closes.

effectively using *fractal scaling analysis* (see §4.6).

As a test of the effectiveness of these new disordered voice analysis tools, this chapter reports the replication of the "hoarseness" diagram [90] illustrating the extent of voice disorder, and demonstrates, using a simple pattern classifier, how these new measures may be used to automatically classify voices as normal or disordered from a large database of subjects.

6.2 Review of Traditional Classification Approaches

The goal of this chapter is to test the effectiveness of new nonlinear signal processing methods for voice disorder characterisation. In order to illustrate how this is achieved currently, we will review three studies that make use of traditional perturbation measures and signal processing tools based around LTI systems theory for automatically classifying voices into normal or disordered cases.

The method of [90] investigates the use of six different classical perturbation and noise measures, varying some of the parameters used to calculate these measures. This results in a 22 element *feature vector* for sustained vowels, with one vector for each of 447 disordered and 88 normal subjects. Using *principal components analysis* (PCA), this vector was projected down onto the two directions in this feature space with the largest variance. The validity of this two-dimensional projection was tested using a reduced, minimally-redundant four-dimensional subset of this vector found using mutual information analysis. These two projected directions for each subject were then used to construct a two-dimensional hoarseness diagram, similar to that shown in figure 6.7 with a horizontal vibrational irregularity and vertical noise axis.

The method of [91] divides the speech signal up into stable segments (in which the pitch period can be reliably extracted), and forms a vector for each segment consisting of nine standard jitter, shimmer, noise and voiced/unvoiced perturbation measures. These vectors are passed on to four different types of classifiers. These are trained on sustained vowels from 150 different normal and disordered subjects and tested on a different set of 250 subjects. The output of these four different classifiers are weighted and combined to obtain a final normal/disordered classification.

Finally, in the study of [92], the speech signal is divided up into frames and noisy or silent frames are removed. For each remaining frame, MFCCs (mel frequency cepstral coefficients), their energy, and their temporal first and second differences form vectors

for both an MLP (Multi-Layer Perceptron) and an LVQ (Learning Vector Quantisation) classifier. The classifiers are trained on 70% and tested on 30% of 135 subjects. Each frame is classified separately, and the whole speech example is classified normal/disordered according to a threshold over the number of frames classified as normal or disordered by the classifier.

We wish to perform a direct comparison of the new, biomechanically-informed, nonlinear signal processing algorithms against traditional perturbation methods, in an experimental setting that brings out their essential differences. Unfortunately, the three studies mentioned above are typical in that they all reach prohibitive levels of complexity, both in terms of the number of measures that are calculated for each subject, and in terms of the classification methods used. There are a very large number of traditional measures (for example, the Kay Multi-Dimensional Voice Program (MDVP) can calculate 33 different measures [85]) that could be combined for each subject, [3] rendering a systematic pairwise comparison largely intractable. Similarly, some of the studies above combine many different and highly complex classification methods. It is not clear that the studies described above represent the most parsimonious approach to evaluating the new methods developed in this chapter.

In order to circumvent these problems, we will select and use just one simple, but nonetheless flexible, classifier: *(Fisher's) quadratic discriminant analysis method* (QDA). Using this classifier we will compare combinations of the new nonlinear signal processing algorithms against combinations of the most widely-used of the traditional measures: Jitter, Shimmer and NHR [89, 11]. This will allow us to focus on the performance of measures, rather than issues related to the classification system. We will next describe the proposed new measures and their algorithms.

6.3 New Practical Analysis Algorithms for Speech Disorder Characterisation

In §4.3 the concept of recurrent orbits was introduced. Using this concept, we can describe nearly periodic speech sounds of Type I as recurrent for some small $r > 0$, with Δn nearly the same for each n. Type II sounds are more irregular than Type I, and for the same r, the Δn will assume a wider range of values than for Type I. Similarly, Type III sounds

[3] For example, choosing pairs of measures from the 33 MDVP system leads to ${}^{33}C_2 = 528$ possible combinations.

that are highly irregular and aperiodic will have a large range of values of Δn again for the same r.

Similarly, in §4.6 the concepts of graph dimension and scaling exponent were introduced. It has also been found experimentally that changes in the statistical time dependence properties of turbulent noise in speech, as measured by a particular fractal graph dimension measure applied to the speech signal, are capable of distinguishing classes of phonemes from each other [23]. Also, it is well known from studies of disordered speech that some voice disorders are accompanied by increased "breathiness", which is due in part to the inability of the vocal folds to close properly, so that air escapes through the partial constriction of the vocal folds creating increased turbulence in the airflow [52]. Thus scaling analysis and/or graph dimension measures could be useful for characterising vocal fold disorders.

Initial pilot studies have shown that recurrence analysis, carried out using the *recurrence probability density entropy* algorithm, and scaling analysis using the *detrended fluctuation analysis* algorithm, both described in the next section, can distinguish healthy from disordered speech on a large database of recordings with high accuracy [102]. These techniques are computationally simple and involve a very small number of arbitrary parameters that have to be chosen in advance, thus leading to increased reproducibility and reliability. We will now describe these algorithms in detail (refer to figure 6.1 for flow chart of these techniques accompanying the description).

6.3.1 Recurrence Probability Density Entropy Algorithm (RPDE)

Measurements of the output of the system (4.1) are assumed to constitute the acoustic signal, s_n:

$$s_n = h(\mathbf{y}_n), \tag{6.1}$$

from which a d-dimensional time-delay reconstruction vector is constructed:

$$\mathbf{s}_n = \left[s_n, s_{n-\tau}, \ldots s_{n-(d-1)\tau}\right]^T. \tag{6.2}$$

Here τ is the reconstruction time delay and d is the reconstruction dimension.

For time-delay reconstruction of stochastic signals such as s_n, techniques such as false-nearest neighbours and minimum time-delayed mutual information discussed in §4.4 for determining the optimal values of d and τ are not applicable. We instead use the approach in [8] of optimising the reconstruction parameters d and τ such that the recurrence analysis produces results as close as possible to analytically derived results upon calibration with

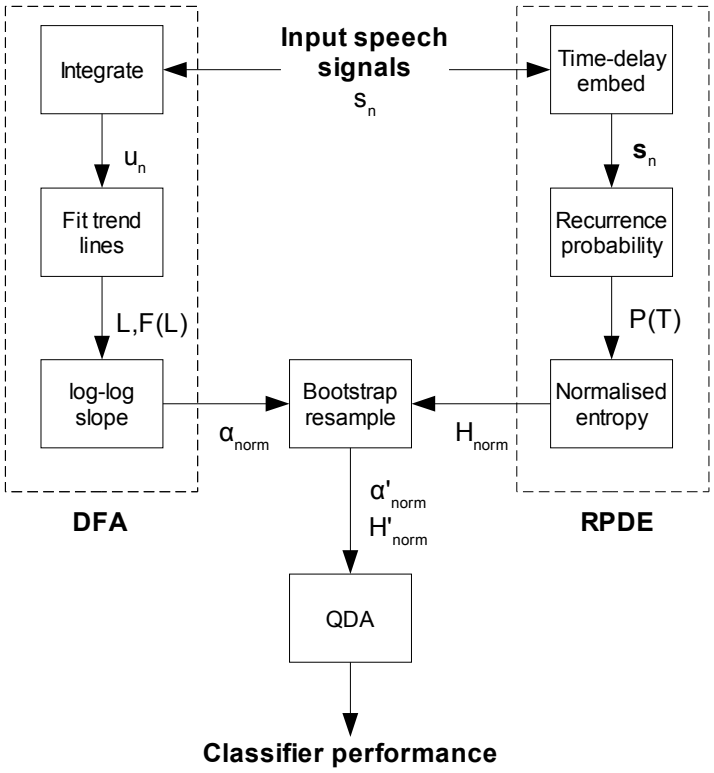

Figure 6.1: Overall flow chart depicting the new voice disorder analysis method described in §6.3, comprising Detrended Fluctuation Analysis (DFA), Recurrence Period Density Entropy (RPDE) and bootstrapped Quadratic Discriminant Analysis (QDA). Each speech signal s_n is passed on to both DFA and RPDE algorithms, which calculate the normalised scaling exponent α_{norm} and recurrence period density entropy H_{norm} measures. The QDA classifier is re-trained on each bootstrap resampled set of measures, and the classifier performance is calculated for each of these sets.

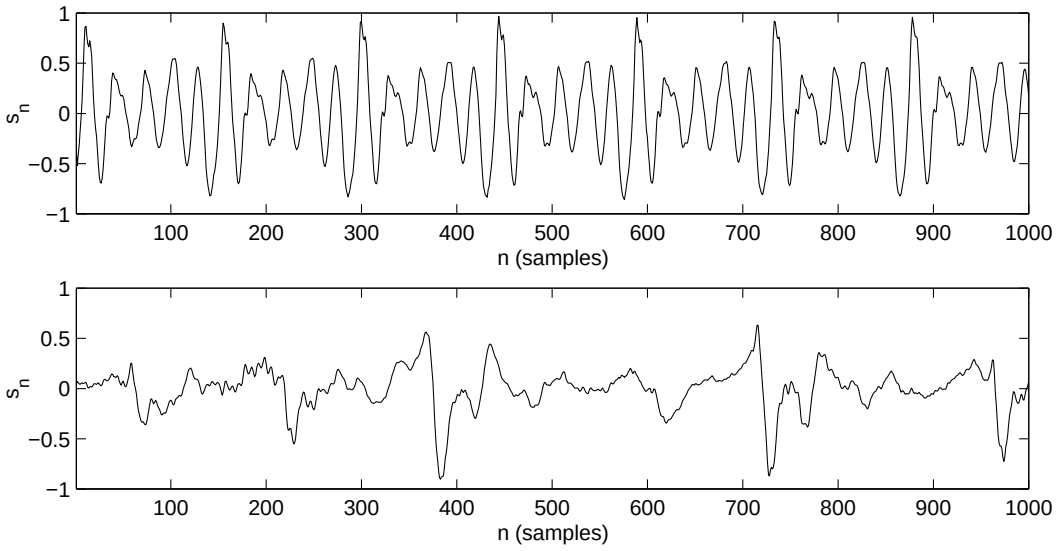

Figure 6.2: Discrete-time signals from (top panel) one normal (JMC1NAL) and (bottom panel) one disordered (JXS01AN) speech signal from the Kay Elemetrics Disordered Voice Database. For clarity only a small section is shown (1500 samples).

known signals. We develop these analytical results in this chapter. This optimisation is carried out by a simple, systematic grid search of values.

Figure 6.2 shows the signals s_n for one normal and one disordered speech example from the Kay Elemetrics Disordered Voice Database. The signals represent sustained, single vowel phonations. Figure 6.3 shows the result of applying the above reconstruction procedure for the same speech signals.

We investigate the recurrence time statistics of both normal and disordered speech using the *method of close returns* [104], an algorithm originally designed to analyse deterministic, chaotic dynamics. In this algorithm, a small, closed ball $B(\mathbf{s}_{n_0}, r)$ of radius $r > 0$ is placed around the embedded data point $\mathbf{s}_{n_0}$. Then the orbit is followed in forward time $\mathbf{s}_{n_0+1}, \mathbf{s}_{n_0+2} \ldots$ until it has left this ball, i.e. until $|\mathbf{s}_{n_0} - \mathbf{s}_{n_0+j}| > r$ for some $j > 0$. Subsequently, the time n_1 at which the orbit first returns to this same ball is recorded (i.e. the first time when $|\mathbf{s}_{n_0} - \mathbf{s}_{n_1}| \leq r$), and the difference of these two times is the (discrete) recurrence time $T = n_1 - n_0$. This procedure is repeated for all the embedded data points $\mathbf{s}_n$, forming a histogram of recurrence times $R(T)$. This histogram is normalised to give the *recurrence time probability density*:

$$P(T) = \frac{R(T)}{\sum_{i=1}^{T_{\max}} R(i)}, \tag{6.3}$$

where $T_{\max}$ is the maximum recurrence time. This fixed parameter is typically chosen in advance such that all empirically-obtained recurrence times for a given finite-length signal

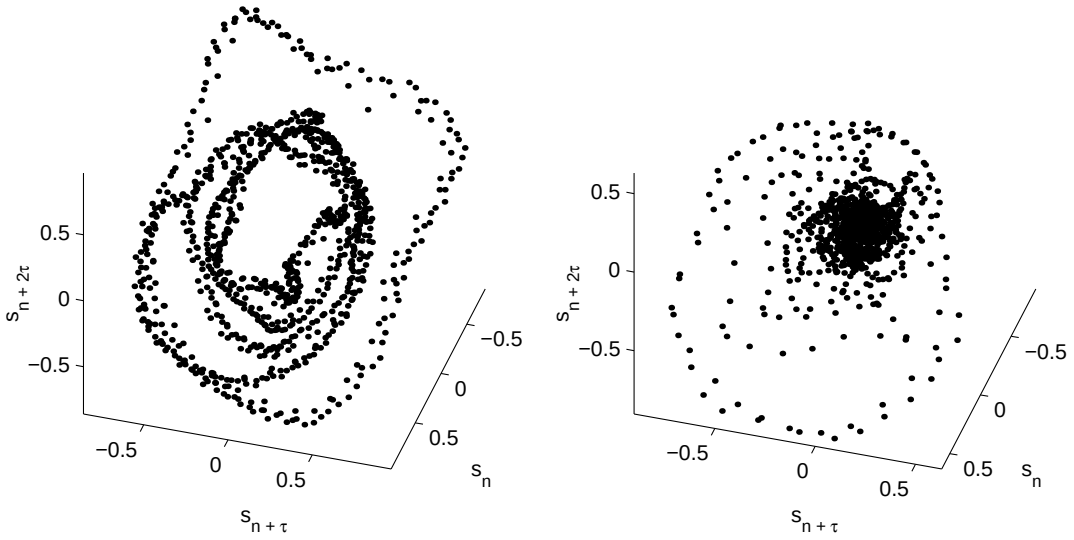

Figure 6.3: Time-delay embedded discrete time signals from (left) one normal (JMC1NAL) and (right) one disordered (JXS01AN) speech signal from the Kay Elemetrics Disordered Voice Database. For clarity only a small section is shown (1500 samples). The reconstruction dimension is $d = 3$ and the time delay is $\tau = 7$ samples.

are less than or equal to this value. The choice of r is important to capture the properties of interest to this study. For example, if the orbit is nearly periodic, we require that r is large enough to capture all the recurrences, but not too large to find recurrences that are due to spurious intersections of $B(\mathbf{s}, r)$ with other parts of the orbit, violating the conditions for proper recurrence. The appropriate choice of reconstruction delay τ has a role to play: selecting τ too small means that any orbit lies close to the diagonal in the reconstructed state space, potentially causing spurious recurrences. Thus τ must be chosen optimally (in this thesis by systematic search).

We consider two extreme forms that the density (6.3) may assume. The first is the ideal limiting case in which the recurrence distance r tends to zero for a periodic orbit. The recurrence time probability density is:

$$P(T) = \begin{cases} 1 & \text{if } T = K \\ 0 & \text{otherwise} \end{cases}, \tag{6.4}$$

where K is the period of the orbit. See Appendix §A.2.6 for a proof of this result. In the second extreme case we consider a purely random, uniform i.i.d. signal which is normalised to the range $[-1, 1]$. The recurrence probability density is approximately uniform:

$$P(T) \approx \frac{1}{T_{\max}}. \tag{6.5}$$

For a proof of this result see Appendix §A.2.7.

We optimise d, τ and r such that for a synthetic signal of perfect periodicity, $P(T)$ is determined using the close returns method such that it is as close as possible to the theoretical expression (6.4). This optimisation is carried out by a straightforward systematic (grid) search of values of these parameters $d = 2, 3 \ldots 10$, $\tau = 2, 3 \ldots 50$, and for $r = 0.02, 0.04, \ldots 0.5$, on a perfectly periodic test signal.

All voice signals will lie somewhere in between the extremes of perfect periodicity and complete randomness. Thus it will be useful to create a sliding scale so that voice signals can be ranked alongside each other. This depends upon a simple characterisation of the recurrence probability density $P(T)$. One such measure that we can use is the entropy of the recurrence probability density, which can rank disordered speech signals according to the uncertainty in the period of the disordered speech signal in the following way. For perfectly periodic signals the *recurrence probability density entropy* (RPDE) is:

$$H_{\text{per}} = -\sum_{i=1}^{T_{\max}} P(i) \ln P(i) = 0. \tag{6.6}$$

since $P(K) = 1$ and the rest are zero. Conversely, for the purely stochastic, uniform i.i.d. case, as shown in the appendix, the uniform density can be taken as a good approximation, so that the RPDE is:

$$H_{\text{iid}} = -\sum_{i=1}^{T_{\max}} P(i) \ln P(i) = \ln T_{\max}, \tag{6.7}$$

in units of nats. The entropy scale H therefore ranges from H_{per}, representing perfectly periodic examples of Type I sounds, to H_{iid} as the most extreme cases of noise-like Type III sounds. In practice, all sounds will lie somewhere in between these extremes.

Because the entropy of a probability density is maximum for the uniform density, H_{iid} is the maximum value that H can assume. For different sampling times Δt the value $T_{\max}$ will change. Therefore, the RPDE is normalised for subsequent calculations:

$$H_{\text{norm}} = \frac{-\sum_{i=1}^{T_{\max}} P(i) \ln P(i)}{H_{\text{iid}}}. \tag{6.8}$$

Figure 6.4 shows the result of this recurrence analysis, applied to a synthetic, perfectly periodic signal created by taking a single cycle from a speech signal and repeating it end-to-end many times. It also shows the analysis applied to a synthesised, uniform, i.i.d. random signal on the range $[-1, 1]$ after optimising d, τ and r. Even though exact results are impossible to obtain due to the approximation inherent to the algorithm and only finite-length signals, the figure shows that a close match is obtainable between the theoretical, predicted results and the simulated results.

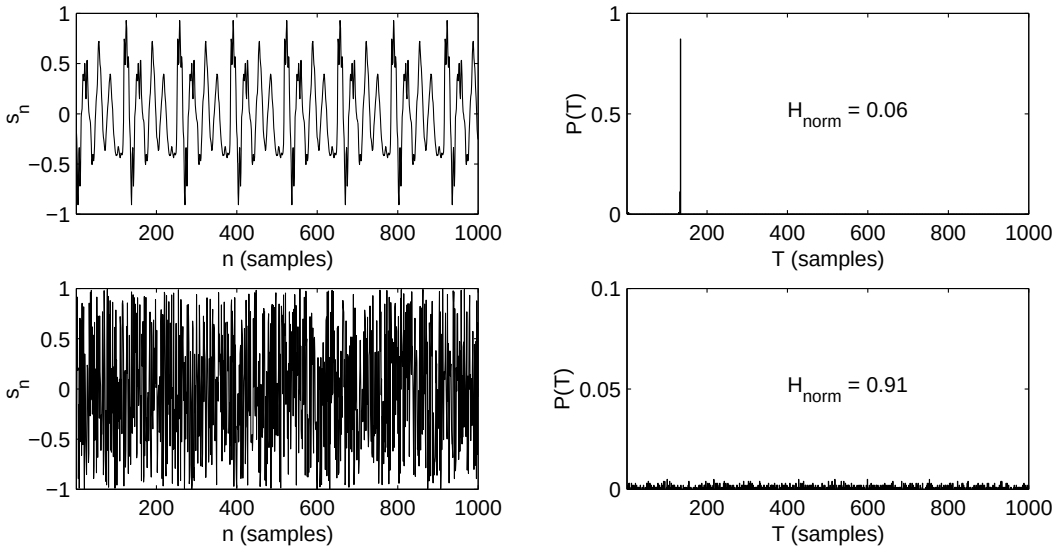

Figure 6.4: Demonstration of results of time-delayed state space recurrence analysis applied to (top row) a perfectly periodic signal created by taking a single cycle (period $K = 134$ samples) from a speech signal and repeating it end-to-end many times. The signal was normalised to the range $[-1, 1]$. All values of $P(T)$ are zero except for $P(133) = 0.1354$ and $P(134) = 0.8646$ so that $P(T)$ is properly normalised. The bottom row shows the analysis applied to a synthesised, uniform i.i.d. random signal on the range $[-1, 1]$. The density $P(T)$ is fairly uniform. For clarity only a small section of the time series (1000 samples) and the recurrence time (1000 samples) is shown. Here, $T_{\max} = 1000$. The length of both signals was 18088 samples. The optimal values of the recurrence analysis parameters were found at $r = 0.12$, $d = 4$ and $\tau = 35$.

6.3.2 Detrended Fluctuation Analysis Algorithm (DFA)

Whilst there exists stationary, self-similar stochastic processes exhibiting power-law scaling $P_{xx}(k) = k^{-\beta}$ of their power spectra, where β is a positive *power-law exponent*, these processes do not, in general, have a corresponding representation as a finite P-order memory Gaussian AR system such as (3.9) – see [105] for further details. Thus measuring the scaling properties of such processes cannot be carried out using Linear Prediction Analysis (LPA), and although power spectral analysis may be possible for statistically self-similar processes, the speech signals we encounter in this study, produced by the model (4.1), will contain both a rapidly varying stochastic component due to the forcing term $\mathbf{x}_n$, and also slower variation due to the nonlinear, deterministic function $\mathbf{F}$. We wish to be able to characterise the scaling exponent of the stochastic component of the model alone, but this slowly varying component will be prominent in the power spectrum precluding measurement of the scaling exponent of the graph of the signal using spectral methods.

As a solution to this, we turn to one straightforward and practical algorithm for estimating the scaling exponent of the graph of a signal: *detrended fluctuation analysis* (DFA) [106]. This method has been shown to be effective for signals exhibiting slowly varying trends [107].

The DFA algorithm is based around fitting straight lines (trends) over intervals of size L to the (integrated) signal, and measuring the average root-mean-square deviation $F(L)$ (fluctuation) around the trend lines. The first step in the algorithm is an integration-like processing of the original time series by summation:

$$u_n = \sum_{i=1}^{n} s_i, \tag{6.9}$$

for $n = 0, 1 \ldots N-1$ where N is the length of the signal s_n. The motivation for this step is to induce self-similarity into signals which have a finite maximum amplitude scale, which is true for the majority of signals we will encounter in this study. For example, a bounded realisation of a Gaussian, i.i.d. signal s_n will result in a self-similar, Gaussian random walk u_n, so that the original signal can be characterised in terms of an associated scaling exponent value.

The next step in the algorithm is the division of the signal u_n into non-overlapping intervals of length L. For each interval a best-fit straight line trend for u_n is calculated, [4] producing a new, piecewise linear trend signal for this interval length – we denote this as

[4] In this study we use least-squares estimation for the slope and intercept parameters for the straight line.

u_n^L. Then the fluctuation for this time scale is calculated:

$$F(L) = \left[\frac{1}{N} \sum_{n=0}^{N-1} \left(u_n - u_n^L \right)^2 \right]^{1/2}. \tag{6.10}$$

The final step is to fit a straight line of slope α to the set of points $\{\log L, \log F(L)\}$ over all interval lengths [5] L. Note that due to the earlier integration step, this will be a different α than the scaling exponent for the original time series, and this must be taken into account in subsequent analysis [106]. For example, a Gaussian, i.i.d. signal s_n will result in a DFA scaling exponent of $\alpha = 1/2$.

The signal s_n represents a combination of deterministic and stochastic dynamics. The deterministic part of the dynamics, dictated by the function $\mathbf{F}$ in equation (4.1) will result in slower changes in the signal s_n taking place over a relatively long time scale. Similarly, the stochastic fluctuations in the signal indicated changes taking place over a much shorter time scale. Since the goal of DFA is to analyse the stochastic properties of the signal, only a limited range of interval lengths is investigated, over which the stochastic component of the signal exhibits self-similarity as indicated by a straight line on the log–log graph of interval length against fluctuation.

The resulting scaling exponent can assume any number on the real line. However, it would be more convenient to represent this scaling exponent on a finite scale from zero to one. Thus it is necessary to find a mapping function $g : \mathbb{R} \to [0, 1]$. One such function finding common use in statistical and pattern recognition applications is the *logistic function* $g(x) = (1 + \exp(-x))^{-1}$ [87], so that the normalised scaling exponent becomes:

$$\alpha_{\text{norm}} = \frac{1}{1 + \exp(-\alpha)}. \tag{6.11}$$

Therefore, each sound will lie somewhere between the extremes of zero and one on this scale, according to the self-similarity properties of the stochastic part of the dynamics. As will be shown later, speech sounds for which α_{norm} is closer to one are characteristic of general voice disorder.

6.3.3 Application of Algorithms to Normal and Disordered Examples

Figure 6.5 shows the normalised RPDE value H_{norm} calculated on the same two speech signals from the Kay Elemetrics database as shown in figure 6.2. Note that the second,

[5] Again, in this study we use least-squares regression.

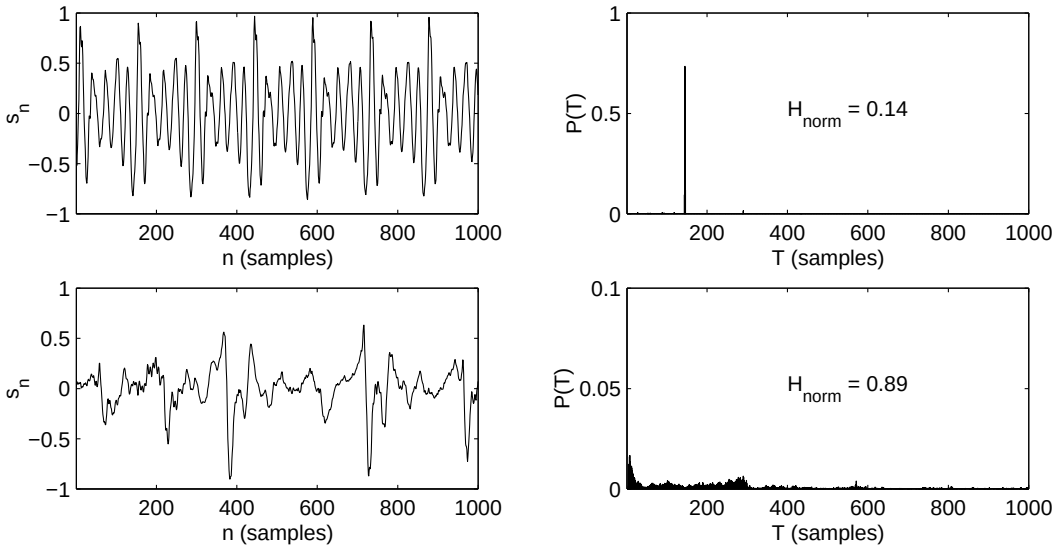

Figure 6.5: Results of RPDE analysis carried out on the two example speech signals from the Kay Elemetrics database as shown in figure 6.2. Top row is normal speech (JMC1NAL), bottom row is disordered speech (JXS01AN). The values of the recurrence analysis parameters were the same as those in the analysis of figure 6.4.

disordered example is of Type III and shows significantly irregular vibration, which is detected by an increase in H_{norm}.

Similarly, figure 6.6 shows two more speech examples, one normal and one disordered from the same database and the corresponding values of the scaling exponent α and α_{norm}. In these cases, the disordered example is extremely "breathy", and this turbulent noise is detected by an increase in the scaling exponent.

6.3.4 Quadratic Discriminant Analysis (QDA)

In order to test the effectiveness of these two measures in practice, the approach taken in this study is to set up a classification task to separate normal control subjects from disordered subjects using these measures alone. We choose one of the simplest approaches, quadratic discriminant analysis, which allows separation of the classes by (hyper)-conic section boundaries. This is achieved by modelling the data conditional upon each class, here the normal (class C_1) and disordered (class C_2) cases, using joint Gaussian probability density functions [87]. For a $I \times J$ data matrix $\mathbf{v} = v_{ij}$ of observations consisting of the measures $i = 1, 2$ for RPDE and DFA respectively, and all subjects j, these likelihood densities are parameterised by the mean and covariance matrices of the data sets:

$$\boldsymbol{\mu} = E\left[\mathbf{v}\right], \quad \mathbf{C} = E\left[\left(\mathbf{v} - \boldsymbol{\mu}\right)\left(\mathbf{v} - \boldsymbol{\mu}\right)^T\right], \tag{6.12}$$

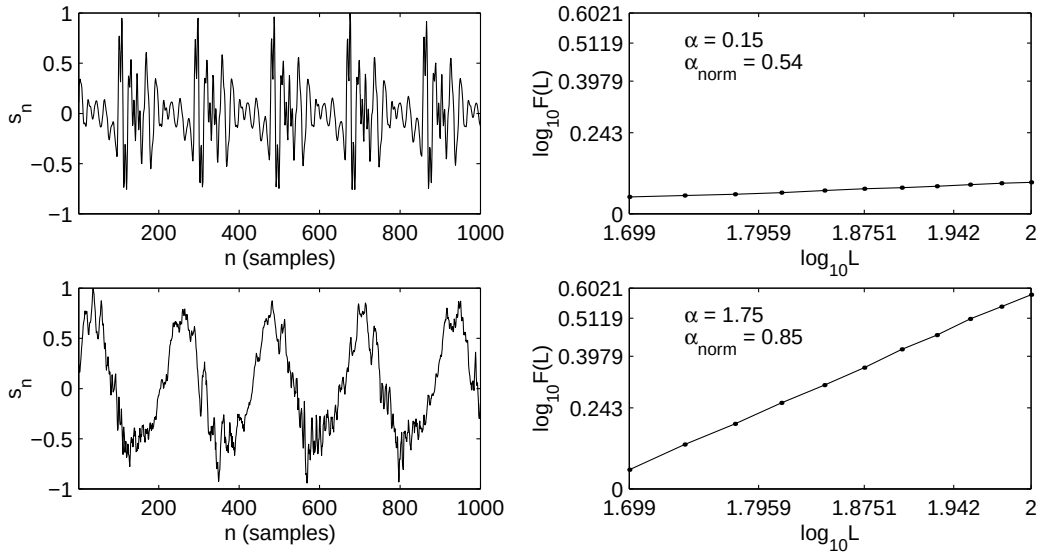

Figure 6.6: Results of scaling analysis carried out on two more example speech signals from the Kay database. Top row is normal voice (GPC1NAL), bottom row is disordered voice (RWR14AN). Left column are the discrete time signals s_n over a limited range of n for clarity. The right column shows the logarithm of scaling interval lengths L against the logarithm of fluctuation size $F(L)$. The values of L ranged from $L = 50$ to $L = 100$ in steps of five.

where E is the expectation operator, and $\boldsymbol{\mu}$ is the mean vector formed from the means of each row of $\mathbf{v}$. The class likelihoods are:

$$f_C(\mathbf{w}|C_k) = (2\pi)^{-I/2} |\mathbf{C}_k|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_k)^T \mathbf{C}_k^{-1} (\mathbf{w} - \boldsymbol{\mu}_k)\right], \tag{6.13}$$

for classes $k = 1, 2$ and an arbitrary observation row vector $\mathbf{w}$. It can be shown that, given these Gaussian class models, the maximum likelihood regions of the space $\mathbb{R}^I$ are separated by a *decision boundary* which is a (hyper-)conic section calculated from the difference of log-likelihoods for each class, which is the unique set of points where each class is equally likely [87]. The maximum likelihood classification problem is then solved using the decision rule that $l(\mathbf{w}) \geq 0$ assigns $\mathbf{w}$ to class C_1, and $l(\mathbf{w}) < 0$ assigns it to class C_2, where:

$$l(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^T \mathbf{A}_2 \mathbf{w} + \mathbf{A}_1 \mathbf{w} + A_0, \tag{6.14}$$

$$\mathbf{A}_2 = \mathbf{C}_1^{-1} - \mathbf{C}_2^{-1}, \mathbf{A}_1 = \boldsymbol{\mu}_1^T \mathbf{C}_1^{-1} - \boldsymbol{\mu}_2^T \mathbf{C}_2^{-1}, \tag{6.15}$$

$$A_0 = -\frac{1}{2}\ln|\mathbf{C}_1| + \frac{1}{2}\ln|\mathbf{C}_2| - \frac{1}{2}\boldsymbol{\mu}_1^T \mathbf{C}_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \mathbf{C}_2^{-1} \boldsymbol{\mu}_2. \tag{6.16}$$

In order to avoid overfitting, the *generalisation performance* of the classifier can be tested using *bootstrap resampling* [87]. The classifier is trained on J cases selected at random with replacement from the original data set of J cases. This trial resampling processes

is repeated many times and the mean classification parameters $E[\mathbf{A}_2], E[\mathbf{A}_1], E[A_0]$ are selected as the parameters that would achieve the best performance on entirely novel data sets.

Bootstrap training of the classifier involves calculating H^j_{norm} and α^j_{norm} (the observations) for each speech sample j in the database (where the superscript j denotes the measure for the j-th subject). Then, J random selections of these values with replacement H'^j_{norm} and α'^j_{norm} form the entries of the vector $v_{1j} = H'^j_{\text{norm}}$ and $v_{2j} = \alpha'^j_{\text{norm}}$. Then the mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\mathbf{C}_1, \mathbf{C}_2$ for each class are calculated. Next, for each subject, the decision function is evaluated:

$$l(\mathbf{w}_j) = l([H^j_{\text{norm}}, \alpha^j_{\text{norm}}]^T). \tag{6.17}$$

Subsequently, applying the decision rule assigns the subject j into either normal or disordered classes. Then the performance of the classifier can be evaluated in terms of percentage of true positives (when a disordered subject is correctly assigned to the disordered class C_1) and true negatives (when a normal subject is correctly assigned to the normal class C_2). The overall performance is the total number of correct classifications. This bootstrap trial process of creating random selections of the measures, calculating the class mean vectors and covariance matrices, and then evaluating the decision function on all the measures to obtain the classification performance is repeated many times. Assuming that the performance percentages are normally distributed, then the 95% confidence interval of the classification performance percentages can be calculated. The best classification boundary is taken as the mean boundary over all the trials.

6.4 Data

This study makes use of the Kay Elemetrics Disordered Voice Database [85], which contains 707 examples of disordered and normal voices from a wide variety of organic, neurological and traumatic voice disorders. This database contains many examples of all three types of disordered speech signals (Types I, II and III). There are 53 control samples from normal subjects. Each speech sample in the database was recorded under controlled acoustic conditions, and is on average around two seconds long, 16 bit uncompressed PCM audio. Some speech samples were recorded at 50kHz and then downsampled with anti-aliasing to 25kHz. Used in this study are sustained vowel phonations, since this controls for any significant nonstationarity due to changes in the position of the articulators such as the

Table 6.1: Summary of disordered voice classification task performance results, for several different combinations of the new measures and traditional perturbation measures, Jitter (Jitt), Shimmer (Shim) and Noise-to-Harmonics Ratio (NHR). The RPDE parameters were the same as for figure 6.4, and the DFA parameters were the same as for figure 6.6. Since the distributions are not precisely Gaussian, some of the extremes of the confidence intervals may be larger than 100%.

Combination	Measures (I)	Subjects (J)	True Positive	True Negative	**Overall**
RPDE/DFA	2	707	95.4±3.2%	91.5±2.3%	**91.8±2.0%**
Jitt/Shim/NHR	3	684	91.5±7.3%	80.5±4.5%	**81.4±3.7%**
Jitt/Shim	2	685	86.9±6.9%	81.0±4.7%	**81.4±3.9%**
Shim/NHR	2	684	91.4±5.9%	79.8±4.7%	**80.7±4.0%**
Jitt/NHR	2	684	93.2±7.4%	75.0±5.5%	**76.4±4.8%**

tongue and lips in running speech, which would have an adverse effect upon the analysis methods.

6.5 Results

Figure 6.7 shows the hoarseness diagram of [90] constructed using the speech data and the RPDE and DFA measures. For direct comparison, it also shows an attempt to construct the same diagram using three other combinations of three traditional perturbation measures, Jitter, Shimmer and NHR (Noise-to-Harmonics Ratio) [11]. The normalised RPDE and DFA scaling exponents are calculated for each of the $J = 707$ speech signals. Where the traditional perturbation algorithms did not fail, the traditional perturbation values were also calculated for a smaller subset of the subjects, see [11] for details of these algorithms. Also shown in figure 6.7 is the result of the classification task applied to the dataset; the best classification boundary is calculated using bootstrap resampling over 1000 trials. Table 6.1 summarises all the classification performance results for the classification tasks on the hoarseness diagrams of figure 6.7. The RPDE parameters were the same as for figure 6.4, and the DFA parameters were the same as for figure 6.6.

6.6 Discussion of Results

As shown in table 6.1, of all the combinations of the new and traditional measures, the highest overall correct classification performance of $91.8 \pm 2.0\%$ is achieved by the RPDE/DFA pair. The combination of Jitter, Shimmer with NHR leads to the next highest performance. These results confirm that, compared under the same, simple classifier

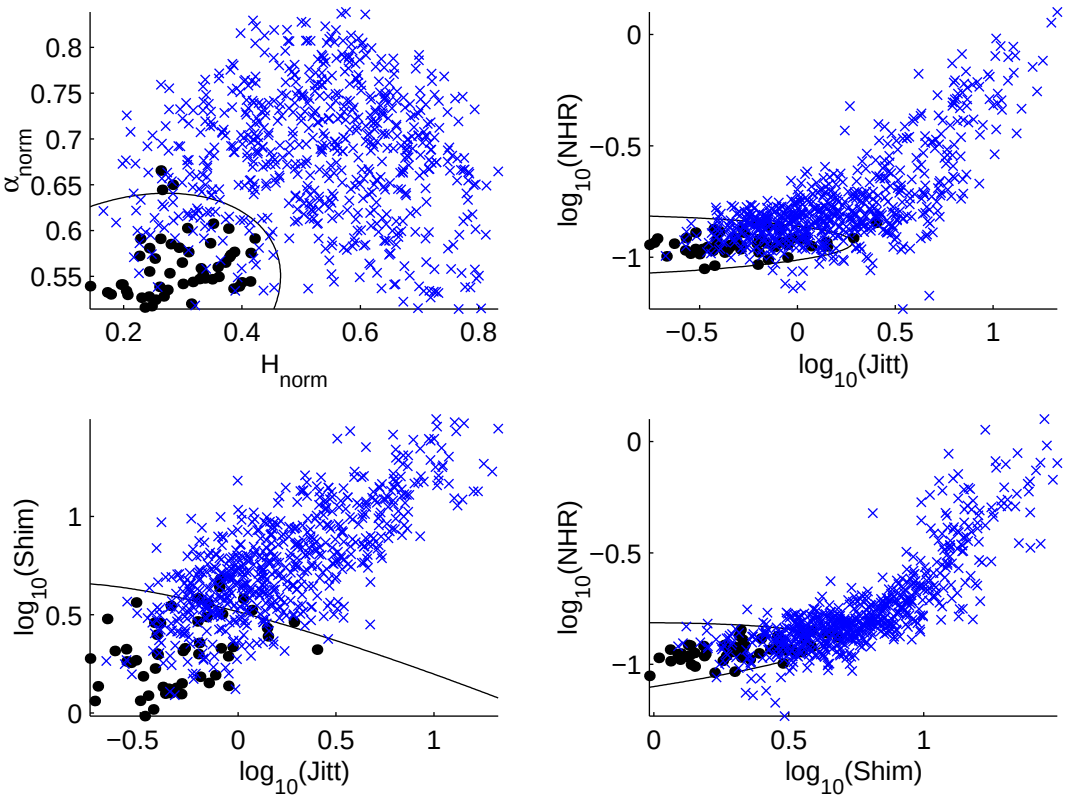

Figure 6.7: Hoarseness diagrams constructed using the new measures and traditional perturbation measures. (Top left) normalised RPDE and DFA measures, (top right) logarithms of NHR (Noise-to-Harmonics Ratio) and Jitter percentage, (bottom left) logarithms of Jitter and Shimmer percentages and (bottom right) logarithms of Shimmer and NHR perturbation measures. The blue crosses are the disordered subjects, the black dots the normal subjects. The black line is the average QDA classification boundary calculated over 1000 bootstrap resampling trials.

approach, the new nonlinear measures are more accurate on average than traditional measures. We will now discuss particular aspects of these results in comparison with traditional measures.

6.6.1 Feature Dimensionality

The *curse of dimensionality* afflicts all challenging data analysis problems [87]. In pattern analysis tasks such as automated normal/disordered separation, increasing the size of the feature vector (in this case, the number of measures I in the classifier vector $\mathbf{v}$) does not necessarily increase the performance of the classifier in general. This is because the volume of the *feature space* (the space spanned by the possible values of the measures) grows exponentially with the number of features. Therefore, the limited number of examples available to train the classifier occupy an increasingly small volume in the feature space, providing a poor representation of the mapping from features to classes that the classifier must learn [87]. For this study, combining only two measures obtains better performance than combining three traditional measures. Therefore the new measures help to mitigate this problem of dimensionality.

6.6.2 Feature Redundancy – Information Content

It is also important to use as few features as possible because in practice, increasing the number of features causes excessive data to be generated that may well contain redundant (repeated) information. The actual, useful information contained in these vectors has a much smaller dimensionality. For clinical purposes, it is important that only this useful data is presented. This effect of redundant information for the traditional measures can be clearly seen in figure 6.7, where combinations of pairs of (the logarithms of) measures are seen to cluster around a line or curve in the feature space, indicating positive correlation between these measures. Traditional measures create an effectively one-dimensional object in this two-dimensional space. This is not seen for the new measures which are spread evenly over the feature space.

6.6.3 Arbitrary Parameters – Reproducibility

Minimising the number of arbitrary parameters used to calculate these measures is necessary to avoid selecting an excessively specialised set of parameters that leads, for example, to good normal/disordered separation on a particular data set but does not generalise well

to new data.

Many parameters are required for the algorithms used in calculating traditional perturbation measures [90, 91, 93]. For example, the waveform matching algorithm [11] requires the definition of rough markers, upper and lower pitch period limits, low-pass filter cutoff frequencies, bandwidth and order selection parameters, and the number of pitch periods for averaging should these pitch period limits be exceeded [41]. Similarly, in just one of the noise measures (glottal-to-noise excitation ratio) used in [90], we can determine explicitly at least four parameters relating to linear prediction order, bandpass filter number, order, cutoff selection, and time lag range parameters. There are two additional parameters for the length and starting sample of the part of the signal selected for analysis.

Our new measures require only five arbitrary parameters that must be chosen in advance: the length of the speech signal N, the maximum recurrence time $T_{\max}$, and the lower value, upper value and increment of the DFA interval lengths L. We have also shown, using analytical results, that we can calibrate out the dependence upon the state space close recurrence radius r, the time-delay reconstruction dimension d and the reconstruction delay τ.

6.7 Interpretation of Results

We have found, in agreement with Titze [94] and Carding [89], that perturbation measures cannot be obtained for all the speech sounds produced by subjects (see table 6.1). This limits the clinical usefulness of these traditional measures. By contrast, the new measures presented in this chapter do not suffer from this limitation and are capable of measuring, by design, all types of speech signals.

Taking into account the number of these measures that need to be combined to create the hoarseness diagram and achieve effective normal/disordered separation, the number of arbitrary parameters used to calculate the measures, and the independence of these measures, traditional approaches are seen to be considerably more complex than the new measures developed in this chapter. The results of the classification comparison with traditional measures suggest that, in order to reach the classification performance of the new measures, we will either need much more complex classifiers, or need to combine many more classical features together [91, 92, 93]. It is therefore not clear that traditional approaches capture the *essential biomechanical differences* between normal and disordered voices in the most parsimonious way, and an excessively complicated relationship exists

therefore between the values of these measures and extent of the voice disorder.

6.8 Limitations of the New Measures

There are certain limitations to the new measures in clinical practice. These measures rely upon sustained vowel phonation, and sometimes subjects experience difficulty in producing such sounds, which limits the applicability. Also, at the beginning of a sustained vowel phonation, the voice of many subjects may require some time to settle into a more stable vibration. As such, discarding the beginning of the phonation is usually a prerequisite (but this does not adversely affect the applicability of these methods). Nonetheless, the extent of breathiness in speech is not usually affected in this way. In practice we require that the subject maintains a constant distance from the microphone when producing speech sounds; this can be achieved, for example, with the use of head-mounted microphones.

6.9 Possible Improvements and Extensions

There are several improvements that could be made to these measures. Firstly, every arbitrary parameter introduces extra variability that affects the reliability of the results. Much as it has been possible to calibrate out the dependence upon the RPDE parameters using analytical results, a theoretical study of the DFA interval lengths based upon typical sustained phonation recurrence periods could reveal values that would be found for all possible speech signals. These would be related to the sampling time Δt. The particular choice of normalisation function g for the scaling exponent might affect the classification performance, and better knowledge of the possible range of α values using theoretical studies of the DFA algorithm would be useful for this. It should also be possible to increase the recurrence time precision of the RPDE analysis by interpolating the state space orbits around the times of close recurrence n_0, n_1. It should then be possible to achieve the same high-resolution as waveform matching techniques [108], which would make RPDE competitive for the detailed analysis of Type I periodic sounds.

6.10 Chapter Summary

In this chapter, to test the usefulness of the new nonlinear, stochastic model of speech production developed earlier in the thesis, we have introduced two measures: the novel recurrence period density entropy and detrended fluctuation analysis, an existing technique,

to analyse and characterise normal and disordered voices. The results show that, when the assumptions of the new speech production model hold under experimental conditions (in that the speech examples are sustained vowels recorded under quiet acoustic conditions), we can construct a hoarseness diagram showing the extent of normality/disorder in a speech signal. The results also show that on average these two measures alone are capable of distinguishing normal from disordered voices of all types, with overall classification performance superior to traditional, LTI-based measures, when compared using a simple classifier.

This chapter demonstrates that the evidence of nonlinearity/non-Gaussianity in speech signals produced in earlier chapters, that motivated the introduction of a new signal model of speech production, is not only of pure scientific interest. Incorporating information from the biomechanics of speech production has practical consequences because it can inform the design of nonlinear/non-Gaussian signal analysis methods and screening systems that are better able characterise the wide variety of complex biomechanical changes arising from voice disease and disorder. This is because, ultimately, the underlying biomechanics are responsible for the widely varying phenomenology.

CHAPTER 7

Discussion and Conclusions

This thesis has addressed the central argument that nonlinear signal processing methods are valuable for digital speech analysis. In this final chapter, we will summarise briefly the results of the thesis and discuss critically the results in relation to comparable studies and the concepts presented in the introduction. We will then draw conclusions about the work, summarising the original contributions of the thesis and forming tentative generalisations to domains outside speech signal processing. Finally, we will discuss future directions that the results in this thesis suggest.

7.1 Thesis Summary

We will now briefly summarise the thread of the argument of the thesis. Linear signal processing methods based around LTI (Linear, Time-Invariant) systems theory have a substantial historical precedent in digital speech processing (see Chapter 3). The linear acoustic, source-filter theory of speech production (described in §2.2.3) provides ready biomechanical justification for the use of LTI techniques, since the vocal tract can be modelled as a passive, linear acoustic resonator (driven by the vocal fold oscillation during vowel production, and vortex sound generating mechanisms during consonants). Despite the successes of this linear model, the nonlinear, deterministic dynamics of the driving signal during vowel production (and the complex, nonlinear vibration characteristic of vocal fold pathologies) is incompatible with the assumptions underlying the tools of LPA (Linear Prediction Analysis) and PSD (Power Spectral Density) estimation, as shown in §3.3.

These theoretical considerations, combined with empirical evidence from digital speech signals (see Chapter 5), lead to the interpretation that healthy vowel sound production cannot be comfortably explained within the LTI framework, although (somewhat counter-intuitively to some speech scientists), the stochastic vortex sound generation mechanisms of consonant phonemes result in speech signals for which the LTI framework could not be

rejected. Pathological disordered vowels require more subtle analysis: the more "breathy" examples (where the vocal fold oscillation ceases altogether) can be described within LTI assumptions – the rest cannot. It was concluded (§5.3.3) that the most parsimonious model for speech production, that can explain all these findings in one unified framework, is a stochastic nonlinear, non-Gaussian model, which subsumes both the classical linear model and purely deterministic nonlinear models.

Therefore, for healthy vowel sounds, the use of nonlinear time series analysis methods based upon purely deterministic modelling assumptions (Chapter 4) was not ruled out (recent research using such techniques being reviewed in §7.2.1). Nonetheless, for healthy speech, nonlinear time series analysis techniques have yet to make a significant impact on speech technology, and one explanation for this was elaborated in §5.3.3. It was concluded that testing the practical value of nonlinear signal processing techniques, and new methods based upon the unified model proposed earlier, would require an application context in which the critical differences could be brought into sharp relief, but on a small scale. This motivated the choice of a case study in biomedical engineering (see Chapter 6), that of the clinical measurement of disordered voices.

For detecting voice disorders, the new RPDE (Recurrence Period Density Entropy) nonlinear signal processing method was devised in Chapter 6, which measures the uncertainty in the recurrence period of nonlinear, stochastic signals. This was then combined with the existing method of DFA (Detrended Fluctuation Analysis), which measures the fractal scaling properties of stochastic signals. A simple pattern classifier was able, using these two methods, to outperform all combinations of classical LTI methods for the detection of voice disorders on a large database of pathological and healthy vowel sounds. This demonstrated that such new nonlinear signal processing techniques, implemented in software, and based upon model choices informed by biomechanics, do indeed have practical value.

7.2 Discussion

7.2.1 Comparison with Similar Studies

Investigations of speech by nonlinear time series analysis methods have been conducted by a number of researchers, and such studies date back to the first half of the 1990s. Initial work focused on the measurement of invariant set dimension (see Chapter 4) from speech

signals by reconstruction [109, 110, 42, 111] (also see [112] and references therein). Following on from this work, attempts were made to reproduce speech signals using nonlinear predictors [113, 114] for speech coding and speech synthesis applications.

Many of these early results were drawn together in the study of Mann [22]. Focusing on applications to speech pitch modification and speech synthesis, the study introduced a novel technique for finding the particular instant of vocal fold closure. It then investigated the use of a data-driven *local linear predictor* in attempting to find a piecewise linear model of the system function $\mathbf{F}$ in an equation such as (4.3). It was found that such local linear models do not *generalise* adequately from the speech data such that any attempts to apply time-scale modifications for synthesis applications lacked stability. Local linear models were therefore not found to be parsimonious models for speech production. The study therefore moved on to investigate global (rather than piecewise local) models for $\mathbf{F}$, and found that stable speech could be reproduced, however only with the use of a *regularisation* technique.

The early studies mentioned above were motivated by the possibility that speech vowel sounds might be chaotic and so exhibit sensitive dependence upon initial conditions (see Chapter 4). Whilst some of the earlier studies claimed to find evidence for positive Lyapunov exponents and non-integer attractor dimensions, the later study of Banbrook [115] concluded otherwise. Simultaneously, the application of *local projective noise reduction* [8] to speech signals has been tested by researchers from the nonlinear physics community [116].

Recent work in the use of nonlinear predictors for nonlinear speech processing has incorporated innovations such as Bayesian training [117], and while these improvements have lead to stable synthesis of vowel sounds, they fail on consonants, due to the apparently stochastic nature of such signals. In order to circumvent this problem, the state-of-the-art involves the use of novel hybrid stochastic/deterministic predictors [118].

A different line of investigation was taken in the study of Fackrell [24]: that of *higher-order statistics*, and particularly the *bispectrum* (please see [119] for more details). Such techniques go beyond the restrictions of second-order moments and Gaussian processes implied by the use of classical linear signal processing of Chapter 3, and therefore can be used, as with nonlinear time series analysis techniques, to characterise nonlinearity in speech signals. The main conclusion of the study was that speech signals are, however, not usefully processed using such techniques.

Most nonlinear signal processing studies of speech have taken the approach that the

signal originates in a deterministic, nonlinear dynamical system. The recent studies of Maragos [23] have begun to explore the possibility of stochastic dynamics in speech, particularly due to turbulent airflow-induced sound sources in both vowels and consonants. By characterising the (multi-scale) dimension of the graph of the speech signal (see Chapter 4) using a nonlinear signal processing technique, they were able to improve the discrimination performance of a speech recognition system [23].

In summarising, these studies can be grouped into foundational work (surrogate data analysis), first-principles modelling, data-driven modelling (constructing nonlinear predictors), statistical (measuring Lyapunov exponents, attractor and graph dimensions) and practical work (testing data-driven modelling and statistical methods in speech applications). However, because all these aspects have been studied separately, the conclusions are somewhat fragmentary, and the results obtained in one study are not readily applicable to others. This thesis therefore builds on these previous studies by bringing together modelling, foundational, statistical and practical aspects in one place, and tying them together in one coherent framework. The use of each technique has been justified at each step, and the coverage of empirical speech sounds is the largest to date. However, this thesis does not address state-of-the-art nonlinear predictors because the results of other studies have suggested that the practical advantages of nonlinear signal processing techniques in general could not be reasonably tested on a small scale using such techniques.

7.2.2 Mathematical Models in Nonlinear Signal Processing

We now turn to a discussion of the wider implications of this work. The general approach of this thesis, as summarised in §7.1 above, can be tentatively generalised in the following way.

LTI, Gaussian signal processing methods are well understood and, for many signals, appropriate. However, ideally tests should be applied to these signals to determine whether they consist of dynamics that might better be analysed using more sophisticated tools, such as nonlinear time series analysis methods. Surrogate data methods offer one convenient test for these properties, but the limitations of the range of null hypotheses that can be tested using these methods require us to invoke information from relevant, first-principles models. Such information helps to narrow down the modelling possibilities. Assuming that stochastic non-Gaussianity can be justified by a combination of hypothesis tests and first-principles knowledge, then standard nonlinear time series analysis methods, and classical linear methods as well, are fundamentally unsuitable. At worst these methods can

produce meaningless analysis results. In these situations, appropriate new signal analysis tools should be preferred.

This methodology can therefore be applied to other signal processing problems where classical digital signal processing has previously been used. Crucially, we were able to produce more reliable and robust signal processing methods with higher performance on a specific analysis task by taking account of evidence from first-principles, mathematical models of the phenomenon under study, here the phenomenon of speech production. In this way, we could also produce more reliable and robust methods than standard nonlinear time series analysis techniques. Furthermore, since first-principles models are applicable to a vast range of phenomena from domains of technological importance, there is, in principle, no obvious reason why, with access to signals from these phenomena, we cannot apply a similar approach to problems in these domains as we have done in this study.

Nonetheless, there are certain limitations to the wholesale application of this methodology. Firstly, the design of techniques for the analysis of voice disorders needed to take into account the specific nature of the problem, that is, we were interested in detecting voice disorders, as opposed to performing speech compression, for example. These are different tasks for which entirely different techniques are appropriate. The nature of the problem we solved determined the choice and design of nonlinear signal processing methods. Therefore this thesis does not describe a "one-size-fits-all" approach that will work for every problem. Secondly, it is necessary to have some prior knowledge about the physical phenomena – this might limit us to certain situations where there is considerable information in addition to the availability of digital signals. An alternative approach that avoids these limitations, it could be argued, are *machine learning* techniques, where the goal is to turn the analysis and processing tasks over in their entirety to general software algorithms that can perform all the tasks automatically [87].

The counter-argument is that such techniques generally lack the critical property of *transparency*: when they work, it is not clear exactly *why* they work. This is of course only a problem, from an engineering point of view, when they fail. Nonetheless, due to the sheer number and diversity of different machine learning techniques that could be applied to any given problem, it is rarely clear from the start which of these methods will be successful, and often such techniques will fail to produce useful or meaningful results. In the face of failure, without knowledge of why a technique fails, it is difficult to know exactly what to do to remedy the situation. We argue that, by referring to specific physical knowledge about the problem, we can diagnose and iteratively improve our techniques because at

each stage we can trace the failure back to the underlying assumptions. For example, in this thesis, we could trace the failure of power spectral density estimation to distinguish between chaotic dynamics and stochastic noise to the assumption of linear superposition. Knowing from the biomechanics of the phenomena that vocal fold dynamics are nonlinear and do not obey the superposition principle then gave us immediate understanding that we required a new technique that did not embody this assumption. In this way, we used physical information to guide our mathematical choices.

7.3 Conclusions

We now return to the introduction and address the central argument of the thesis: that nonlinear approaches are valuable for digital speech signal analysis, barring important limitations. Taking each supporting hypotheses in turn, we will identify the particular place in the thesis that justifies the claim.

- **Based upon knowledge in speech science and evidence from speech signals themselves, the mathematical assumptions of LTI systems theory cannot represent all the dynamics of all speech.** We have shown in Chapter 2 that nonlinearity is an important feature of vocal fold dynamics. Similarly, Chapter 3 demonstrated the limitations of LTI systems-based digital signal processing methods for analysing nonlinear, chaotic dynamics. Chapter 5 showed that a large proportion of speech signals are unlikely to be parsimoniously represented by LTI systems models.

- **LTI systems theory is only appropriate for some limited cases of speech phonemes.** Chapter 5 showed that consonants and highly breathy disordered speech sounds may be most parsimoniously represented by LTI systems approaches.

- **Nonlinear, non-Gaussian stochastic assumptions are particularly important to some speech phonemes, and some disordered speech.** Again, Chapter 5 showed that normal and some disordered vowels sounds are unlikely to be best represented by LTI systems models.

- **Appropriate nonlinear signal processing methods are, in some aspects, better than LTI systems approaches in voice disorder detection.** Chapter 6 demonstrated that, using a simple classifier, appropriately chosen and designed

nonlinear/non-Gaussian signal processing methods (RPDE and DFA) were able to outperform classical LTI-systems based approaches in separating normal from disordered voices, although as yet these new methods are not as accurate in analysing highly periodic speech sounds.

- **Nonlinear, non-Gaussian assumptions for speech signals offer a simplified, mathematical framework that explains more phenomena with fewer assumptions than classical LTI assumptions, and as such can offer improvements in engineering reliability, robustness and performance.** In Chapter 5, it was discussed how a new signal model for speech production, incorporating both nonlinear and stochastic elements, was able to subsume both the Gaussian linear models appropriate for consonants and breathy disordered speech, and the deterministic nonlinear models suitable for vowel sounds. This new signal model required the design and choice of nonlinear signal processing methods (RPDE and DFA) (Chapter 6) with fewer arbitrary parameters (increased *reliability*) than classical LTI-based methods, and applicability to a wider range of speech signals (increased *robustness*). The use of these new methods lead to increased classification *performance* for disordered voice signals.

- **Not all the standard, nonlinear algorithms are robust enough to be of practical value to speech processing, so that new, nonlinear algorithms are required.** As discussed in Chapter 5, a substantial fraction of all speech signals (consonants and some disordered voice signals) could not be parsimoniously modelled with a deterministic, nonlinear dynamical system, the critical assumptions underlying many of the more popular nonlinear time series analysis methods (e.g. Lyapunov exponent measurement, attractor dimension estimation). Being forced to accept the inherently stochastic nature of speech signals, we required new nonlinear/non-Gaussian signal analysis methods (RPDE and DFA) to characterise all speech signals in one single approach.

These supporting arguments justify the claim that nonlinear (and non-Gaussian) signal analysis methods are valuable in speech processing. The limitation to which we refer is the qualification that linear signal processing methods still have value in certain restricted speech analysis applications, and that many nonlinear time series analysis techniques are not appropriate.

7.3.1 Summary of Contributions

We will now briefly summarise the contributions made to the state-of-the-art in the discipline of nonlinear digital signal processing:

- **The systematisation and improvement of a statistical surrogate data test for nonlinearity/non-Gaussianity in digital signals.** This is the subject of §4.7.
- **Application of this test to the largest database assembled to date, assessing the evidence for and against nonlinearity/non-Gaussianity in the predominant classes of speech phonemes and in disordered speech.** See Chapter 5.
- **The introduction and justification for a new, parsimonious, nonlinear/non-Gaussian model for speech signals.** This is the final part of Chapter 5.
- **The development of a novel method for characterising the nonlinear/non-Gaussian dynamics represented in a signal, and the case study application of this method to the automated detection of voice disorders.** This is addressed in Chapter 6.

7.3.2 Suggested Future Directions

Since the year 2000 and the instigation of the (now completed) Europe-wide collaborative research network COST277, nonlinear speech signal processing has gained momentum as an increasingly self-contained area of research. Indeed, the recent announcement of a new research network, COST2103, involves over 30 researchers from nine different European countries. The participants come with a diverse set of interests, including speech coding and synthesis (engineering) to clinical voice disorder assessment (clinical practice). Helping to confirm the results of this thesis, it is recognised by these initiatives that speech signal processing by nonlinear means has much to offer, and is therefore a growth area of scientific and practical interest. How might the results of this thesis influence future work in this area, therefore?

Given that nonlinear/non-Gaussian approaches are valuable, consensus amongst those who have used these new techniques on critical points of contention such as whether speech signals are Gaussian linear, chaotic, deterministic or fractal is yet to emerge. This lack of consensus hinders the adoption of these new techniques by the majority of speech

scientists and engineers. These practitioners do not readily see an inherent advantage over classical linear, time-invariant signal processing techniques with which they are familiar. There is therefore some resistance to the introduction of these new techniques, despite the advantages they offer.

This thesis suggests that this lack of consensus stems mostly from the conflicting mathematical assumptions that are adopted, unexamined, by practitioners. Where their assumptions differ, their conclusions will inevitably clash. We suggest in this thesis that a synthesis of the classical mathematical assumptions of linear digital signal processing with those of nonlinear time series analysis is not only more parsimonious with respect to the evidence (both empirical and theoretical) than either set of assumptions alone, but leads to the design of more reliable, robust and better-performing signal analysis methods for practical applications.

The results of this thesis which assesses the appropriateness and limitations of classical LTI and nonlinear time series analysis techniques for speech analysis, should help to serve as a cautionary example that classical digital signal processing of speech is not necessarily the best approach, but that neither is the uncritical use of nonlinear time series analysis methods. It is better to assess each analysis problem separately, and then to select or design appropriate techniques for that problem, taking into account the nature of the evidence.

One very promising extended study that could build on the results in this thesis is the application of the new speech analysis techniques (RPDE and DFA) to the detection of *Parkinson's disease*, a crippling neurological disorder [120]. The typical symptoms include physical tremors, muscular rigidity and postural abnormalities, but also increasingly disordered voice. The early, correct diagnosis of this disease can be critical to attempt to arrest the neurological degeneration by new neuroprotective and surgical techniques. However, there is currently no biological test that can be applied to correctly diagnose Parkinson's before the tremor symptoms become clearly noticeable and the prognosis poor. However, a recent study [120] found that, interestingly, the voices of Parkinson's sufferers shows changes indicative of very early signs of the disease, due to degeneration that affects the very fine motor control abilities required to articulate speech sounds and maintain a controlled vocal fold oscillation. It is quite possible then that the techniques developed in this thesis could indeed have value in the early diagnosis of this disease, and, in fact, be the *only* viable method for such early detection.

In wider, practical technological applications, speech signals cannot be considered to

satisfy the constraints of stationarity, to which this thesis has largely been confined. A possible future extension to this work would therefore attempt to relax the mathematical requirement of time-invariance, both for linear, deterministic nonlinear, and stochastic non-Gaussian models. This could lead to the design and use of non-stationary techniques that would be able to cope naturally with the ever-present articulatory dynamics of running speech.

The biomechanics introduced in Chapter 2 has discussed the main components of speech production and presented examples of vocal fold dynamics that are highly irregular. Evidence from disordered voice samples and other modelling studies show that vocal fold disorders tend to produce such irregular vibrations. Simulating vocal fold disorders could be valuable for a number of purposes, including testing new disordered voice analysis methods and therapeutic feedback in clinical settings. Early pilot studies by the author have shown that it is not necessary to produce vocal fold models of the full detail of those presented in Chapter 2 in order to reproduce quite convincing disordered voice sounds. The output of a simple nonlinear dynamical system that is capable of chaotic dynamics, after appropriate processing, is passed through a linear resonator in order to simulate the effect of the vocal tract and the radiative lip opening. This results in a simulated digital speech pressure signal.

The nonlinear signal processing methods introduced in Chapter 6 for speech analysis have been shown to enable the detection of voice disorders. However, early studies by the author of normal voices recorded under quiet acoustic conditions shows that not only can these measures distinguish normal from disordered voices, they are also capable of distinguishing one individual from another. This implies that the measures reflect, to a certain extent, the unique character of an individual's voice. This raises the possibility of an extension to this work that uses these measures for *biometric identification*: distinguishing one individual from another on the basis of their speech signal.

Outside the area of speech processing, early pilot work by the author has suggested that the combination of RPDE and DFA may be valuable for the detection of life-threatening cardiac abnormalities. This is because the heart, which can be modelled from first principles as a nonlinear dynamical system, in some severe pathological cases appears to fall into patterns of vibration that look very similar to chaos. Ventricular fibrillation (VF) is a classic example [121], and RPDE is designed to detect changes in the complexity of the vibration pattern (with regular, sinus rhythm at one extreme and irregular VF at the other). Furthermore, on a longer time scale, heart disease is often accompanied by

changes in heart rate that are detectable in the stochastic fractal scaling properties of the heart-rate time series [121]. Thus the combination of new nonlinear measures, which are designed to characterise both deterministic and stochastic nonlinear properties, might be valuable for this detection problem.

Finally, in Chapter 4 a novel calibration approach was taken to account for the inaccuracy in measuring mutual information. An extension to this study could look at other methods for estimating the probability density functions upon which the entropy calculations are based. Of particular interest here are kernel density estimation methods [87], since this can produce smoother density estimates than discrete histograms. In theory, mutual information estimation errors using this technique could therefore be smaller. Combining the proposed calibration method with kernel density estimation might lead to an improved method for entropy-based signal processing techniques such as Independent Components Analysis (ICA).

Appendix

A.1 Numerical Solution to Vocal Tract Tube Model

This section describes the implementation of the numerical solution to the varying cross-sectional area vocal tract model of Chapter 2, using finite differences. The full length L of the vocal tract model is divided into equal intervals of length $\Delta x = L/N$ where N is the number of discretisation intervals. The boundary value problem to be solved is:

$$U''(x,\omega) - \frac{A'(x)}{A(x)}U'(x,\omega) + \frac{\omega^2}{c^2}U(x,\omega) = 0, \tag{A.1}$$

$$U(0,\omega) = 1, \tag{A.2}$$

$$U'(x,\omega)\mid_{x=L} = \frac{\omega A(L)}{i\rho c^2}Z(\omega)U(L,\omega), \tag{A.3}$$

where the prime denotes differentiation with respect to x. Using forward differences, the above problem is discretised into the following implicit scheme:

$$\frac{u^\omega_{n+1} - 2u^\omega_n + u^\omega_{n-1}}{\Delta x^2} - \frac{A'(n\Delta x)}{A(n\Delta x)}\left(\frac{u^\omega_n - u^\omega_{n-1}}{\Delta x}\right) + \frac{\omega^2}{c^2}u^\omega_n = 0, \tag{A.4}$$

$$u^\omega_0 = 1, \tag{A.5}$$

$$\frac{u^\omega_N - u^\omega_{N-1}}{\Delta x} = \frac{\omega A(x)}{i\rho c^2}Z(\omega)u^\omega_{N-1}, \tag{A.6}$$

where u^ω_n denotes the acoustic flow rate at spatial position $n\Delta x$, at a given frequency ω, for $n = 1, 2 \ldots N-1$. The Struve function used in expression (2.9) is numerically integrated using the trapezoidal iteratively convergent Romberg method and the following identity:

$$\mathbf{H}_1(x) = \frac{2x}{\pi}\int_0^1 \sqrt{1-t^2}\sin(xt)dt. \tag{A.7}$$

The scheme is formulated as a matrix problem:

$$\mathbf{C}^\omega\mathbf{u}^\omega = \mathbf{D}^\omega, \tag{A.8}$$

with $\mathbf{C}^\omega$ an $N \times N$ matrix, $\mathbf{u}^\omega$ the acoustic flow rate solution row vector of size N, and $\mathbf{D}^\omega$ the right-hand row vector of size N. All the entries in $\mathbf{C}^\omega$ are zero apart from the

following, representing the boundary conditions:

$$\mathbf{C}^{\omega}_{0,0} = 1,\ \mathbf{C}^{\omega}_{N,N} = \frac{1}{\Delta x},\ \mathbf{C}^{\omega}_{N,N-1} = -\frac{1}{N} - \frac{\omega A(L)}{i\rho c^2} Z(\omega), \tag{A.9}$$

and the following entries:

$$\mathbf{C}^{\omega}_{n,n-1} = \frac{1}{\Delta x^2},\ \mathbf{C}^{\omega}_{n,n} = -\frac{2}{\Delta x^2} + \frac{A'(n\Delta x)}{A(n\Delta x)\Delta x} + \frac{\omega^2}{c^2},\ \mathbf{C}^{\omega}_{n,n+1} = \frac{1}{\Delta x} - \frac{A'(n\Delta x)}{A(n\Delta x)\Delta x}, \tag{A.10}$$

for $n = 1, 2 \ldots N-1$. All the entries of $\mathbf{D}^{\omega}$ are zero except $\mathbf{D}^{\omega}_0 = 1$. The matrix problem (A.8) is solved using Gaussian elimination with partial pivoting. Finally, the transfer function $H(\omega)$ evaluated at $\omega = 2\pi f$ is:

$$H(\omega) = u^{\omega}_N. \tag{A.11}$$

A.2 Miscellaneous Proofs

A.2.1 Linear Combinations of Gaussian Random Variables

The linear combination of any number of Gaussian random variables is also a Gaussian random variable. Let $u_n, v_n \sim \mathcal{N}(0, \sigma^2)$ be i.i.d. Gaussian random variables with density functions $f(x)$. The density function $g(z)$ for the linear sum $w_n = a_1 u_n + a_2 v_n$, where a_1, a_2 are arbitrary real constants is [54]:

$$\begin{aligned} g(z) &= \int_{-\infty}^{\infty} \frac{1}{a_1 a_2} f\left(\tfrac{x}{a_1}\right) f\left(\tfrac{z-x}{a_2}\right) dx && \text{(A.12)} \\ &= \frac{1}{2\pi\sigma^2 a_1 a_2} \int_{-\infty}^{\infty} \exp\left(-\tfrac{x^2}{2\sigma^2 a_1^2}\right) \exp\left(-\tfrac{(z-x)^2}{2\sigma^2 a_2^2}\right) dx && \text{(A.13)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2\left(a_1^2 + a_2^2\right)}} \exp\left(-\tfrac{z^2}{2\sigma^2\left(a_1^2+a_2^2\right)}\right) && \text{(A.14)} \end{aligned}$$

which is also Gaussian with zero mean and variance $\sigma^2(a_1^2 + a_2^2)$. This result can be extended by induction to the linear sum of any number of Gaussian random variables, so that the output of a Gaussian AR system is also Gaussian as required [54].

A.2.2 Autocorrelation of Gaussian i.i.d. Signals

We prove equation (3.30). The autocorrelation is related to the covariance[1] of the signal w_n by [12, 54]:

$$\begin{aligned} r_{ww}(l) &= \sum_{n=-\infty}^{\infty} w_n \overline{w}_{n-l} = \operatorname{cov}(w_n, w_{n-l}) && \text{(A.15)} \\ &= E\left[w_n \overline{w}_{n-l}\right] - E\left[w_n\right] E\left[\overline{w}_{n-l}\right]. && \text{(A.16)} \end{aligned}$$

[1] This holds for the normalised sum definition of autocorrelation.

Assume w_n to be a real-valued, zero mean, Gaussian i.i.d., strongly stationary discrete time stochastic process of variance σ^2. Then $E[w_n] = E[w_{n-l}] = 0$. Therefore:

$$r_{ww}(l) = E[w_n w_{n-l}]. \tag{A.17}$$

There are two different cases to consider. Firstly, for $l = 0$:

$$r_{ww}(0) = E[w_n w_n] = E\left[w_n^2\right] = \sigma^2, \tag{A.18}$$

since w_n is a real-valued signal. Secondly, for $l \neq 0$, w_n and w_{n-l} are independent. Therefore the joint density of w_n and w_{n-l} factorises [54]:

$$r_{ww}(l) = E[w_n w_{n-l}] = E[w_n] E[w_{n-l}] = 0. \tag{A.19}$$

Therefore, a compact way of writing the autocorrelation is $r_{ww}(l) = \sigma^2 \delta_l$ as required.

A.2.3 Wiener-Khintchine Theorem for Finite Length Signals

We prove equation (3.49) making use of the circular cross-correlation property of the DFT.

$$P_{xx}(k) = \mathcal{F}[\tilde{r}_{xx}] = X(k)\overline{X}(k) = |X(k)|^2 \tag{A.20}$$

A.2.4 IIR Filters and Forced Nonlinear Systems

The linear IIR filter system (3.9) can be written in the form of (4.1) by associating the elements of the vector $\mathbf{y}_n = [y_{1,n}, y_{2,n} \ldots y_{P,n}]^T$ with time-delayed copies of the univariate signal y_n. Let $y_{k,n} = y_{n-k+1}$ for $k = 1, 2 \ldots P$. Also, define the forcing vector as $\mathbf{x}_n = [x_n, 0, 0 \ldots 0]^T$. Then, defining the system function $\mathbf{F}$ as an appropriate matrix equation operating on the vector $\mathbf{y}_n$ and expanding out the equation (4.1) gives:

$$\begin{bmatrix} y_n \\ y_{n-1} \\ y_{n-2} \\ \vdots \\ y_{n-P+1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_P \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_{n-1} \\ y_{n-2} \\ y_{n-3} \\ \vdots \\ y_{n-P} \end{bmatrix} + \begin{bmatrix} x_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{A.21}$$

which is the system of (3.9).

A.2.5 TDMI for Gaussian Linear Signals

We prove that the equation (4.18) holds. Using the definition of mutual information we get:

$$I[s](\tau) = H[s_n] + H[s_{n-\tau}] - H[s_n, s_{n-\tau}], \tag{A.22}$$

so that, using the definitions of differential entropy for Gaussians:

$$\begin{aligned} H[s_n] + H[s_{n-\tau}] - H[s_n, s_{n-\tau}] &= \frac{1}{2}\ln(2\pi e\, r_{ss}(0)) + \frac{1}{2}\ln(2\pi e\, r_{ss}(0)) - \ln(2\pi e) - \frac{1}{2}\ln|\mathbf{C}| \\ &= \frac{1}{2}\left[\ln r_{ss}(0) + \ln r_{ss}(0) - \ln\left(r_{ss}(0)r_{ss}(0) - r_{ss}(\tau)r_{ss}(\tau)\right)\right] \\ &= \frac{1}{2}\ln\left[\frac{r_{ss}^2(0)}{r_{ss}^2(0) - r_{ss}^2(\tau)}\right], \end{aligned} \tag{A.23}$$

as required.

A.2.6 Periodic Recurrence Probability Density

We consider the purely deterministic case, i.e. when the model of equation (4.3) applies. Thus the measured time series is purely deterministic and points in the time series follow each other in an exactly prescribed sequence. When the measured, time-delay reconstructed orbit $\mathbf{s}_n$ is a purely periodic orbit of finite period K steps, there is an infinite sequence of points $\{\mathbf{r}_n\}, n \in \mathbf{Z}$ in the reconstructed state space with $\mathbf{r}_n = \mathbf{r}_{n+K}$, and $\mathbf{r}_n \neq \mathbf{r}_{n+j}$ for $0 < j < K$.

Picking any point $\mathbf{s}$ in the reconstructed state space, there are two cases to consider. In the first case, if $\mathbf{s} = \mathbf{r}_n$ for some n, then $\mathbf{s}$ is not the same as any other points in the periodic orbit except for $\mathbf{r}_{n+K}$, so that the orbit returns with certainty for the first time to this point after K time steps. This certainty, with the requirement that the probability of first recurrence is normalised for $T = 1, 2\ldots$ implies that:

$$P_{\mathbf{s}}(T = r) = \begin{cases} 1 & \text{if } r = K \\ 0 & \text{otherwise} \end{cases}. \tag{A.24}$$

In the second case when $\mathbf{s} \neq \mathbf{r}_n$ for any n, the orbit never intersects the point so that there are also never any first returns to this point. All the points in the reconstructed space form a disjoint partition of the whole space. Thus the probability of recurrence to the whole space is the sum of the probability of recurrence to each point in the space separately, appropriately weighted to satisfy the requirement that the probability of first recurrence to the whole space is normalised However, only the K distinct points of the periodic orbit contribute to the total probability of first recurrence to the whole space. Therefore, the probability of first recurrence is:

$$P(T) = \frac{1}{K}\sum_{i=0}^{K-1} P_{\mathbf{r}_i}(T = r) = \begin{cases} 1 & \text{if } r = K \\ 0 & \text{otherwise} \end{cases}. \tag{A.25}$$

A.2.7 Uniform i.i.d. Stochastic Recurrence Probability Density

Consider the purely stochastic case when the nonlinear term **F** in equation (4.1) is zero and the stochastic forcing term is a uniform, i.i.d. random vector. Then the time-delay reconstructed orbit $\mathbf{s}_n$ is also a stochastic, uniform i.i.d. random vector. Since all the time series are normalised to the range $[-1, 1]$ then each member of the measurement takes on a value from this range. Then the orbits $\mathbf{s}_n$ occupy the reconstructed state space which is the region $[-1, 1]^d$, and each co-ordinate s_n is i.i.d. uniform. We form an equal-sized partition of this space into N^d (hyper)-cubes, denoting each cubical region R. The length of the side of each cube R is $\Delta s = 2/N$. Then the probability of finding the orbit in this cube is $P_R = \Delta s^d/2^d$. Since the co-ordinates s_n are uniform i.i.d., then the probability of first recurrence of time T to this region R is geometric [66]:

$$P_R(T) = P_R[1 - P_R]^{T-1} = \frac{\Delta s^d}{2^d}\left[1 - \frac{\Delta s^d}{2^d}\right]^{T-1}. \tag{A.26}$$

This is properly normalised for $T = 1, 2\ldots$. However, we require the probability of first recurrence to all possible cubes. The cubes are a disjoint partition of the total reconstruction space $[-1, 1]^d$. Thus the probability of recurrence to the whole space is the sum of the probability of recurrence to each cube separately, appropriately weighted to satisfy the requirement that the probability of recurrence to the whole space is normalised. Since the probability of first recurrence to each cube R, $P_R(T)$ is the same, the probability of recurrence to all cubes is:

$$P(T) = \sum_{i=1}^{N^d} \frac{\Delta s^d}{2^d} P_R(T) = N^d \frac{\Delta s^d}{2^d} P_R(T) \tag{A.27}$$

$$= \frac{2^d}{\Delta s^d}\frac{\Delta s^d}{2^d} P_R[1 - P_R]^{T-1} = \frac{\Delta s^d}{2^d}\left[1 - \frac{\Delta s^d}{2^d}\right]^{T-1}. \tag{A.28}$$

For small cube side lengths Δs and close returns algorithm radius r, the first recurrence probability determined by the close returns algorithm is then:

$$P(T) = \frac{\Delta s^d}{2^d}\left[1 - \frac{\Delta s^d}{2^d}\right]^{T-1} \approx \frac{r^d}{2^d}\left[1 - \frac{r^d}{2^d}\right]^{T-1}. \tag{A.29}$$

Similarly, for small close returns radius r and/or for large reconstruction dimensions d, $1 - r^d/2^d \approx 1$ so that:

$$P(T) \approx \frac{r^d}{2^d}. \tag{A.30}$$

Note that for fixed d and r this expression is constant. Since the close returns algorithm can only measure recurrence periods over a limited range $1 \leq T \leq T_{\max}$, and we normalise

the recurrence histogram $R(T)$ over this range of T, then the probability of first recurrence is the uniform density:

$$P(T) \approx \frac{1}{T_{\max}}, \tag{A.31}$$

which is proportional to the expression $r^d/2^d$ above. Thus, up to a scale factor, the uniform i.i.d. stochastic recurrence probability density is itself uniform.

A.3 Derivation of Corrected TDMI Estimator

The probability densities $p_\tau(u_i, v_j)$ and $p_0(u_i)$ required to calculate the TDMI expression (4.16) are estimated by first binning the signal s_n into equal-width intervals $\Delta v = \Delta u = (\max(s_n) - \min(s_n))/(Q-1)$ where Q is the number of intervals needed to cover the full range of the signal. These bin counts are then normalised by the number of samples N used to estimate the density (note that in the case of p_τ this is $N - \tau$), and normalised by the length (in the case of p_0) and area (for p_τ) of each bin. Let $u_i = \min(s_n) + i\Delta u$ and $v_j = \min(s_n) + j\Delta v = \min(s_n) + j\Delta u$. Then the (uncorrected) estimator for the TDMI is:

$$I_{\mathrm{E}}[s](\tau) = 2H_{\mathrm{N}}(0) - H_{\mathrm{N}}(\tau), \tag{A.32}$$

using the strong stationarity property of the signal s_n. The above entropy expressions are estimated using the two-point trapezoidal rule:

$$H_{\mathrm{N}}(0) = -\frac{1}{2}\Delta u \sum_{i=0}^{Q-2} \left[q_i + q_{i+1}\right], \tag{A.33}$$

where $q_i = p_0(u_i) \ln p_0(u_i)$ and

$$H_{\mathrm{N}}(\tau) = -\frac{1}{4}\Delta u^2 \sum_{i=0}^{Q-2}\sum_{j=0}^{Q-2} \left[q_{i,j} + q_{i+1,j} + q_{i,j+1} + q_{i+1,j+1}\right], \tag{A.34}$$

where again the shorthand $q_{i,j} = p_\tau(u_i, v_j) \ln p_\tau(u_i, v_j)$ has been used.

The bias introduced by finite length signals, binned probability density estimation and numerical integration is dominated by additive errors, and can therefore be substantially corrected using calibration against the TDMI of the known special case of the zero mean, i.i.d., strongly stationary Gaussian stochastic process w_n, shown in equation (4.19). Because w_n is i.i.d. and strongly stationary, as shown in §A.2.5 for all lags not equal to zero, the TDMI is zero. Therefore, the dependence of any additive error on the parameters Q, N and τ can be explored using w_n as a test signal, see figure A.1.

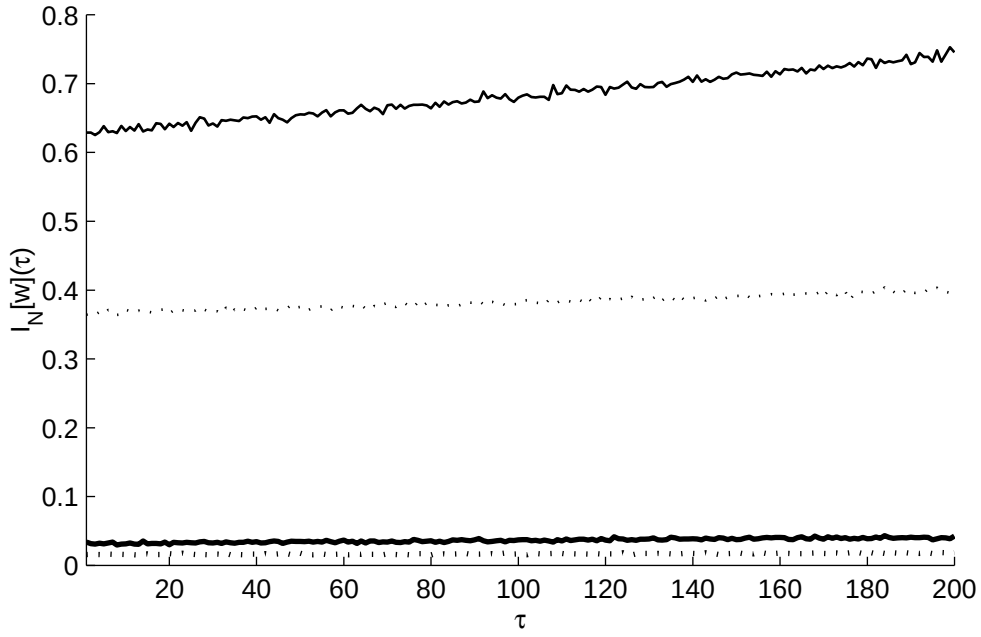

Figure A.1: Parametric dependence of TDMI statistic $I_{\mathrm{N}}[w](\tau)$. Parameters are N: the length of the zero mean, unit variance, Gaussian, strongly stationary i.i.d. signal w_n, the number of binning intervals Q use to estimate the required probability densities, and the time lag τ. From the top down: $Q, N = 50, 1000$, $Q, N = 50, 2000$, $Q, N = 10, 1000$ and $Q, N = 10, 2000$. TDMI shown are averages over 20 realisations of w_n for each set of parameters.

As can be seen from this figure, the error increases approximately linearly with the time lag τ, with a slope and intercept that depends upon Q and N. Therefore for each set of Q, N we fit a straight-line model $I_{\mathrm{adj}}(\tau) = a\tau + b$ to the mean of this error $I_{\mathrm{N}}[w](\tau)$ over 20 realisations of w_n. The parameters a, b were estimated using least-squares. Subsequently, the corrected TDMI estimation is:

$$I_{\mathrm{N}}[s](\tau) = I_{\mathrm{E}}[s](\tau) - I_{\mathrm{adj}}(\tau). \tag{A.35}$$

Glossary

General Mathematical and Engineering Terms

AAFT	Amplitude-Adjusted Fourier Transform surrogate generation
ADC	Analogue-to-Digital Convertor
AR	Auto-Regressive model
CELP	Code-Excited Linear Prediction
DAC	Digital-to-Analogue Convertor
DFA	Detrended Fluctuation Analysis
DFT	Discrete Fourier Transform
FIR	Finite Impulse Response filter
FFT	Fast Fourier Transform
IAAFT	Iterative Amplitude-Adjusted Fourier Transform surrogate generation
IDFT	Inverse Discrete Fourier Transform
i.i.d.	independent, identically distributed
IIR	Infinite Impulse Response filter
LPA	Linear Prediction Analysis
LTI	Linear, Time-Invariant
NHR/HNR	Noise-to-Harmonics (Harmonics-to-Noise) ratio
PCA	Principle Components Analysis
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
ROC	Region of Convergence
RPDE	Recurrence Period Density Entropy
TDMI	Time-Delayed Mutual Information

Mathematical Notation

Generally, each mathematical symbol in this thesis has a meaning unique to each chapter. An important subset of these symbols though have a meaning that spans one or more chapters; these are listed below.

t	Continuous time in seconds
Δt	Signal sampling interval, time discretisation interval
n	Discrete time index
$\tau, l, \Delta n, T$	Discrete time delay, time lag and recurrence time
f	Frequency in Hertz
ω	Angular frequency in radians per second
z	Complex variable
$p(t), p(x,t), p_n$	Acoustic pressure functions
$u(t), u(x,t)$	Acoustic flow rate functions
$H(\omega), H(z)$	Linear system transfer functions
c	Speed of sound in air
ρ	Constant equilibrium air pressure
$s(t)$	Continuous time signal
s_n	Discrete time signal and projected state space signal variable
$x_n, y_n, \mathbf{x}_n, \mathbf{y}_n$	Discrete time system input and system output signals
u_n	Discrete time surrogate signal
e_n	Discrete time error signal
w_n	Discrete time stochastic signal
$X(k)$	DFT of signal x_n at frequency index k
$P_{xx}(k)$	Power spectrum of signal x_n at frequency index k
$r_{xx}(l)$	Autocorrelation of signal x_n at time lag l
$a_k, \mathbf{a}$	Parameters of AR system and nonlinear systems models
$\sigma^2, \mathbf{C}$	Gaussian i.i.d. uni- and multi-variate random variable (co)-variance
P	AR system model order and system state space dimension
Q	System parameter vector size
$\mathbf{F}(\mathbf{y}_n, \mathbf{a})$	System function
L, N, M	Discrete time signal and interval length
$P(X), p(x), P(X,Y), P(X\|Y)$	Single, joint and conditional probability density functions
H	Entropy
I	Mutual information
$F(L)$	DFA fluctuation size with interval length
α	DFA scaling exponent
$P(T)$	Recurrence time discrete probability density
S	Significance probability
H_0, H_1	Null and alternative hypotheses
$B(\mathbf{y}, r)$	Closed ball of radius r around point $\mathbf{y}$

Bibliography

[1] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[2] K. R. Popper. *Conjectures and refutations: the growth of scientific knowledge*. Routledge, London; New York, 2002.

[3] J. A. Simpson, E. S. C. Weiner, and Oxford University Press. *The Oxford English dictionary*. Oxford University Press, Oxford; New York, 2nd edition, 1989.

[4] A. C. Fowler. *Mathematical models in the applied sciences*. Cambridge University Press, Cambridge; New York, 1997.

[5] J. R. Ockendon. *Applied partial differential equations*. Oxford University Press, Oxford; New York, 2003.

[6] S. Howison. *Practical applied mathematics: modelling, analysis, approximation*. Cambridge University Press, New York, 2005.

[7] N. D. Fowkes and J. J. Mahony. *An introduction to mathematical modelling*. Wiley, Chichester; New York, 1994.

[8] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, Cambridge; New York, 2nd edition, 2004.

[9] L.A. Smith. Disentangling uncertainty and error: On the predictability of nonlinear systems. In A. I. Mees, editor, *Nonlinear dynamics and statistics*, pages 31–64. Birkhuser, Boston, 2001.

[10] P. Ladefoged. *A course in phonetics*. Harcourt College Publishers, Fort Worth, 4th edition, 2001.

[11] R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Singular Thomson Learning, San Diego, 2nd edition, 2000.

[12] J. G. Proakis and D. G. Manolakis. *Digital signal processing: principles, algorithms, and applications*. Prentice Hall, Upper Saddle River, N.J., 3rd edition, 1996.

[13] J. D. Markel and A. H. Gray. *Linear prediction of speech*. Springer-Verlag, Berlin; New York, 1976.

[14] K. Johnson. *Acoustic and auditory phonetics*. Blackwell Pub., Malden, Mass., 2nd edition, 2003.

[15] T. F. Quatieri. *Discrete-time speech signal processing: principles and practice*. Prentice Hall, Upper Saddle River, NJ, 2002.

[16] M. B. Priestley. *Spectral analysis and time series*. Academic Press, London; New York, 1981.

[17] T. M. Cover and J. A. Thomas. *Elements of information theory*. J. Wiley, Hoboken, N.J., 2nd edition, 2005.

[18] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6):1098–1109, 2005.

[19] P. Maragos, A. Dimakis, and I. Kokkinos. Some advances in nonlinear speech modeling using modulations, fractals, and chaos. In *Proceedings of the 14th International Conference on Digital Signal Processing, DSP 2002*, volume 1, pages 325–332, 2002.

[20] I. Tokuda, T. Miyano, and K. Aihara. Surrogate analysis for detecting nonlinear dynamics in normal vowels. *Journal of the Acoustical Society of America*, 110(6):3207–17, 2001.

[21] T. Miyano, A. Nagami, I. Tokuda, and K. Aihara. Detecting nonlinear determinism in voiced sounds of Japanese vowel /a/. *International Journal of Bifurcation and Chaos*, 10(8):1973–1979, 2000.

[22] I. N. Mann. *An investigation of nonlinear speech synthesis and pitch modification techniques*. PhD thesis, Edinburgh University, 1999.

[23] P. Maragos and A. Potamianos. Fractal dimensions of speech sounds: computation and application to automatic speech recognition. *Journal of the Acoustical Society of America*, 105(3):1925–32, 1999.

[24] J.W.A. Fackrell. *Bispectral analysis of speech signals*. PhD thesis, Edinburgh University, 1996.

[25] B. H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustic Science and Technology*, 23(4):195–206, 2002.

[26] W. von Kempelen. *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*. F. Frommann, Stuttgart-Bad Cannstatt, 1970.

[27] L. E. Kinsler and A. R. Frey. *Fundamentals of acoustics*. Wiley, New York, 2d edition, 1962.

[28] P. M. Morse and K. U. Ingard. *Theoretical acoustics*. Princeton University Press, Princeton, N.J., 1986.

[29] R. M. Aarts and A. J. E. M. Janssen. Approximation of the Struve function H-1 occurring in impedance calculations. *Journal of the Acoustical Society of America*, 113(5):2635–2637, 2003.

[30] A. R. Greenwood, C. C. Goodyear, and P. A. Martin. Measurements of vocal-tract shapes using magnetic-resonance-imaging. *IEE Proceedings-I Communications Speech and Vision*, 139(6):553–560, 1992.

[31] J. L. Flanagan. *Speech analysis, synthesis and perception*. Springer-Verlag, Berlin; New York, 2nd edition, 1972.

[32] M. D. LaMar, Y. Y. Qi, and J. Xin. Modeling vocal fold motion with a hydrodynamic semicontinuum model. *Journal of the Acoustical Society of America*, 114(1):455–464, 2003.

[33] M. P. de Vries, H. K. Schutte, A. E. P. Veldman, and G. J. Verkerke. Glottal flow through a two-mass model: Comparison of Navier-Stokes solutions with simplified models. *Journal of the Acoustical Society of America*, 111(4):1847–1853, 2002.

[34] I. R. Titze. Human vocal cords - mathematical-model .1. *Phonetica*, 28(3-4):129–170, 1973.

[35] D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer. Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *Journal of the Acoustical Society of America*, 95(6):3595–3604, 1994.

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

[37] H. Herzel, D. Berry, I. Titze, and I. Steinecke. Nonlinear dynamics of the voice - signal analysis and biomechanical modeling. *Chaos*, 5(1):30–34, 1995.

[38] I. Steinecke and H. Herzel. Bifurcations in an asymmetric vocal-fold model. *Journal of the Acoustical Society of America*, 97(3):1874–1884, 1995.

[39] J. J. Jiang, Y. Zhang, and J. Stern. Modeling of chaotic vibrations in symmetric vocal folds. *Journal of the Acoustical Society of America*, 110(4):2120–2128, 2001.

[40] D. W. Jordan and P. Smith. *Nonlinear ordinary differential equations; an introduction to dynamical systems*. Oxford University Press, Oxford; New York, 3rd edition, 1999.

[41] I. Titze, R. Baken, and H. Herzel. Vocal fold physiology: frontiers in basic science. In Ingo R. Titze, editor, *Vocal fold physiology series*, pages 143–188. Singular Pub. Group, San Diego, Calif., 1993.

[42] H. Herzel, D. Berry, I. R. Titze, and M. Saleh. Analysis of vocal disorders with methods from nonlinear dynamics. *Journal of Speech and Hearing Research*, 37(5):1008–1019, 1994.

[43] M. di Bernardo, C. J. Budd, and A. R. Champneys. Normal form maps for grazing bifurcations in n-dimensional piecewise-smooth dynamical systems. *Physica D*, 160(3-4):222–254, 2001.

[44] R. W. Chan. Constitutive characterization of vocal fold viscoelasticity based on a modified Arruda-Boyce eight-chain model. *Journal of the Acoustical Society of America*, 114(4):2458, 2003.

[45] S. McLaughlin and P. Maragos. Nonlinear methods for speech analysis and synthesis. In S. Marshall and G. Sicuranza, editors, *Advances in nonlinear signal and image processing*, EURASIP Book Series on Signal Processing and Communications. Hindawi, 2006.

[46] D. J. Acheson. *Elementary fluid dynamics*. Oxford University Press, Oxford; New York, 1990.

[47] K. J. Falconer. *Fractal geometry: mathematical foundations and applications*. Wiley, Chichester; New York, 1990.

[48] M. S. Howe. *Theory of vortex sound*. Cambridge University Press, New York, 2003.

[49] D.J. Sinder. *Synthesis of unvoiced speech sounds using an aeroacoustic source model*. PhD thesis, Rutgers University, 1999.

[50] G. Richard, M. Liu, D. Sinder, H. Duncan, Q. Lin, J. Flanagan, S. Levinson, D. Davis, and S. Simon. Vocal tract simulations based on fluid dynamic analysis. *Journal of the Acoustical Society of America*, 97(5):3245–3245, 1995.

[51] W. Zhao, C. Zhang, S. H. Frankel, and L. Mongeau. Computational aeroacoustics of phonation, part I: Computational methods and sound generation mechanisms. *Journal of the Acoustical Society of America*, 112(5 Pt 1):2134–46, 2002.

[52] M. H. Krane. Aeroacoustic production of low-frequency unvoiced speech sounds. *Journal of the Acoustical Society of America*, 118(1):410–427, 2005.

[53] R. S. McGowan. An aeroacoustic approach to phonation. *Journal of the Acoustical Society of America*, 83(2):696–704, 1988.

[54] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, Oxford; New York, 3rd edition, 2001.

[55] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

[56] P. Kroon and W. Kleijn. Linear-prediction based analysis-by-synthesis coding. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 79–119. Elsevier, Amsterdam; New York, 1995.

[57] D. J. DeFatta, J. G. Lucas, and W. S. Hodgkiss. *Digital signal processing: a system design approach*. Wiley, New York, 1988.

[58] W. Kleijn and K. Paliwal. An introduction to speech coding. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 1–47. Elsevier, Amsterdam; New York, 1995.

[59] R.V. Cox. Speech coding standards. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 49–78. Elsevier, Amsterdam; New York, 1995.

[60] W. Fisher, G. Doddington, and K. Goudie-Marshall. The DARPA speech recognition research database: Specifications and status. In *Proceedings of the DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

[61] B. Yegnanarayana and R. N. J. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6(4):313–327, 1998.

[62] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London; New York, 1993.

[63] L. Arnold. *Random dynamical systems*. Springer, Berlin ; New York, corr. 2nd print. edition, 2003.

[64] J. Guckenheimer and P. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. Springer, New York, corr. 5th edition, 1997.

[65] M. Kac, K. Baclawski, and M. D. Donsker. *Mark Kac: probability, number theory, and statistical physics: selected papers*. MIT Press, Cambridge, Mass., 1979.

[66] E. G. Altmann and H. Kantz. Recurrence time analysis, long-term correlations, and extreme events. *Physical Review E*, 71(5):–, 2005.

[67] V. Balakrishnan, G. Nicolis, and C. Nicolis. Recurrence time statistics in deterministic and stochastic dynamical systems in continuous time: A comparison. *Physical Review E*, 61(3):2490–2499, 2000.

[68] M. C. Casdagli. Recurrence plots revisited. *Physica D*, 108(1-2):12–44, 1997.

[69] J. Stark, D. S. Broomhead, M. E. Davies, and J. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis-Theory Methods and Applications*, 30(8):5303–5314, 1997.

[70] J. Stark, D. S. Broomhead, M. E. Davies, and J. Huke. Delay embeddings for forced systems. II. Stochastic forcing. *Journal of Nonlinear Science*, 13(6):519–577, 2003.

[71] M. S. Roulston. Estimating the errors on measured entropy and mutual information. *Physica D*, 125(3-4):285–294, 1999.

[72] T. Schurmann. Bias analysis in entropy estimation. *Journal of Physics A-Mathematical and General*, 37(27):L295–L301, 2004.

[73] W. H. Press. *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, Cambridge; New York, 2nd edition, 1992.

[74] M. Little, P. McSharry, I. Moroz, and S. Roberts. Testing the assumptions of linear prediction analysis in normal vowels. *Journal of the Acoustical Society of America*, 119(1):549–558, 2006.

[75] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142(3-4):346–382, 2000.

[76] D. Kugiumtzis. On the reliability of the surrogate data test for nonlinearity in the analysis of noisy time series. *International Journal of Bifurcation and Chaos*, 11(7):1881–1896, 2001.

[77] P. E. McSharry, L. A. Smith, and L. Tarassenko. Prediction of epileptic seizures: are nonlinear methods relevant? *Nature Medicine*, 9(3):241–242, 2003.

[78] M. Small, D. J. Yu, and R. G. Harrison. Surrogate test for pseudoperiodic time series. *Physical Review Letters*, 8718(18):–, 2001.

[79] M. Barahona and C. S. Poon. Detection of nonlinear dynamics in short, noisy time series. *Nature*, 381(6579):215–217, 1996.

[80] M. Palus. Testing for nonlinearity using redundancies - quantitative and qualitative aspects. *Physica D*, 80(1-2):186–205, 1995.

[81] T. Nakamura, X. D. Luo, and M. Small. Testing for nonlinearity in time series without the Fourier transform. *Physical Review E*, 72(5):–, 2005.

[82] D. Kugiumtzis. Test your surrogate data before you test for nonlinearity. *Physical Review E*, 60(3):2808–2816, 1999.

[83] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC, Boca Raton, 3rd edition, 2004.

[84] J. Theiler. On the evidence for low-dimensional chaos in an epileptic electroencephalogram. *Physics Letters A*, 196(5-6):335–341, 1995.

[85] KayPENTAX. Kay elemetrics disordered voice database, model 4337, 1996-2005.

[86] G. Kubin. On the nonlinearity of linear prediction. In *IXth European Signal Processing Conference EUSIPCO'98*, Rhodes, Greece, 1998.

[87] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford; New York, 1995.

[88] P. M. B. Vitanyi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464, 2000.

[89] P. N. Carding, I. N. Steen, A. Webb, K. Mackenzie, I. J. Deary, and J. A. Wilson. The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical Otolaryngology*, 29(5):538–544, 2004.

[90] D. Michaelis, M. Frohlich, and H. W. Strube. Selection and combination of acoustic features for the description of pathologic voices. *Journal of the Acoustical Society of America*, 103(3):1628–1639, 1998.

[91] B. Boyanov and S. Hadjitodorov. Acoustic analysis of pathological voices. *IEEE Engineering in Medicine and Biology Magazine*, 16(4):74–82, 1997.

[92] J. I. Godino-Llorente and P. Gomez-Vilda. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384, 2004.

[93] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer. Automatic detection of pathologies in the voice by HOS based parameters. *EURASIP Journal on Applied Signal Processing*, 4:275–284, 2001.

[94] I. R. Titze. Workshop on acoustic voice analysis: Summary statement. NVCS Report Series, National Center for Voice and Speech, Iowa, USA, 1995.

[95] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. *Journal of Voice*, 19(4):519–528, 2005.

[96] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J. J. Jiang. Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps. *Journal of the Acoustical Society of America*, 115(5):2270–2277, 2004.

[97] A. Giovanni, M. Ouaknine, and J. L. Triglia. Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis. *Journal of Voice*, 13(3):341–354, 1999.

[98] Y. Zhang, J. J. Jiang, S. M. Wallace, and L. Zhou. Comparison of nonlinear dynamic methods and perturbation methods for voice analysis. *Journal of the Acoustical Society of America*, 118(4):2551–2560, 2005.

[99] A. Behrman and R. J. Baken. Correlation dimension of electroglottographic data from healthy and pathologic subjects. *Journal of the Acoustical Society of America*, 102(4):2371–2379, 1997.

[100] I. Hertrich, W. Lutzenberger, S. Spieker, and H. Ackermann. Fractal dimension of sustained vowel productions in neurological dysphonias: An acoustic and electroglottographic analysis. *Journal of the Acoustical Society of America*, 102(1):652–654, 1997.

[101] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. *IEEE Transactions on Biomedical Engineering*, 45(3):300–313, 1998.

[102] M. Little, P. McSharry, I. Moroz, and S. Roberts. Nonlinear, biophysically-informed speech pathology detection. In *2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006. ICASSP-2006.*, volume 2, pages II–1080–II–1083, Toulouse, France, 2006. IEEE Press.

[103] P. E. McSharry, L. A. Smith, and L. Tarassenko. Prediction of epileptic seizures: are nonlinear methods relevant? *Nature Medicine*, 9(3):241–2, 2003.

[104] D. P. Lathrop and E. J. Kostelich. Characterization of an experimental strange attractor by periodic-orbits. *Physical Review A*, 40(7):4028–4031, 1989.

[105] S. Plaszczynski. Fast 1/f alpha noise generation. *ArXiv*, pages astro–ph/0510081, 2005.

[106] C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time-series. *Chaos*, 5(1):82–87, 1995.

[107] Z. Chen, P. C. Ivanov, K. Hu, and H. E. Stanley. Effect of nonstationarities on detrended fluctuation analysis. *Physical Review E*, 65(4):–, 2002.

[108] I. R. Titze and H. X. Liang. Comparison of F(O) extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research*, 36(6):1120–1133, 1993.

[109] N. Tishby. A dynamical systems approach to speech processing. In *1990 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990. ICASSP-1990.*, volume 1, pages 365–368, 1990.

[110] A. Kumar and S. K. Mullick. Attractor dimension, entropy and modeling of speech time-series. *Electronics Letters*, 26(21):1790–1792, 1990.

[111] I. Tokuda, R. Tokunaga, and K. Aihara. A simple geometrical structure underlying speech signals of the Japanese vowel a. *International Journal of Bifurcation and Chaos*, 6(1):149–160, 1996.

[112] G. Kubin. Nonlinear processing of speech. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 557–610. Elsevier, Amsterdam; New York, 1995.

[113] G. Kubin and W. B. Kleijn. Time-scale modification of speech based on a nonlinear oscillator model. In *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.*, volume 1, pages I/453–I/456, 1994.

[114] G. Kubin. Synthesis and coding of continuous speech with the nonlinear oscillator model. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96.*, volume 1, pages 267–270, 1996.

[115] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Transactions on Speech and Audio Processing*, 7(1):1–17, 1999.

[116] R. Hegger, H. Kantz, and L. Matassini. Denoising human speech signals using chaoslike features. *Physical Review Letters*, 84(14):3197–3200, 2000.

[117] E. Rank. Application of Bayesian trained RBF networks to nonlinear time-series modeling. *Signal Processing*, 83(7):1393–1410, 2003.

[118] E. Rank and G. Kubin. An oscillator-plus-noise model for speech synthesis. *Speech Communication*, 48(7):775–801, 2006.

[119] C. L. Nikias and A. P. Petropulu. *Higher-order spectra analysis: a nonlinear signal processing framework*. Prentice Hall, Englewood Cliffs, N.J., 1993.

[120] B. Harel, M. Cannizzaro, and P. J. Snyder. Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: A longitudinal case study. *Brain and Cognition*, 56(1):24–29, 2004.

[121] M. Malik and A. J. Camm. *Heart rate variability*. Futura Pub. Co., Armonk, NY, 1995.

Index