Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning

Athanasios Tsanas

University of Oxford, St. Cross College

Thesis in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Supervisors: Dr. M.A. Little and Dr. P.E. McSharry

June 2012

To my family, with love

To my teachers, with gratitude

Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning

Athanasios Tsanas

University of Oxford, St. Cross College

Summary of thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

This study focuses on the development of an objective, automated method to extract clinically useful information from sustained vowel phonations in the context of Parkinson's disease (PD). The aim is twofold: (a) differentiate PD subjects from healthy controls, and (b) replicate the Unified Parkinson's Disease Rating Scale (UPDRS) metric which provides a *clinical impression* of PD symptom severity. This metric spans the range 0 to 176, where 0 denotes a healthy person and 176 total disability. Currently, UPDRS assessment requires the physical presence of the subject in the clinic, is *subjective* relying on the clinical rater's expertise, and logistically costly for national health systems. Hence, the practical frequency of symptom tracking is typically confined to once every several months, hindering recruitment for large-scale clinical trials and under-representing the true time scale of PD fluctuations.

We develop a comprehensive framework to analyze speech signals by: (1) extracting novel, distinctive signal *features*, (2) using robust *feature selection* techniques to obtain a parsimonious subset of those features, and (3a) differentiating PD subjects from healthy controls, or (3b) determining UPDRS using powerful *statistical machine learning* tools. Towards this aim, we also investigate 10 existing fundamental frequency (F_0) estimation algorithms to determine the most useful algorithm for this application, and propose a novel ensemble F_0 estimation algorithm which leads to a 10% improvement in accuracy over the best individual approach. Moreover, we propose novel feature selection schemes which are shown to be very competitive against widely-used schemes which are more complex. We demonstrate that we can successfully differentiate PD subjects from healthy controls with 98.5% overall accuracy, and also provide *rapid*, *objective,* and *remote* replication of UPDRS assessment with clinically useful accuracy (approximately 2 UPDRS points from the clinicians' estimates), using only simple, self-administered, and non-invasive speech tests.

The findings of this study strongly support the use of speech signal analysis as an objective basis for practical clinical decision support tools in the context of PD assessment.

Acknowledgements

First and foremost I would like to thank my family whose continuous support for all those years has always been well beyond the call of duty. I would like to extend my thanks to a truly remarkable individual, Angeliki, for making me smile and being a constant source of resourcefulness and inspiration. I was extremely fortunate to have been amongst talented students and inspiring scholars during the course of my previous higher education in Athens, Liverpool, and Newcastle. Although not directly linked to this work, their influence was keenly felt. My previous supervisors, Theofanis Maniatis (Athens) and John Goulermas (Liverpool), had been great mentors and I am indebted to them in more than one way. A special thanks to my colleagues Mohammed Aziz, Trevor Wood, Yiming Zhong, Denis Lapitski, Chris Lustri and Sofia Piltz in DH27 with whom we shared the same office over the past years, and the members of the SAMP group Siddharth Arora, Paul Moore, Shilan Mistry, and Ada Lau from whom I learned so much and often failed to give adequate contribution.

I am grateful to my supervisors Max Little and Patrick McSharry for offering me this CASE studentship, and for their never-ending support, patience, and advice. Max in particular has been an exceptional mentor and we often got engaged in stimulating discussions, not to mention the literally thousands of emails we exchanged. Gari Clifford was an endless source of inspiration, and encouraged me to join his excellent group meetings and interact with many of his colleagues. My college advisor, James Taylor, was particularly helpful in college matters. The staff in the Mathematical Institute has been very kind to me: especially Ruth Preston, Sandy Patel, Val Timms, Keith Gillow, and Waldemar Schacklow. Talks on the phone and over emails with Michael Deisher, Bill Deleeuw and Sangita Sharma from Intel Corporation have helped me keep focused and shaped my research. The data used in this work was collected as part of a large multi-centre study in the US; I am obliged to all those involved. Matias Zañartu generated the speech signals I used to validate the fundamental frequency estimation algorithms in Chapter 5. Peter Robbins and Philip Maini kindly offered advice and pointed out useful references early in my research. I extend my thanks to the many researchers who have replied to my emails explaining concepts, shared source code, provided valuable feedback and made useful suggestions. At the risk of undeservedly leaving someone out, I want to acknowledge: Lorraine Ramig, Jennifer Spielman, Declan Costello, Eric Hunter, Alexander Ihler, Hideki Kawahara, Yijun Sun, Laurens van Maaten and Paul Boersma. Finally, I would like to thank all my examiners for their constructive feedback and valuable suggestions, in particular Philip Maini, Pedro Gomez-Vilda, Gari Clifford, and Steve Roberts.

This project is made possible through the generous funding of an Intel/EPSRC scholarship.

Abbreviations

AHTD	Intel’s At Home Telemonitoring Device
BG	Basal Ganglia
CART	Classification and Regression Trees
EGG	Electroglottography
FS	Feature Selection
GCI	Glottal Closure Instance
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
mRMR	Minimum Redundancy Maximum Relevance
OAA	One Against All
OAO	One Against One
PD	Parkinson’s Disease
PDA	Pitch Detection Algorithm
PWP	People with Parkinson’s
RF	Random Forests
RRCT	Relevance Redundancy and Complementarity Trade-off
SVM	Support Vector Machine
SPL	Sound Pressure Level
UPDRS	Unified Parkinson’s Disease Rating Scale

Frequently used notation

The following mathematical notational conventions are used throughout this thesis:

Vectors are written in bold lower case letters, for example **x**; matrices are written in bold capital letters, for example **X**. $\{\cdot\}_{i,j}$ denotes the *i*th row, *j*th column matrix entry. The subscript *n* in the form x_n indicates the *n*th element of a vector. The expectation and the conditional expectation operators used are: $E[\cdot]$ and $E[(\cdot\,|condition)]$. Also, $(\cdot)^{\mathrm{T}}$ denotes the transpose of a matrix, and $\frac{df}{dx}$ represents the differentiation of a function f with respect to x. The covariance of two random variables X, Y is defined by $Cov(X,Y) = E[X,Y] - E[X] \cdot E[Y]$. The convolution operator is denoted by $\otimes$, and the distance metric is represented by $\|\cdot\|$. Unless otherwise specified, the Euclidean distance is used. In the context of this thesis we work with (a) real numbers (represented with $\mathbb{R}$), (b) natural numbers (represented with $\mathbb{N}$), and (c) integer numbers (represented with $\mathbb{Z}$).

List of figures

List of tables

Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Appendices

CHAPTER 1

Introduction

This study addresses the pertinent problem of monitoring neurological disorders and in particular *Parkinson's disease* (PD). Using speech signals as a measurement, we develop clinically useful tools for (a) differentiating healthy controls from people with Parkinson's (PWP), and (b) monitoring *accurately*, and *remotely*, average PD symptom severity as defined by the clinical metric *Unified Parkinson's Disease Rating Scale* (UPDRS).

1.1 Historical overview of Parkinson's disease[1]

The oldest description of *parkinsonism* symptoms goes as far back as 5,000 B.C. in India allegedly described in *The Four Vedas*. Other possible references to parkinsonian indications include descriptions in the Bible and Iliad of Homer, and later descriptions can be found in the works of Leonardo DaVinci and the plays of William Shakespeare during the 16th century.

However, it was the milestone work of James Parkinson in 1817 reported in "*An essay on the shaking palsy*", which provided an overview of the disease in its medical context based on anecdotal observations (Parkinson, 1817). Parkinson himself referred to it as *paralysis agitans* and the term *Parkinson's Disease* (PD) was coined later by Jean-Martin Charcot in 1876. Charcot was a highly influential PD researcher, adapting the sphygmograph (originally designed for recording arterial pulse) to record tremor at the wrist, prescribing early drugs and

[1] This is necessarily kept brief; the interested reader may want to consult Chapter 1 of Pahwa and Lyons (2007) and the website www.movementdisorders.org for a more detailed historical overview.

developing methods to alleviate symptoms. Attempts to cure PD were conducted by Benjamin Duchenne who used electrotherapy as early as 1855 (Duchenne, 1855).

The cornerstone of contemporary treatment of PD is the manipulation of pharmacological pathways in the form of levodopa (L-dopa), which alleviates some symptoms of the disease. This was based on the work of Nobel Prize winner Arvid Carlsson, who demonstrated in the 1950s that dopamine is a neurotransmitter in the brain, and George Cotzias, who administered L-dopa in patients with successful outcomes (Cotzias, 1968). Numerous developments in pharmaceuticals and surgical techniques have followed in recent years as remedies against PD, but to the present day there is no available cure and PD is eventually fatal.

1.2 Statement of the problem

Neurological disorders affect people profoundly and claim lives at an epidemic rate worldwide, with PD being the second most common neurodegenerative disorder after Alzheimer's (de Rijk et al., 2000). *Incidence* rates and *prevalence* rates[2] of PD in different studies vary, with a large recent study reporting incident rates of 20/100,000 (Rajput et al., 2007). It is believed that there are more than one million PWP in North America alone (Lang and Lozano, 1998), whilst a large meta-analysis study in Europe reported prevalence rates approximately 108-257/100,000 and incidence rates 11-19/100,000 (Campenhausen et al., 2005). Furthermore, Schrag et al. (2002) report that an estimated 20% of PWP go undiagnosed. Most sources claim greater PD prevalence in men than women (Baldereschi et al., 2000; Haaxma et al., 2007) and the lifetime risk, considering current global average life expectancy, is estimated to be 4.4% (men) and 3.7% (women) (Elbaz et al., 2002).

[2] *Incidence rate* is the fraction of newly diagnosed patients per year in the population, usually quoted in cases per 100,000 (Rajput et al., 2007). *Prevalence rate* refers to the fraction of people in the population diagnosed at any given time.

Aging is associated with a number of detrimental effects on a person's health impinging on, amongst others, the nervous system. Thus, the aforementioned statistics are bound to increase due to worldwide population aging. In fact, all studies suggest that age is the single most important risk factor for PD onset, which increases steeply after age 50 (Elbaz et al., 2002). The disease is *progressive,* where symptoms get worse with time and PD progression cannot be stopped; however pharmaceutical and surgical intervention can mitigate the effect of some of the symptoms and prolong the patient's life.

Clinicians have devised a number of methods to quantify PD symptom severity, and the most widely used metric is the *Unified Parkinson's Disease Rating Scale* (UPDRS) (Ramaker et al., 2002), which reflects the presence and severity of symptoms (but does not measure their underlying causes). Monitoring PD progression is critical because this enables improved patient-directed treatment. At present, PD monitoring has many shortcomings:

1) It requires the *patient's frequent physical presence* in the clinic, which may be logistically and financially difficult both for the patients and their carers, especially in the later stages of the disease.
2) It requires the availability of *expert clinical staff* to do the tests and assess the patient's symptoms in order to determine the UPDRS score.
3) The UPDRS assessment is *subjective* and different expert clinical raters often do not agree on the reported scores (*inter-rater variability*) (Rajput et al., 1991; Hughes et al., 1993; Ramaker et al., 2002; Post et al., 2005).
4) It is *costly* for national health systems, which need to provide facilities to accommodate patients and allocate expensive human resources.
5) It is *time-consuming*, since the UPDRS examination normally lasts more than two hours (when assessing PD severity both 'off' and 'on' medication)

For all these reasons, currently, most PWP will only have UPDRS assessed once every three to six months, if at all, because of the scarcity of resources available to patient, carers, and clinical staff. Therefore, frequent, *remote* monitoring emerges as a compelling solution to accurately and efficiently follow PD progression at more frequent intervals with less cost and minimal waste of resources. Noninvasive telemonitoring is an emerging option in general medical care, potentially affording reliable, cost-effective screening of PWP, and potentially alleviating the burden of frequent, and often inconvenient, visits to the clinic. This also relieves national health systems from excessive additional workload, decreasing the cost and increasing the accuracy of clinical evaluation of the subject’s condition.

Speech disorders have been linked to PD (Darley et al., 1969a; Gamboa et al., 1997; Ho et al., 2008), and there is strong supporting evidence of degrading performance in voice with PD progression (Harel et al., 2004; Skodda et al., 2009). Speech signals fit ideally the purpose of telemonitoring, because they are non-invasive, can be self-recorded, and are easy to obtain from a subject who is not expected to perform any special kinds of actions in order to record his voice. Differentiating PWP from healthy controls using speech has attracted interest in the research community (Harel et al., 2004; Sapir et al., 2010; Cnockaert et al., 2008; Little et al., 2009); in this study we also extend this concept to map the severity of voice-based PD symptoms to UPDRS. We also wanted to determine the feasibility of remote PD clinical trials on large scale voice data recorded in typical home acoustic environments, where previous studies have been limited to controlled acoustic environments and relatively small numbers of recordings (Little et al., 2009). Recent studies have raised the important topic of finding a statistical mapping between speech properties and UPDRS as an issue worthy of further investigation, but had not addressed it explicitly (Skodda et al., 2009; Goetz et al., 2009).

In this study, we will focus on both discriminating PWP from healthy controls, and also determining UPDRS using speech signals alone.

1.3 First-principles models versus data-driven models

Approaches to the mathematical modeling of data can be roughly divided into two categories: *first-principles* and *data-driven* (Little et al., 2006). Other terms have been introduced which essentially amount to the same thing: *system model* and *signal model* in the context of speech synthesis (Sinder, 1999), as well as *white-box* and *black-box* (common terminology in control applications)[3]. In all these cases, the first category employs physical principles that are believed to govern the modelled system, whereas the second develops some mathematical relationship, whose only constraint is that it must approximate as well as possible the measured data, without reference to any physical principles. Mathematical models have been used in practically all disciplines and there is vast literature for both categories; see for example Howison (2005) and Hastie et al. (2009).

First-principle models are increasingly popular in biology and medicine. Modelling specific organs and their interactions has attracted enormous interest aiming to discover the underlying mechanisms of certain physiological functions of the human body. Standard reference works for mathematical modelling in biology and physiology include Keener and Sneyd (1998), Ottesen et al. (2004), and Hoppensteadt and Peskin (2002). In the words of Ottesen: *“Statistical analysis may discover correlations but may fail to provide insight into the mechanisms responsible for these correlations. However, when it is combined with mathematical modelling of the dynamics, new insights into physiological mechanisms may be revealed”* (Ottesen et al., 2004). Most importantly, the results of first-principle models can be more easily interpreted and understood by specialists who are not necessarily mathematically oriented and thus provides the means for multi-disciplinary interaction.

[3] There is also the possibility of hybrid approaches, informally known as *grey-box* models.

Nevertheless, the data-driven approach has its own importance and complements the first-principles approach. Data-driven models do not usually reveal insights into biological causes and functions with quite the same transparency as first-principles modelling, but often, it is the only *practical* thing that can be done, given the typical level of noise and other unknown sources of physiological and environmental variability that affect data recorded in real-world clinical experiments. Data-driven modelling can infer interesting structure in the data, which can sometimes have a meaningful tentative physiological interpretation. The discipline of data-driven inference is more widely known as *statistical machine learning*, and has led to many exciting discoveries. Stark and Hardy (2003) in their paper in *Science* conclude: *"By combining the best features of these two approaches in models that incorporate the main mechanisms underlying specific applications, today's researchers can make far more progress with practical problems than was hitherto possible. Perhaps modern biology, which argues about a choice between hypothesis- and data-driven research should heed this lesson. Neither approach provides the complete picture, and only the synergy between them is likely to lead to solutions to real world problems in an increasingly complex world."*

An additional, major point of controversy in statistical machine learning is whether one should be aiming to impose a *parametric* mathematical structure (e.g. a standard linear/nonlinear model) or simply allow the data itself define the structure (nonparametric). Both approaches are useful: a lively discussion on the topic can be read in Breiman (2001a).

1.4 Scope and structure of the thesis

This study is an investigation of signal processing and machine learning techniques for the extraction of clinically useful information from speech signals. The aim is both to differentiate PWP from healthy controls, and also to map average PD symptom severity to the standard reference clinical metric UPDRS. We aim to infer properties of the speech signals, extracting useful distinguishing features which are altered as the orchestrated muscle movements involved in voice production become hindered due to the deterioration of neurological control attributed to dopaminergic neuron loss in the basal ganglia. The means by which we achieve this include: (a) developing novel speech signal processing algorithms, (b) the investigation of robust feature selection algorithms to identify the most useful feature subset, and (c) the subsequent exploitation of the feature subset to estimate UPDRS.

The thesis begins, in Chapter 2, with a concise description of the physiology of the nervous system and the systems responsible for the production of speech. It focuses particularly on the essential physiological concepts which are later addressed in the thesis in the data-driven signal processing methods. Chapter 3 provides a comprehensive literature review of the most popular clinical speech signal processing algorithms used in biomedical applications. In this chapter, we also develop novel extensions to known approaches and present some new algorithms to characterize some patterns that were not previously captured with existing methods. Chapter 4 identifies some mathematical and statistical tools widely-used in this context, and provides a short review of the machine learning techniques used later in this study. Moreover, it describes an effective machine learning methodology which is applicable to a wide range of problems dealing with *high-dimensional* data, the *curse of dimensionality*, and the *principle of parsimony*[4]. Chapter 5 compares fundamental frequency estimation algorithms, using artificially generated speech signals where the ground truth fundamental frequency is known. Moreover, we demonstrate the potential of a novel ensemble approach which is, on average, 10% more accurate compared to the best individual fundamental frequency estimation algorithm. In addition, this chapter presents a comparison of the feature selection techniques described in Chapter 4 using widely used datasets in the literature.

[4] The terms *high-dimensional* data, *curse of dimensionality* and *principle of parsimony* are defined in Chapter 4.

Chapter 6 describes the speech-PD database used in this study, and identifies the most important confounding factors that need to be considered when inferring PD severity from speech signals. Chapter 7 brings together the information from the previous chapters: we use the signal processing algorithms introduced in Chapter 3, the machine learning methodology (presented in Chapter 4), and the findings in Chapter 5 to study the speech-PD databases presented in Chapter 6. The aim is to (a) study the binary discrimination of healthy controls from PWP, and (b) determine a functional relationship between speech and UPDRS. Chapter 8 draws conclusions and suggests areas of potential interest for future work.

1.5 Summary of contributions

This section summarizes the contributions of this study, and refers to the particular sections of the thesis where they can be found:

1. Development of an ensemble fundamental frequency estimation scheme, which gives 10% more accurate estimates compared to the best individual fundamental frequency estimator amongst the ten popular algorithms investigated in this study (§ 3.2.1.10).
2. Development of novel speech signal processing algorithms, which reveal additional pathophysiological characteristics in the voice of PWP which were not previously captured by the available state of the art algorithms (§ 3.2.4).
3. Development of novel feature selection algorithms (§ 4.2.3). The first of these algorithms, which we refer to as Relevance, Redundancy, and Complementarity Trade-off (RRCT), is a fast correlation-based approach invoking some information theoretic concepts. RRCT is shown to outperform popular feature selection algorithms of comparable complexity in the literature. We also extend known feature

selection algorithms, for example, approaches which were originally proposed for binary classification problems to tackle multi-class classification problems.

4. Empirical evaluation of a wide range of state of the art algorithms for (a) fundamental frequency estimation in sustained vowels (§ 5.1), and (b) feature selection (§ 5.3).
5. This study reports results suggesting that it is possible to discriminate healthy controls from PWP with almost 99% accuracy (the current state of the art results are about 93% accuracy). This improvement is attributed to the novel pool of features proposed in this study (§ 7.1).
6. This study has shown, for the first time, that telemonitoring of average PD symptom severity (quantified using UPDRS) can be achieved *remotely*, *objectively*, and *accurately* using speech signals (§ 7.3). We demonstrate that we can replicate UPDRS within about 1.6 points from the clinicians' estimates.
7. Development of two Matlab toolboxes for: (a) speech signal processing algorithms, and (b) statistical machine learning techniques. The speech signal processing toolbox includes implementations of a wide range of known, and novel, speech signal processing algorithms drawing on the methods described in Chapter 3. These speech signal processing algorithms were previously scattered across the research literature and some algorithms were made available in different software platforms; here they are presented for the first time in Matlab. The machine learning toolbox focuses on the techniques described in Chapter 4, including data visualization, feature selection, and mapping features to the response using an automated process. Both toolboxes are heavily annotated facilitating easy experimentation: good default values are automatically provided, but the annotations suggest a range of parameter values which can be optimized for specific applications.

Chapter 2

Essential physiological background

This chapter presents a concise synopsis of the physiological systems relevant to PD telemonitoring by speech signal processing, exploring the nervous system and speech production mechanisms. It also discusses critical aspects of life-span changes in speech, providing the basis for a fair comparison between age- and gender-matched PWP and healthy controls. Therefore, this chapter provides the critical physiological link between speech and PD, which is used to interpret the results of the later analyses in this study.

2.1 Nervous system

The nervous system consists of a sophisticated network of dedicated cells (neurons) that coordinate actions and transmit signals between different parts of the body. It is exceptionally complicated, and intensive research has revealed only a fraction of its functionality. The abundance of uncharted areas and speculative theories for regions of the brain and its various functional interconnections, suggest that we are still a long way from truly understanding how the nervous system works. The nervous system is responsible for processing sensory input (from the senses), coordinating movements towards the desired goal, and apparently all other cognitive functions.

On a large scale, neurologists often divide the nervous system into two parts, the *central nervous system* (CNS) and the *peripheral nervous system* (PNS). The CNS consists of the brain and the spinal cord, the rest of the neuronal circuitry belongs to the PNS. For the purposes of this thesis we are primarily interested in the brain, since this is the pathophysiological locus of PD. We shall briefly describe the physiology of the basic unit of the brain (the neuron), and the basal ganglia, which is the brain structure believed to be affected in PD; the reader may wish to refer to the textbook of Guyton and Hall (2006) for a more elaborate discussion of nervous system physiology.

2.1.1 Physiology of the basic functional unit of the brain: the neuron

The nervous system comprises two main types of specialized cells, the *neurons* and the *glia*. The neuron consists of the cell body (*soma*), which contains the *nucleus*, the *axon* (neuron output) which is an electrically conducting fiber and leads to the nerve terminals, and the *dendrites* (neuron input) which receive signals from other neurons (see Fig. 2.1). The functionality of the glia is to assist, support, and protect neurons.

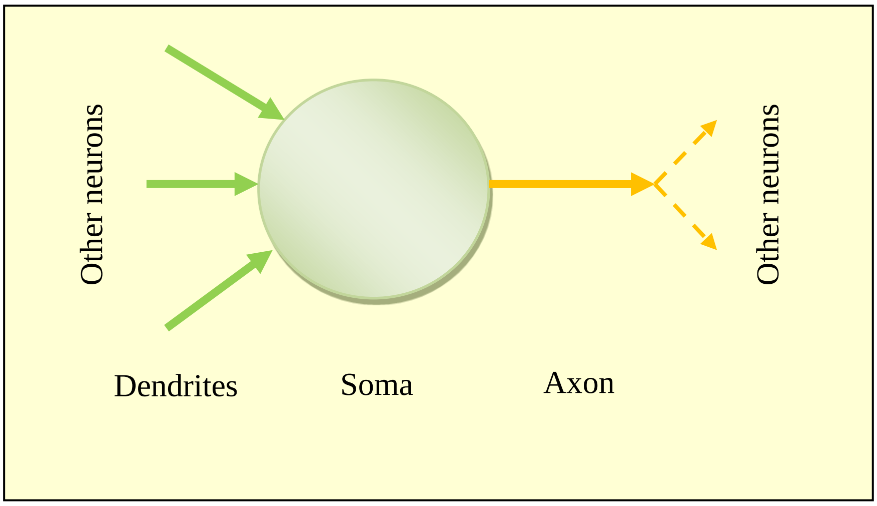

Fig. 2.1: Schematic diagram of a neuron, showing its main anatomical parts.

The main communication system between neurons is achieved through electrical impulses (also known as *nerve impulses*, spikes and more commonly *action potentials*), which is a small amount of current travelling across the axon. The nerve impulse is the result of electrical discharge due to the *sodium-potassium pump*, and occurs when the neuron's membrane has been sufficiently depolarised (15 mV above the normal resting voltage level of -70 mV).

2.1.2 The basal ganglia

The brain comprises billions of neurons, which apparently form functional and anatomical structures at various levels of organisation. The inter- and intra-interaction of these structures, as well as their detailed functional capacity is still a controversial subject and much is speculative. The *basal ganglia* (BG) is a group of highly interconnected anatomical structures positioned approximately in the middle of the brain, and is critically involved in muscle and cognitive control. The BG nuclei appear in two sets, in the left and right cerebral hemispheres. Most BG-directed research effort is motivated by its direct link to a wide range of disorders, including PD, Huntington's disease and schizophrenia. In addition, there is the possibility of extracting (invasive) data, known as *Local Field Potentials* (LFP), which quantify neuronal activity in regions within the brain.

The BG consist of the *striatum*, the *globus pallidus* (GP) (subdivided into the internal segment GPi and the external segment GPe), the *subthalamic nucleus* (STN) and the *substantia nigra* (which is further subdivided into the *pars compacta* (SNc) and *pars reticulata* (SNr)). The BG receive input into the striatum from the *cortex* and another brain region called *thalamus*, and project their output into the thalamus and the brainstem through the SNr and GPi. The interconnections between each pair of these BG nuclei are either *inhibitory* or *excitatory* and are facilitated by the neurotransmitter *dopamine*, a substance produced by the *dopaminergic cells*. Fig. 2.2 shows the schematic functional architecture diagram and the interconnections of the BG according to Gurney et al. (2001a).

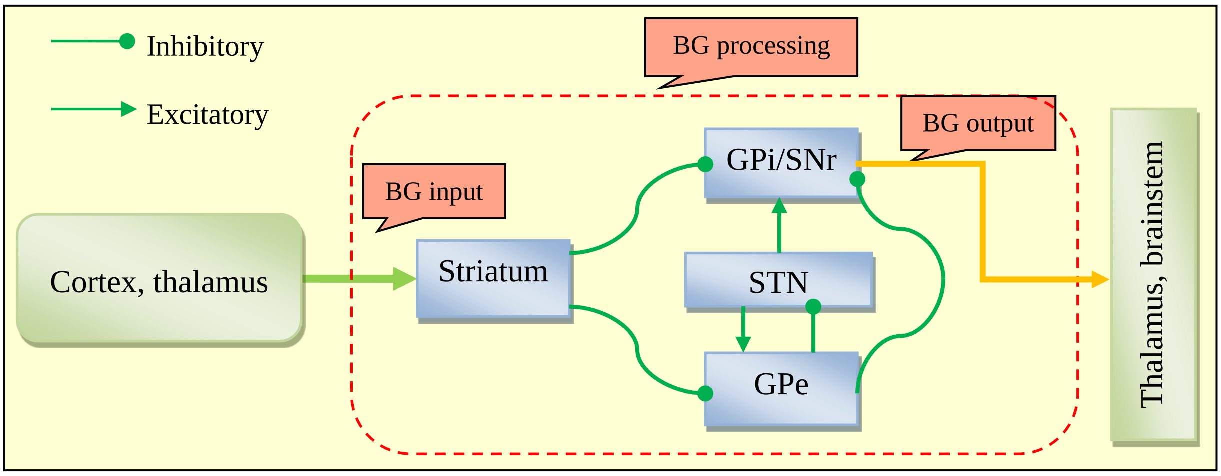

Fig. 2.2: Schematic diagram of the basal ganglia (BG) architecture, showing the input into the BG from the cortex and the thalamus, the BG processing, the BG output, and the interconnections between the BG nuclei. The abbreviations of the BG nuclei are explained in the text.

2.1.3 Parkinson's disease: mechanisms, symptoms, diagnosis, and management

The aetiology (underlying cause) of PD is largely unknown (Lang and Lozano, 1998), but the symptoms are caused by substantial dopaminergic neuron reduction, leading to dysfunction of the BG which mediates motor and some cognitive abilities (Singh et al., 2007). The dopaminergic cells assist in neurotransmission (transmission of information between neurons); consequently their decline leads to malfunction of the CNS which can no longer co-ordinate muscle movements appropriately and delicately. The clinically noticeable symptoms appear when the disease has progressed considerably and about 60-80% of the dopaminergic cells have already died (Bernheimer et al., 1973); by that time it is too late to intercept the degradation. The evolution of the disease involves progressive dopaminergic loss which results in gradually more severe symptoms such as tremor and loss of muscle control.

The main symptoms are tremor, rigidity and movement disorders. Vocal impairment is also common (Hanson et al., 1984; Ho et al., 1998) and is met in approximately 70-90% PWP (Logemann, 1978; Hartelius and Svensson, 1994; Ho et al., 1998). Moreover, it may be one of the earliest indicators (Duffy, 2005) and 29% of patients consider it one of their greatest hindrances associated with the disease (Hartelius and Svensson, 1994). Typically, the symptoms initially appear *unilaterally* (on either the left or right side, indicating that dopaminergic loss is more pronounced in the BG of one of the brain hemispheres) but in time proceed bilaterally.

There is no consensus for diagnosing a patient with PD, which is the cause of many misdiagnoses (Lang and Lozano, 1998; Rajput et al., 2007). According to de Rijk et al. (1997) a patient should be diagnosed with PD if they fulfill at least two of the following three criteria: bradykinesia (slow movement), rigidity, and tremor. Additionally, if the individual is known to suffer from chronic essential tremor (kinetic tremor mostly in the arms, neck and jaw which is apparent during voluntary movement), then a PD diagnosis should be made if all three criteria are present (Rajput et al., 1993). The term *idiopathic* PD (Rajput et al., 1984), which means that the underlying cause of the observed symptoms is unknown, has been introduced to differentiate PD from other neurological disorders eliciting Parkinsonian characteristics[5]. These are known as *Parkinsonism*, and may be due to, for example, drugs or neurotoxins (Rajput et al., 1984; Bower, 1999; Baldereschi et al., 2000). Although accurate *pathophysiological classification* (diagnosis of PD or some form of Parkinsonism) of subjects

[5] Some studies separate idiopathic PD from the remaining Parkinsonism variants referring to it as IPD. In the context of this thesis, PD coincides with IPD.

is extremely difficult, it has clinical importance and facilitates better treatment (Rajput et al., 2007).

Thanks to the use of pharmacopathological manipulation (drug treatment of PD), the mean life expectancy of PWP disease has increased significantly over the previous decades. Currently, it is estimated that a patient diagnosed with PD at the age of 62 is expected to live for about 20 more years (Rajput et al., 2007). Pharmaceutical (combinations of levodopa and other agents) and surgical interventions such as Deep Brain Stimulation (DBS) (Benabid et al., 2009) are documented to improve motor functionality and reduce tremor, delaying disease progression and offering reasonably good quality of life (Singh et al., 2007). Of relevance to this study, however, the impact of treatment on speech is inconclusive (Larson et al., 1994; Ho et al., 2008).

Management of PD involves the administration of physical examinations applying tests assessing the subject's ability to perform a range of tasks, and these tests are designed to enable the quantification and monitoring of disease progression. The UPDRS is the standard reference scale (Ramaker et al., 2002), approved by the Movement Disorders Society (MDS), and has lately been revised as the MDS-UPDRS scale (Goetz et al., 2008). This revision addresses some deficiencies of the current version, which were previously discussed in Goetz et al. (2003). UPDRS tests along with the indications the medical rater will use to score the subject's symptoms appear in Appendix II. The UPDRS metric consists of 44 *sections*[6], where each section addresses different symptoms in different parts of the body and spans the range 0-4, with 0 denoting no symptoms and 4 severe impairment or problem. Summing up these 44 sections gives rise to the *total-UPDRS* score, which spans the range 0-176, with 0 representing perfectly healthy individual and 176 total disability.

[6] Note that the UPDRS scale discussed in this thesis is for *untreated* patients because that is the kind of data used in this study (see Chapter 6); the UPDRS has additional sections for treated patients.

The UPDRS metric can be divided in three major parts, which we will henceforth refer to as *components*: (1) Mentation, Behavior and Mood (MBM); (2) Activities of daily living (ADL); (3) Motor. The *motor component* (we will refer to it as *motor-UPDRS*) is comprised of sections 18-44 and ranges from 0-108, with 0 denoting symptom free and 108 severe motor impairment, and encompasses tasks such as speech, facial expression, tremor and rigidity. This component contributes most of the points in the UPDRS scale and many studies focus exclusively on that, because motor symptoms are often the most problematic and the most prominent aspect of PD. In this study we deal with both 'motor-UPDRS' and 'total-UPDRS'. Alternative metrics monitoring PD progression may also be used, such as the *Hoehn-Yahr* (H&Y) *stage* (Hoehn and Yahr, 1967), and recent studies have shown that it is possible to map UPDRS onto H&Y (Tsanas et al., 2012c).

As discussed above, dopaminergic depletion within the BG is the hallmark of PD, and clinicians often rely on brain scans in order to *noninvasively* reveal the subject's brain pathophysiology (structural and functional operational condition) so that they can augment their PD diagnosis. For a recent review of the current imaging methods refer to Brooks (2007). Nevertheless, although imaging biomarkers are measuring the relevant physiological process, they do not measure dopamine density, and therefore cannot be used as a monitoring tool (Ravina et al., 2005).

2.2 Speech related organs

Before proceeding with the discussion of the organs either directly or indirectly involved with the production of speech, it is appropriate to clarify some recurring concepts. Two terms which regularly occur in this thesis are *speech* and *voice*. They are often used interchangeably and in fact I. Titze asserts "*in the broader sense voice is synonymous with speech*". However, he also mentions that there *is* in fact a subtle difference and the term *voice* in its narrow definition refers only to the sound produced by the vocal organs (Titze, 2000). For the purposes of this thesis, they will be considered synonymous. Fig. 2.3 displays a simple schematic diagram of the major anatomical parts involved in the production of speech. The following sections described these anatomical parts in some detail.

2.2.1 Pulmonary system

The pulmonary system comprises the lungs and the *respiratory airways* (tubes which allow the passage of air from the atmosphere to the lungs and vice-versa). The lungs consist of millions of *alveoli* (air sacks connected together) and their primary role is to assist metabolism through respiration, i.e. the exchange of oxygen and carbon dioxide for the oxygenation of cells in the whole body. The lungs also provide the driving energy for the speech production system, the *lung pressure*, which is between 0.3-1.2 KPa in conversational speech (Titze, 2000). Speech production is dependent on air flow along the respiratory tract, originating in the lungs and travelling along the *trachea*.

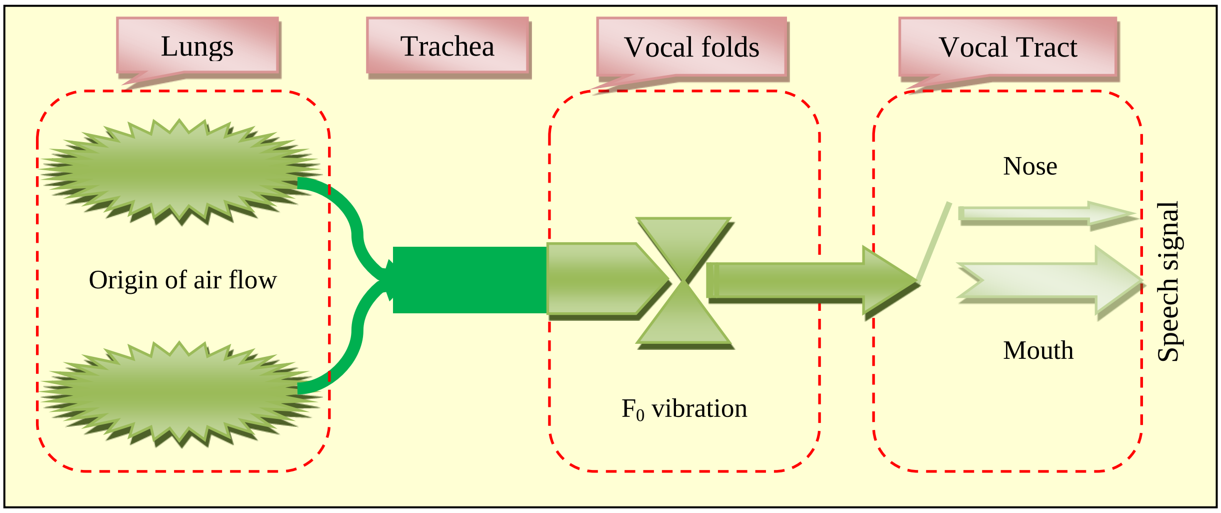

Fig. 2.3: Schematic diagram of the major parts involved in the production of speech.

During *inhalation* (the inspiration phase) the lungs expand and air flows into the lungs; during *exhalation* (the expiration phase) the lungs collapse and the air flows out. These delicate movements of the inspiration and expiration phases are controlled by the *diaphragm muscles* (muscle structure at the lower part of the thorax, underneath the lungs), which expand and collapse[7]. Of direct relevance to this study, respiratory muscle control is known to be compromised in PWP (Apps et al., 1985), which partly explains why those subjects often fail to be able to produce prolonged vocal effort, by comparison to age- and gender-matched healthy speakers.

2.2.2 Vocal folds

The *vocal folds* (also known as *vocal cords*) are located above the trachea and across the larynx. The *vagus nerve* innervates (controls) the larynx muscles, and this nerve originates in the brain structure called the *brainstem* that is itself connected to the BG (Guyton and Hall, 2006). The vocal folds move backward and forward forming a *self-sustained oscillator* and thereby modulate the airflow from the lungs in the process of phonation as it travels through the *glottis* (airspace between the vocal folds). For a detailed explanation of the self-sustained vocal fold oscillation mechanism we refer to Titze (2000). We can examine vocal fold movement using electroglottography (EGG): this provides a signal recorded from a device that is placed externally to the larynx and detects glottal cycles while the subject is speaking.

Men and women have different vocal fold size, which causes different patterns of vibration, in particular the number of times the vocal folds vibrate during a second, i.e. the *frequency* (in Hertz) of oscillation (Titze, 2000). The time taken for the vocal folds to complete one oscillation (*cycle*) is known as *pitch period T*, and the *fundamental*

[7] For a detailed introduction of the physiology of speech production please refer to Titze (2000).

frequency $F_0 \overset{\text{def}}{=} 1/T$. We will see that it is not possible to have an entirely rigorous definition of T because voice signals are never *exactly* periodic[8], when we revisit the vocal fold vibration pattern in sustained vowels (see § 2.2.4). The time varying motion of the vocal folds can be described in the frequency domain, and consists of many harmonics in addition to F_0. Sub-multiples of the true F_0 are known as *sub-harmonics*, and in healthy voices are kept to a minimum. The signal is often represented in the frequency domain to identify its main frequency components; the following chapter presents approaches to exploiting these. For now, we define two additional commonly occurring terms in the speech science discipline: *semitone difference* and *octave*. The *semitone difference* between two arbitrary frequencies f_1 and f_2 is $12 \cdot \log_{10}(f_2/f_1)$, and doubling the frequency is equivalent to rising by 12 semitones, where 12 semitones make up one *octave* (Baken and Orlikoff, 2000).

Research has shown that some PWP exhibit incomplete vocal fold closure and increased *breathiness* during phonation (that is, their voice becomes dominated by noticeable breath noise) (Ho et al., 1998). The incomplete closure causes airflow vortex shedding to occur throughout the entire vocal fold vibration cycle, rather than just after the moment of vocal fold closure, causing increased *turbulent noise*. However, incomplete vocal fold closure and breath noise are not necessarily caused by some neurological deficit: they could be, for example, the result of a vocal fold tissue problem (such as a nodule). Vocal fold closure, which is required for normal phonation, is more complete in patients with PD symptoms confined to one side of the body only (Hanson et al., 1984). This may be one of the reasons why speech symptoms are less prominent in the earlier stages of the disease than in later stages. As we will see in the following chapter, most algorithms focus on the analysis of vocal fold-related problems because mathematically, algorithmically and computationally it is easier to extract *signal characteristics* related to the vocal fold vibration pattern.

[8] We will define *periodicity* mathematically in § 2.2.4

2.2.3 Vocal Tract

The nose, mouth, tongue and lips are collectively referred to as the *vocal tract.* Whereas the vocal folds can be viewed as an *oscillator*, the vocal tract can be described as a *resonator* that amplifies certain acoustic frequencies and attenuates others. Depending on its shape, the vocal tract enhances certain harmonics in the oscillation of the vocal folds which are known as *formants*, and these can be seen in a spectral analysis of voice recordings (Titze, 2000). The vocal fold-vocal tract interaction has often been referred to as the s*ource-filter coupling in phonation* (Titze, 2008), where the sound source is the vocal folds and the filter is the vocal tract. At least since the important work of Fant (1960), and probably even earlier, the source-filter relationship was assumed to be *linear*, i.e. the resulting speech signal was considered to be the result of the convolution of the vocal folds signal and the vocal tract signal. The linear source-filter theory gives interesting insights and is the basis for the vast majority of speech signal processing algorithms (see Chapter 3). However, relatively recently the linear assumption of the source-filter theory has been challenged, and these days research is focused on the *nonlinear* interaction between the vocal folds and the vocal tract (Titze, 2008). During phonation, for example, a portion of the air in the vocal tract is reflected back to the vocal folds when the vocal folds collide, due to the sudden *supraglottal pressure* (pressure just above the glottis) drop. The reflected air towards the vocal folds depends on the vocal tract shape; for further details we refer to the theory of *acoustic wave propagation* (Titze, 2000). Nevertheless, particularly for *sustained vowels* (see the following section) the source-filter theory may often be adequate (at least in healthy voices), which partly explains the widespread use of this theory in vocal quality assessment (Titze, 2000).

Although the correlation of vocal tract changes with PD has already been reported (Hanson et al., 1984; Logemann, 1978), evidence is fairly scarce compared to the investigation of the effect of PD on the vocal folds. As we will see in Chapter 7, analysing vocal tract-related signal characteristics provides clinically useful information about PD status.

2.2.4 Sustained vowels

The use of *sustained vowel phonations* to assess the extent of vocal symptoms, where the subject is requested to hold the frequency of phonation steady for as long as possible, is common in general speech clinical practice (Titze, 2000) and has shown promising results in separating healthy controls from PWP (Cnockaert et al., 2008; Little et al., 2009, Tsanas et al., 2012b), and PD monitoring (Tsanas et al., 2010a; Tsanas et al., 2011a). Maximum phonation duration carries clinically useful information, and a healthy adult should be able to sustain his voice for about 20 seconds on average, although this depends on factors such as age, gender, body stature and general health condition (McNeil, 1997). Sustaining vowels builds on the idea that a healthy subject can elicit a *stationary* phonation, whereas a subject with some form of vocal impairment cannot (Titze, 2000). Informally, a stationary process does not change when shifted in time or place, and implies periodicity. In addition, using sustained vowels circumvents some of the confounding articulatory effects and linguistic components of *running speech* (Schoentgen and Guchteneere, 1995), i.e. the recording of standard phrases. We adopt the typical convention in the speech science literature to represent sustained vowels using the vowel between *slashes*, e.g. for the sustained vowel 'ahh…' we write /a/.

The sustained vowels /a/, /i/ and /u/[9] are used in some clinical applications to assess vocal performance; however, most studies focus solely on the sustained vowel /a/ because this is the simplest sound to produce, and empirically has been found to convey the most clinically useful information (Titze, 2000). Physiologically, /a/ involves the delicate combination of a

[9] These three vowels are known as *corner* vowels because of the extreme placements of the tongue; the reader is referred to Titze (2000) for further discussion.

variety of muscles in the vocal folds and the vocal tract, so it increases the probability that a neurological problem can be identified. Also, in /a/ the mouth is maximally open compared to other vowels, which minimizes the reflected air pulse back to the vocal folds; therefore the recorded SPL at the lips is maximized (Titze, 2000). Vowel sounds in speech have particular formant patterns, and are typically characterised by the two lowest frequency formants labeled F_1 and F_2, which can be plotted on an $F_1 - F_2$ *chart*. In particular for the vowel /a/, F_1 ranges between 600-1,300 Hz and F_2 between 900-1,600 Hz, but this is somewhat subject-dependent (Peterson and Barney, 1952).

As an illustration of the concepts introduced above, Fig. 2.4 shows a typical sustained vowel /a/ phonation. Qualitatively, we note that the overall speech signal amplitude (difference of maximum and minimum values of the signal during a pitch period) is decaying towards zero (Fig. 2.4a). During the last seconds of the phonation, the amplitude tends to shrink, which is the result of lung collapse. Zooming in on the signal (Fig. 2.4b), we can extract the pitch period and the F_0 by observing the peaks between cycles (repetitions of the same pattern in the signal). This makes the tacit assumption that we can define T as the cycle-to-cycle interval, which corresponds to the *exact periodicity* of the signal. But periodicity is a formal mathematical concept, and if we represent the speech signal as $\mathbf{s} = f(t)$, where $t \in \mathbb{R}$ is time, T should satisfy $f(t+T) = f(t)$ for all $t \in \mathbb{R}$. In fact, periodicity does not actually apply to any real speech signal (Titze, 2000) since successive cycles are never *exactly* the same (Fig 2.4b), but this terminology pervades the speech science literature and will be used throughout this thesis. Slight disturbances in the pitch period are attributed to physiologic tremor in the laryngeal muscles, and are known as (smooth) vocal *vibrato*, suggesting that even speech signals from healthy people are not exactly periodic. We introduce the commonly used terms *nearly periodic*, to describe signals that deviate slightly from periodicity, and *aperiodic*, to describe signals which do not exhibit any obvious oscillating pattern.

Three additional, regularly recurring terms in the thesis are *variability*, *perturbation* and *fluctuation*. These are general terms and apply in various fields, but in the context of this study they are typically used to describe the signal amplitude and F_0. *Variability* is the ability of a quantity to vary, which in this study is used to discuss *amplitude variation* and *F_0 variation*, and is usually considered between successive cycles. *Perturbation* is a minor disturbance or deviation from the expected norm (behavior) of a system, where the expected norm is typically considered to be the mean of some quantity. *Fluctuation* is a more severe disturbance than perturbation, and reflects an inherent instability in a system.

The F_0 and amplitude perturbations quantify departure from periodicity, and give rise to some of the most important methods for extracting clinically relevant information from speech signals, which are called *dysphonia measures* in the literature (or simply *measures*). However, since there is sufficient evidence suggesting that speech disorders are commonplace in PWP, caution should be exercised in assuming vocal cycle periodicity. As we shall see in Chapter 3, the most successful measures are those which do not assume periodic signals.

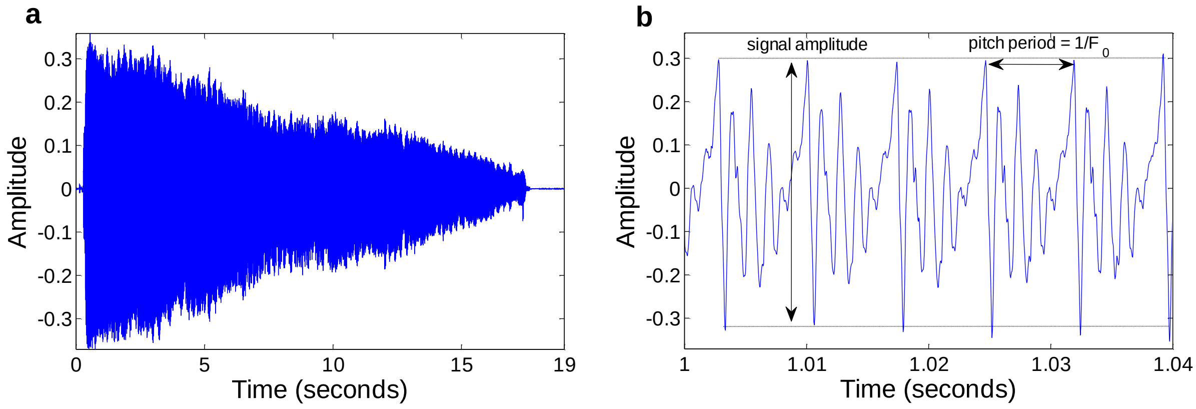

Fig. 2.4 (a) Typical sustained vowel /a/ phonation. The overall amplitude decays over the duration of the phonation (usually 10-30 seconds). (b) Magnified version of the same sustained vowel /a/ phonation to illustrate the signal amplitude and the signal period. The magnified signal is not exactly periodic, a concept we revisit later.

2.3 Life-span changes in physiology

A rigorous assessment of any kind of medical disorder will entail comparing a pathological group of patients with a control group that will typically be age- and gender-matched. Other factors such as subject profession and demographics may also be relevant. Aging is associated with a number of physiological changes of the major organs and organ systems, including the nervous system (Fearnley and Lees, 1991) and the speech production system (Titze, 2000).

Fearnley and Lees (1991) reviewed the physiological changes in neuronal loss with age, and compared healthy subjects and PWP examining the regional substantia nigra of the BG. They confirmed previous findings of neuronal loss in healthy controls with advancing age, and also found that PWP exhibited further decreased neural population numbers compared to their matched controls. These findings were later verified in neuro-imaging studies (Pirker et al., 2002). In addition, loss of muscle control in advanced age is very well documented (Laughton et al., 2003; Lewis, 2006).

Our voices also tend to alter with age due to changes in laryngeal and vocal tract size, muscles which tend to *ossify* (turn to bone), various hormonal changes, and reduced nervous system control (Titze, 2000). Probably the most characteristic change of acoustic variable with age is in the F_0. Fig. 2.5 shows the changes in F_0 with respect to gender for the ages 20-90 years. The trend is different in males and females, with F_0 monotonically decreasing in women, whereas F_0 decreases approximately until age 40 and subsequently increases for men. A recent paper confirms these findings, and stresses that reference F_0 intervals could be useful markers of laryngeal malfunction (Nishio and Niimi, 2008).

It is sensible to assume that the combination of neuronal and muscle control loss with age explains many of the general speech changes observed in the elderly, as suggested by Titze (2000). Furthermore, since control loss is exacerbated in PD due to dopaminergic neuron reduction, it is reasonable to expect speech to be more severely affected in PWP compared to healthy controls. This brings us to the link between speech and PD.

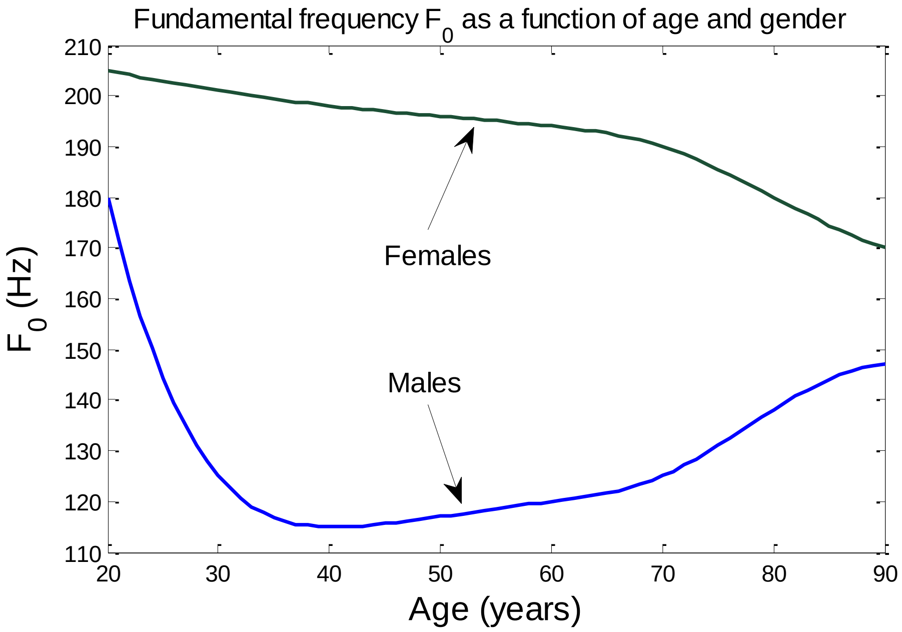

Fig. 2.5: Life-span changes of the fundamental frequency F_0 as a function of gender for the ages 20-90 years old (after Titze, 2000).

2.4 Speech and Parkinson's disease

Neurons orchestrate all muscle movement, managing the delicate co-ordination needed to successfully complete a given task, e.g. walking or lifting something. Similarly, there are neurons controlling the speech-related organs which have to co-operate in the production of speech. Loss of neurons associated with the task of controlling some of the speech-related organs, leads invariably to speech disorders. The speech disorders associated with PD are termed *hypokinetic dysarthrias* (Darley et al., 1975) and lead to reduced speech intelligibility. They can be broadly categorised as:

1) *Hypophonia* (reduced voice volume)

2) *Dysphonia* (breathy, hoarse voice)
3) *Hypokinetic articulation* (imprecise articulation due to reduced articulatory movement range)
4) *Hypoprosodia/monotonous speech* (reduced speech *pitch*[10] variability)
5) *Palilalia* (dys-fluent or hesitant speech)

The relationship between speech and PD has been studied systematically at least as far back in the 1970's by Darley et al. (1969a; 1969b; 1975). They reported reduced loudness, monotonous voice, breathy and hoarse voice, and imprecise articulation in PD subjects. Other studies followed, confirming that PWP exhibit breathiness, hoarseness and articulation problems (Logemann, 1978). More recent studies validate and extend these results comparing healthy controls and PWP (Gamboa et al., 1997), with the majority of patients exhibiting laryngeal tremor during normal or loud phonation (Perez et al., 1996). A 40% reduction in vocal loudness was reported in Fox and Ramig (1997), further endorsed by Ho et al. (2001) who attribute this finding to *symptomatic frontostriatal circuit*[11] *dysfunction*. Interestingly, however, the vocal sound pressure level (SPL) during sustained vowel phonation is no different from that of the healthy controls (Rosen et al., 2005). PWP show signs of increased vibrational aperiodicity (the vocal folds' oscillating pattern departs from periodicity) and increased breathiness (noise) (Michaelis et al., 1998). A decrease in F_0 and F_0 variability in speech is also documented (King et al., 1994; Homes et al., 2000) and these decreases mirror the severity of the disease (Metter and Hanson, 1986). In general, hypophonia and dysphonia

[10] *Pitch* is the perceived F_0 (i.e. pitch is the psychoacoustic equivalent of the physical measure fundamental frequency). It should not be confused with *pitch period*, which we defined earlier as the inverse of the actual F_0. Pitch can be measured by asking a listener to compare speech signals with a pure sinusoid for which the frequency can be adjusted. The pitch of the original speech signal is then by definition the adjusted frequency of the sinusoid that the listener has determined that gives the same tone output. In general, F_0 and pitch correlate well.

[11] The reader will remember the *striatum*, part of the basal ganglia (BG). The *frontostriatal circuit* consists of neural pathways connecting the BG with other regions of the brain, which are also involved in higher mental functions.

precede the rest of the disorders (Logemann, 1978; Ho et al., 1998), and 98% of hypokinetic dysarthritic speech pathologies are related to PD (Berry, 1983). PD appears to affect men and women differently with respect to their vocal performance. For example, Cnockaert et al. (2008) report that average F_0 increases in male PWP and decreases in female PWP compared to matched healthy controls. Recent studies have not taken into account this male-female distinction, e.g. (Little et al., 2009; Tsanas et al., 2010a), and as we will see in Chapter 7 some measures are very highly correlated with UPDRS in one gender, and almost negligibly correlated in the other, confirming the gender selectivity of PD effects on speech.

Interestingly, a pilot study in 2004 revealed that speech impairment could be detectable as early as five years prior to diagnosis of PD (Harel et al., 2004). The speech recordings of two people (one of whom was eventually diagnosed with PD) of similar demographic characteristics (age, gender, and profession) had been examined for 11 years (including 7 years prior to diagnosis). Although that study consisted in comparing the voices of only two people and the authors caution regarding the interpretation of their findings, it is reasonable to assume that some early symptoms of the disease could be traceable before the patient is diagnosed. Similarly, it is plausible that movement disorders could be discovered prior to diagnosis as well but it is difficult to recruit people into such longitudinal studies. As soon as some dopaminergic cells die, an imperceptibly subtle difference in muscle control might be detectable with sensitive equipment and appropriate tests[12]. Monitoring speech signals from patients who have been recently diagnosed with PD could facilitate understanding about the progression of the disease and give rise to improved diagnostic and treatment methods.

[12] As we have noted, typically 60% or more of the dopaminergic cells have already died by the time clinical PD symptoms become measurable.

Chapter 3

Clinical speech signal processing algorithms

This chapter provides a literature survey of some of the most widely used approaches to processing speech signals in order to extract clinically useful information. Many of these tools originate in the speech signal processing research community and have been developed for various purposes, including speaker identification, speech encoding, and, as is the case with the present study, for extracting clinically-relevant information. Other tools have been developed within distinct but related disciplines such as time series analysis. By definition, *time series* refers to a quantity changing in time, and speech falls into this category. Before surveying the various speech signal processing algorithms, we briefly summarize some recurring mathematical concepts.

3.1 Recurring mathematical concepts

3.1.1 Data discretization

Most signals in nature are *continuous time-varying* quantities, and can be represented as $\mathbf{s}(t),\ t \epsilon \mathbb{R}$. In order to process such a signal on a computer we need to discretise it, a process known as *sampling*. This is achieved with an analog-to-digital converter (ADC) which samples the signal at (small, positive) Δt intervals, and produces the *discrete time signal* $\mathbf{s} = \mathbf{s}(n\Delta t)$ where $n \epsilon \mathbb{Z}$ denotes the time index in samples. The sampling time interval

Δt is typically given indirectly via the *sampling frequency* F_s, which is defined as $F_s = \Delta t^{-1}$ (typically given in Hz). The ADC associates $\mathbf{s}$ with a finite number of *quantisation levels* $Q = 2^k$, where k is the ADC *bits of precision* (or simply ADC bits). Thus, the discretisation process is mainly characterised by the sampling frequency and the number of ADC bits. For example, if we assume a bounded signal $-1 \leq \mathbf{s} \leq 1$ and $k = 16$, there will be 65536 different possible values to cover the range -1 to 1.

3.1.2 Linear signal processing tools: autocorrelation and cross-correlation

Many classical signal processing methods are largely based on techniques such as *short-time autocorrelation,* which for a signal $\mathbf{x}$ at lag l is defined as:

$$R_{xx}(l) = \sum_{n=-\infty}^{\infty} x_n \cdot x_{n-l}^* = x_l \otimes x_{-l}^* \tag{3.1}$$

where $(\cdot)^*$ denotes the complex conjugate. Some authors refer to Eq. (3.1) as *autocovariance*, and reserve the term *autocorrelation* for the case where $\mathbf{x}$ is normalized to its z-score (i.e. zero mean, standard deviation equal to 1). The autocorrelation is a tool to find repeating patterns in a signal which may be embedded in noise, and expresses the similarity between samples as a function of the difference of their indices. It has a global maximum at $l = 0$, and dividing $R_{xx}(l)$ by $R_{xx}(0)$ normalises it to the range -1 to 1. Henceforth, we always assume $-1 \leq R_{xx}(l) \leq 1$. A similar concept to autocorrelation is *cross-correlation*, which is a measure of similarity between *two* signals as a function of time lags in one of them:

$$R_{xy}(l) = \sum_{n=-\infty}^{\infty} x_n \cdot y_{n-l}^* = x_l \otimes y_{-l}^* \, . \tag{3.2}$$

Similarly to autocorrelation, we work with a normalized estimate, that is $-1 \leq R_{xy}(l) \leq 1$.

3.1.3 Frequency analysis

An important tool used extensively in signal processing is *frequency analysis*: the original time domain signal is represented in the frequency domain using a linear combination of complex exponential signals. The frequencies present in the time-domain signal constitute the *spectrum*. The representation of a discrete-time signal in the frequency domain is achieved using the *discrete time Fourier transform* $\mathcal{F}(\mathbf{x})$, which is defined as:

$$\mathcal{F}(\mathbf{x}) = \mathrm{X}(\omega) = \sum_{n=-\infty}^{\infty} x_n \cdot \exp(-j\omega n). \tag{3.3}$$

The function $\mathrm{X}(\omega)$ is periodic with period 2π, and the frequency range of the signal is bounded in the region $0 \leq \omega < 2\pi$ due to sampling effects. In practice, the spectrum $\mathrm{X}(\omega)$ is evaluated at N frequency points denoted by $\omega_\kappa = 2\kappa\pi/N$ where $1 \leq \kappa \leq N$. Given that all practical signals are of finite length, the time domain signal is expressed as a function of κ:

$$\mathcal{F}(\mathbf{x}) = \mathrm{X}(\kappa) = \sum_{n=1}^{N} x_n \cdot \exp\left(-j\frac{2\kappa\pi}{N}n\right). \tag{3.4}$$

We can recover the original time domain signal from the frequency domain signal using the inverse Fourier transform:

$$\mathcal{F}^{-1}\big(\mathrm{X}(\kappa)\big) = \mathbf{x} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{X}(\kappa) \cdot \exp\left(j\frac{2\kappa\pi}{N}n\right). \tag{3.5}$$

In practice, there is a fast method to compute Eq. (3.4), using the *Fast Fourier Transform* (FFT) (Proakis and Manolakis, 1996). Interestingly, there is an algorithmic relationship between FFT and autocorrelation: the autocorrelation can be computed from the inverse FFT of the *power spectral density* $|\mathrm{X}(\omega)|^2$:

$$R_{xx}(l) = \frac{1}{N}\sum_{n=1}^{N} |\mathrm{X}(\omega)|^2 \cdot \exp\left(-j\frac{2\kappa\pi}{N}n\right), \qquad 1 \le \kappa \le N. \tag{3.6}$$

3.1.4 Probabilities and probability density estimation

The ubiquitous concept of *probability* appears often in everyday life. It expresses the possibility that an event (or even group of outcomes) A occurs, and is denoted by $\mathrm{P}(A)$. The *event* A is one out of the total possible outcomes of the *sample space* Ω, with $0 \le \mathrm{P}(A) \le 1$, and $\mathrm{P}(\Omega) = 1$. Then, a *random variable* X is defined as the function of each possible outcome $a \epsilon \Omega$ so that $X(a) = x \epsilon \mathcal{H}$, where $\mathcal{H}$ is known as the *alphabet* of possible real numbers[13]. Similarly, we can define the *conditional probability* of event A occurring given another event B as $\mathrm{P}(A|B)$. At this time, we need to define the *probabilities* of the possible x of X, that is, the *distribution* of potential values x to determine how probable they are. This distribution is known as a *probability density function* $p(x)$[14] and informally expresses how likely x is when observing X. Now, if the random variable X is *normally distributed*, we get a bell-shaped curve, which is known as the normal or *Gaussian* probability density function (Fig. 3.1).

The Gaussian probability density function has many properties which facilitate mathematical analysis, and is therefore very commonly used. It has the analytic form $p(x) =$

[13] We use the common mathematical convention of assigning capital letters (e.g. X) to denote the random variables, and lower case (e.g. x) to denote their numerical values. Note that in the context of this thesis, the random variables always take real values, i.e. $X \epsilon \mathbb{R}$.

[14] Strictly speaking, for discrete random variables the distribution is known as a *probability mass function*, but we will adopt the general term throughout this thesis.

$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ and can be conveniently characterized unambiguously using solely the first and second order *central moments*, that is the *mean* μ and the *variance* σ^2, and is typically represented in the form $X\sim\mathcal{N}(\mu,\sigma^2)$. The positive square root of the variance is known as the *standard deviation* σ.

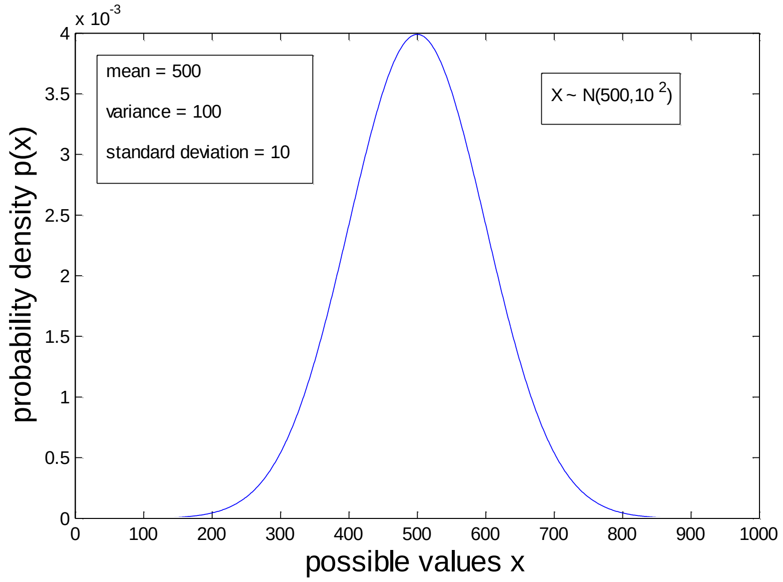

Fig. 3.1: Gaussian probability distribution function $\mathrm{X}\sim\mathcal{N}(500,\ 10^2)$.

The mean, the variance and in general the central moments of a random variable X are:

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x \cdot p(x)dx \tag{3.7}$$

$$Var(X) = \sigma_X^2 = E[(X - \mathrm{E}[X])^2] \tag{3.8}$$

$$Moment_X^{(m)} = E[(X - \mathrm{E}[X])^m]. \tag{3.9}$$

The expectation operator $E[\cdot]$ is computed from the possible values in X multiplied by their probabilities. The mean denotes the average and the variance denotes the dispersion of values around the mean. It is often useful to study the relationship between two random variables

X, Y. Similarly to the single random variable case where we used central moments, this is now achieved using the *central joint moments* of X, Y. Thus, one standard measure is the *covariance* between X, Y which is:

$$Cov(X,Y) = \mathrm{E}[(X - \mathrm{E}[X]) \cdot (Y - \mathrm{E}[Y])] = \mathrm{E}[X,Y] - \mathrm{E}[X] \cdot \mathrm{E}[Y]. \tag{3.10}$$

In general, probability density estimation is essential in analysing data (more about data analysis in Chapter 4), because it succinctly presents the probability of all possible values the random variable can have. If the realisations of the random variable follow a distribution known beforehand, then we can write down the analytic form of the distribution and fit the parameters using the data: this setting is known as *parametric* density estimation. More often than not, however, there is no prior knowledge and little can be assumed about the form of the distribution. Thus, typically probability density estimation uses a *nonparametric* setting, where the distribution is determined by the data itself (i.e. there is no prior analytic form where we try to fit parameters in a pre-specified equation).

A simple nonparametric approach to estimate the probability density of a random variable X from a collection of observations $\{x_i\}_{i=1}^{N}$ is to use a *histogram*, introducing an arbitrary number K of (usually equally spaced) bin edges $\mathbf{b} = (b_1, b_{2,} \dots, b_K)$, and counting the number of data observations x_i that are lying within the bin edges. Typically we set $b_1 = \min(\{x_i\}_{i=1}^{N})$ and $b_K = \max(\{x_i\}_{i=1}^{N})$. Then, the probability density estimate for each bin is $\frac{\# x_i \in \mathbb{Q}(x_0)}{Nw}$, with $\mathbb{Q}(x_0)$ being a small region (bin) around each x_0 of width w (equal to the difference between two successive bin edges, which is constant when using equally-spaced bin edges), and N is the total number of samples. The greatest virtue of histograms is their simplicity, making few prior assumptions about the distribution of the data. On the other hand, one problem with histograms is how to determine the width and bin edges, which affects considerably the final probability density estimate. In addition, the probability density estimate is usually not smooth, and for this reason *kernels* are often used, which can be visualized as smoothing windows (Hastie et al., 2009).

Fitting a kernel over each observed point x_i and subsequently adding these kernels provides a smoother estimate of the probability density, and is known as *kernel density estimation*. Typically, *Gaussian kernels* are used which have the form presented in Fig. 3.1, where the peak of the kernel is the point of the observation, and the bandwidth (denoted through the kernel's standard deviation σ) remains to be determined. The use of Gaussian kernels is associated with E. Parzen (1962), and the terms *Parzen window* and *Parzen density estimate* are commonly also used. Effectively, the Parzen density estimate adds independent Gaussian noise to each data point observation x_i, and is defined as:

$$\hat{p}(x_0) = \frac{1}{N\sqrt{2\pi\sigma^2}} \sum_{\iota=1}^{N} \exp\left[-\frac{(\|x_i - x_0\|)^2}{2\sigma^2}\right] \tag{3.11}$$

where ||·|| denotes the distance between the observation sample x_i from the point x_0 where we want to obtain a local density estimate, and σ is the bandwidth of the kernel. The set where $\hat{p}(x_0) > 0$ is known as the *support set* of X. Generalising Eq. (3.11) to estimate the *joint* probability density estimate of ρ random variables (thus working in $\mathbb{R}^{\rho}$) gives:

$$\hat{p}(x_0) = \frac{1}{N(2\pi\sigma^2)^{\frac{\rho}{2}}} \sum_{i=1}^{N} \exp\left[-\frac{(\|x_i - x_0\|)^2}{2\sigma^2}\right]. \tag{3.12}$$

In both (3.11) and (3.12) the bandwidth of the Gaussian kernel σ is a free parameter. There are various approaches to select it, some of which are based on assumptions about the underlying distribution and length of the data, e.g. Silverman's rule of thumb (Silverman, 1986). For a recent paper that describes different methods to determine the bandwidth of the kernel, refer to Shimazaki and Shinomoto (2010). It is possible to use a different kernel (not Gaussian), but practice has shown that density estimates are considerably more sensitive to the bandwidth of the kernel rather than the kernel type (Gray and Moore, 2003). Density estimation is an interesting area of research, but is beyond the scope of this study. As we shall see, speech signal processing algorithms that rely on density estimates typically use histograms for simplicity.

3.1.5 Uncertainty and entropy

An additional important notion in various disciplines is the concept of *entropy*, which is one way of representing the *uncertainty* in a quantity. The entropy H quantifies the average *information content* of a probability density function of a random variable X and is often associated with Shannon's pioneering work, establishing the discipline of information theory (Shannon, 1948). Some texts use different terms to refer to the entropy of a continuous random variable (*differential entropy*) and a discrete random variable (*discrete entropy*). In this thesis, we will not use separate terms: the following equations are defined for continuous random variables and use integrals; for discrete random variables we would use summations instead. The entropy of a continuous random variable is defined as:

$$H(X) = -\int_S p(x) \cdot \log_b p(x)\, dx \tag{3.13}$$

where S is the support set of X. If the base b in the expression of the logarithm is 2, the entropy is measured in units of *bits*, whereas if $b = e$ (i.e. the natural logarithm) gives units of *nats*. The entropy spans the range $0 \leq H(X) \leq \log_b N$, where $H(X) = 0$ implies no uncertainty and $H(X) = \log_b N$ denotes the maximum uncertainty regarding the expected outcome in a given trial. Eq. (3.13) can be generalized to compute the *joint entropy* of X, Y:

$$H(X,Y) = -\iint_S p(x,y) \cdot \log_b p(x,y)\, dxdy \tag{3.14}$$

where S is the support set of X, Y. Often, we are interested in expressing the uncertainty of a variable Y with respect to X, i.e. the *conditional entropy*:

$$H(Y|X) = -\iint_S p(x,y) \cdot \log_b p(y|x)\, dxdy. \tag{3.15}$$

Now, we turn our attention to speech signals, and review the speech signal processing algorithms which are widely used to extract clinically useful information.

3.2 Speech signal processing algorithms

In principle, any signal processing tool and any time series analysis tool could be used to extract *characteristics* from the speech signal which might be useful for clinical applications. As mentioned in the previous chapter, we will refer to the techniques used to extract characteristics from speech signals as *dysphonia measures*, or simply as *measures*. Many dysphonia measures require the prior computation of fundamental frequency, so we will first describe approaches to estimate this intricate characteristic of speech signals. Subsequently, we will describe linear speech signal processing algorithms which are well developed and established, and currently dominate speech signal analysis. We will then describe nonlinear speech signal processing algorithms that have, relatively recently, attracted great interest and demonstrated their value as useful clinical tools.

3.2.1 Fundamental frequency estimation

The accurate estimation of the fundamental frequency is critical to characterize speech signals (Christensen and Jakobsson, 2009), and has led to the development of various F_0 estimation algorithms (some researchers prefer the term *pitch detection algorithms* abbreviated as PDA[15] and these terms will be used interchangeably throughout this thesis). In general, there may be different requirements in F_0 estimation depending on the application, or the assumptions researchers are willing to make. For example, in some cases it is necessary to consider the presence or not of *voiced speech* (i.e. glottis-induced movement), computational complexity, or the specific needs of the particular application (e.g. singing, speech coding in digital communications, and clinical assessment). Roark (2006) highlights the existence of more than 70 methods to extract F_0, which reflects both the importance and difficulty of the problem. Roark argued there is no single 'correct' method for F_0 extraction, because, fundamentally, there is no definition of what F_0 means if it does not just mean "signal period". In most practical applications we are interested in computing the time-varying F_0, i.e. observing how F_0 changes during the phonation. This F_0 time series is better known in the speech signal processing literature as the F_0 *contour*.

Typically, most PDAs have three main components (Talkin, 1995): (a) pre-processing, (b) identification of possible F_0 candidates, and (c) post-processing, to decide on the final F_0 estimate. The pre-processing step depends on the actual PDA requirements. One example for

[15] The reader may recall that *pitch* was defined as the psychoacoustic equivalent of the fundamental frequency (see Chapter 2). In practice, PDA and F_0 *estimation* are often interchangeably used, although strictly speaking PDA is a misnomer since inherently estimating F_0 is not a detection problem, but rather an estimation problem. Hence, the term *pitch estimation algorithm* is also met in practice.

pre-processing is low-pass filtering of the speech signal to remove formants, but this is a double-edged sword: reducing the bandwidth increases the inter-sample correlation and could be detrimental to PDAs which detect periodicity using correlations. Post-processing is typically used to select the most likely candidate from the pool of F_0 candidates or to refine the F_0 contour estimates by smoothing sudden jumps in successive F_0 estimates. Large changes in neighboring F_0 estimates may not be physiologically realistic in most applications, although this is not universally true which further complicates this step. One straightforward and simple approach is to use *median smoothing* or *dynamic programming*, and we will see both approaches later in this chapter when we describe specific PDAs.

There is no single best PDA for *all* applications, and here we will mention some of the most interesting algorithms which have gained wide acceptance in the speech signal processing community. These algorithms work either in the time domain (mostly using autocorrelation and some using cross-correlation approaches), or in the frequency domain (frequency spectrum and cepstral approaches). In a few cases, PDAs combine both time-domain and frequency-domain information to obtain more reliable F_0 estimates. A further division for time domain approaches can be on the grounds of whether PDAs work on short-time average windows (*local* estimates) or detect single glottal cycles (*instantaneous* estimates). PDAs operating on short-time windows are typically applied to a small, pre-specified segment of the signal, and the F_0 estimates are obtained for adjacent non-overlapping windows (e.g. 10 ms). A further differentiation of PDAs can be made on the strategy used to estimate F_0: the most common are *peak picking* (for example identifying successive negative or positive peaks), and *waveform matching* (matching cycle to cycle waveforms). The overall consensus is in favour of waveform matching because of its improved robustness against noise (Titze and Liang, 1993; Boersma, 2009). For a more detailed background on F_0 estimation we refer to Talkin (1995), and Gerhard (2003).

Table 3.1 Summary of the F_0 estimation algorithms used in this study

Algorithm	Brief explanation	Implementation used
DYPSA (Naylor et al., 2007)	Identifies glottal closure instances, time domain approach	http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
PRAAT (Boersma, 1993)	Time-domain approach, using autocorrelation	http://www.fon.hum.uva.nl/praat/
YIN (de Cheveigne and Kawahara, 2002)	Time-domain approach, using autocorrelation	http://audition.ens.fr/adc/
RAPT (Talkin, 1995)	Time-domain approach, using cross-correlation	http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
SHRP (Sun, 2002)	Frequency domain approach, using sub-harmonics to harmonics ratio, aims to determine pitch	http://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm
SWIPE (Camacho and Harris, 2008)	Frequency-domain approach, aims to determine pitch	http://www.cise.ufl.edu/~acamacho/publications/swipep.m
TEMPO (Kawahara et al., 1999)	Frequency-domain approach, using Gabor filter banks, instantaneous F_0 estimates	Source code provided by H. Kawahara (not publicly available)
NDF (Kawahara et al., 2005)	Combines time-domain and frequency domain cues, instantaneous F_0 estimates	Source code provided by H. Kawahara (not publicly available)
XSX (Kawahara et al., 2008)	Frequency domain approach	Source code provided by H. Kawahara (not publicly available)
Ensemble approach (this study)	Combines the F_0 algorithms to obtain an improved outcome	http://people.maths.ox.ac.uk/tsanas

Table 3.1 summarizes the F_0 estimation algorithms studied here, presents their main characteristics, and refers to the implementation used. In the following sections we present these PDAs in more detail. The selection of the ten PDAs investigated in this study is partly guided by the availability of open source-code implementations and their extensive use in the

speech signal processing literature. In all cases we used the default parameter settings for the PDAs, choosing, where appropriate, the F_0 search range to be $F_{0,\min} = 50$ Hz and $F_{0,\max} = 500$ Hz. Although the expected physical maximum F_0 cannot, realistically, be so high in the case of comfortably-produced sustained vowel /a/ signals (the exclusive focus of this study), we wanted to test the full range of inputs to the PDAs. Since this study only deals with voiced speech and there is no need to identify whether parts of the speech signal are voiced or unvoiced *frames* (segments of the original speech signal, usually pre-specified within an algorithm with a duration of a few milliseconds), that interesting part of the PDAs will not be addressed at all. To avoid putting F_0 estimation algorithms that use the voice/unvoiced detection step as part of the estimation process at a disadvantage, this option was disabled when possible. The segmentation of the speech signal is achieved using an appropriate window function to mitigate the effects of *spectral leakage*[16].

3.2.1.1 DYPSA

The Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) was proposed by Kounoudes et al. (2003) and later refined in Naylor et al. (2007). Contrary to the PDAs which will be described in the following sub-sections, DYPSA aims at identifying the glottal cycles directly instead of providing F_0 estimates at pre-specified time intervals. From the identified glottal cycles, we can infer F_0. Specifically, DYPSA aims at detecting the glottal closure instances (GCI – collision of the vocal folds) in voiced speech, since the glottal opening

[16] By selecting a segment of a signal (also known as *truncating*) and taking its Fourier transform (FT), frequencies which did not exist in the original signal appear in the frequency domain – this phenomenon is known as *spectral leakage*. Spectral leakage occurs due to the clash between truncation and the periodic continuation assumption of the FT, which breaks down since there is always a finite length signal. Therefore, window functions are used to reduce the effects of frequencies other than those in the initial signal appearing in the signal spectrum. In short, spectral analysis involves a trade-off between resolving signals with comparable strength and similar frequencies, and resolving signals with disparate strength and dissimilar frequencies. This trade-off of resolution and sensitivity guides the selection of the window function; typically the Hamming window is a good choice. For more details refer to Proakis and Manolakis (2006).

instances (GOI – separation of the vocal folds) are more difficult to detect because the excitation energy is typically both weaker and more dispersed in time compared to GCIs (Backstrom et al., 2002). In short, DYPSA identifies many GCI candidates, and a post-processing step using dynamic programming recovers the most likely "true" GCIs.

The first step in DYPSA is to compute the linear prediction residual signal $\mathbf{u}$ from the original speech signal (this step aims to minimize the autocorrelation), and subsequently segment the residual using a sliding Hamming window in order to obtain $\mathbf{a}$. Then, using the FFT of the segmented signal we obtain $\mathrm{A}(\omega)$, and the phase-slope $\tau_n(\omega)$ which is defined as:

$$\tau_n(\omega) = \frac{darg\big(\mathrm{A}(\omega)\big)}{d\omega}. \tag{3.16}$$

Effectively, the phase-slope is the average slope of the phase spectrum of the FFT of a segmented part of the linear prediction residual signal. Then, $\tau_n(\omega)$ can be sampled at specific frequencies. The GCIs are identified as positive-going zero crossings in the phase slope. In addition, Naylor et al. (2007) reported that DYPSA is able to recover some additional true GCIs by including GCI candidates where the phase slope failed to cross zero. These additional GCI candidates are obtained by taking the mid-point between a local minimum and a local maximum in the phase slope, and projecting this point with unit slope in the time axis. The reasoning is that the inclusion of additional (potentially faulty) GCIs is preferable to missing true GCIs: dynamic programming as a post-processing step can eliminate candidates, but cannot recover true GCIs not included in the candidate list.

The dynamic programming in DYPSA is effectively an optimization problem, aiming to determine the most likely GCIs from the list of GCI candidates by penalizing a number of attributes. These attributes include a *speech waveform similarity cost* (successive cycles are expected to have similar excitation characteristics), a *pitch deviation cost* (the lag times at which successive GCIs occur is expected not to differ massively), a *projected candidate cost* (GCI candidates resulting from projections are penalized as less probable GCIs compared to GCIs that crossed zero), and a *normalized energy cost* (to minimize spurious noise events).

3.2.1.2 PRAAT

The PRAAT algorithm uses the autocorrelation function to obtain F_0 estimates (Boersma and Weenink, 2009) and has been originally proposed by Boersma (1993). Typically, an overlapping *time window* of 40-80 ms is applied to the speech signal **x** breaking it in N segments (also known as *frames*), where $N \geq \frac{signal\ duration}{time\ window}$ (the equality holds for non-overlapping windows). For each of these N segments we subtract its average value and the resulting signal **s** is multiplied by an appropriate window **w** (PRAAT uses the Hanning or Gaussian window) and gives rise to N *windowed signals* **a**. The pitch period is identified for each of the windowed signals by taking the mid-sample within each window signal and finding the most correlated time instant l (with the exception of $l = 0$, which is by definition the global maximum), which we shall call l_{max}. PRAAT computes the autocorrelation of the signal segment **s** as $\mathbf{r}(l) \cong \frac{\mathrm{R_{aa}(l)}}{\mathrm{R_{ww}(l)}}$, where $\mathrm{R_{aa}}(l)$ is the autocorrelation of the windowed signal, and $\mathrm{R_{ww}}(l)$ is the autocorrelation of the window **w**. Boersma (1993) attributes the success of PRAAT in providing accurate F_0 estimates to this normalization where the signal autocorrelation is divided by the autocorrelation of the window, noting "*…the need for this correction seems to have gone by unnoticed in the literature*". Moreover, Boersma compared different window functions and found that the Gaussian window is complicated but usually provides very good performance. Most PDAs in the literature working with frames of the speech signal use Hanning windows, so in our study we decided to test both PRAAT's default with the Hanning window (this will be referred to as PRAAT1) and using Boersma's suggestion with the Gaussian window (this will be referred to as PRAAT2).

The index l which lies between some minimum and maximum boundaries (typically the F_0 range $([F_{0,min}\ F_{0,max}])$ chosen is $F_{0,min} = 50$ Hz and $F_{0,max} = 500$ Hz, so $F_s/500 \leq l_{max} \leq F_s/50$). The value of the index l that maximizes $\mathbf{r}(l)$ (the l_{max}) provides a crude estimate of the pitch period in signal samples. Inspired by the Nyquist sampling theorem (Proakis and Manolakis, 2006), PRAAT then uses an interpolation based on the $sinc(x)$ function to refine the final estimate around $\mathbf{r}(l_{max})$. Effectively, this aims to overcome problems such as spurious spikes due to noise. We used a Matlab wrapper[17] to access the PRAAT program by Boersma and Weenink (PRAAT, 2009).

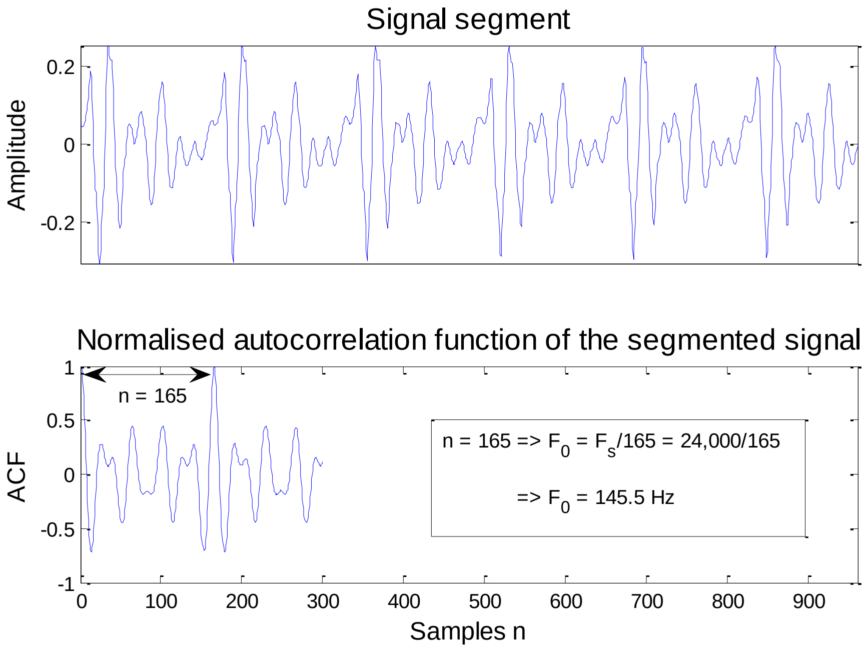

Fig. 3.2 Sustained vowel phonation and the PRAAT algorithm to determine F_0. For clarity in presentation we skip the last step of PRAAT interpolating around the point with the maximum autocorrelation. The sampling frequency of the present signal is 24 kHz.

[17] The Matlab wrapper for PRAAT was originally developed by Max Little.

PRAAT is one of the most popular software packages for speech signal processing and F_0 estimation amongst speech scientists. Fig. 3.2 presents the main idea of the PRAAT F_0 estimation algorithm in action.

3.2.1.3 YIN

Conceptually, YIN (de Cheveigne and Kawahara, 2002) is similar to PRAAT and also relies on the autocorrelation function to provide F_0 estimates in pre-specified time intervals. Assuming that the speech signal is periodic with period T, then by definition:

$$x_n - x_{n+T} = 0, \qquad \forall n. \tag{3.17}$$

We can then square Eq. (3.17) and average the function over a window with length w:

$$\sum_{j=n+1}^{n+w} \left(x_j - x_{j+T}\right)^2 = 0. \tag{3.18}$$

The unknown period of the signal can then be determined using Eq. (3.18) by finding the smallest value of τ where $d(\tau) = 0$:

$$d(\tau) = \sum_{j=1}^{W} \left(x_j - x_{j+\tau}\right)^2. \tag{3.19}$$

Now, expanding Eq. (3.19) we can express it in terms of the autocorrelation function:

$$d(\tau) = R_{xx}(0) + R_{xx+\tau}(0) - 2R_{xx}(\tau). \tag{3.20}$$

Using Eq. (3.18) de Chaveigne and Kawahara (2002) have shown a large improvement over simply using the autocorrelation. The justification for this improvement according to de Chaveigne and Kawahara is that Eq. (3.20) is more robust to amplitude perturbations due to the second energy term that varies with τ. To overcome the practical problem of possible overlapping of the first formant, F_1, with F_0, YIN uses a modified version of Eq. (3.20), effectively normalizing $d(\tau)$ with the average $d(\tau)$ over shorter time lags:

$$d'(\tau) = \left\{ \begin{matrix} 1, \text{if } \tau = 0 \\ d(\tau) \Big/ \left[(1/\tau) \cdot \sum_{\xi=1}^{\tau} d(\xi) \right], \text{if } \tau \neq 0 \end{matrix} \right\}. \tag{3.21}$$

Then, YIN uses thresholding of $d'(\tau)$ to overcome the effect of strong sub-harmonics (set to 0.1 according to the developers of YIN) and identify F_0 local estimates. YIN uses as a post-processing step which is similar to median smoothing, to avoid wild F_0 jumps.

3.2.1.4 RAPT

The Robust Algorithm for Pitch Tracking (RAPT) was proposed by Talkin (1995) and relies on the cross-correlation function to estimate F_0, but conceptually it is not very different from PRAAT. RAPT works in segmented time windows of the original signal, exactly as explained in PRAAT to obtain **s**, and provides F_0 estimates for each of those frames. Practically, it compares the original speech signal with a generated *sub-sampled* version of the original signal with new sampling frequency F_{ds}, and attempts to identify the maxima where the cross-correlation is close to 1 (with the exception of the point at zero lag). The F_{ds} of the resampled signal is a function of the F_s of the raw signal, and $F_{0,max}$:

$$F_{ds} = \frac{F_s}{round\left(F_s/4 \cdot F_{0,max}\right)}. \qquad (3.22)$$

RAPT computes the cross-correlation of the lower sampled signal for lags that fall within the F_0 range $\left(\left[F_{0,min}\ F_{0,max}\right]\right)$ and records the positions of up to 20 maxima within the examined frame which have a cross-correlation value above a certain threshold. Then, the cross-correlation is computed on the original raw speech signal in the vicinity (7 lags) of the promising maxima identified in the previous step. This two-stage approach is designed to reduce the computational load of RAPT. The identified peaks from this high resolution (second-step) cross-correlation computation correspond to the F_0 candidates for this frame.

Having identified the lag times where the potential F_0 candidates for each frame are located, RAPT then uses post processing to decide the most likely candidates. Specifically, RAPT uses dynamic programming (Ney, 1983) to penalize some quantities, such as F_0 transition cost for successive frames. Once the most likely lag point is determined, RAPT refines the F_0 estimate by parabolic interpolation of the three lag points around the determined peak.

3.2.1.5 SHRP

The SHRP algorithm (Sun, 2002) has been endorsed by speech scientists (Hunter, 2009) as perhaps one of the most accurate methods for F_0 estimation (E. Hunter, personal communication) alongside PRAAT. As we have seen, PRAAT works in the time domain and is based on autocorrelation; SHRP works in the frequency domain using the sub-harmonics to harmonics ratio (SHR), which is defined as the ratio between the amplitude of sub-harmonics and harmonics (a formal definition is provided later). Practice has shown that simply computing the spectrum and finding the lowest component (harmonic) does not provide a reliable F_0 estimate (Hess, 1991), which has prompted researchers to use the entire harmonic structure (Sun, 2002). The main motivation behind SHRP is to address problems that appear in F_0 estimation due to *alternate cycles* (adjacent vocal cycles fluctuating in amplitude and/or in period), which are manifested in the frequency domain as strong sub-harmonics. Effectively, SHRP is motivated by human pitch perception in identifying F_0: low SHR does not affect pitch, whereas relatively large SHR (which in itself indicates some sort of vocal pathology) leads to pitch perception of one octave lower than the actual F_0, which corresponds to the lowest sub-harmonic.

As with the other algorithms already discussed, SHRP operates on windowed versions of the signal (e.g. providing estimates every 10 ms). The starting point is to define the *sum of harmonic amplitude* (SHA) at multiples of F_0:

$$\text{SHA} = \sum_{n=1}^{S_{max}} \text{A}(n \cdot F_0) \tag{3.23}$$

where S_{max} is the maximum number of harmonics in the spectrum, and $\text{A}(f)$ represents the spectral amplitude at frequency f. Frequencies above the investigated upper threshold $F_{0,max}$ (as we noted in PRAAT this threshold is usually set to 500 Hz) are set to zero, i.e. $\text{A}(f) = 0$ if $f > F_{0,max}$. The number of harmonics is computed as $S_{max} = floor\left(f_{max}/F_{0,min}\right)$, where f_{max} is the maximum frequency of $\text{A}(f)$, and SHRP uses the default $f_{max} = 1250$ Hz.

The next step in SHRP is to define the sub-harmonics, assuming the lowest sub-harmonic is half the F_0. Then, the *sum of sub-harmonic amplitude* (SSA) is defined as:

$$\text{SSA} = \sum_{n=1}^{S_{max}} \text{A}\big((n - 0.5) \cdot F_0\big). \tag{3.24}$$

In healthy voices, SSA should be practically zero. Then, SHR is defined simply using the ratio of Eqs. (3.23) and (3.24):

$$\mathrm{SHR} = \mathrm{SHA}/\mathrm{SSA}. \tag{3.25}$$

However, Sun (2002) remarked that computing SHR from Eq. (3.25) is not trivial, and therefore proposed log-transforming the originally linear frequency scale before computing SHA and SSA, without, however, providing any particular justification. The log-transformation of F_0 is common in speech signal processing to smoothen vibrato by compressing the spectrum, and we will revisit this concept later in this chapter when we define some dysphonia measures (for example see § 3.2.3.4). Moreover, Sun's approach has a solid physiological basis in the context of F_0: pitch is perceived on a *logarithmic* scale because that is how the cochlea in the human ear works (Baken and Orlikoff, 2000).

Sun (2002) then proposed a criterion for deciding whether harmonics or sub-harmonics should be used to provide pitch estimates. Based on physiological experiments with human listeners who provided pitch estimates in synthesized vowels, he found that a value of SHR<0.2 indicates that sub-harmonics are weak, and that pitch and F_0 effectively coincide in those circumstances. He has shown that when SHR is very low, then pitch should be determined using the global maximum of the log-spectrum, whereas if SHR is relatively high (SHR>0.2), then pitch should be estimated by the next local maximum of the log-spectrum in the vicinity of the global maximum. Sun also proposed using a larger value of SHR when researchers prefer to emphasize the use of harmonics to detect F_0.

SHRP has the option to use median smoothing for post-processing the computed F_0 estimates in order to avoid large, spurious deviations across successive frames.

3.2.1.6 SWIPE

The Sawtooth Waveform Inspired Pitch Estimator (SWIPE) algorithm was recently proposed by Camacho and Harris (2008), and similarly to SHRP is a frequency domain approach. We should note that the aim of the algorithm is to detect *pitch*, which as we know (see Chapter 2), is not always equivalent to F_0. The main idea in SWIPE is similar to SHRP, but instead of focusing solely on the harmonic locations, it uses the information available in the entire spectrum using kernels. We need to clarify that the following description refers to SWIPE’ (swipe prime), an extension of SWIPE proposed in the original paper (Camacho and Harris, 2008), but the prime is omitted here for simplicity.

SWIPE works on frames of the speech signal, exactly as defined in PRAAT. On each frame, the square root of the spectrum is computed: using the square root instead of the square of the spectrum (this is the auto-correlation – see PRAAT for example) or the logarithm of the spectrum (for example in SHRP) avoids certain shortcomings. For example, PDAs using the square of the spectrum are prone to fail in the presence of salient harmonics (Rabiner, 1977), whereas PDAs using the logarithm of the spectrum are problematic in cases of missing harmonics. Then, SWIPE identifies the harmonics in the square root of the spectrum and imposes kernels with harmonically decaying weights. Camacho and Harris (2008) empirically found that the optimal results were obtained when using kernels with weight $1/\sqrt{k}$, where k corresponds to the kth harmonic. They reported that using Gaussian or cosine kernels did not markedly affect their findings. This finding is consistent with the literature in kernel density estimation, where the width of the kernel is a considerably more important parameter compared to the actual shape of the kernel (Silverman, 1986). Perhaps not surprisingly, transforming the frequency using the equivalent rectangular bandwidth (ERB) scale prior to imposing the kernels in the square root of the spectrum boosted the performance of SWIPE in estimating pitch. The ERB transformation of the frequency is inspired by the way the cochlea in the ear works, and is defined in Glasberg and Moore (1990):

$$\mathrm{ERB}(f) = 21.4 \cdot \log_{10}(1 + f/229). \tag{3.26}$$

Camacho and Harris (2008) had also experimented with alternative frequency transformation ideas (e.g. using the Mel scale as used in MFCCs which will be described in § 3.2.2.5), but have found the ERB transformation leads to more accurate pitch estimates.

3.2.1.7 TEMPO

The TEMPO algorithm was proposed by Kawahara et al. (1999) and works on the log-transformed frequency domain. A filter bank of equally spaced band-pass Gabor filters is imposed on the log-frequency axis, and is used to map the central filter frequency to the instantaneous frequency of the filter outputs. The original algorithm used 24 Gabor filters covering an octave. The instantaneous angular frequency $\boldsymbol{\omega}(t)$ is computed using the Hilbert transform, which is, effectively, the convolution of the signal $\mathbf{x}(t)$ with $1/t$; hence promoting the local properties of $\mathbf{x}(t)$.

Specifically, the instantaneous frequency is computed as follows:

$$\boldsymbol{\omega}(t) = \frac{d\boldsymbol{\theta}(t)}{dt} \tag{3.27}$$

where the phase of the signal $\boldsymbol{\theta}(t)$ is computed by combining and solving a system comprising the following three equations with three unknowns:

$$\left\{\begin{array}{c} \boldsymbol{\theta}(t) = \tan^{-1}\mathbf{y}(t)/\mathbf{x}(t) \\ \mathbf{x}(t) + j\mathbf{y}(t) = \mathbf{a}(t)\cdot\exp\big(j\boldsymbol{\theta}(t)\big) \\ \mathbf{a}(t) = \sqrt{\mathbf{x}^2(t) + \mathbf{y}^2(t)} \end{array}\right\}. \tag{3.28}$$

Each of the Gabor filters provides an instantaneous frequency estimate; then a selection mechanism is used to identify the filter that actually corresponds to the F_0 estimate. This is achieved using a carrier to noise (C/N) ratio estimation procedure, where the filter that has the largest C/N ratio ultimately provides the F_0 estimate. Kawahara et al. (1999) proposed further refinement of the F_0 estimate using parabolic time warping: effectively this is similar to the ideas we have already seen in SHRP and SWIPE where the PDA makes use of information distributed in all harmonics. Practically, this means that the filters with the highest C/N are allowed to "vote" for the F_0 estimate with weights determined by their C/N values. TEMPO, along with the following two PDAs presented below, is part of a software package for speech signal analysis known as STRAIGHT developed by H. Kawahara.

3.2.1.8 NDF

The Nearly Defect-free (NDF) F_0 estimation algorithm was proposed by Kawahara et al. (2005) and combines information from both time-domain and frequency-domain to provide estimates. NDF was originally conceived as an extension of TEMPO for demanding applications where it is difficult to identify F_0 such as in some forms of expressive speech and for musical instrument sounds. The algorithm combines an interval based extractor and an instantaneous frequency based extractor to determine the F_0 candidates, which are subsequently refined in a post-processing step.

The instantaneous frequency based extractor is similar to the extractor in TEMPO, the main difference being that it generates multiple F_0 candidates at each step instead of a single F_0 estimate. The frequency domain is scanned using channel pitch synchronized Gaussian filters to cover the user specified F_0 range $([F_{0,min}\ F_{0,max}])$. The interval based extractor computes autocorrelations at each frequency band using FFT, where the power spectra were initially normalized by their spectral envelope prior to the computation of the autocorrelations. The F_0 candidates are generated using the average autocorrelation at each frequency band, where the average autocorrelations are weighted as a function of the local signal to noise ratio.

Then, the F_0 candidates from the instantaneous based extractor and the interval based extractor are mixed using the normalized empirical distribution of side information to determine the most likely candidates. The final refinement of F_0 is identical to that described in TEMPO.

3.2.1.9 XSX

The eXcitation Structure eXtractor (XSX) was fairly recently proposed by Kawahara et al. (2008), and is the latest PDA in the software package STRAIGHT. They wanted to provide a fast alternative to NDF (see the preceding section), which their experiments demonstrated to be very accurate, but also computationally demanding. XSX relies on spectral division using two power spectral representations. The conceptual idea is that when a power spectrum that has periodic information is divided by its envelope, the result promotes F_0 estimates at pre-assigned F_0 candidates. XSX uses a set of F_0 detectors equidistantly placed on the log-frequency axis which cover the user specified F_0 range $([F_{0,min}\ F_{0,max}])$. We define the fluctuation spectrum, $P_C(\omega, t)$ according to Kawahara et al. (2008) as:

$$P_C(\omega,t) = \frac{P_T(\omega,t)}{P_{ST}(\omega,t)} - 1 \tag{3.29}$$

where
$$P_T(\omega,t) = \frac{|S(\omega,t-T_0/4)|^2 + |S(\omega,t+T_0/4)|^2}{2} \tag{3.30}$$

$$P_{ST}(\omega,t) = \exp\left[k_1 \cdot \left(L(\omega+\omega_0,t) + L(\omega-\omega_0,t)\right) + k_2 \cdot L(\omega,t)\right] \tag{3.31}$$

and
$$\begin{cases} L(\omega,t) = \ln[C(\omega+\omega_0/2,t) - C(\omega-\omega_0/2,t)] \\ C(\omega,t) = \int_{\omega_L}^{\omega} P_T(\lambda,t)d\lambda \end{cases} \tag{3.32}$$

where $P_T(\omega,t)$ is the TANDEM spectrum, $P_{ST}(\omega,t)$ is the STRAIGHT spectrum (interference-free spectrum), $S(\omega,t)$ is the FT of the segmented speech signal using a windowing function (e.g. Hanning window as we have seen in PRAAT), $\omega_0 = 2\pi F_0$ is the angular fundamental frequency, and k_1, k_2 are constants to ensure the positivity of the STRAIGHT spectrogram and depend on the window function used to segment the signal.

The TANDEM spectrum is by definition computed using the average of two power spectra of the same signal using time windows which are separated by half fundamental period T_0 (in practice the estimate is centrally placed ¼ before and after the time instant t we are interested in obtaining estimates).

3.2.1.10 A novel ensemble approach for fundamental frequency estimation

In the preceding sections (§ 3.2.1.1 – 3.2.1.9) we have presented ten popular PDAs, which are both freely available and have been shown to be competitive in the research literature. It is possible that *combining* the outputs of these PDAs could lead to improved F_0 estimation. *Ensemble learning* (combining individual estimators) is an active area of research in machine learning, where typically many classifiers are combined to produce a superior classifier (Kuncheva, 2004). We defer detailed discussion about classifiers and ensembles of classifiers for Chapter 4. Inspired by the many successes of ensembles in machine learning (Polikar, 2006), we recently proposed a simple F_0 ensemble approach (Tsanas et al., 2011a), remarking that this is a topic worth investigating in further detail in its own right. Here, we will investigate PDA ensembles more thoroughly.

The two simplest ensemble PDAs use the mean and the median estimate from all individual PDAs. This corresponds to the situation where all PDAs have equal weight in the final output, and does not take into account the accuracy of each PDA. This simple approach can produce quite competitive ensembles, against which more sophisticated ensemble schemes can be benchmarked. An alternative approach to the simple scheme which allows all PDAs to vote for the F_0 is to determine the most successful individual PDA subset according to some criterion, and allow only this subset to contribute to the final F_0 estimate. In its simplest form, all the PDAs in the selected subset are given equal weight in the ensemble. Selecting a PDA subset is effectively equivalent to the problem of *feature selection* (in this case the features are the PDAs), which will be discussed in detail in § 4.2.2. As we will see in that section, it is generally better to use a small fraction of all PDAs; reducing the number of PDAs may or may not increase the final accuracy of the ensemble, but it always saves on computational resources and simplifies the required computations. Here, we have experimented with two well-established feature selection methods: (a) Least Absolute Shrinkage and Selection Operator (LASSO) (see § 4.2.2.1), and minimum Redundancy Maximum Relevance (mRMR) (see § 4.2.2.2).

A more sophisticated ensemble approach consists of each PDA contributing with a different weight towards the final F_0 estimate. The weights for each PDA can be determined, for example, using a performance score with respect to the ground truth F_0. Alternatively, optimization algorithms can be used to determine the weights that minimize an error term. Two simple approaches which fall within this category use (a) ordinary least squares (OLS), or (b) iteratively reweighted least squares (IRLS) to determine the weights (Bishop, 2007). The choice of the criterion over which to optimize (e.g. to minimize the mean squared error of the mean absolute error), may be critical and promote different PDA ensembles. Moreover, prior feature selection can be used, before computing the weights for the selected PDAs.

Overall, we have investigated 12 PDA ensembles: (1) the mean F_0 from the ten PDAs, (2) the median F_0 from the ten PDAs, (3) the mean F_0 from the K best PDAs (the methodology to compute the optimal K is described in the following paragraph) using LASSO, (4) the median F_0 from the K best PDAs using LASSO, (5) the mean F_0 from the K best PDAs using mRMR, (6) the median F_0 from the K best PDAs using mRMR, (7) ensemble weights optimized for all 10 PDAs to minimize the squared error from the ground truth using OLS, (8) weights optimized for the ten PDAs to minimize the squared error term using IRLS, (9) weights optimized using the best K PDAs selected using LASSO, to minimize the squared error term using OLS, (10) weights optimized using the best K PDAs selected with LASSO, to minimize the squared error term using IRLS, (11) weights optimized using the best K PDAs selected with mRMR to minimize the squared error term using OLS, and (12) weights optimized for the best K PDAs using mRMR to minimize the squared error term using IRLS.

In § 5.1 we evaluate the performance of the PDAs and the ensemble PDAs in accurately detecting F_0 in a database consisting of 92 sustained vowel /a/ phonations, where the 'true' F_0 is known *a priori*. Further details about the validation setting of the ensemble PDAs, and how to optimize the weights and number of contributing PDAs in the ensembles will be provided in that section. For now, we remark that a 10% reduction in the mean absolute deviation from the ground truth F_0 was achieved when using an ensemble approach over the best individual PDA.

3.2.2 Classical dysphonia measures and minor algorithmic variants

Linear signal processing is a well-established discipline which dominates speech analysis. The term *linear* refers to a method where the output is proportional to a linear combination of the inputs, a property known as the *superposition principle*, which for a system S is expressed as:

$$S(a_1\mathbf{f} + a_2\mathbf{g}) = a_1 \cdot S(\mathbf{f}) + a_2 \cdot S(\mathbf{g}) \tag{3.33}$$

where a_1 and a_2 are arbitrary constants, and $\mathbf{f}, \mathbf{g}$ are arbitrary digital signals.

Conversely, *nonlinear* methods have more general relationships between the inputs and the output, that is, changes in the inputs produce complex effects in the output.

Because nonlinearity can introduce mathematical complexities, a common approach in the mathematical modelling or analysis of signals and systems in engineering contexts is to *linearize* the system. A nonlinear system with input $\mathbf{x}$ and output $\mathbf{y} = f(\mathbf{x})$, where f is a *smooth function*[18], can be expanded using the Taylor series close to the operating point of interest x_0, y_0:

$$\mathbf{y} = f(\mathbf{x}) = f(x_0) + \frac{df}{dx}(x - x_0) + \frac{1}{2!} \cdot \frac{d^2 f}{dx^2}(x - x_0)^2 + \cdots \tag{3.34}$$

where the derivatives are evaluated at x_0. Now, assuming $x - x_0$ is sufficiently small, the higher order terms of the Taylor expansion series in Eq. (3.34) can be eliminated and the system becomes linear. If this holds in reality, then linear relationships are an excellent

[18] *Smooth* functions have bounded derivatives of all orders.

approximation, which drastically facilitates the analysis. This is because linear signal processing is a well understood, mature field and includes powerful tools such as autocorrelation, cross-correlation, auto-covariance, cross-covariance, power spectrum analysis, linear prediction analysis, and power spectral density estimation, to name only a few. A concise discussion can be found in Little et al. (2006) and Little (2011); another standard general reference work is Proakis and Manolakis (2006). Thus, speech disorders have traditionally been assessed using classical, linear speech measures, a trend which has only recently begun to change. In the following sections, we review the most popular classical dysphonia measures. Although most of the following measures have fragmentally appeared previously in the literature, we summarize them along with their detailed algorithmic implementations in Tsanas et al. (2011a).

3.2.2.1 Jitter variants

By definition, *jitter* aims to quantify cycle-to-cycle F_0 *perturbations* (small deviations from exact periodicity), but lacks a rigorous, unequivocal formal definition (Titze, 2000) which has led to the development of many *jitter variants* (Schoentgen and de Guchteneere, 1995; Baken and Orlikoff, 2000). Jitter can be computed using either the F_0 contour, or the inversely proportional *pitch period* $T_0 = 1/F_0$ contour; researchers typically focus only on the latter. In Tsanas et al. (2011a) we investigated whether there would be noticeable differences in the quantification of the information in the speech signal using either the F_0 contour or the T_0 contour, and we found that neither approach led to improved quantification of vocal severity in PD. Specifically, the jitter variants we used are:

1) The mean absolute difference of F_0 estimates between successive cycles:

$$Jitter_{F_{0,\mathrm{abs}}} = \frac{1}{N}\sum_{i=1}^{N-1}\left|F_{0,\mathrm{i}} - F_{0,\mathrm{i+1}}\right| \tag{3.35}$$

where N is the number of F_0 computations. Eq. (3.35) is also the discrete *total variation*.

2) F_0 mean absolute difference of successive cycles divided by the mean F_0, expressed in percent (%):

$$Jitter_{F_{0,\%}} = 100 \cdot \frac{\frac{1}{N}\sum_{i=1}^{N-1}\left|F_{0,\mathrm{i}} - F_{0,\mathrm{i+1}}\right|}{\frac{1}{N}\sum_{i=1}^{N}F_{0,\mathrm{i}}}. \tag{3.36}$$

3) Perturbation quotient measures using K cycles (we used $K = 5$):

$$Jitter_{F_{0,\mathrm{PQ1,K}}} = \frac{\frac{1}{N-K+1}\sum_{i=K_1}^{N-K_2}\left[\frac{1}{K}\sum_{j=\mathrm{i}-K_2}^{i+K_2}\left|F_{0,\mathrm{j}} - F_{0,\mathrm{i}}\right|\right]}{\frac{1}{N}\sum_{i=1}^{N}F_{0,\mathrm{i}}} \tag{3.37}$$

where $K_1 = \mathrm{round}(K/2)$ and $K_2 = K - K_1$.

$$Jitter_{F_{0,\mathrm{PQ2,K}}} = \frac{\frac{1}{N-K+1}\sum_{i=K_1}^{N-K_2}\left[\frac{1}{K}\sum_{j=\mathrm{i}-K_2}^{i+K_2}F_{0,\mathrm{j}} - F_{0,\mathrm{i}}\right]}{\frac{1}{N}\sum_{i=1}^{N}F_{0,\mathrm{i}}}. \tag{3.38}$$

4) Perturbation quotient using an autoregressive model:

$$Jitter_{F_{0,\mathrm{PQ3,K}}} = \frac{\frac{1}{N-p}\sum_{i=p+1}^{N}\left|\sum_{j=\mathrm{i}-p}^{i} a_j \cdot \left(F_{0,\mathrm{j}} - \frac{1}{N}\sum_{i=1}^{N}F_{0,\mathrm{i}}\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}F_{0,\mathrm{i}}} \tag{3.39}$$

where $\left\{a_j\right\}_{j=1}^{p}$ are the autoregressive model coefficients, which were estimated from the F_0

contour using the Yule-Walker equations (Chatfield 2004). We used $p = 5$ coefficients, following Schoentgen and de Guchteneere's (1995) suggestion. Eq. (3.39) is effectively the generalization of Eq. (3.36), quantifying the absolute (weighted) average difference between the mean F_0 estimate and the F_0 estimate of the previous p time windows, instead of quantifying only the average absolute difference between two successive F_0 estimates. Conceptually, higher order differences are used to smooth vibrato (we will revisit the concept of smoothing vibrato in some other dysphonia measures later, for example § 3.2.3.4).

5) Mean absolute and normalized mean squared perturbations:

$$Jitter_{\mathrm{F_{0,p1}}} = \frac{1}{N}\sum_{i=1}^{N-1}\left|F_{0,\mathrm{i}} - \frac{1}{N}\sum_{j=1}^{N} F_{0,\mathrm{j}}\right| \tag{3.40}$$

$$Jitter_{\mathrm{F_{0,p2}}} = \frac{\frac{1}{N}\sum_{i=1}^{N-1}\left(F_{0,\mathrm{i}} - F_{0,\mathrm{i+1}}\right)^2}{\left(\frac{1}{N}\sum_{i=1}^{N} F_{0,\mathrm{i}}\right)^2}. \tag{3.41}$$

Additional jitter-like measures were computed using the standard deviation of the F_0 contour (which can be computed with any of the PDAs described in § 3.2.1). We also calculated the difference between the mean F_0 from the F_0 estimation algorithm with the average F_0 of age- and gender- matched healthy controls: this information was summarized in Fig. 2.5. In addition, we computed *frequency modulation* (FM) (Titze, 2000):

$$\mathrm{FM} = \frac{\max\left(F_{0,i}\right)_{i=1}^{N} - \min\left(F_{0,i}\right)_{i=1}^{N}}{\max\left(F_{0,i}\right)_{i=1}^{N} + \min\left(F_{0,i}\right)_{i=1}^{N}}. \tag{3.42}$$

We also calculated the range of F_0 using the 5th and 95th percentiles: $F_{0,r} = F_{0,95\,percentile} -$

$F_{0,5\ percentile}$: this way we do not take into account the entire range of F_0 values, and hence $F_{0,r}$ is more robust to outliers (spurious occasional F_0 estimates which appear in the form of spikes in the F_0 contour).

We also analyzed the F_0 contour using the nonlinear *Teager-Kaiser energy operator* (TKEO) Ψ (Kaiser 1990), and computed the mean, standard deviation and 5th, 25th, 75th and 95th percentile values of $\Psi(F_0)$. Ψ is defined as:

$$\Psi(x_n) = x_n^2 - x_{n+1} \cdot x_{n-1}. \tag{3.43}$$

TKEO quantifies the *amplitude modulation* (AM) and the *frequency modulation* (FM) content of an oscillating signal $x_n = A_n \cdot \cos(\omega_n)$: $\Psi(x_n) \cong A_n^2 \cdot \sin^2(\omega_n)$, where $\omega = 2\pi f$ is the frequency of the signal in rad/s and f is the frequency in Hz. Consequently, TKEO is proportional to the instantaneous amplitude and instantaneous frequency of the analysed signal, and has found wide applicability in speech signal processing (Maragos et al., 1993). Similarly to other generic operators such as energy and entropy, TKEO can be applied to any time series. Here, we simply fed the F_0 contour into Eq. (3.43) to compute the nonlinear energy of the fundamental frequency. Ψ can be directly contrasted to the standard linear signal processing approach of the *instantaneous* energy of a signal based on the *squared energy operator* (SEO) x_n^2. Both SEO and TKEO have been used in parallel in this study to directly compare their effect as part of the dysphonia measures. We also refer the reader to Dimitriadis et al. (2009) for a recent study which compared SEO and TKEO applied to both synthetic and real speech signals. Overall, they have found that TKEO outperforms SEO in most cases, a finding that is in agreement with ours, reported in Tsanas et al. (2011a).

The jitter variants developed so far used the F_0 contour in the computations. As mentioned earlier, we can substitute F_0 with T_0 and recast the corresponding measures: it can easily be verified that the algorithmic results in that case are not simply inversely proportional to those based on F_0. Thus, in principle, measures based on T_0 could provide independent information about the speech signal relevant to the purposes of this study.

3.2.2.2 Shimmer variants

In the preceding section we defined jitter as the cycle-to-cycle F_0 perturbations. *Shimmer* is the analogue of jitter for the *amplitude* of the speech signal, rather than F_0. We have used the same calculations presented in the preceding section for the jitter variants, but using the amplitude A_0 contour instead of the F_0 contour in Eqs. (3.35) - (3.42) to derive the *shimmer variants*. For the computation of the A_0 contour we first used DYPSA to obtain the glottal cycles (see § 3.2.1.1 for a description of DYPSA). Then, we defined the A_0 contour using the maximum amplitude value within each glottal cycle. Alternatively, we can define the A_0 contour by focusing on signal segments (e.g. 40 ms) instead of within glottal cycles, or using the minimum amplitude values. The only difference of the shimmer variants compared to the jitter variants is that we have used $K = 3, 5,$ and 11 in Eqs. (3.37) - (3.40) to conform with traditional amplitude perturbation quotient measures as used by standard reference software programs such as PRAAT. An additional shimmer-variant acoustic measure that we computed is shimmer in decibels (dB), since this has often been previously used:

$$Shimmer_{\mathrm{dB}} = \frac{1}{N}\sum_{i=1}^{N-1} 20 \cdot \left|\log_{10}\frac{A_{0,\mathrm{i}}}{A_{0,\mathrm{i+1}}}\right|. \tag{3.44}$$

This concludes the presentation of the two most popular perturbation algorithms (jitter and shimmer) in speech signal processing.

3.2.2.3 Harmonics to Noise Ratio (HNR) and Noise to Harmonics Ratio (NHR)

Harmonics-to-Noise Ratio (HNR) and *Noise-to-Harmonics Ratio* (NHR) are also commonly used measures, aiming to express the amount of noise in the speech signal. They can be considered part of the third large group of dysphonia measures which aim to characterise the signal using *signal to noise ratio* (SNR) approaches. The motivation behind HNR and NHR is to quantify noise in the speech signal, which is caused mainly as a result of incomplete vocal fold closure. Similarly to jitter and shimmer, HNR and NHR have many variants which have been reviewed in Ferrer et al. (2006); we used the definition by Boersma and Weenink (2009) in this study. We computed both the mean and standard deviations of HNR and NHR (typically, only the mean values are used in the literature but expressing also the spread of the noise estimates might reveal additional useful information).

Specifically, similarly to PRAAT defined in § 3.2.1.2, we start with the computation of the (normalized) autocorrelation. There, we defined l_{max} to correspond to the sample that provided the global maximum of the autocorrelation (with the exception of zero lag). Conceptually, for a signal without noise, the autocorrelation at the instant $R_{xx}(l_{max})$ should be 1. Then, the PRAAT definition for HNR is:

$$\mathrm{NR(dB)} \qquad \left[R \quad (l \qquad)/\left(1 \qquad (l \qquad)\right)\right]. \tag{3.45}$$

Similarly, NHR is defined as:

$$\left[\left(1 \qquad (l \qquad)\right)/R \quad (l \qquad)\right]. \tag{3.46}$$

3.2.2.4 Linear Predicting Coding Coefficients (LPCC)

In many time series analysis applications, it is often desirable to express future samples of a time series signal as a linear combination of previous values of the same signal. This tool is more widely known as the *auto-regressive model* (AR model) (Chatfield, 2004). Linear Predictive Coding (LPC), a widely used speech analysis technique in applications such as low-bit rate speech compression, is effectively the application of an AR model in speech signals. Conceptually, LPC leans heavily on the linear source-filter theory of voice production. Specifically, a sample of the speech signal s_n is predicted from the past p samples:

$$\hat{s}_n = \sum_{l=1}^{p} a_l \cdot s_{n-l}. \tag{3.47}$$

The parameters $\{a_l\}_{l=1}^{p}$ are the Linear Predictive Coding Coefficients (LPCCs) which are typically obtained using the least squares method or the Yule-Walker equations in order to minimize the squared error between predicted and actual values (Chatfield, 2004). The difference of the actual value s_n and the predicted value $\hat{s}_n$ is known as the *residual*: $e_n = s_n - \hat{s}_n$.

The reliability of LPCCs as accurate markers in characterizing disordered speech signals depends on how valid the linear source-filter theory is in vocal disorders. Given that the linear source-filter theory has lost many of its adherents even for healthy voices (Titze, 2000), their use in analysing voice disorders may be questionable. Nevertheless, LPCCs are widely used in speech signal processing (for example in speech coding), and hence were also studied here.

3.2.2.5 Mel Frequency Cepstral Coefficients (MFCC)

So far the measures described are targeted mainly at characterizing the vocal fold dynamics as the organs of speech production affected in PD. However, as we have noted in Chapter 2, research into PD has shown that co-ordination of the articulators of the vocal tract, such as the tongue, jaw, and lips, are also affected in addition to the vocal folds (Ho et al., 1998). The linear source-filter theory suggests that the sound produced by the vocal organs is the result of the convolution of the output of the vocal folds and the vocal tract *impulse response*. Therefore, the recorded signal **x** needs to be *deconvolved* to distinguish the vocal fold and the vocal tract parts in order to analyze them separately. *Deconvolution* is the inverse of convolution, is *ill-posed* having no unique solution, and there are a number of techniques for deconvolving two signals using linear or nonlinear algorithms. The deconvolution of speech signals is often performed using *cepstral domain* analysis because of its simplicity. The cepstral domain is defined as the inverse Fourier transform of the logarithm of the power spectrum of the speech signal. Specifically, for a signal $\mathbf{x} = (x_1 \ldots x_N)$ we have:

$$c_n = \frac{1}{N}\sum_{k=1}^{N} \log(|\mathrm{X}(\omega)|) \cdot \exp\left(-j\frac{2\kappa\pi}{N}n\right) \tag{3.48}$$

where $\mathrm{X}(\omega)$ is the DFT of the original signal.

In the field of speech processing, cepstral analysis is often combined with spectral domain partitioning, using *filterbanks*. Although it is possible to analyse the signal by dividing the spectrum into linear bands, often the *mel frequency scale* is used, which is defined as:

$$mel(f) = 2595\log(1 + f/5000)\ , \qquad 0 \le f \le F_s. \tag{3.49}$$

This scale approximates the human auditory system's response to sounds of different frequencies, emulating the effective filtering properties of the human ear (in this respect, it is similar to the ERB scale we have seen earlier in this chapter). The Mel frequency scale is a nonlinear transformation of the linear frequency scale, which divides the frequencies into (overlapping) frequency bands. Then, cepstral analysis takes place in each of these bands. The Mel frequency scale in combination with cepstral analysis gives rise to *Mel-Frequency Cepstral Coefficients* (MFCCs) (Mermelstein, 1976), which are the reference standard feature for speaker identification and automatic speech recognition (Murty and Yegnanarayana, 2006). MFCCs compute the contribution of the energy of the speech signal at each frequency band:

$$MFCC_n = \sum_{k=1}^{K} E_k \cos[n(k-0.5)\,\pi/K]\ , \qquad n = 0, \dots, L \tag{3.50}$$

where L is the number of MFCC coefficients (typically 12-16 are used), and E_k is the mean energy of the kth filter (typically $K = 20$ to 30) (Davis and Mermelstein, 1980). Comparing the MFCC coefficients in Eq. (3.50) and the cepstral coefficients in Eq. (3.48), the difference is that the former are derived using a narrower spectral sampling and the computation of energy takes place for each spectral bin rather than the square of the spectrum. The mean energy is the average of the squared amplitude of the Fourier transform, taking into account the triangular mel filters. The 0th MFCC coefficient represents the signal energy. As with other nonlinear techniques (reviewed later in this chapter), MFCCs circumvent the difficult task of F_0 extraction, which is always challenging, particularly for pathological voices.

Often, the first and second derivatives in time (over successive frames) of the MFCCs are also used, which are known as *delta-coefficients* and *delta-delta coefficients*, respectively.

MFCCs have traditionally been used for speaker identification (automatic speaker recognition), but have recently been successfully adopted for *voice quality* assessment (Godino-Llorente et al., 2006; Fraile et al., 2009; Tsanas et al., 2011a). We extracted 14 MFCCs including the *0th coefficient* and the *log-energy* of the signal, along with their associated *delta* and *delta-delta coefficients*, using the implementation in M. Brookes's Matlab Toolbox (Brookes 2006).

3.2.3 Modern dysphonia measures

Whereas all biological systems are inherently nonlinear, classical signal analysis approaches have often provided reasonably accurate and useful results, and linear models have shed some light on the underlying physiological mechanisms. Although linear methods still find wide applicability and are immensely useful, current research has explored nonlinear modelling which can often represent characteristics of a system more accurately than linear approaches (Stark and Hardy, 2003). This nonlinear approach has also been applied to speech signals (Teager, 1980; Titze, 2008), where recent explorations of nonlinear signal processing tools have shown very promising results (Little et al., 2006; Little et al., 2009; Tsanas et al., 2011a; Tsanas et al., 2011b, Tsanas et al., 2012b).

More recently, researchers have turned their attention in using various nonlinear tools. *Nonlinear time-series analysis* is a general approach applicable to speech data, and provides new methods for characterising disordered voices more accurately than the classical perturbation methods we have seen in the preceding section. Standard reference works in the field of nonlinear time series analysis include the works of Kantz and Schreiber (2004), and Small (2005). Although some of the widely used methods such as *Lyapunov exponents* and *correlation dimension* have shown some promising results in speech signal analysis (Zhang et al., 2005; Giovanni et al., 1999), these tools are sensitive to noise, and have numerical and algorithmic problems (Little et al., 2007). Therefore, care needs to be exercised when applying and interpreting some of these nonlinear techniques. In the following sections we review some of the most promising nonlinear tools applied to speech signal processing.

3.2.3.1 Glottal to Noise Excitation ratio (GNE)

The *Glottal-to-Noise Excitation Ratio* (GNE) belongs to a family of measures that aim to quantify the extent of noise in the signal (they can be viewed as SNR-like measures). GNE builds on the premise that vocal fold collision events lead to synchronous excitation of different frequency bands, whereas turbulent noise, which is mainly caused by incomplete vocal fold closure, leads to asynchronous excitation. Michaelis et al. (1997) proposed the following steps for the computation of GNE: 1) downsample the speech signal to 10 kHz, 2) Inverse filtering of the signal to detect glottal cycles, and subsequently work with each of those glottal cycles (signal segments), 3) Compute the Hilbert envelopes of different frequency bands using a specified bandwidth for each glottal cycle, 4) Compute the cross-correlation of pair-wise envelopes where the central frequencies of the bands are greater than half the bandwidth, 5) Choose the maximum value amongst the correlations between pairs of the frequency bands, 6) Choose the maximum of step 5, which is the GNE value for the detected glottal cycle. 7) Compute the mean of the resulting vector GNE_{mean} (in this study we also compute the standard deviation GNE_{std}). We have chosen to scan the frequency range using shifts of 500Hz to determine the central frequency, and used 500Hz for the bandwidth following the suggestion of the originator of the algorithm (Michaelis et al., 1997). Godino-Llorente et al. (2010) experimented with different bandwidth values; here we used the default settings to compare GNE with the new measure introduced in § 3.4.2.3.

3.2.3.2 Detrended Fluctuation Analysis (DFA)

DFA is a scaling analysis method used to quantify long range *power-law*[19] autocorrelations in signals which are non-stationary, thus overcoming some of the problems of scaling analysis techniques which are only suitable for stationary signals (Chen et al., 2002). Here, we describe the DFA method using the definitions used in Little et al. (2007). In the speech signal applications context, DFA characterises turbulent noise, quantifying the stochastic self-similarity of the noise caused by turbulent air-flow in the vocal tract. Conceptually, it focuses on the stochastic component of the speech signal (like RPDE which will be discussed in § 3.2.3.3), aiming to characterise its scaling exponent. This is achieved by fitting straight lines over small time intervals of length L, and measuring the average fluctuation $F(L)$ of the signal against the straight lines within that interval, using the root mean squared metric (therefore '*de-trending*' the signal). Then, the algorithm fits a straight line to the set $\{logL, logF(L)\}$ over different values of L using standard *linear least squares regression*[20].

Initially, the algorithm integrates the signal samples to induce self-similarity in the signal: $s_n = \sum_{i=1}^{n} x_i$, where $n = 1 \dots N$, and N is the length of the speech signal $\mathbf{x}$. Then, $\mathbf{s} = (s_1 \dots s_n)$ is divided into non-overlapping intervals of length L, and for each interval a best fit in the least-squares sense is determined for the window with length L: $\operatorname{argmin}_{a,b} E^2 = \sum_{n=1}^{L} (s_n - a_n - b)^2$, where a is the slope and b is the intercept. Then, the fluctuation is:

$$F(L) = \left[\frac{1}{L} \sum_{n=1}^{L} (s_n - a_n - b)^2 \right]^{1/2}. \tag{3.51}$$

[19] *Power law* refers to a scale-invariant relationship between two quantities, where the dependent quantity varies as a power of the independent quantity.

[20] *Least squares regression* is a simple method of mapping **x** to *y*, and will be defined mathematically in § 4.2.1.

The scaling exponent γ is determined from the log-log plot of L versus $F(L)$ by fitting a straight line, and is normalized to lie in the range 0-1, by using the *logistic function* $f(x) = \left(1 + \exp(-x)\right)^{-1}$.

$$DFA = \frac{1}{1 + \exp(-\gamma)}. \tag{3.52}$$

Breathiness or other similar dysphonias caused by, e.g. incomplete vocal fold closure can increase the DFA value because the fluctuations around the fitted lines will be greater compared to healthy phonations. We refer to Little et al. (2007) for further specifics of DFA.

3.2.3.3 Recurrence Period Density Entropy (RPDE)

RPDE addresses the ability of the vocal folds to sustain stable vocal fold oscillation, quantifying the deviations from exact periodicity. The underlying concept is that the speech signal is composed of a (nonlinear) *deterministic*[21] and a *stochastic* component and the method tries to bring out the latter. This measure is based on the notion of *recurrence* (Kantz and Schreiber, 2004), which can be seen as a generalization of periodicity. For the purposes of this study, recurrence can be informally expressed as the amount of time (number of samples/F_s) before a small m-dimensional segment of the speech signal (length m to be determined) is within an arbitrary constant $r > 0$ from another m-dimensional segment of the speech signal forward in time. To formally define this algorithmically, we need two m-dimensional vectors $\mathbf{x}$. Starting from the sample n_0 and moving forward in time, we identify the sample k at which the Euclidean distance $\|\cdot\|$ between the two vectors $\mathbf{x}_{n0} = [x_{n0}, \dots, x_{n0+m}]$ and $\mathbf{x}_{n0+k} = [x_{n0+k}, \dots, x_{n0+k+m}]$ is $\|\mathbf{x}_{n0} - \mathbf{x}_{n0+k}\| > r$. Next, we want to

[21] *Deterministic* refers to a signal or system which can be defined with mathematical equations precisely; that is, for a given input the output can be predicted exactly. *Stochastic* signals or systems are not simply governed by algorithmic expressions, i.e. they are not deterministic; there is a certain degree of randomness in their output.

determine the sample n_1 $(n_1 > n_{0+k})$ forward in time to determine $\mathbf{x}_{n1} = [x_{n1}, \dots, x_{n1+m}]$, where $\|\mathbf{x}_{n0} - \mathbf{x}_{n1}\| \leq r$, which gives rise to the recurrence time $T = n_1 - n_0$. This procedure is repeated for the entire speech signal to form a histogram of the recurrence times, $R(T)$, which is normalized to get the recurrence time probability density $p(T) = \frac{R(T)}{\sum_{i=1}^{T_{max}} R(i)}$, where T_{max} is the maximum recurrence time (maximum time found over the segment of speech analysed). RPDE is then determined from the entropy of the distribution of $p(T)$, normalized by the entropy of a purely stochastic signal, which is $\ln(T_{max})$, in order to provide an output in the range 0 to 1. Thus, RPDE takes the form (Little et al., 2007):

$$RPDE \equiv \frac{-\sum_{i}^{T_{max}} p(i) \cdot \ln\big(p(i)\big)}{\ln(T_{max})}. \tag{3.53}$$

The two free parameters m and r were optimized using grid search on synthetic signals in Little et al. (2007). Speech dysphonias typically cause an increase in RPDE because of the increased uncertainty in the period of the speech signal (RPDE is zero for perfectly periodic signals and close to 1 in the purely stochastic case).

3.2.3.4 Pitch Period Entropy (PPE)

PPE measures the impaired control of stable pitch during sustained phonations (Little et al., 2009), a symptom common to PWP (Cnockaert et al., 2008). The novelty of this measure is that it uses a logarithmic pitch scale and is robust to ubiquitous confounding factors such as smooth vibrato which is present in both healthy and dysphonic voices. Initially, the estimated F_0 contour (which can be extracted using any PDA, such as those described in § 3.2.1) is converted to the logarithmic (perceptual) semitone (Baken and Orlikoff, 2000): $F_{0,\mathrm{per}} =$

$12\log_2(F_0/127)$[22]. The perceptual F_0 contour, $F_{0,\mathrm{per}}$, is filtered to flatten the spectrum of the semitone series and remove the effect of the mean semitone (which is gender- and subject-specific), giving rise to the series which characterizes the occurrence of semitone variations **r**. Next, the procedure is similar to the computation of RPDE: the probability density of the semitone values $p(r)$ is obtained and is expressed using the concept of entropy:

$$PPE \equiv \frac{-\sum_i^{L_{\mathrm{PPE}}} p(i)\cdot\ln(p(i))}{\ln(L_{\mathrm{PPE}})} \tag{3.54}$$

where L_{PPE} is the length of points used to calculate the spread measure. For more specifics on the PPE algorithm please refer to Little et al. (2009).

3.2.4 Novel speech measures

We have already touched upon one of the most challenging tasks in the computation of the dysphonia measures: the accuracy of F_0 estimation given the speech signal. In addition, some of the dysphonia measures already presented above rely on *linear signal processing methods* for *stationary* signals (for example HNR relies on the computation of the Fourier transform (Boersma and Weenink, 2009)). Thus, they are inherently limited because emerging evidence strongly suggests the existence of non-negligible nonlinearity and non-stationarity in the speech production mechanism (Little et al., 2006; Titze, 2008). Therefore, we wanted to develop measures which may be able to overcome the shortcomings of the algorithms presented in the preceding sections.

In Tsanas et al. (2010c) we introduced the use of wavelets to study the F_0 contour: although

[22] The value 127 was chosen because it was the average F_0 in the initial study (Little et al., 2009), and is adopted here for compatibility with that study.

wavelet analysis is a linear technique, it is appropriate for the analysis of *non-stationary* signals. Details of the computation of measures based on wavelet analysis are provided in § 3.2.4.1. Subsequently, in Tsanas et al. (2011a) we proposed a range of novel nonlinear dysphonia measures which are summarized in § 3.2.4.2 – 3.2.4.4. We wanted these measures to be robust for general speech signal analysis (i.e. minimize assumptions and possible confounding factors) and hence we rely neither on F_0 estimation, nor linear voice production assumptions. Conceptually, the aim is to quantify SNR by building on the fact that the energy in the high frequency bands is generally increased in pathological voices (Godino-Llorente et al., 2006). This is caused by incomplete vocal fold closure, resulting in the creation of vortices and turbulent noise (Titze, 2000). Although the neurological mechanisms that control the vocal folds are not fully understood, the vibration pattern of the vocal folds is known to be affected in PD (Titze, 2000). In the following sub-sections, we indicate the characteristics in the sustained vowel phonations that the new dysphonia measures attempt to quantify.

3.2.4.1 Wavelet measures

DFT expresses the time-domain signal in the frequency domain, which can potentially reveal useful properties of the signal. However, DFT assumes that the signal is stationary, and hence does not provide information regarding the observed frequencies as a function of time (the frequency content of stationary signals does not change in time). The discrete wavelet transform (DWT) expresses the time-domain signal in the wavelet domain, and provides a time-frequency representation: it has the property of quantifying regularity effects (scale aspects) and transient processes (time aspects), qualities which make them well suited for detecting scale and time deviations. The DWT analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation

(approximation coefficients) and detail information (detail coefficients). The *wavelet decomposition* can be thought of as an extension of the DWT, successively expressing the approximation coefficients using subsequent layers (known as *levels*) to extract new approximation and detail signals (see Fig. 3.3). Moreover, the wavelet decomposition is well adapted to the study of fractal properties and self-similarity of signals, characteristics of speech signals used previously in developing dysphonia measures (Little et al., 2007). Practically speaking, the resulting wavelet coefficients can be thought of as similarity (resemblance) indices between the selected wavelet and the signal in each level, where large coefficients represent large resemblance. The rationale for the developed measures based on wavelet decomposition is that a healthy person is expected to be able to sustain a vowel with minimal deviation from exact periodicity, whilst people with pathological voices cannot (Titze, 2000). For more details regarding wavelets we refer to Mallat (2009).

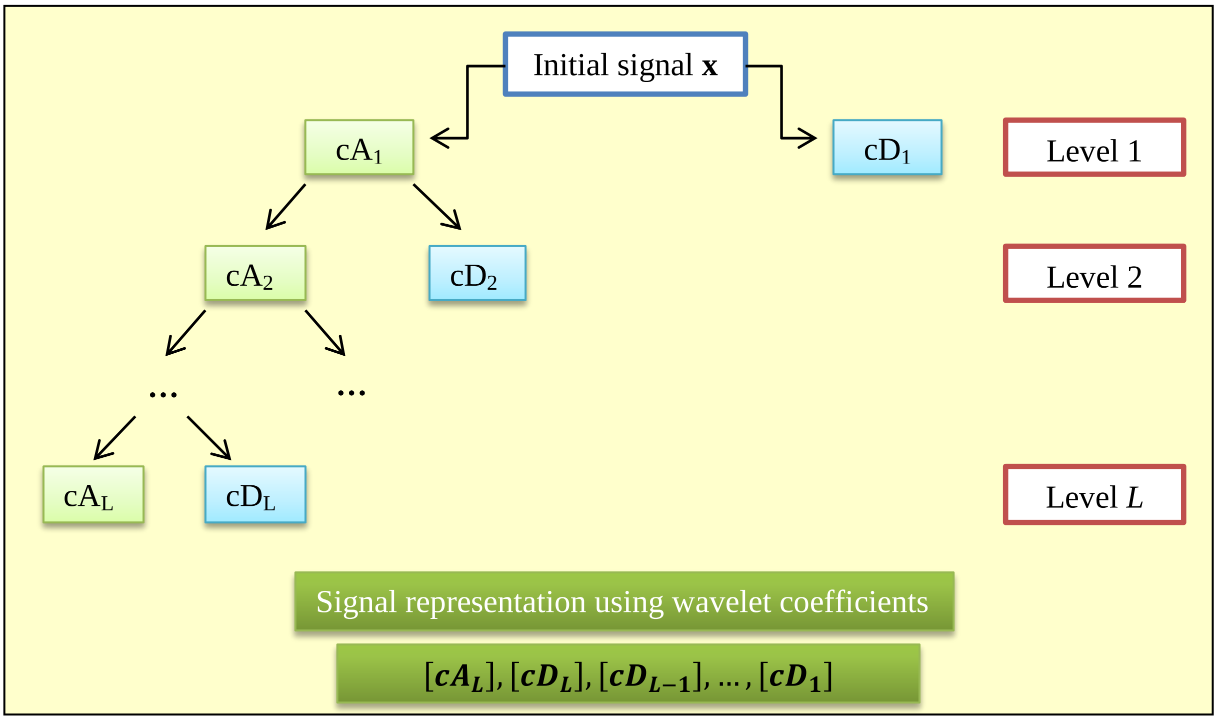
Initial signal **x**
cA_1
cD_1
Level 1
cA_2
cD_2
Level 2
…
…
cA_L
cD_L
Level *L*
Signal representation using wavelet coefficients
$[\boldsymbol{cA_L}], [\boldsymbol{cD_L}], [\boldsymbol{cD_{L-1}}], \ldots, [\boldsymbol{cD_1}]$

Fig. 3.3: Schematic diagram of wavelet decomposition. "A" corresponds to approximation coefficients, and "D" to detail coefficients. The number in each box denotes the level.

As a first step to compute the wavelet measures introduced in Tsanas et al. (2010c), we extract the F_0 contour from the speech signals. Then, the F_0 contour is decomposed into 10 wavelet levels (10 was chosen arbitrarily) and for each level we obtain a vector with approximation coefficients, and a vector with detail coefficients. For each vector that contains the approximation and wavelet coefficients we computed the energy, entropy (using both Shannon's and the log energy definitions), and the TKEO at all levels. Our experiments with three commonly used wavelet families (Daubechies, Symlets, and Coiflets) did not reveal any practically useful difference in the quantification of PD symptom severity (Tsanas et al., 2010c). Recently, Little et al. (2009) have shown that the transformation of the fundamental frequency into the logarithmic perceptual semitone scale can enhance robustness to confounding factors such as smooth vibrato prior to further processing (see the PPE dysphonia measure, § 3.2.3.4). Therefore, in addition to the features extracted using the raw F_0 contour, we also suggest computing the log transform of the F_0 contour and then follow the methodology already outlined to obtain additional features. The wavelet dysphonia measures defined here reduce the initial vector space with elements equal to the length of the F_0 contour, to a reduced space equal to the number of computed features.

3.2.4.2 Empirical Mode Decomposition Excitation Ratios (EMD-ER)

The *Empirical Mode Decomposition Excitation Ratios* (EMD-ER) are a family of dysphonia measures that build on physiological evidence suggesting that turbulent noise is increased in pathological voices due to incomplete vocal fold closure (Titze, 2000). As we have remarked in the discussion about MFCCs, the vocal tract articulators are affected in PD, shifting the resonant frequencies and altering the expected (healthy) energy distribution of the speech signal. The EMD-ER family aims to quantify this new energy distribution. Effectively, EMD-ER can be seen as another approach to quantify the signal to noise ratio in voice production.

The empirical mode decomposition (EMD) was proposed by Huang et al. (1998) as a promising nonlinear tool for time-series analysis. Conceptually, EMD decomposes a multi-component signal (signal composed of multiple superimposed signals) into elementary signal components with AM-FM contributions, which are known as *intrinsic mode functions* (IMFs). Each of the IMFs contributes both in amplitude and frequency towards generating the observed signal, which contains all the superimposed AM and FM components. EMD is similar to Fourier and wavelet analysis conceptually: however, in both these methods the basis functions are pre-determined (exponential function and wavelet function, respectively), which may be a limitation for some applications, whereas EMD uses adaptive basis functions (i.e. the basis functions are determined from the data). The algorithm can be summarized in five steps: 1) identify the minima and maxima of the signal, 2) Use cubic spline interpolation to join the minima together to generate the *lower envelope*, and similarly join the maxima to generate the *upper envelope*, 3) Compute the mean time series of the envelopes, 4) Subtract the mean time series from the data to obtain the IMF component, 5) Repeat steps 1-4 using as the starting signal the *residue* (the original signal minus the IMF component) from the previous step. This process stops when the residue satisfies a given stopping criterion. We have used the implementation of Rilling and Flandrin (2008) for computation of the IMFs.

The first few IMFs are the time-varying high frequency components of the signal, which can be considered to be the noise in the signal. Here, we define the first few IMFs to represent the noise in the signal, and the latter IMFs to represent the actual useful information in the signal, in order to build SNR measures. Then, we quantify the typically increased noise in the higher frequencies of pathological voices, without having to make rigid pre-specification of the frequency bands (which would be required e.g. in Fourier analysis). Specifically, we compute the SEO, TKEO and Shannon's entropy for each IMF. We compute the mean values of the IMFs, and define three IMF-SNR measures (using for each of the three IMF_{SNR} measures the SEO, TKEO, or Shannon's entropy):

$$\text{IMF}_{\text{SNR}} = \frac{\sum_{d=4}^{D} \mu_d}{\sum_{d=1}^{3} \mu_d} \qquad (3.55)$$

where μ_d represents the mean values of SEO, TKEO, or Shannon's entropy for each of the $d = 1, 2 \ldots D$ IMFs, and D is the total number of extracted IMFs.

The log-transformation of dysphonia measures may be a useful pre-processing step before feeding a learning algorithm (Tsanas et al., 2010b) because it normalizes the measures; hence we also investigate whether log-transforming all the IMFs might convey additional information over and above the raw IMF analysis. We computed the SEO, TKEO and Shannon's entropy for each log-transformed IMF (log-transforming all the IMF components, and setting any negative entries to zero, i.e. using the convention $log(0) = 0$). Similarly to Eq. (3.55), we define the IMF_{NSR} (Noise-to-Signal Ratio) using the log-transformed IMFs; the difference is that only the first two IMFs are used to represent the noise in the signal:

$$\text{IMF}_{\text{NSR}} = \frac{\sum_{d=1}^{2} \mu_{d,L}}{\sum_{d=2}^{D} \mu_{d,L}} \qquad (3.56)$$

where $\mu_{d,L}$ represents the mean values of SEO, TKEO or Shannon's entropy for each of the $d = 1, 2 \ldots D$ of the log-transformed IMFs.

The use of the first four IMFs in the raw case and the first two IMFs after log-transformation to represent the noise in the signal was decided after experimentation, following visual inspection of the results in phonations with low, mild and severe UPDRS (Tsanas et al., 2011a), but this definition might need to be optimized for another application.

3.2.4.3 Vocal Fold Excitation Ratios (VFER)

The *Vocal Fold Excitation Ratios* (VFER) are another family of dysphonia measures with similar conceptual justification to GNE: glottal cycles lead to synchronous excitation of different frequency bands and turbulent noise leads to uncorrelated excitation. As with the reasoning behind MFCCs and EMD-ER, the energy distribution of the phonation is altered because of the placement of the vocal tract articulators. However, contrary to MFCCs and EMD-ER, the VFER family works directly on the vocal fold cycles to quantify energy ratios during each cycle. The objective is to quantify nonlinear, interacting physiological phenomena in speech production of the vocal folds and the vocal tract as a result of (a) pathological vocal fold vibration pattern (incomplete vocal fold closure leading to the creation of vortices and turbulent noise) and (b) positioning of the articulators (affecting the energy distribution).

Michaelis et al. (1997) suggested down-sampling the signal to 10 kHz in GNE, which implicitly assumes that frequencies over 5 kHz do not carry clinically useful information. Our experiments contradict this view, and we have found that down-sampling the 24 kHz signals may lead to loss of clinically useful information, so we dismiss this pre-processing step. The fact that speech scientists recommend the use of sampling frequencies higher than 20 kHz (Titze, 2000) suggests that practice has taught them there may be useful information in frequencies at least up to 10 kHz. A plausible explanation for the necessity to use signals sampled at high F_S is that pathological voices are characterized by high frequency noise (Godino-Llorente et al., 2006) which has clinical value, and down-sampling the signals may potentially destroy relevant information. Moreover, in the VFER family we substitute the 2nd step of GNE (where the glottal cycles are determined using inverse filtering) with DYPSA (see § 3.2.1.1). We then integrate the concepts of SEO and TKEO that have been previously also used when processing the IMFs in the EMD, to compute the energy ratios of different frequency bands during a vocal fold cycle. Specifically, for each of the time windows between GCIs and GOIs determined by DYPSA (i.e. when the vocal folds are apart), we scan the entire frequency range up to 11.5 kHz (close to half of the sampling frequency). We used frequency shifts of 500Hz to determine the central frequency and 500Hz for the bandwidth, and compute the SEO and TKEO of the signal bandpass-filtered with that central frequency and bandwidth. The choice of the frequency shift and the bandwidth was decided a-priori, following the suggestion of Michaelis et al. (1997) from the original GNE algorithm. More recently, Godino-Llorente et al. (2010) have tried to optimize those free parameters in GNE; it would be possible to do something similar in VFER. We used the Hanning window to process the appropriate signal segment (the time periods indicated by DYPSA) for further processing, similarly to Michaelis et al. (1997).

Then, we define the VFER_{SNR} measures similarly to Eq. (3.55). We have used the first five frequency bands (1 Hz — 2.5 kHz) to denote the 'signal energy', and the remaining frequency bands (2.5 kHz — 11.5 kHz) to denote the 'noise' bands. These choices were made after experimentation on the speech-PD database presented in § 6.3, following visual inspection of the results in phonations associated with low, mild and severe UPDRS (similarly to the selection of the IMF components to denote signal energy and noise in the preceding section). We have used both SEO and TKEO to compute the VFER_{SNR} measures, as in the EMD-ER family of dysphonia measures. The corresponding VFER_{NSR} measures are defined using a form similar to Eq. (3.56) where we have used the log-transformed SEO and TKEO values. Finally, we have followed steps 3-7 from GNE to extract the $\text{VFER}_{\text{mean}}$ and VFER_{std}. The latter two measures differ from GNE only in that the signal was not downsampled and DYPSA was used instead of inverse filtering to extract the glottal cycles.

The steps used to define the VFER_{SNR} and VFER_{NSR} measures (after the estimation of the glottal cycles with DYPSA) were also integrated into GNE: after the GNE algorithm's first two steps, we follow the same procedure as described in VFER, forming the $\mathrm{GNE}_{\mathrm{SNR}}$ and $\mathrm{GNE}_{\mathrm{NSR}}$ measures. That is, the $\mathrm{GNE}_{\mathrm{SNR}}$ and $\mathrm{GNE}_{\mathrm{NSR}}$ measures use down-sampling to 10 kHz and detection of the glottal cycles with inverse filtering (the first two steps in GNE); subsequently we follow the $\mathrm{VFER}_{\mathrm{SNR}}$ and $\mathrm{VFER}_{\mathrm{NSR}}$ methodology: scan the entire frequency range up to 11.5 kHz with frequency shifts of 500Hz to determine the central frequency and 500Hz for the bandwidth, compute the SEO and TKEO, and use the ratio of the frequency ranges.

3.2.4.4 Glottal Quotient (GQ)

The DYPSA algorithm was also used to derive a new set of measures, taking into account the length of time that vocal folds are apart (glottis is open) or in collision (glottis is closed). This family of dysphonia measures bears close resemblance to jitter conceptually, and has a similar rationale. The difference is that instead of working with F_0 estimates computed using one of the windowed PDAs, we work with the glottal cycles estimated using DYPSA.

Specifically, we computed the standard deviation of the duration when the glottis is open (vocal folds are apart, denoted by $\mathrm{GQ}_{\mathrm{open}}$) and when the glottis is closed (vocal fold collision, denoted by $\mathrm{GQ}_{\mathrm{open}}$). In addition, we define $\mathrm{GQ}_{5-95\,percentile}$ as the difference between the 5th and 95th percentile values of the duration that the vocal folds are apart, over the same percentile range of the duration of the vocal fold cycle. The rationale is that in healthy voices, which are *almost* periodic (Titze, 2000), the vocal fold cycles should not differ considerably, and the interval that the glottis is open or closed should remain roughly equal across all vocal cycles of the sustained vowel phonations. However, pathological voices are characterised by increased aperiodicity, because the normal vibration of the vocal folds is affected. The use of the 5th and 95th percentiles instead of the range makes this measure more robust to outliers.

3.3 Overview of the signal processing methods

This section summarizes the speech signal processing algorithms described in the preceding sections. Linear signal processing is a mature, well established branch of engineering and has been applied with some notable successes to a wide variety of physiological topics. We have seen that the majority of dysphonia measures currently in use stem from linear signal processing methods. These methods are well understood and computationally very simple, which makes them attractive and easily understandable to clinicians who are not necessarily mathematically oriented. Nevertheless, linear signal processing techniques make some implicit assumptions (see Eq. 3.34) which may not be necessarily true in practice, particularly for disordered voices.

Nonlinear modelling and signal analysis techniques complement the classical approaches and can often explain the data better than linear models, uncovering complex underlying relationships that may be obscured when using the linear prism alone. Indicative of this is the successful adoption of measures originating from nonlinear time series analysis, and methods we discussed previously such as RPDE, DFA, PPE, EMD-ER and VFER. However, these measures require fine-tuning of parameters and are often mathematically more complex than the linear measures, which is a burden for wider adoption by the clinical community. Table 3.2 succinctly summarizes the key information of all the dysphonia measures used in this study.

Ongoing research is expected to lead to enhanced linear and nonlinear speech signal processing techniques, and their conjunction could lead to hybrid models assisting in our quest for understanding and interpreting physiological systems.

Table 3.2: Summary and key information of the dysphonia measures used in this study

Measure	Motivation	Number of features
Jitter & Jitter variants	The vocal folds are affected in PD, and jitter aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in *fundamental frequency*	One for each variant
Shimmer & shimmer variants	The vocal folds are affected in PD, and shimmer aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in *amplitude*	One for each variant
Harmonics to Noise Ratio (HNR) & Noise to Harmonics Ratio (NHR)	In speech pathologies there is increased noise due to turbulent airflow, resulting from incomplete vocal fold closure. HNR and NHR quantify the ratio of actual signal information over noise.	2
Linear Predicting Coding Coefficients (LPCC)	Quantify deviations of the prediction of the current data sample as a function of the preceding samples. In pathological voices this deviation is expected to be larger.	10
Mel Frequency Cepstral Coefficients (MFCC)	PD affects the articulators (vocal tract) in addition to the vocal folds, and the MFCCs attempt to analyse it separately from the vocal folds	12-42, depends on additional components
Glottal to noise excitation (GNE)	Extent of noise in speech using energy and nonlinear energy concepts	6
Detrended Fluctuation Analysis (DFA)	Quantify the stochastic self-similarity of the noise caused by turbulent airflow	1
Recurrence Period Density Entropy (RPDE)	Quantify the stochastic component of the deviation of vocal fold periodicity	1
Pitch Period Entropy (PPE)	In speech disorders it is very difficult to sustain stable pitch due to incomplete vocal fold closure. PPE quantifies the impaired control of stabilised pitch.	1
Wavelet measures	Quantify deviations in F_0 (obtained using any F_0 estimation algorithm)	180
Empirical mode decomposition excitation ratio (EMD-ER)	Signal to noise ratios using EMD-based energy, nonlinear energy and entropy	6
Vocal fold excitation ratio (VFER)	Extent of noise in speech using energy, nonlinear energy, and entropy concepts	9
F_0-related measures	Summary statistics of F_0, differences from expected F_0 in age- and gender- matched controls, variations in F_0	Three for each F_0 estimation algorithm

Chapter 4

Methodology for data analysis

Quantitative empirical modelling usually relies on a multidisciplinary approach to combining data analysis, mathematical modelling and information processing. This chapter provides a succinct overview of various aspects of data-driven statistics and some of the available machine learning techniques which are used in the following chapters.

In many practical applications, we record signals and a result (*outcome*) which is computed or provided by other means (e.g. by human experts). Typically, we believe there may be an association between the signals and the outcome (for example it is sensible to believe that the electrocardiogram may be a good indicator of cardiovascular status). The aim is then to identify useful characteristics (*patterns*) in the signal so that the result can be accurately predicted from the computed patterns without resorting to direct measurement of the result (which may be very difficult and/or costly to obtain, e.g. invasive recordings). The extraction of useful information from the data in the form of identifying patterns is known as *pattern recognition*: those patterns can be conveniently presented in a row vector format $\mathbf{x}$, where the premise is that these patterns may be indicative of the outcome (typically a scalar, conveniently presented as y). Repeating this process for a number of observations (for example recording the electro-cardiogram of 100 people), we can summarize the patterns in a matrix format $\mathbf{X}$, where each row contains the patterns $\mathbf{x}$ for each observation. Similarly we can concatenate the results in a column vector format $\mathbf{y}$. In algorithmic terms, we want to determine the function $f(\mathbf{X}) = \mathbf{y}$, which relates $\mathbf{X}$ and $\mathbf{y}$, and this is known in machine learning contexts as the *supervised learning problem*[23]. If $\mathbf{y}$ takes discrete values, such as might occur when separating the data into two or more *groups* (also known as *classes*), e.g. health and disease, determining f is known as a *classification* problem. Conversely, if $\mathbf{y}$ takes values on the real axis, determining f is known as a *regression* problem. In practice, we use a *training set* (where both the patterns and the outcome are known) to determine the function f which relates patterns and outcome, and a *testing set* (only the patterns are known) to predict the outcome for new observations using f.

The patterns which are used as inputs into f are also known as *features*, *predictors*, *input variables*, *explanatory variables*, *covariates*, *dimensions,* or *independent variables*. The function f is the *prediction model* or *learner*, and $\mathbf{y}$ is called the *outcome measurement*, *response variable* (or simply *response*), *target, label,* or *dependent variable*. The term *learner* is often substituted with either *classifier* or *regressor* which describes the application explicitly. This abundance in terminology stems from the fact that different disciplines (statistics, engineering, computer science) have studied essentially the same problems with different tools and each discipline adopts its own terms. To minimize confusion we will use the terminology from machine learning: the computed characteristics will be referred to as *features*, and the outcome quantity of interest as *response* or *response variable*.

This chapter presents a methodological approach to *supervised* data analysis, presuming that the features have already been computed (in the case of speech signals this would be achieved by applying the dysphonia measures of Chapter 3) and they are associated with a response. Then, the most common steps are (a) explore the data using statistical analysis, (b) find a compact representation of the data selecting or transforming the features, (c) map (associate) the compact set of features to the response, and (d) validate the model using statistical hypothesis tests and surrogate tests. We look into each of these steps in detail.

[23] Parenthetically, in *unsupervised learning* we infer properties from the patterns without using $\mathbf{y}$. One example of unsupervised learning is the computation of densities, which we have briefly described in § 3.1.

4.1 Data exploration and statistical properties

So far, we have informally indicated the setup of the data in supervised learning applications. The most common formalization of supervised learning is the following:

$$\mathbf{X} = \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NM} \end{bmatrix}}_{Design\ matrix}, \qquad \mathbf{y} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{Response}$$

where M represents the number of features and N the number of observations (*samples*). We define the *design matrix* (*data matrix*) $\mathbf{X} \epsilon \mathbb{R}^{N\times M}$ and the *response variable* $\mathbf{y} \epsilon \mathbb{R}^{N\times 1}$. Each feature is an N-dimensional column vector of $\mathbf{X}$, and will be represented as $\mathbf{f}_j = \left(x_{1j} \ldots x_{Nj}\right)^T$. Each feature expresses a characteristic of the original signal in a different domain compared to the original domain where the signal lies (for speech signals, this is the time series). Each sample is an M-dimensional row vector of $\mathbf{X}$, and will be represented as $\mathbf{x}_i = (x_{i1} \ldots x_{iM})$. Each sample $\mathbf{x}_i$ has a corresponding response y_i. More often than not, the representation of the data in the design matrix is subject to further processing as we shall see in the following sections.

4.1.1 Density plots and scatter plots

The first step in data analysis is the exploration of some statistical properties of the data, and producing plots in order to get a feel for the data structure. Initially the probability densities of the features can be plotted, which gives an overview of the response and the features. Typically, for visualization purposes, some kind of prior normalization is used (for example normalizing values between 0 and 1 or -1 and 1), so that all the features have the same scale which assists in the visual interpretation of any obvious relationships. The approach for the estimation of the densities depends on the data type (discrete or continuous variables) and the available computational resources. For continuous variables, it is usually preferable to use kernel density estimation (see § 3.1) because this provides a more realistic and often more accurate representation of the data distribution (Hastie et al., 2009). In addition to *density plots*, we suggest using *scatter plots* to visualize whether there is any obvious relationship between each feature and the response. Scatter plots present all the $\{\mathbf{f}_j, \mathbf{y}\}$ N points in a figure, where $j = 1 \ldots M$ refers to the feature used.

This first step presenting the density plots and the scatter plots could, for example, suggest a useful transformation of some of the features. In circumstances where there are too many features, we suggest screening for the most correlated features (see the following section) and plotting the most strongly correlated.

4.1.2 Correlation analysis

The inspection of density plots and scatter plots is usually followed by formal statistical tests in order to determine qualitatively and quantitatively how well the features are related to the response variable. Using *correlation analysis* gives a preliminary indication of the association between features and the response variable, and between features. However, correlation does not necessarily imply *causation* (change in the values of the feature affect the response) in all contexts. One example where correlation analysis would give erroneous results is the following: let us consider the scenario where the feature is the frequency of measuring blood pressure and the response is the patient condition in the intensive care unit. In practice, this would only take place once every couple of hours. However, for some ill patients this would be repeated considerably more often, and it is possible that the greater the severity of the patient's condition the more frequent the blood pressure measurement would be. Hence, the "raw" interpretation of correlation analysis between the frequency of measurements and patient condition would suggest that there is indeed a positive relationship between the two quantities, a finding which would be false. Thus, caution is needed in the careful interpretation of the observed findings.

The strength of association between two random variables X, Y can be estimated using *correlation coefficients*, and this is one measure of dependence between two variables upon which subsequent analysis could be directed. One simple method to express the dependence between X and Y is by *covariances* (denoted by Cov): then the *Pearson* correlation coefficient is defined as (Stirzaker, 2003):

$$R(X,Y)_{Pearson} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X) \cdot (y_i - \mu_Y)]}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_Y)^2}} \quad (4.1)$$

where N is the number of realisations of the random variables X, Y (i.e. N samples), and can be written as $\{x_i, y_i\}_{i=1}^{N}$

However, if the first and second-order central moments do not suffice to characterize the dependence between the two variables, the *Spearman* correlation coefficient can be used, which is effective in quantifying general *monotonic* relationships. In Spearman's method, the measurements $\{x_i, y_i\}_{i=1}^{N}$ of the random variables X, Y are ranked (sorted) in increasing order, and tied ranks are substituted by their average values to give the sorted $\{x_i, y_i\}_{i=1}^{N}$. For simplicity we do not introduce superfluous notation here to denote the sorted values. Using the sorted values $\{x_i, y_i\}_{i=1}^{N}$, the Spearman correlation coefficient is defined as:

$$R(X,Y)_{Spearman} = 1 - \frac{6\sum_i (x_i - y_i)^2}{N \cdot (N^2 - 1)}. \tag{4.2}$$

The Spearman rank correlation coefficient and the linear correlation coefficient lie in the numeric range $[-1 \dots 1]$, and the relationship between X and Y is interpreted using (a) the sign, which represents the direction of the relationship, and (b) the magnitude. Negative sign suggests the direction of the relationship between the two variables is opposite: the increase in the values of one variable leads to the decrease in the values of the other. The larger the magnitude of the correlation coefficient, the stronger the statistical relationship is. There is no universal guideline to determine when a relationship is *statistically strong*; it depends merely on the application (Cohen et al., 2002). In this study, we will refer to statistically strong relationships when $\left|R(X,Y)_{CorrType}\right| > 0.3$ (the value is arbitrarily set), where $CorrType$ denotes Pearson or Spearman.

4.1.3 Statistical hypothesis tests

Statistical hypothesis tests are commonly used in data analysis applications to determine, informally, whether the observed result conforms to a particular hypothesis, which in statistical terminology is known as the *null hypothesis*. Often, the null hypothesis is the opposite of what we aim to demonstrate; therefore in practice the objective is often met when we can *reject* the null hypothesis in favour of the *alternative hypothesis*. Statistical hypothesis tests compute significance values, the '*p-values*', which can be interpreted as the probability of obtaining a similar result by chance if the null hypothesis is true. The null hypothesis is rejected when the '*p-value*' is lower than a pre-specified *significance level*, typically 0.05 or 0.01, and the result is then deemed to be *statistically significant*. Thus, for example, $p < 0.05$ denotes a statistically significant result at the 5% significance level (i.e. there is less than 5% probability that the observed values are due to chance). Contrary to the visualization tools (§ 4.1.1), and the correlation analysis (§ 4.1.2) which are known as *exploratory data analysis* approaches (analysing the data to formulate plausible hypotheses for further investigation), statistical hypothesis testing is *confirmatory data analysis* (accept or reject the null hypothesis).

There are many statistical hypothesis tests, depending upon the null hypothesis we want to investigate. Here, we use the statistical hypothesis test to assess whether the relationship between two random variables (e.g. expressed using the Spearman correlation coefficient) is statistically significant. Specifically, the null hypothesis is that there is no correlation between the two random variables against the alternative hypothesis that there is a nonzero correlation. A simple way to compute the *p-value* of the Spearman correlation coefficient is by using *permutations*[24]. The computation starts by redefining the pairs of realization of the examined random variables $\{x_i, y_i\}_{i=1}^{N}$ using all possible permutations ($N!$), i.e. create $N!$ new combinations with the realizations of the random variables where each ith realization of X corresponds to a different jth ($j = 1 \dots N$) realization of Y each time (i, j refer to the indices of the original realizations of the random variables). Then, we compute the probabilities (via histograms) of the summation of the squared difference between each sample x_i with each y_j, for all possible combinations. The final *p-value* is computed as twice the smaller value of the tail area above and below the observed value.

We will revisit again the concept of statistical hypothesis testing when referring to model validation (in § 4.4.3).

[24] In the case of using the Pearson correlation coefficient there are simpler approaches to assess statistical significance, which rely on the Gaussianity of the data. However, the method of assessing statistical significance using permutations described here is more general and can also be used to compute the statistical significance of both the Spearman and the Pearson correlation coefficient.

4.1.4 Divergences and mutual information

In § 4.1.2 we have presented the linear (Pearson) correlation coefficient and the Spearman correlation coefficient to quantitatively express the association strength between two random variables X, Y. An alternative, more general method of expressing the information shared between two random variables X, Y is by using their marginal probability densities and their joint probability density. Methods which quantify differences in probability densities are known as *divergences*.

A divergence between two probability densities p, q is represented as $D(p||q)$, where D has a form quantifying the differences in the two distributions, for example using the absolute difference, or the squared difference between samples in the densities. By definition, the divergences share the following properties: a) they are always non-negative, b) a divergence is zero *iff* [25] the compared densities are identical, c) they need not necessarily satisfy the triangle inequality, and d) they are not necessarily symmetric (hence divergences are a weaker form of *distances*), that is $D(p||q) \neq D(q||p)$ in general. Those divergences which *are* symmetric can also be referred to as *distances*. Some commonly used divergences appear in Table 4.1. Each of these divergences has special properties, which suggests some may be more suitable in a given domain or application (Cover and Thomas, 2006). Also, note that the presented list in Table 4.1 is indicative and by no means exhaustive.

The Kullback-Leibler (KL) divergence is most commonly used, since it stems directly from Shannon's (1948) foundations on information theory and the definition of entropy. It can be leveraged to define the *mutual information* $I(X, Y)$, which attempts to characterise the information in Y (in many applications this random variable is the response) also contained in X (which typically represents a feature). The mutual information (MI) is symmetric, i.e.

[25] Commonly used abbreviation, which means "if and only if".

$I(X,Y) = I(Y,X)$, and can be used as a powerful tool to generalize the association strength between each of the features $\mathbf{f}_j$ with the response $\mathbf{y}$, and also between features. Contrary to the correlation coefficients presented in the previous section, MI can express any arbitrary (potentially nonlinear) relationship between two random variables.

Table 4.1: Commonly used divergences to express differences between densities.

Divergence name	Mathematical expression	Comments
Kullback-Leibler	$D_{KL}(p\|\|q) = \int p(x) \cdot \log(p(x)/q(x))\, dx$	Most commonly used divergence, stems directly from Shannon's theory of information
Quadratic	$D_Q(p\|\|q) = \int (p(x) - q(x))^2\, dx$	Special case of the more general Kapur divergence, similar to the L_2 norm
Total variation	$D_{TV}(p\|\|q) = \int \|p(x) - q(x)\|\, dx$	Conceptually, this is similar to the L_1 norm, also known as variational distance
Hellinger	$D_H(p\|\|q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$	The square root is a variance stabilizing transform, and the Hellinger divergence is closely related to the geodesic distance
Bhattacharyya	$D_B(p\|\|q) = \int \sqrt{p(x) \cdot q(x)}\, dx$	Also known as Bhattacharyya distance.
F-divergence	$D_F(p\|\|q) = \int p(x)/q(x)\, dx$	Also known as Csiszár f-divergence.
Jensen-Shannon	$D_{KL2}(p\|\|q) =$ $\int \begin{bmatrix} p(x) \cdot \log(p(x)/g(x)) + \\ q(x) \cdot \log(q(x)/g(x)) \end{bmatrix} /2\, dx$ where $g(x) = (p(x) + q(x))/2$	Extension of the Kullback-Leibler divergence (smoothed version)
a-divergence	$D_a(p\|\|q) = \int p^a \cdot g^{(a-1)}, a \in \mathbb{N} - \{1\}$	Also known as Renyi generalized divergence, typically $a = 2$

To ease notation the integrals appear indefinite. In practice the integration is computed over the range of values the densities span.

MI is defined as follows for two random variables X, Y:

$$I(\mathbf{x}, \mathbf{y}) = \int\int p(x, y) \cdot \log_{\mathrm{b}} \frac{p(x, y)}{p(x)p(y)} dxdy \qquad (4.3)$$

where $p(x, y)$ represents the joint probability density of X, Y, and $p(x)$, $p(y)$ are the marginal probability densities. Eq. (4.3) is the KL divergence in Table 4.1, where the two terms denoted above by p, q are (a) the joint probability density, and (b) the multiplication of the marginal probability densities. Similarly to the definition of entropy (see § 3.1), the base of the logarithm b (in the expression $\log_{\mathrm{b}}$) defines the units of the mutual information. We remark that the divergences in Table 4.1 can be generalized in more than one dimension by integrating over the number of dimensions of the random variables.

When the random variables X, Y are discrete, the computation of the divergences is straightforward since the integrals become summations. However, in most applications at least one variable is continuous, and this presents problems in the computation of the divergences because the computation of probability densities is more problematic for continuous variables. One solution is to pre-process the variables in order to discretize them, and some sophisticated algorithms have been proposed for accurate discretization (Kurgan and Cios, 2004; Tsai et al., 2008). Alternatively, density estimation methods can be used such as histograms or Parzen windows, which were outlined in § 3.1. Then, the computation of the divergences is achieved using numerical integration over the range of values the densities span using the equations presented in Table 4.1.

For more information on statistical analysis, we refer to Webb (2002), and to Ross (2009). Further details on divergences and information theory can be found in the textbook by Cover and Thomas (2006).

4.2 Curse of dimensionality and dimensionality reduction techniques

A problem often encountered in regression and classification settings when using a large number of features is the *curse of dimensionality*[26]: reducing the number of features could potentially lead to a more accurate model (Bellman, 1961). This occurs because it is impossible to adequately populate the feature space with limited data (the number of required samples grows exponentially with the number of features). The problem is exacerbated when the number of features is substantially larger than the number of samples (*fat dataset*), for example, in microarray data analysis problems (Hastie et al., 2009).

Practice has shown that features can often be highly correlated, contributing little additional information to predicting the response. According to the general *principle of parsimony*, which simply states that the model with the least number of features with predictive power should be given preference, we would like to reduce the dimensionality of the input space. This approach is known as *dimensionality reduction*, and can be achieved either by *feature transformation* (transforming the features to populate a new, lower dimensional space), or by *feature selection* (choosing a subset of features from the original feature set)[27].

There is vast literature on the topic of dimensionality reduction for both feature selection and feature transformation; a good starting reference is Guyon et al. (2006). We will briefly discuss feature transformation and focus in greater detail on feature selection in the following sections.

[26] The curse of dimensionality is also known as *Occam's razor*.

[27] Sometimes the terms *variable selection* or *gene selection* are used in specific disciplines to refer to the same concept.

4.2.1 Feature transformation

Feature transformation aims to combine the originally computed features to create a new dimensional feature set: then a subset of those new features may be more predictive of the response compared to the original feature set. That is, feature transformation techniques represent the original M-dimensional feature space by some combination amongst the original features to obtain the new K features, producing a compact representation of the information that may be distributed across several of the original features. The premise is that a few of the new features (*latent variables*) could account for the properties observed in the dataset, allowing a condensed representation of the information content existing in the data. Two of the most commonly used feature transformation methods are *principal component analysis* (PCA) and *factor analysis* (FA). Both methods form a *linear* combination of the original features to construct the new feature space. PCA constructs new features (typically referred to as *components* in the PCA setting) defined in such a way so as to capture as much of the variability in the data as possible. By design, the resulting components are *uncorrelated* with each other (but not necessarily *independent*, i.e. PCA does not take into account joint moments higher than second order). FA is typically used when we are interested in the *interpretability* of the resulting new features, that is, when we are interested in explaining the relationship of the K new features with the response. The underlying concept in FA is that the features $\mathbf{f}_j$ ($j = 1 \dots M$), are affected by *common factors*. Specifically, the method combines the features $\mathbf{X} = (\mathbf{f}_1 \dots \mathbf{f}_M)$ into common factors $\mathbf{C} = (\mathbf{c}_1 \dots \mathbf{c}_K)$, where $\mathbf{C}$ can now be used as the new design matrix to be presented to the learner. Each feature is assumed to be a function of a linear combination of the common factors, and the coefficients associated with each common factor are known as *loadings* $\mathbf{a}_i = (a_{1i} \dots a_{Mi})$. This representation is formally written as:

$$\begin{aligned}\mathbf{x}_1 &= a_{11}\cdot\mathbf{c}_1+\cdots+a_{1K}\cdot\mathbf{c}_K+\varepsilon_1\\ &\vdots\\ \mathbf{x}_M &= a_{M1}\cdot\mathbf{c}_1+\cdots+a_{MK}\cdot\mathbf{c}_K+\varepsilon_K.\end{aligned} \tag{4.4}$$

The ε_i are zero-mean disturbances (deviations from the actual value) and are unique to each original feature $\mathbf{f}_j$. The common factors are assumed to be uncorrelated Gaussian variables, and the loadings are determined by *maximum likelihood* [28].

Both PCA and FA are widely used linear feature transformation techniques, and FA is commonly used in medical settings; however, FA is not very popular amongst statisticians because there is no unique representation of the original features (we can rotate the transformed space to find a convenient representation) (Hastie et al., 2009). There is a considerable body of research on extensions for these techniques, including a large number of *nonlinear* approaches (i.e. the transformed features are not a linear combination of the original features). Discussion of all these methods is beyond the scope of this thesis, and we refer to Bishop (2007) for a brief introduction, and to van der Maaten et al. (2009) for a more extensive review.

Although feature transformation has shown promising results in many applications (Torkkola, 2003; Hastie et al., 2009), it is not easily interpretable because the physical meaning of the original features is obscured. In addition, it does not save on resources required during the feature calculation (or data collection) process since *all* the original features still need to be measured or computed. Moreover, in very high dimensional settings where the number of irrelevant features may exceed the number of relevant features, reliable feature transformation can be problematic (Torkkola, 2003).

[28] Maximum likelihood estimation aims to determine the most reasonable values for the parameters to maximise the probability to obtain the observed or measured values of the response(s). It is the more general principle upon which the least squares technique is based. For more details, the reader is referred to Hastie et al. (2009).

4.2.2 Feature selection: introduction and known approaches

Feature selection (FS) is particularly desirable in many disciplines because the features typically quantify some characteristic which is interpretable to experts in that domain, and feature selection simply chooses a subset of the original features. There has been extensive research on FS; after motivating the topic, we will review only a small fraction of the most commonly used FS algorithms. We will point out the limitations of existing approaches, and how we propose to tackle these limitations with the new FS algorithms. For a more detailed introduction to FS we refer to Guyon and Elisseeff (2003), Liu et al. (2005) and Guyon et al. (2006).

FS algorithms can be broadly categorized into *wrappers* and *filters*, while some researchers use an additional category, the *embedded* FS algorithms. Wrappers incorporate the learner in the process of selecting the feature subset, and may improve the overall machine learning algorithm performance (Tuv et al., 2009; Torkkola, 2003). However, there are at least four major issues with wrappers: a) increased computational complexity (compared to filters), which is exacerbated as the dataset grows larger, b) the selected feature subset for a specific learner may be suboptimal for a different learner, a problem known as *feature exportability* (that is, the selected feature subset is not 'exportable' to other learners), c) controlling internal parameters (parameter fine-tuning) of the learner requires experimentation, expertise, and is time-consuming, and d) inherent learner constraints, for example some learners do not handle multi-class classification or regression problems. The problem with feature exportability is that the selected feature subset may not reflect the global properties of the original dataset, so that wrapper-selected feature subsets may not generalize to alternative learners (Hilario and Kalousis, 2008). Embedded FS algorithms incorporate FS as part of the learning process. One example of an embedded FS algorithm is *ensembles of decision trees*, which we shall discuss in some detail in § 4.3.3. Filters attempt to overcome these limitations of wrapper methods and commonly evaluate feature subsets based on their information content (for example using statistical tests) instead of optimizing the performance of specific learners, and are computationally more efficient than wrappers. For all these reasons, filters are more popular and will be the main focus of this thesis.

Given the data matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ and the response $\mathbf{y} \in \mathbb{R}^{N \times 1}$ where N is the number of samples and M is the number of features, the FS algorithms aim to reduce the input feature space M down to m features, where $m < M$ (m can be chosen based on prior knowledge and possible constraints of the application, or can be determined via cross validation). That is, we want to select a feature set S comprising m features $\{\mathbf{f}_j\}$, $j \in (1 \dots M)$, where each $\mathbf{f}_j$ is a column vector in the design matrix $\mathbf{X}$. The optimal feature subset maximizes the combined information content of all features in the feature subset with respect to the response variable. However, this is a complicated combinatorial optimization problem, and the optimal solution can only be found by a brute force search. Since a brute force search is usually computationally intractable for datasets of any meaningful size (e.g. more than 10 features), sub-optimal alternatives must be sought. Although in principle combinatorial optimization methods (such as simulated annealing and genetic algorithms) can be applied to the FS problem, these techniques are computationally expensive (the computational cost depends heavily on the optimizing criterion, which typically involves a learner).

As an approximate solution to the combinatorial one, researchers often consider the suitability of each feature individually, in order to determine the overall information content of the feature subset from each individual feature in the subset. There are two FS strategies: a) sequential forward selection (features are sequentially added to the selected feature subset), and b) sequential backward elimination (starting from the entire feature set and eliminating one feature at each step). Forward FS is often used in many filter applications (Peng et al, 2005; Sun et al., 2010), and is particularly suitable for those problems where a small feature subset is required.

One of the simplest FS algorithms is to use only those features which are maximally related to the response, where the association strength of the features with the response can be quantified using a suitable criterion $I(\cdot)$ (not necessarily a distance metric in the mathematical sense). One straightforward criterion is the Pearson correlation coefficient: this assumes that the association strength between the response and each of the features can be characterized using the mean and covariance (first two joint statistical moments) alone, and that the higher order moments are zero, or at least sufficiently small that they can be neglected. Alternatively, the Spearman rank correlation coefficient, which is a more general criterion, can be used to quantify the relationship between each feature and the response. More complicated criteria can also be used to characterize potentially nonlinear (and non-monotonic) relationships between the features and the response, such as the MI. In fact, MI has attracted extensive and systematic interest in the FS literature (Battiti, 1994; Peng et al., 2005; Meyer et al., 2008; Estevez et al., 2009). However, we have already noted that the computation of MI is not trivial (particularly in domains with continuous variables), which hinders its widespread use (Torkkola, 2003).

Conceptually, the simple approach discussed thus far, which relies solely on the association strength between individual features and the response variable, works well in the presence of *independent* (orthogonal) features. It is now well established that in most practical applications a good feature subset needs to account for *overlapping* information shared amongst features useful in predicting the response. That is, the *relevance* (association strength of a feature with the response variable) needs to be counter-weighted with the *redundancy* (overlapping information shared amongst features in the feature subset useful in predicting the response) (Battiti, 1994; Yu and Liu, 2004; Guyon et al., 2006). This is the general rationale behind most contemporary FS algorithms. The following sections present a concise summary of some of the most commonly used algorithms.

4.2.2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

The *Least Absolute Shrinkage and Selection Operator* (LASSO) (Tibshirani, 1996) is a popular FS method, which is particularly efficient in *sparse*[29] contexts (Donoho, 2006) and in contexts where the features are not too highly correlated (Meinshausen and Yu, 2009). The LASSO is based on the concept of the L_1-norm, which acts as a *sparsity promoting* function (Candes et al., 2008). It has the desirable characteristic of simultaneously minimising the prediction error whilst producing some coefficients that are effectively zero (thus reducing the number of input variables). This is achieved using an adjustable *shrinkage parameter*: decreasing its value causes additional coefficients to shrink towards zero, further reducing the number of contributing features. Then it becomes a matter of experimentation to determine the number of features m to be selected (this is typically achieved using cross-validation – see § 4.4.1).

Specifically, the LASSO induces the *sum of absolute values penalty* (the L1-norm): $\hat{\mathbf{b}}_{LASSO} = \arg\min_{\mathbf{b}} \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} x_{ij} b_j\right)^2$, subject to $\sum_{j=1}^{M}\left|b_j\right| \leq t$, where $\mathbf{b} = (b_1, \dots, b_M)$ represents the ordinary least squares parameters, and t is the shrinkage parameter. The constraint $\sum_{j=1}^{M}\left|b_j\right| \leq t$ can be expressed in Lagrangian form via a regularization parameter λ and used in the computation of the least squares coefficients. Thus, imposing the penalty $\lambda\sum_{j=1}^{M}\left|b_j\right|$ on the residual sum of squares yields:

[29] *Sparse* data means that many features do not contribute toward the prediction of the response. The number of contributing components (i.e. features associated with non-zero coefficients) is known as *sparsity level*.

$$\hat{\mathbf{b}}_{LASSO} = \arg\min_{\mathbf{b}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{M} x_{ij} b_j \right)^2 + \lambda \sum_{j=1}^{M} \left| b_j \right| . \qquad (4.5)$$

Various extensions of the LASSO have been recently proposed, for example Zou (2006). Other penalties are possible, including the combination of L_1-norm and L_2-norm penalty (elastic net), which may be successful in specific applications but overall have not shown superior performance to LASSO. Efron et al. (2004) have designed an efficient algorithm to determine the entire LASSO *regularization path* (that is, the values of the coefficients as λ is varied), increasing the popularity of the method, since this obviates the need for the user to search manually for the best λ by varying across the entire range of the regularization parameter. The LASSO has been shown extremely effective in environments where the features are not highly correlated (Donoho, 2006), and more recent research endorses its use even under those circumstances (Meinshausen and Yu, 2009). We have used K. Skoglund's implementation to determine the entire LASSO regularization path[30].

4.2.2.2 Minimum Redundancy Maximum Relevance (mRMR)

We have briefly indicated in the introduction of FS (§ 4.2.2) that accounting only for the relevance of the features in predicting the response often fails to account for overlapping information amongst the features. This has prompted the investigation of *pre-filtering* to reduce the number of features: this method combines pairs of features and computes correlation coefficients; when the correlation is above a high threshold (for example 0.95, one of the pair of features is removed (Little et al., 2009). The process continues until no more coefficients can be eliminated. Although this approach addresses the problem of *collinearity*

[30] The Matlab source code for computing the LASSO path is available at http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897

(presence of highly correlated features), it fails to remove all the non-contributing features towards predicting the response. Moreover, the feature that is removed between two correlated features is random, raising further questions about the effectiveness of this approach.

Intuitively, combining features with maximum relevance and minimum overlapping information could offer a near-optimal solution. Battiti (1994) proposed a compromise between relevance and redundancy:

$$\mathrm{FS}_{\mathrm{Battiti}}(\beta) = \max_{j \in Q-S}\left[\underbrace{I(\mathbf{f}_j, \mathbf{y})}_{relevance} - \beta \underbrace{\sum_{s \in S} I(\mathbf{f}_j, \mathbf{f}_s)}_{\text{redundancy}}\right] \tag{4.6}$$

where $\mathbf{f}_j$ denotes the j^{th} variable in the initial M-dimensional feature space, $\mathbf{f}_s$ is a variable that has been already selected in the feature index subset S (s is an integer, Q contains the indices of all the features in the initial feature space, that is $1 \ldots M$, S contains the indices of selected features and $Q - S$ denotes the indices of the features not in the selected subset), β is a parameter chosen to achieve a desired balance between the relevance and redundancy terms, and $I(\cdot)$ is the criterion used to quantify the relevance or redundancy. Battiti's (1994) algorithm is an incremental (*greedy*) search solution, which consists of the steps summarized in Table 4.2.

A major problem with the approach formalized by Eq. (4.6) is that it requires the specification of the free parameter β (which can be achieved using grid search and cross validation). Moreover, the optimal value of β may vary with the size of the feature subset. Peng et al. (2005) modified the criterion in Eq. (4.6) to avoid the fine tuning of the free parameter, proposing the *minimum redundancy maximum relevance* (mRMR) algorithm:

$$\text{mRMR} \stackrel{\text{def}}{=} \max_{j \in Q-S}\left[I(\mathbf{f}_j, \mathbf{y}) - \frac{1}{|S|}\sum_{s \in S} I(\mathbf{f}_j, \mathbf{f}_s)\right] \tag{4.7}$$

where $|S|$ is the cardinality of the selected subset. As in Battiti's (1994) study, Peng et al. (2005) used MI for relevance and redundancy, and the greedy search solution follows the same steps described above. In practice the mRMR filter approach is highly successful in many applications (Peng et al., 2005; Meyer et al., 2008), thereby justifying the intuitive concept that selecting features based on the compromise between relevance and redundancy may be more appropriate than relying solely on the naïve idea of selecting features only on the basis of strong association with the response.

Table 4.2: Incremental feature selection steps suggested by Battiti

1. (Selecting the first feature index) include the feature index j: $\max_{j \in Q}\left(I(\mathbf{f}_j, \mathbf{y})\right)$ in the initially empty set S, that is $\{j\} \rightarrow S$
2. (Selecting the next $m-1$ features, one at each step, by repeating the following) apply the criterion in Eq. (4.6) to select the next feature index j, and include it in the set: $S \cup \{j\} \rightarrow S$
3. obtain the feature subset by selecting the features $\{\mathbf{f}_j\}_{j=1}^{m}$, $j \in S$ from the original data matrix **X**.

More recently, Estevez et al. (2009) refined the criterion used in mRMR by dividing through the redundancy term with the minimum of the entropy $H(\cdot)$ of the two features. The argument for this adjustment is founded on the fact that the MI is bounded ($0 \leq \text{MI}(\mathbf{f}_j; \mathbf{f}_s) \leq \min\{H(\mathbf{f}_j), H(\mathbf{f}_s)\}$), and the use of the normalized version of the redundancy term compensates for the MI bias. The MI bias occurs due to finite number sampling, and is a common problem in MI estimation (Quinlan, 1986).

$$\mathrm{mRMR}_{\mathrm{normalized}} \overset{\mathrm{def}}{=} \max_{j \in Q-S}\left[I(\mathbf{f}_j, \mathbf{y}) - \frac{1}{|S|}\sum_{s \in S} NI(\mathbf{f}_j, \mathbf{f}_s)\right] \tag{4.8}$$

where $NI(\mathbf{f}_j, \mathbf{f}_s) = I(\mathbf{f}_j, \mathbf{f}_s)/\min\{H(\mathbf{f}_j), \mathrm{H}(\mathbf{f}_s)\}$.

Vinh et al. (2010) argued that the criterion $\mathrm{mRMR}_{\mathrm{normalized}}$ creates an imbalance in the relevance-redundancy relationship, and proposed normalizing the relevance term in addition to the redundancy term by using a similar transformation:

$$\mathrm{mRMR}_{\mathrm{normalized2}} \overset{\mathrm{def}}{=} \max_{j \in Q-S}\left[NI(\mathbf{f}_j, \mathbf{y}) - \frac{1}{|S|}\sum_{s \in S} NI(\mathbf{f}_j, \mathbf{f}_s)\right] \tag{4.9}$$

where $NI(\mathbf{f}_j, \mathbf{y}) = I(\mathbf{f}_j, \mathbf{y})/\min\{H(\mathbf{f}_j), \mathrm{H}(\mathbf{y})\}$.

They proceeded to demonstrate that for multi-class classification problems this adjustment is beneficial, whereas for binary classification problems the results did not differ compared to the approach endorsed by Estevez et al. (2009). We remark that the normalization of the relevance and the redundancy terms was empirically shown to be useful in this application because MI is not strictly upper bounded to a predefined value (as for example with correlation coefficients) but rather to the minimum entropy of the two random variables.

So far, we focused on two very important aspects of FS: relevance and redundancy. A further aspect of FS that is often underestimated or ignored is *feature complementarity*. Feature complementarity (also known as *conditional relevance*) quantifies the extent to which two or more features are strongly associated with the response variable *jointly*, whilst the same features may be only moderately associated with the response *individually*. This issue has been explicitly addressed in a number of recent studies, for example Meyer et al. (2008) and Brown et al. (2012). Meyer et al. (2008) extended mRMR to include up to second order interactions because in general this keeps algorithm complexity low, although in principle the interactions could be generalized to higher order. They demonstrated that their algorithm has the potential to outperform mRMR in some datasets, although it was not universally superior. This suggests that second order complementarity proves quite useful in some datasets, and their results may indicate that including higher order interactions could further improve the performance of the FS filter scheme. However, the evaluation of high order interactions is both computationally expensive and difficult to be accurately estimated, for example generalizing criteria such as MI (e.g. using total correlation). In § 4.2.3.3 we suggest one way to tackle the computation of high order interactions very efficiently (albeit compromising on accuracy) in a novel FS algorithm.

This section was reviewed thoroughly because many of these concepts will be used in novel FS algorithms discussed later (see § 4.2.3.1 and 4.2.3.3).

4.2.2.3 Gram-Schmidt Orthogonalisation (GSO)

The Gram-Schmidt Orthogonalization (GSO) is also a sequential forward FS algorithm, where a feature is selected at each step on the basis of being maximally correlated to the response and minimally correlated to the existing feature subset, so conceptually it similar to mRMR (Stoppiglia et al., 2003). The GSO algorithm projects the candidate features for selection at each step onto the *null space* of those features already selected in previous steps: the feature that is maximally correlated with the target in that projection is selected next. The procedure iterates until the number of desired features has been selected. Further details of the GSO algorithm used for FS can be found in Stoppiglia et al. (2003) and in Guyon et al. (2006). We have used the implementation of Guyon (2008).

4.2.2.4 RELIEF

RELIEF was proposed as a heuristic FS algorithm by Kira and Rendell (1992), and selects features that contribute to the separation of samples from different classes. Originally, RELIEF was limited to binary classification applications, but was extended to multi-class classification applications by Kononenko (1994) and to regression applications by Robnik-Sikonja and Kononenko (1997). RELIEF is a *feature weighting* algorithm, where each feature is assigned a weight depending on how "useful" it is in the context of predicting the response. Conceptually, features which do not contribute towards predicting the response will be associated with very low weights. Ultimately, the user selects a cut-off for the weight values, effectively deciding on the number of features which will be used (this corresponds to selecting m, and can be optimized by cross-validation).

The principle of RELIEF is similar to the k-nearest neighbour classifier (see § 4.3.1), making use of the concept of *Nearest Hit* (NH) and *Nearest Miss* (NM). Given a data sample, NH refers to that sample's nearest neighbour which belongs to the same class, and NM refers to the nearest neighbour which belongs to a different class. RELIEF aims to select features which contribute to the separation of samples into differing classes, and therefore takes a very different approach to addressing the problem of the curse of dimensionality, by comparison to the preceding FS algorithms. RELIEF takes the algorithmic form in Eq. (4.10):

$$W(\mathbf{f}_j) \overset{\text{def}}{=} \frac{1}{q}\sum_{i=1}^{q}\left\{\underbrace{-\frac{1}{|\mathrm{NH}(\mathbf{x}_i)|}\cdot\sum_{\mathbf{x}_n\,\in\,\mathrm{NH}(\mathbf{x}_i)}\left\|x_{i,j}-x_{n,j}\right\|}_{\mathit{Nearest\ hit\ term\ distance}} + \underbrace{\sum_{y_l\neq y_i}\frac{1}{|\mathrm{NM}(\mathbf{x}_i)|}\cdot\frac{\mathrm{P}(y=y_l)}{1-\mathrm{P}(y=y_i)}}_{\mathit{Normalizing\ factor\ with\ prior\ probabilities}}\cdot\underbrace{\sum_{\mathbf{x}_n\,\in\,\mathrm{NM}(\mathbf{x}_i)}\left\|x_{i,j}-x_{n,j}\right\|}_{\mathit{Nearest\ miss\ term\ distance}}\right\} \quad (4.10)$$

where $W(\mathbf{f}_j)$ refers to the weight associated with the j^{th} feature, q represents the number of instances randomly sampled from the data (potentially we can use $q = N$ to exhaustively search the entire data sample space)[31], $\mathbf{x}_i$ refers to a data sample (row in the design matrix **X**), $|\cdot|$ refers to the size of nearest hits or nearest misses, $\|\cdot\|$ is a distance metric (the Euclidean distance or the Manhattan distance are often used). Typically, the size of nearest hits $|\mathrm{NH}(\mathbf{x}_i)|$ and the size of nearest misses $|\mathrm{NM}(\mathbf{x}_i)|$ are fixed to some pre-specified value, e.g. 10 according to Kononenko (1994).

There exist attempts to generalize RELIEF to regressions settings, but these are beyond the scope of this work. More recently, there has been extensive research interest to theoretically justify RELIEF's successful performance in many practical settings. For example, Gilad-Bachrach et al. (2004) reported that RELIEF is related to hypothesis margin maximization (we will see more about this concept in the section on support vector machines in § 4.3.2). The RELIEF family of algorithms has applications beyond FS; it has been very successful in a broad spectrum of machine learning applications, including split selection in decision trees and inductive logic programming. For a general overview of RELIEF in machine learning we refer to Robnik-Sikonja and Kononenko (2003) and the references therein.

4.2.2.5 Local Learning Based Feature Selection (LLBFS)

The Local Learning Based Feature Selection (LLBFS) was originally inspired by RELIEF and was proposed by Sun et al. (2010). Its developers only demonstrated how the algorithm works in binary classification problems, and the analysis in this section focuses only on cases where the response variable is binary; we will later extend LLBFS to the multi-class

[31] In this study we set $q = N$ to obtain a deterministic version of RELIEF (using a subset of the available samples leads to stochastic results): using all samples in RELIEF to infer the feature weights was referred to as Relieve*d* by Kohavi and John (1997).

classification scenario in § 4.2.3.2. The algorithm aims to decompose the intractable, exhaustive combinatorial problem of FS into a set of locally linear problems through *local learning*. The original features are assigned feature weights which denote their importance to the classification problem, and the features with the maximal weights are then selected (similarly to RELIEF, the user needs to set a threshold). The local linearization of the global problem of selecting the most appropriate features for predicting the response stems from the use of a margin function which focuses on the neighbourhood of the investigated data samples. LLBFS is based on the definition of *margin* ξ (which is used implicitly in RELIEF):

$$\xi_i = \underbrace{\|\mathbf{x}_i - \mathrm{NM}(\mathbf{x}_i)\|}_{Nearest\ miss\ sample\ distance} - \underbrace{\|\mathbf{x}_i - \mathrm{NH}(\mathbf{x}_i)\|}_{Nearest\ hit\ sample\ distance} . \tag{4.11}$$

Sun et al. (2010) proposed generalizing Eq. (4.11) by introducing a non-negative weight vector **w**, which scales each feature to obtain a *weighted* feature space onto which the margins are computed:

$$\xi_i(\mathbf{w}) = \underbrace{\|\mathbf{x}_i - \mathrm{NM}(\mathbf{x}_i)|\mathbf{w}\|}_{Nearest\ miss\ sample\ distance} - \underbrace{\|\mathbf{x}_i - \mathrm{NH}(\mathbf{x}_i)|\mathbf{w}\|}_{Nearest\ hit\ sample\ distance} . \tag{4.12}$$

We can extend the idea of Eq. (4.12) to estimate the expectation of the margin $\mathrm{E}(\xi_i(\mathbf{w}))$:

$$\begin{aligned}\mathrm{E}(\xi_i(\mathbf{w})) = \mathbf{w}^{\mathrm{T}} \Bigg(& \underbrace{\sum_{n \in \mathcal{M}_i} \mathrm{P}(\mathbf{x}_n = \mathrm{NM}(\mathbf{x}_i)|\mathbf{w}) \cdot \{x_{i,j} - x_{n,j}\}_{j=1}^{M}}_{Nearest\ miss\ samples} \\ & - \underbrace{\sum_{n \in \mathcal{H}_i} \mathrm{P}(\mathbf{x}_n = \mathrm{NH}(\mathbf{x}_i)|\mathbf{w}) \cdot \{x_{i,j} - x_{n,j}\}_{j=1}^{M}}_{Nearest\ hit\ samples} \Bigg) = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{z}_i\end{aligned} \tag{4.13}$$

where $\{x_{i,j} - x_{n,j}\}_{j=1}^{M}$ is a vector containing the element-wise differences between $\mathbf{x}_i$ and $\mathbf{x}_n$, $\mathcal{M}_i$ contains the indices of the data samples belonging to a different class compared to sample

$\mathbf{x}_i$, and $\mathcal{H}_i$ contains the indices of the data samples belonging to the same class as sample $\mathbf{x}_i$. For convenience in notation we defined $\mathbf{z}_i = \sum_{n\,\in\,\mathcal{M}_i} \mathrm{P}(\mathbf{x}_n = \mathrm{NM}(\mathbf{x}_i)|\mathbf{w}) \cdot \left\{x_{i,j} - x_{n,j}\right\}_{j=1}^{M} - \sum_{n\,\in\,\mathcal{H}_i} \mathrm{P}(\mathbf{x}_n = \mathrm{NH}(\mathbf{x}_i)|\mathbf{w}) \cdot \left\{x_{i,j} - x_{n,j}\right\}_{j=1}^{M}$. The probabilities of hit or miss are obtained from probability density functions, which in turn are computed using kernel density estimation (see § 3.1 for a brief introduction). Finally, the weights **w** which reflect the importance of the features are computed using optimization of a logistic regression problem, where an additional regularization parameter λ is introduced to promote sparsity (in principle, the induced penalty is similar to LASSO, penalising the absolute value of the weight coefficients $\|\mathbf{w}\|_1$):

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \log\left(1 + \exp(-\mathbf{w}^{\mathrm{T}} \cdot \mathbf{z}_i)\right) + \lambda \cdot \|\mathbf{w}\|_1 \text{, subject to } \mathbf{w} \geq 0. \qquad (4.14)$$

The free parameters in LLBFS are the sparsity parameter λ (by default $\lambda = 1$), the distance metric (the authors of the algorithm used the Manhattan distance), and the kernel width for the computation of the probability densities (optimized using cross-validation). Quite importantly for this application, Sun et al. (2010) demonstrated that LLBFS is fairly robust to the choice of the free parameters.

4.2.3 Novel feature selection algorithms

The preceding sections have described some widely used FS algorithms, and highlighted some of their shortcomings. In this section we extend the available mehods and propose an entirely new FS algorithm which is computationally simple and addresses many of the deficiencies of the current algorithms.

4.2.3.1 A minimum redundancy maximum relevance approach for feature selection using alternative criteria to mutual information

We have already discussed the concepts of relevance and redundancy, and how we can build a practical compromise between them in the definition of mRMR. In particular, Peng et al. (2005) and most of the literature on the FS topic uses MI to quantify the statistical relationships between features and response, and between features. Here, we investigate the use of different *metrics* in the same mRMR type format. The use of alternative divergences suggests itself (since MI is the application of KL divergence using the joint probability density between two random variables, and the product of their marginal probability densities) and is amenable to experimentation. The motivation is that differences in probability density functions might be more appropriately expressed using a different criterion to the KL divergence. Therefore, we used Eq. (4.7) substituting MI with the divergences which were summarized in Table 4.1. Each of the resulting new algorithms will be referred to using the subscript of the corresponding divergence. In all cases, the probability densities for the computation of the divergences were computed using kernel density estimation with Gaussian kernels. The bandwidth of the kernel was computed using *likelihood cross-validation*, which is commonly used for bandwidth selection, e.g. Gray and Moore (2003).

In addition, we used the Spearman correlation coefficient instead of the MI to account for relevance and redundancy in Eq. (4.7), and we call this FS algorithm $\text{mRMR}_{\text{Spearman}}$. The $\text{mRMR}_{\text{Spearman}}$ can be thought of as a simple, computationally efficient alternative to the standard mRMR which relies on MI, and was endorsed as a practical and simple FS algorithms as recently discussed in Tsanas et al. (2012a). Although using correlation coefficients as a criterion in mRMR is mentioned in passing by Peng et al. (2005), to the best of our knowledge this is the first time this idea has been explored in practice.

4.2.3.2 Extending LLBFS to multi-class classification problems

The LLBFS algorithm was originally proposed to tackle binary classification problems. Sun et al. (2010) briefly mentioned an approach to generalizing their algorithm to multi-class classification settings without actually providing sufficient details or validating its effectiveness. Here, we suggest generalizing LLBFS to multi-class classification problems by decomposing the multi-class classification problem into several binary sub-problems. The suggested approach is inspired by the literature on support vector machines, which will be described in § 4.3.2. Although the context is different (support vector machines are classifiers, not FS algorithms) the generalization is identical in that it uses binary sub-problems: a) we can work with data from every possible pair of classes in the original dataset, and treat the problem as binary classification (this approach is known as One-Against-One, abbreviated as OAO), or b) compare each class in **X** against all the remaining classes, which are treated as a single class (this approach is known as One-Against-All, abbreviated as OAA). These approaches to treating multi-class classification problems as binary classification sub-problems will be described in detail in the section on support vector machines in § 4.3.2.

We have seen that contrary to many competing FS algorithms, LLBFS is a feature weighting approach, where we set some threshold below which we discard features. Splitting the original multi-class classification problem to many binary sub-problems, creates an additional difficulty: how to set the threshold for each of the sub-problems. A sensible approach would be to include those features which appear with relatively large weights in each of the sub-problems, and also those features associated with lower weights which appear in *many* of the binary sub-problems. Additionally we could take into account the number of samples for each binary sub-problem; the premise being to emphasize good discrimination amongst classes with large numbers of data samples. More formally, we need to develop an algorithm which will rank the features in descending order of importance similarly to mRMR. Based on empirical experimentation (see Appendix I for details), we propose the following approach to select features in multi-class classification settings:

$$\mathbf{w} = \mathbf{n}^{\mathrm{T}} \cdot \mathbf{W} \tag{4.15}$$

where $\mathbf{n}$ is an L-dimensional vector, with L denoting the number of binary sub-problems, and the i^{th} entry ($i = 1 \dots L$) is a scalar which is equal to: (a) the number of samples of the i^{th} class for the i^{th} binary sub-problem in OAA, or (b) the number of samples for the two classes under investigation in OAO. $\mathbf{W}$ is an $L \times M$ dimensional matrix (i.e. number of binary sub-problems × dimensionality of the original design matrix), which contains the weights for each binary sub-problem in each of the L rows. The resulting weights of the features are then summarized in the M dimensional $\mathbf{w}$. Through empirical validation, we have found that the OAO generalization may be preferable in this application (interestingly, Hsu and Lin (2002) reported that OAO is the simplest and probably preferable approach to generalize the support vector machine to multi-class classification problems).

4.2.3.3 Relevance Redundancy and Complementarity Trade-off (RRCT)

We propose a new FS algorithm which attempts to include all the major components outlined above for efficient FS: relevance, redundancy and complementarity. The proposed correlation-based filter builds on the $\mathrm{mRMR}_{\mathrm{Spearman}}$ discussed previously, by incorporating a complementarity term. It relies on the computation of correlation coefficients, which are subsequently transformed using a function inspired by *information theoretic* (IT) *concepts*. We invoke these IT concepts under the assumption that the features are normal, which is common in diverse machine learning applications and often works well in practice (Bishop, 2007). This assumption greatly facilitates analysis since important IT concepts that are of central importance to this new algorithm are simple to compute and to work with analytically. The features and the response variable are standardized to have zero mean and unit standard deviation before further processing. This is also a common pre-processing step in machine learning applications, facilitating subsequent analysis: for example, it finds use in LASSO (Tibshirani, 1996) and in mRMR (Peng et al., 2005).

First, we compute the Spearman rank correlation coefficient between the features and the response variable to obtain the vector of rank correlations $\mathbf{r} = [r_1, r_2 \dots r_M]$, where each entry denotes the correlation of each feature with the response. We used the Spearman rank correlation coefficient over the linear correlation coefficient, as a more general method to express the relationship between variables. Then, we compute the covariance matrix $\mathbf{\Sigma}$, and denote its entries ρ_{ij}: these entries are the Spearman rank correlation coefficients computed between the features $\mathbf{f}_i$ and $\mathbf{f}_j$, where $i, j \in (1 \dots M)$.

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{12} & 1 & \cdots & \rho_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1M} & \rho_{2M} & \cdots & 1 \end{bmatrix}. \tag{4.16}$$

For the Gaussian distribution, there is an analytic expression for MI that depends only on the linear correlation coefficient ρ (Cover and Thomas, 2006) (note that MI also relies on the variance, but this is 1 due to the standardization step):

$$MI = -0.5 \cdot \ln(1 - \rho^2). \tag{4.17}$$

Eq. (4.17) leads to an IT quantity (MI) that is obtained using the linear correlation coefficient:

here we will use the same notion to define an IT quantity exactly as in Eq. (4.17), only that this time the Spearman correlation coefficient will be used. For convenience, we will use the notation $r_{\mathrm{IT}}(X,Y) = -0.5 \cdot \log[1 - r_{XY}^2]$ to refer to the non-linearly transformed rank correlation coefficient r_{XY} between two random variables X, Y. Now, we can write in compact vector form all the *relevance* terms using the IT inspired transform in Eq. (4.17):

$$\mathbf{r}_{\mathrm{ITL}} = -0.5 \cdot \log[1 - r_1^2 \quad \cdots \quad 1 - r_M^2]. \tag{4.18}$$

Similarly, using the covariance matrix $\mathbf{\Sigma}$ and Eq. (4.17), the *redundancy* between pairs of features can be conveniently expressed as a matrix, where each (i,j) entry denotes the information that two features share in predicting the response:

$$\mathbf{\Sigma}_{\mathrm{IT}} = -0.5 \cdot \log \begin{bmatrix} 1 & 1-\rho_{12}^2 & \dots & 1-\rho_{1M}^2 \\ 1-\rho_{12}^2 & 1 & \cdots & 1-\rho_{2M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho_{1M}^2 & 1-\rho_{2M}^2 & \dots & 1 \end{bmatrix}. \tag{4.19}$$

Now, inserting the relevance terms in Eq. (4.18) across the main diagonal of $\mathbf{\Sigma}_{\mathrm{IT}}$ in Eq. (4.19), we obtain a matrix which will be used to compute the compromise between relevance and redundancy:

$$\mathbf{D} = -0.5 \cdot \log \begin{bmatrix} 1-r_1^2 & 1-\rho_{12}^2 & \dots & 1-\rho_{1M}^2 \\ 1-\rho_{12}^2 & 1-r_2^2 & \cdots & 1-\rho_{2M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho_{1M}^2 & 1-\rho_{2M}^2 & \dots & 1-r_M^2 \end{bmatrix}. \tag{4.20}$$

The matrix $\mathbf{D}$ is essentially a compact form of mRMR relying on the IT quantity of Eq. (4.17) which alleviates the need for repeated computation of the relevance and complementarity terms in the iterative steps (therefore this expedites the incremental FS process in large datasets). Conceptually, the IT transformation of the rank correlation coefficient assigns greater weight to coefficients above the absolute value 0.5 (see Fig. 4.1). The effect is that weak associations (between a feature and the target or between features) are penalized; conversely strong associations (large absolute correlation coefficients) are enhanced. Compared to MI, the new IT quantity is bounded unless the correlation coefficient has an absolute value 1 which is highly unlikely in practice; therefore no additional normalization, such as dividing by the entropy, is necessary (e.g. see the mRMR extension in Eq. (4.8) and Eq. (4.9)). If absolute value of the rank correlation coefficient is 1, we set the MI quantity to a very large value (we chose 1000).

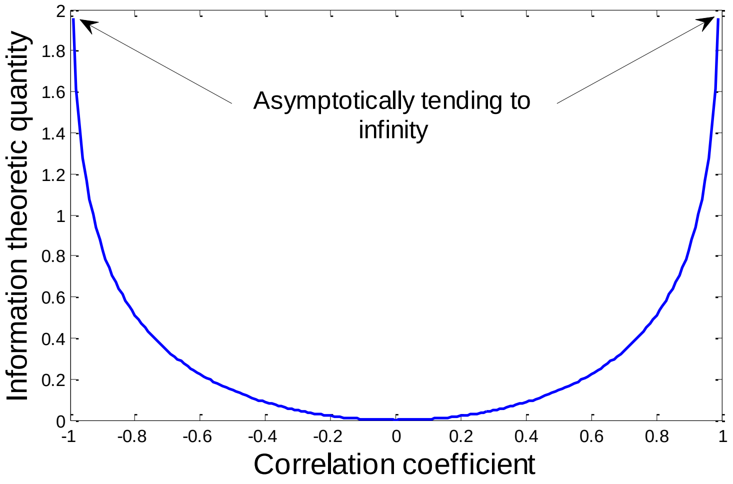

Figure 4.1: Information theoretic (IT) quantity (relevance or redundancy) as a function of the rank (Spearman) correlation coefficient ρ, computed as $\mathrm{I}(\rho) = -0.5 \cdot \log(1-\rho^2)$. Asymptotically, as the absolute value of the correlation coefficient tends to ± 1, the IT quantity becomes infinite (in practice we set this to a very large value). We demonstrate that this IT nonlinear transformation of the correlation coefficients is valuable in feature selection.

The proposed algorithm developed thus far can be seen as an extension of the classical mRMR using an *information theoretic* inspired transformation, and for this reason we call it mRMR$_{\text{ITL}}$. Thus, the mRMR$_{\text{ITL}}$ is conveniently calculated in terms of the matrix **D**, where for the computation of the new candidate feature $\mathbf{f}_j$ (which corresponds to a feature not in the existing feature subset) we focus on the i^{th} row. The relevance of the feature $\mathbf{f}_j$ lies on the main diagonal of the matrix **D**, and the redundancy is computed from the average of the terms that appear in the column s (the $\mathrm{D}_{i,s}$ entries) where s corresponds to features in the already selected subset ($s \epsilon S$).

Now, we embrace the concept of quantifying the *conditional relevance* (complementarity) of a feature as the usefulness of that feature in predicting the response *conditional upon the already selected feature subset*. This is achieved using the rank *partial correlation coefficient*, which quantifies the statistical association between two random variables X, Y whilst controlling for the effect of a set of a conditioning random variable Z. This is defined as:

$$r_{\mathrm{p}}(X,Y|Z) = \frac{N \cdot \sum_{i=1}^{N} r_{X,i} \cdot r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \cdot \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \cdot \sum_{i=1}^{N} {r^2}_{X,i} - \left(\sum_{i=1}^{N} r_{X,i}\right)^2} \cdot \sqrt{N \cdot \sum_{i=1}^{N} {r^2}_{Y,i} - \left(\sum_{i=1}^{N} r_{Y,i}\right)^2}} \tag{4.21}$$

where $r_{X,i}$ and $r_{Y,i}$ denote the residuals of X, Y, respectively, on Z. That is, the partial correlation coefficient is computed by first solving the two associated linear regression problems, and calculating the correlation between their residuals. Alternatively, the partial correlation coefficient can be computed using a recursive formula working directly with correlation coefficients: the n^{th} order partial correlation (that is, the conditioning random variable Z contains n features) is computed from three $(n - 1)$ order partial correlations (the 0th order partial correlations are by definition the correlation coefficients). For the simplest case where the conditioning random variable Z comprises a single feature, this reduces to Eq. (4.22):

$$r_{\mathrm{p}}(X,Y|Z) = \frac{r(X,Y) - r(X,Z)\cdot r(Y,Z)}{\sqrt{r^2(X,Z)}\cdot\sqrt{r^2(Y,Z)}}. \tag{4.22}$$

The partial correlation coefficient expresses the contribution of the independent random variable X over and above the contributions of the conditioning random variable Z for predicting the dependent random variable Y, and accounts for the *additional* explanation of the variance observed in Y as a result of including X in the regression setting. Fig. 4.2 presents a Venn diagram to graphically illustrate this point, where the different regions denote the information captured by each random variable, and the overlapping regions denote the shared information between the random variables.

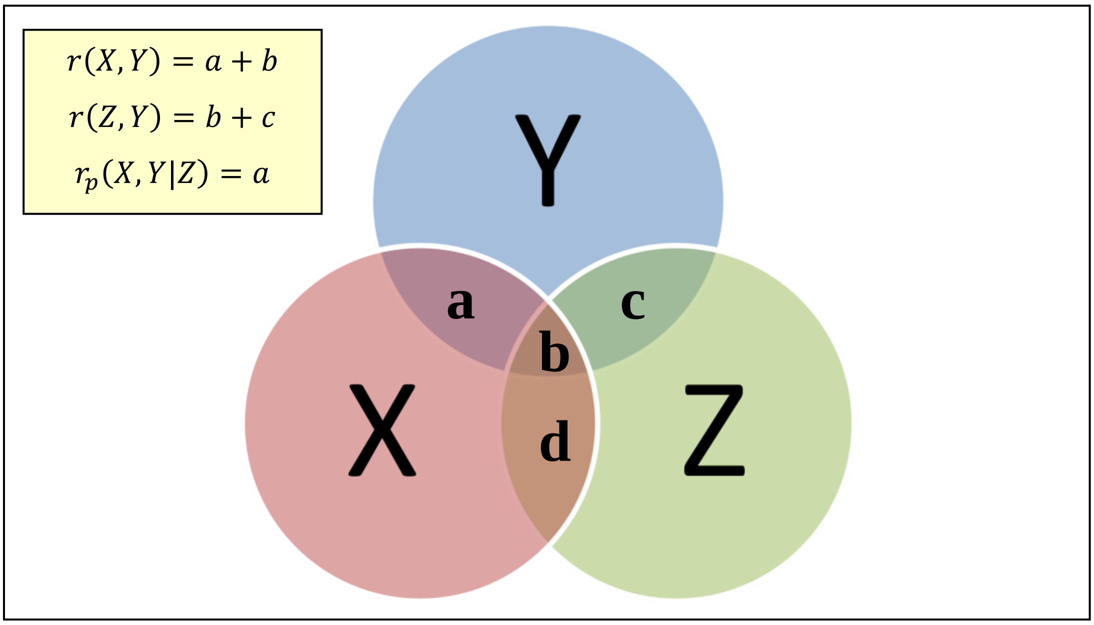

Figure 4.2: Graphical representation of the effect of the partial correlation coefficient. The lower case letters represent the shared information between the random variables.

In the context of the developed FS algorithm, the partial correlation coefficient r_{p} is defined as the rank correlation coefficient between a new candidate feature $\mathbf{f}_j$ and the response $\mathbf{y}$, controlling for the existing features in the subset, i.e. $r_{\mathrm{p}}\left(\mathbf{f}_j, \mathbf{y}|S\right)$. This approach aims to incorporate how well the candidate feature pairs up with the existing features that have already been chosen. Then, we transform the computed partial correlation coefficient using the IT inspired transformation in Eq. (4.17), which gives:

$$r_{\mathrm{p,IT}} = -0.5 \cdot \log\left[1 - r_{\mathrm{p}}^2\right]. \tag{4.23}$$

Since the controlling variables S (whose effect needs to be removed to compute the partial correlation coefficient) are not known and will vary at each step, it is not possible to express this quantity in vector or matrix form as we did above for **D**.

This additional term in Eq. (4.23) is added to mRMR$_{\mathrm{ITL}}$, and we therefore obtain the new FS algorithm which we call *relevance, redundancy and complementarity trade-off* (RRCT):

$$\begin{aligned}\mathrm{RRCT} \overset{\mathrm{def}}{=} \max_{j\,\in\,Q-S} \Bigg[& r_{\mathrm{IT}}\left(\mathbf{f}_j, \mathbf{y}\right) - \frac{1}{|S|}\sum_{s\,\in\,S} r_{\mathrm{IT}}\left(\mathbf{f}_j, \mathbf{f}_s\right) + \mathrm{sign}\left(r_{\mathrm{p}}\left(\mathbf{f}_j, \mathbf{y}|S\right)\right) \\ & \cdot \mathrm{sign}\left(r_{\mathrm{p}}\left(\mathbf{f}_j, \mathbf{y}|S\right) - r\left(\mathbf{f}_j, \mathbf{y}\right)\right) \cdot r_{\mathrm{p,IT}} \Bigg]\end{aligned} \tag{4.24}$$

sign(·) returns +1 if the quantity (·) is positive and -1 if (·) negative, and is used to determine whether $r_{\mathrm{p,IT}}$ is added or subtracted in Eq. (4.24). RRCT follows Battiti's (1994) algorithmic steps (see Table 4.2) using Eq. (4.24) instead of Eq. (4.5) to select features. Care needs to be exercised in the RRCT expression when including the $r_{\mathrm{p,IT}}$ term. Given that this term is non-negative due to the IT transformation, we need to determine whether the inclusion of the candidate feature to the existing subset actually contributes *additional* information conditional on the features in the selected subset (conditionally relevant). Consideration must be made of both the sign of the partial correlation coefficient, and the sign of the difference in magnitudes between $r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S)$ and $r(\mathbf{f}_j, \mathbf{y})$. The sign $\left(r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S) - r(\mathbf{f}_j, \mathbf{y})\right)$ term in Eq. (4.24) is used to determine whether the conditional relevance $r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S)$ is larger than $r(\mathbf{f}_j, \mathbf{y})$ magnitude; that would suggest that including the candidate feature has additional (conditional) relevance given the features in the selected subset. The sign $\left(r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S)\right)$ term is used to make the overall complementarity contribution positive in the case that $r(\mathbf{f}_j, \mathbf{y}) < 0$, $r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S) < 0$ and $\left(r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S) - r(\mathbf{f}_j, \mathbf{y})\right) < 0$, because then the term sign $\left(r_{\mathrm{p}}(\mathbf{f}_j, \mathbf{y}|S) - r(\mathbf{f}_j, \mathbf{y})\right)$ would indicate the additional contribution offered by the complementarity term is negative.

To isolate the advantages of using the partial correlation coefficient from the advantages of using the IT transformation in mRMR$_{\mathrm{ITL}}$, we define an alternative FS algorithm, *RRCT*$_0$. RRCT$_0$ is identical to Eq. (4.24) except that all the terms (relevance, redundancy, and complementarity) have not undergone IT transformation. That is, we use the raw correlation coefficients and the raw partial correlation coefficient instead.

We aim to demonstrate that the simple nonlinear transformation of the correlation coefficients using IT concepts derived under the assumption of Gaussianity, brings a tangible advantage in FS over alternative approaches (for example, over the mRMR$_{\mathrm{Spearman}}$ scheme). Moreover, introducing the conditional relevance term that controls for the existing features in the selected subset at each iteration, combined with the IT transformation, brings additional power in selecting a parsimonious feature subset rich in information content.

So far, the IT approach has assumed that all the distributions of the features and the response are Gaussian. Because this may be substantially inaccurate in some circumstances, we can use the *Box-Cox transform*, which aims to normalize non-Gaussian random variables (Box and Cox, 1964). The Box-Cox transformation (see Eq. 4.25) belongs to a family of *power transformations*, and takes the form:

$$f(x,\lambda) = \begin{cases} \frac{(x^{\lambda} - 1)}{\lambda}, \lambda \neq 0 \\ \log(x), \lambda = 0 \end{cases} \tag{4.25}$$

where λ is determined via optimization to maximize the associated log likelihood function.

There is active research into the determination of the optimal λ (Marazzi and Yohai, 2006) which is beyond the scope of this work, and here we will use a standard maximum likelihood estimate. We apply the Box-Cox transform to the raw data prior to standardization, and compute the RRCT on this transformed data, in addition to RRCT for the non-transformed data. This is indicated as $\text{RRCT}_{\text{Box-Cox}}$ for convenience.

4.2.4 Summary of the feature selection schemes and a methodology for selecting features

We have looked into some detail at a few FS algorithms in the preceding sections. These algorithms are summarized in Table 4.3 to facilitate their comparison in terms of the main FS properties, i.e. relevance, redundancy, complementarity. We should note that in a practical setting it would be wrong to use the entire design matrix to determine the feature set, and then use this feature set to test the performance of the model using, for example, cross-validation (for details regarding cross validation see § 4.4.1). Instead, feature sets need to be selected using cross-validation (CV), which is a more realistic setting (Hastie et al., 2009). Ideally, we should obtain the same feature subset in all cross-validation replications which would clearly indicate which features should be selected in the dataset. However, in practice the selected features for any given FS algorithm may be different across different CV replicates. Hence, we need to develop a strategy to select the features which appear most often under the investigated FS algorithm(s), to select one feature subset for each FS algorithm. Specifically we follow the methodology outlined in Tsanas et al. (2012b), which is summarized in Table 4.4.

The methodology in Table 4.4 is general and can be applied to any greedy FS scheme, i.e. all those FS algorithms which select features one at a time (this includes all the schemes described thus far, with the exception of LASSO). For non-greedy FS algorithms, we need to adapt this methodology to account for the fact that the Kth step does not necessarily include all the features selected in the preceding steps.

Table 4.3: Summary of the properties of the feature selection algorithms used in this study.

	Relevance	Redundancy	Complementarity	Information theoretic transformation	Box-Cox transformation
LASSO	X	X	-	-	-
GSO	X	X	-	-	-
RELIEF	X	-	X	-	-
LLBFS	X	-	X	-	-
mRMR$_{\text{MI}}$	X	X	-	-	-
mRMR$_{\text{Spearman}}$	X	X	-	-	-
mRMR$_{\text{ITL}}$	X	X	-	X	-
mRMR$_{\text{ITL,Box-Cox}}$	X	X	-	X	X
RRCT$_0$	X	X	X	-	-
RRCT	X	X	X	X	-
RRCT$_{\text{Box-Cox}}$	X	X	X	X	X

Table 4.4: Proposed methodology for selecting features using the greedy feature selection algorithms.

1. For the FS algorithm we want to investigate, an empty set S is created which will contain the indices of those features that will be selected.
2. First, randomly select 90% of the data samples from the original design matrix $\mathbf{X}$, along with their corresponding response variable values $\mathbf{y}$.
3. Run the FS algorithm to select features using the 90% of the randomly selected samples. The result is an ordered sequence of features where the first feature is considered the most important for this particular FS algorithm.
4. Repeat the steps 2-3 a number of times, say R_p, and store the results in a matrix $\mathbf{X}_{FS}$. In each of the $1 \ldots R_p$ rows of $\mathbf{X}_{FS}$ we store the selected feature subset.
5. The following voting scheme is then applied, to decide on the final feature subset for the FS algorithm. Feature indices are incrementally included, one at a time, in S. For each step K (K is a scalar taking values $1 \ldots M$) we find the indices corresponding to the features selected in the $1 \ldots K$ search steps for all the repetitions in step 4. That is, we work only on the $1 \ldots K$ columns of $\mathbf{X}_{FS}$ and identify the indices corresponding to the features selected in the first K FS steps.
6. We select the feature index which appears most frequently amongst these $R_p \times K$ elements and which is also not already included in S. This index is now included as the Kth element in S. Ties are resolved by including the lowest index number.
7. Repeat steps 5 and 6 for the number of features we want to ultimately use.

LASSO is not a greedy FS algorithm, since it may remove features in subsequent steps during its incremental FS search. Therefore, for LASSO we endorse repeating the random sample selection process independently for each Kth step, interrogating the algorithm to provide the best K features prior to the voting scheme explained in steps 5-6 of Table 4.4. Potentially, this enables the exclusion of features from the final LASSO feature subset which comprises K features, which may have been selected and removed in prior steps in the LASSO FS process. Once the final selected feature subset S is decided for each FS algorithm, these features can be presented to the learner in the subsequent mapping phase.

4.3 Mapping features to the response

As we have mentioned in the beginning of this chapter, in a wide range of problems we are interested in determining the function f which associates the features with the response, that is $f(\mathbf{X}) = \mathbf{y}$. This can be achieved in two ways: a) we can impose a structure on the functional form of f, and determine the parameters of that form (*parametric* learning), or b) allow the data itself determine the structure and the parameters of that structure (*non-parametric* learning). One example of the parametric setting has the form $y = a_1 x_1 + \cdots + a_M x_M$, where the parameters $\mathbf{a} = (a_1 \ldots a_M)$ need to be estimated. Parametric settings are generally simpler than non-parametric settings, and may be more easily interpretable. If the functional form (*model structure*) is known *a priori*, then parametric settings can be very useful providing a clearly interpretable framework. However, imposing an inappropriate functional form structure may lead to false interpretation of the properties of the data. Hence, in practice, non-parametric learners may often be more appropriate. Nevertheless, the issue on whether parametric or non-parametric learners should be used is not a settled matter, and is still a matter for debate in the statistics literature (Breiman, 2001a; Hand, 2006).

The following section aims to provide a general overview of classification approaches, and subsequently we describe two of the most powerful non-parametric classifiers which are widely used in many practical applications.

4.3.1 Overview of classification approaches

We have already mentioned that a fundamental aim in supervised learning is to train a learner using the data (features and response) available in the training set, so that it can automatically and accurately estimate the unknown response of new samples (testing set). Conceptually, one of the simplest classification approaches for assigning a class label to a new sample (*query point*) is to assign it the class of identified samples from the training set which are "close". The informal term "close" can be formulated in terms of a *distance metric*, for example the Euclidean distance. Thus, it is reasonable to classify the new sample, as having the class label of the closest train sample in the feature space. It is possible to use more than one sample (typically an odd number, for example 3 or 5) from the training set, and use majority voting to assign the class that appears most often amongst those samples closest to the test sample. The training samples which are close to the test sample are known as *neighbours*, and this intuitive and powerful classification method is known as the *k-Nearest Neighbour* (kNN) classifier. The free parameter *k* refers to the number of neighbours used in making the classification for the test sample, and can be optimized using cross-validation. Typically, each feature is standardized (zero mean and standard deviation equal to one) to avoid scaling problems amongst the features. Despite its simplicity, kNN has often demonstrated excellent results in practice (Michie et al., 1994; Hastie et al., 2009).

In general, nearest neighbour methods find wide applicability in diverse topics in the discipline of machine learning, including FS (for example see § 4.2.2.4 and 4.2.2.5), sample selection for the purpose of storage reduction (Marchiori, 2010) and entropy estimation (Kraskov et al., 2004). Amongst the attractive properties of kNN (with $k = 1$) is that the classifier's *asymptotic*[32] error rate is never more than *twice* the Bayes error rate (Cover and Hart, 1967; Ripley, 1996). This result can be tentatively used in providing an estimate of the best possible performance (the Bayes error rate) of a classifier for a given dataset. However, this finding can only provide some preliminary guidance, since in practice the bias could be substantial due to the finiteness of the data which cannot densely populate the entire feature space (Hastie et al., 2009). Therefore, in practice, more than a single neighbour is often used in kNN to avoid the bias problem and its susceptibility to noise. Moreover, there have been various attempts to refine the distance metric (Hastie and Tibshirani, 1996; Paredes and Vidal, 2006), and extend kNN to a probabilistic setting (Holmes and Adams, 2002).

Another intuitive approach for classification of new samples, which is conceptually different to kNN, focuses on creating *decision boundaries* in the feature space, where the decision for the class assignment for a new sample is made depending on the side of the boundary it belongs. The use of *linear boundaries* suggests itself as a simple approach to construct *hyperplanes*[33] which discriminate pairs of classes resulting in a division of the feature space into regions which are assigned to a class. Two popular classification methods, the linear discriminant analysis and the logistic regression fall into this category. For further information about these methods we refer to Hastie et al. (2009). One generalization of the decision boundaries concept will be described in the following section which describes the support vector machines.

The concepts of nearest neighbours and constructing boundaries underlie many of the more sophisticated classification methods (Hastie et al., 2009). We will now describe two of the

[32] *Asymptotic* here refers to the assumption of having an infinite number of samples for training the learner.

[33] Some authors reserve the term *hyperplanes* for boundaries that pass through the origin and use the term *affine sets* for those boundaries which do not; in the context of this work we do not make such a distinction.

most commonly used classifiers which have shown very promising results in many practical applications.

4.3.2 Support vector machines

Support Vector Machines (SVM) were popularized by Vapnik (1995) and have attracted great research interest in the machine learning community over the past decade. The definition of margin, which was introduced in LLBFS in Eq. (4.10), is critical in SVM. Similarly to LLBFS, we focus on binary classification problems, and will generalize the concepts to the multi-class classification setting later. To simplify the representation of the equations we will assume that the classes in the response variable can be either -1 or 1, that is $y \epsilon \{-1, +1\}$. Conceptually, SVM constructs a *decision boundary* (separating hyperplane) in the feature space maximizing the margin between samples which belong to the two different classes. By the law of large margin theory, it is expected this will provide good generalization accuracy for unknown data. Those data samples which form the decision boundary upon which future samples (from a new dataset) will be classified, are called *support vectors* and hence the name of this learner. In its simplest form, and assuming the existence of a linear boundary to separate the two classes, SVM can be written as:

$$\left\{\begin{matrix} \mathbf{w}^T\mathbf{x}_i + b_0 \geq +1, \text{for } y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b_0 \leq -1, \text{for } y_i = -1 \end{matrix}\right\} \Rightarrow y_i \cdot (\mathbf{w}^T\mathbf{x}_i + b_0) \geq 1, \forall\, i \,\epsilon \{1 \dots N\} \qquad (4.26)$$

where $\mathbf{w}$ represents the weight vector, and b_0 is the intercept. The optimal hyperplanes are then denoted by H_1: $\mathbf{w}^T\mathbf{x}_i + b_0 = 1$ and H_2: $\mathbf{w}^T\mathbf{x}_i + b_0 = -1$, and are computed by minimizing $||\mathbf{w}||^2$ subject to the constraint in Eq. (4.26). Generalizing this concept to account for nonlinear boundaries, SVM aim to solve the following optimization problem, known as the *primal problem* L_P:

$$L_P = \min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{i=1}^{N} \xi_i \tag{4.27}$$

$$\text{subject to: } \begin{Bmatrix} y_i \cdot (\mathbf{w}^T\varphi(\mathbf{x}_i) + b_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{Bmatrix}, \forall\, i \,\epsilon\, \{1 \ldots N\}.$$

Here, ξ_i are *slack variables* representing the margin of each data sample from the separating hyperplane, C $(C \geq 0)$ is a regularization parameter compromising between complexity and misclassified data samples, and $\varphi(\cdot)$ is a function which maps (*projects*) the samples from the M-dimensional feature space to a larger (potentially infinite) dimensional space. This is done in order to ensure that the samples are linearly separable in the new feature space. This means that SVM do not attempt to model a nonlinear decision boundary *per se*, but rather that they build a linear decision boundary in the *transformed* feature space. A data sample $\mathbf{x}_i$ is misclassified when $\xi_i > 1$, so it is desirable to bound the slack variables via C. Large values of the regularization parameter C discourage large values of the slack variables, but may lead to *overfitting* the data (SVM learns particular characteristics of the dataset used to train the classifier, and these characteristics may not hold in general); optimizing the value of C is usually achieved by cross-validation (see § 4.4.1).

In practice, we will see that an explicit definition of the function $\varphi(\cdot)$ is not required; instead a *kernel function* $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \cdot \varphi(\mathbf{x}_j)$ needs to be defined (this will become clear below). There are many types of kernel functions, but here we focus on the commonly used *radial basis function* (RBF) $K_{RBF}(\mathbf{x}_i, \mathbf{x}_j)$, which often works well in practical applications (Hsu et al., 2010):

$$K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \cdot \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \gamma > 0. \tag{4.28}$$

For simplicity in the subsequent notation, the subscript RBF will be dropped from the kernel function name. The kernel parameter γ controls the width of the kernel and is the second free parameter in SVM with an RBF kernel (different kernel functions require different parameters to be optimized). In the context of this thesis, we will work only with RBF kernels, and hence the examined SVM will always have just two degrees of freedom.

Using the Lagrangian formulation of the primal optimization problem above, the original L_P formulation is transformed to the *dual problem* L_D, which is what SVM actually solve to determine the support vectors:

$$L_D = \max \sum_{i=1}^{N} \xi_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_i \cdot \xi_j \cdot y_i \cdot y_j \cdot \underbrace{\varphi(\mathbf{x}_i)^T \cdot \varphi(\mathbf{x}_j)}_{K(\mathbf{x}_i, \mathbf{x}_j)} \tag{4.29}$$

$$\text{subject to:} \begin{Bmatrix} \sum_{i=1}^{N} \xi_i \cdot y_i = 0 \\ 0 \leq \xi_i \leq C \end{Bmatrix}, \forall\, i \,\epsilon\, \{1 \dots N\}.$$

We remark that the dual problem does not explicitly depend on $\mathbf{w}$ and b_0. The solution to Eq. (4.28) gives a trained SVM, where the support vectors are those data samples where $\xi_i > 0$. Now, when we want the trained SVM to classify a new data sample $\mathbf{s}$ (similarly to the data samples used for training this sample should also be an M-dimensional vector), the decision to assign it to -1 or $+1$ is:

$$f(\mathbf{s}) = sign\left(\sum_{j=1}^{N} \xi_i \cdot y_i \cdot K(\mathbf{x}_i, \mathbf{s}) + b_0\right). \tag{4.30}$$

SVM are extremely sensitive to the specification of the two free parameters (C, γ), and it is essential they are properly optimized using a grid search since it is not easy to select *a priori* good values. We followed the suggestion of Hsu et al. (2010) for the specification of the grid search (C, γ): $\{C = 2^{-5}, 2^{-3}, \dots, 2^{15}\}$ and $\{\gamma = 2^{-15}, 2^{-13}, \dots, 2^{3}\}$. To speed up computations it is possible to perform a coarse grid search, and subsequently use a finer grid to determine the optimal C, γ values. Finally, it is important to note that in SVM each feature needs to be linearly scaled to lie within the range 0 and 1. This normalization is necessary to avoid numerical difficulties in the SVM computations because features with greater numeric ranges could otherwise turn out to dominate features with smaller numeric ranges.

We have so far focused on binary classification problems. Now, we describe two approaches to generalizing the problem to multi-class classification. As we have already mentioned when extending LLBFS, the idea is to decompose the multi-class classification problem into several binary classification sub-problems. There are two main approaches to split a multi-class problem into many binary sub-problems: OAA and OAO.

The first method, OAA, uses the samples from one class (treated as positive examples) versus the samples collected from all other classes which are pooled together and treated as negative examples. The process is repeated for all classes. Thus, we construct l binary classifiers, where l represents the number of classes in the given problem. SVM proceeds to solve the binary problem between each class against all the other samples which comprise the competing class. The classification of a new data sample $\mathbf{s}$ is achieved by assigning it to the class label of the classifier with the largest output function. One problem with this approach is that the training sets of each of the l binary classifiers may be highly imbalanced (i.e. binary problems with widely different number of samples); another is that the binary SVM classifiers in OAA are trained on different tasks.

The second method, OAO, uses data from every possible pair between two classes. That is, the algorithm identifies all instances in the dataset that belong to the two investigated classes and solves this binary problem. The process repeats for all pairs of classes, which leads to the construction of $l(l-1)/2$ binary SVM classifiers. The classification of a new data sample $\mathbf{s}$ is achieved by assigning it to the class that receives the largest number of "votes" from the individual binary classifiers. Compared to OAA, this approach requires the training of a larger number of classifiers and is computationally more costly to determine the class of new samples. Hsu and Lin (2002) compared three popular methods for generalizing SVM for multi-class classification problems (including OAA and OAO) across a wide range of problems, and reported that OAO was very competitive.

There is a considerable body of research on the topic of generalizing binary sub-problems to multi-class sub-problems, see for example Burges (2003), and Hsu and Lin (2002). For further details on SVM we refer to the detailed tutorial of Burges (2003), and to the textbooks of Hastie et al. (2009), and Bishop (2007). In this study, we used the LIBSVM implementation (Chang and Lin 2011) which is one of the most popular SVM software packages (LIBSVM uses the OAO approach for tackling multi-class classification problems).

4.3.3 Ensembles of decision trees

Ensembles of decision trees (or the more commonly used term *random forests*[34] abbreviated as RF), is a powerful non-parametric learner formed by a combination of many simple *base learners*, the trees. First we will describe how trees work, and then expand on how these trees are combined to produce an enhanced learner.

[34] Strictly speaking, the term "*random forests"* is copyrighted and should be avoided, but has pervaded the machine learning literature and is very commonly used.

The *classification and regression tree* (CART) method is a conceptually simple, yet powerful nonlinear, nonparametric method that often provides excellent results (Hastie et al., 2009). CART finds the best split of the range of one of the features, partitioning the range of this feature into two sub-regions (*nodes*). This partitioning process is repeated on each of the resulting sub-regions, recursively partitioning the original feature space into smaller and smaller, hyper-rectangular sub-regions. This recursive procedure can be represented graphically as a tree that splits into successively smaller branches, where each branch represents a sub-region of the feature space. This tree is “grown” up to T_0 splits, learning a successively detailed mapping between all the available data and the response. So, CART partitions the feature space into hyper-rectangles and assigns each hyper-rectangle a constant value (which is typically the mean or median of the response variables found in that hyper-rectangle – more about that later). Specifically, the algorithm works in the following steps:

1) Decide on the *loss function*, that is, the criterion for minimizing the deviation between the actual y and the predicted $\hat{y} = f(\mathbf{x})$. Typically, we decide to minimize the sum of squares $\sum_{i \, \epsilon \, Q}\left(y_i - f(\mathbf{x}_i)\right)^2$ (alternatively the absolute difference $\sum_{i \, \epsilon \, Q}|y_i - f(\mathbf{x}_i)|$ can also be used), where Q contains the indices i of the data set at each splitting junction.
2) Decide on the minimum *node size* (typically 5-10), which is the minimum number of observations i in every node. In effect, this sets the stopping criterion to halt the splitting process.
3) Having decided the criterion which determines ‘the best’ split at each junction and the node size, we proceed with a greedy algorithm. Starting with all the N samples (all the data), we use each jth variable to split the data into two parts (two nodes) by finding a splitting point s. We repeat this scanning for all the M features and determine the pairs of half-planes $\{R_1(j, s), R_2(j, s)\}$:

$$\begin{cases} R_1(j,s) = \{\mathbf{X} | \mathbf{f}_j \leq s\} \\ R_2(j,s) = \{\mathbf{X} | \mathbf{f}_j > s\} \end{cases} \tag{4.31}$$

Then, we need to determine the optimal feature $\mathbf{f}_j$ and splitting point s according to the loss function we have selected (for example the sum of squares). That is:

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right] \tag{4.32}$$

where c_1 and c_2 are the mean values of the y_i in the node when using the sum of squares as the loss function, and the median values of y_i in the node when using the absolute difference. Since we are using the sum of squares, c_1 and c_2 are expressed as:

$$\begin{cases} c_1 = mean(y_i | x_i \in R_1(j,s)) \\ c_2 = mean(y_i | x_i \in R_2(j,s)) \end{cases} \tag{4.33}$$

When the optimal splitting variable and split point are determined from (4.32), we have the actual split of the data, and proceed to the next step.

4) We repeat the process at step 3 for each node using all the $Q < N$ samples present in the node to further split the data into two nodes, unless the stopping criterion is met.
5) When the tree has grown fully up to T_0 splits (i.e. all the data has been assigned to nodes and the stopping criterion is met), we have fully partitioned the feature subspace into hyper-rectangles. The final nodes in the tree are called *terminal nodes*. When the loss function is the sum of squares, the resulting hyper-rectangle in the feature space is assigned the mean of the y_i responses in that rectangle, whereas when the loss function is

the absolute difference the resulting rectangle takes the median of the y_i responses in that terminal node.

Although this process is in principle very flexible and hence able to produce highly convoluted mappings, it can easily *overfit* the data: that is, become highly sensitive to noisy fluctuations in the input data and fail to generalize to new, unseen data. To address this problem some splits are collapsed (a process known as *pruning*): the amount of split reduction is determined by the *pruning level* $a \geq 0$ (where $a = 0$ is the full tree). The pruning level is set to minimise the prediction error, e.g. in the cross-validation setting (see § 4.4.1) and is subject to trial and error.

Pruning collapses some of the internal nodes to get a tree $T \subset T_0$, aiming to successively collapse those nodes which produce the smallest criterion increase, which is intuitively appealing; moreover for each pruning level a it can be shown there is a unique tree T_a (Hastie et al., 2009). Specifically, if we denote the terminal nodes m, R_m the resulting feature space hyper-rectangles, and $|T|$ the number of terminal nodes in T, pruning seeks to minimize the cost complexity function:

$$C_a(T) = \min_a \sum_{m=1}^{|T|} N_m Q_m(T) + a\,|T| \tag{4.34}$$

where $N_m = \#\{x_i \epsilon R_m\}$, and

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \,\epsilon\, R_m} \left[y_i - \left(\frac{1}{N_m} \sum_{x_i \,\epsilon\, R_m} y_i \right) \right]^2 . \tag{4.35}$$

Ensembles of decision trees combine many *weaker* individual trees (*base learners*), where the premise is that this combination will provide an improved functional form because the noise from the base learners will be smoothed out (Breiman 2001b). The procedure for growing each tree is essentially identical to the procedure described for CART; the only difference is that a random subset of the features is chosen for each tree. The trees are grown fully and there is no pruning, hence there is no need to experiment with the pruning level. The final decision regarding the classification of a new sample is achieved using *majority voting* from the base learners. Breiman (2001b) convincingly demonstrated that ensembles of decision trees are effective in various prediction tasks, whilst they do not overfit as more trees are added to the ensemble. This learner relies on a single tuning parameter, which is the number of randomly selected features to be used for split selection in each tree. Practice has shown that the decision tree ensemble is fairly robust to a wide range of values for this parameter (Breiman, 2001b; Meinshausen and Yu, 2009), which by default is set to the square root of the number of features in the design matrix.

For more details about CART and ensembles of decision trees, refer to Hastie et al. (2009).

4.4 Model evaluation and generalisation

The ultimate aim in training a learner, is to be able to satisfactorily assign appropriate response values to new unobserved samples which have not been used in the training process. Informally, a good learner is able to provide $\hat{y}$ which ideally should be identical to the response value y if this was measured directly. The following sections describe approaches to formally investigate the accuracy of the learning schemes.

4.4.1 Cross-validation

Once the functional form f has been determined using one of the learners, we need to establish how accurate the mapping $f(\mathbf{X}) = \mathbf{y}$ might be expected to be on a novel dataset. This is known as the *generalization* performance of the model which is typically estimated using a) *cross validation*, b) *bootstrapping*, or c) an *additional dataset*, which has not been used to train the model (i.e. in the determination of f). We use cross validation (CV), a well-established statistical re-sampling technique (Webb, 2002) which is commonly used in many settings because often we are limited by a relatively small dataset.

Specifically, in CV the dataset is split into a *training* subset, which is used to determine f, and a *testing* subset, which is used to assess the model's generalization performance. The ratio of the training subset over testing subset (number of samples in each subset) is determined by the modeler and is known as K-fold cross validation. Inherent in the choice of K is the *bias-variance trade-off*: using $K = N$ (leave one sample out) leads to low bias and large variance, whereas low K might lead to large bias due to potentially under-fitting the data. Typically, 5-fold (5:1) and 10-fold (10:1) CV is a good bias-variance trade-off (Hastie et al., 2009).

The model parameters are determined using the *training* subset, and errors are computed using the *testing* subset (*out-of-sample error* or *testing error*). The process should be repeated a large number of times (e.g. 100-1000), where the dataset is randomly permuted in each run prior to splitting into training and testing subsets, in order to obtain statistical confidence in this assessment. Depending on the requirements of the problem, different *loss functions* can be introduced. In all cases, on each repetition we record an error which has the form $L(\{y_i, \hat{y}_i\}_{i=1}^{N_t})$, where N_t represents the number of samples in the training or testing subset. Some widely used metrics are the *mean absolute error* (MAE), the *mean relative error* (MRE – also known as mean percentage error), and the *root mean squared error* (MSE):

$$\text{MAE} = \frac{1}{N_t}\sum_{i\in Q}|y_i - \hat{y}_i| \tag{4.36}$$

$$\text{MRE} = 100 \cdot \frac{1}{N}\sum\nolimits_{i\in Q}(|\hat{y}_i - y_i|/y_i) \tag{4.37}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum\nolimits_{i\in Q}(\hat{y}_i - y_i)^2} \tag{4.38}$$

where Q contains the indices of the training or testing set. Errors from all repetitions are averaged, and the generalization performance of the learner is decided using the *out-of-sample* error. Alternative error metrics are possible, which often depend on the specific requirements of the examined application. In this study, the MAE is mostly used because it is based on the L_1-norm and is known to be more indicative of the learner's prediction accuracy (it is more robust to outliers) compared to the frequently used RMSE. For binary classification problems MAE is equivalent to misclassification, that is, MAE is equal to the number of samples incorrectly assigned to the wrong class. It is worth noting that the RMSE is always equal to or greater than the MAE, and is particularly sensitive to the presence of large errors (hence it finds applicability in scenarios where large errors are particularly unwelcome, such as in the case of evaluating PDAs (Christensen and Jakobsson, 2009). The larger the variability of the errors in the model, the larger the difference between MAE and RMSE. Therefore, these metrics can be considered complementary when evaluating the performance of a model.

4.4.2 Addressing the principle of parsimony

Statistical learning has two fundamental aims: a) *prediction accuracy* (typically defined by the deviation of the estimated response from the true response) and b) *interpretation* (usually by identifying the feature set most predictive of the response). For the first of these, we can use different loss functions depending on the problem at hand (Hastie et al., 2009). As highlighted in the preceding sections, feature selection can aid the production of an interpretable model. Although we have seen a number of diverse algorithms for selecting features, it is not clear what the optimal number of features m is (most FS algorithms only provide a ranking of the features, and the decision where to set the cut-off depends on the researcher). For example, it is possible that presenting most of the features into the learner provides a more accurate model compared to feeding the learner with a few features. Using a large number of features makes interpretation difficult, may be considerably more expensive from a computational point of view, and may fail to generalize well to new data. This is just another statement of the principle of parsimony, which, as we have already mentioned, says that the number of features should be kept as low as possible, given the same prediction accuracy. Hence, in some applications it is desirable to *trade-off* accuracy against complexity (number of features).

One approach to finding a compromise between model complexity and predictive accuracy is to use *information criteria* which induce a penalty on the number of features. Information criteria abound following the introduction of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Hastie et al., 2009). Alternative information criteria include the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), and Mallows' Cp (Mallows, 2000). Each criterion is different from the others in the way in which it penalises complexity relative to the estimated prediction error. There is no clear consensus regarding the optimal criteria, and some authors simply use both AIC and BIC which are the most widely known information criteria (Stergiopulos et al., 1999; Tsanas et al., 2010a).

Another commonly used approach to account for the number of features versus model accuracy is to use the *one standard error rule* (Hastie et al., 2009): we pick the most *parsimonious subset* (subset with the lowest number of features) in which the error metric is no more than one standard deviation above the error of the subset leading to the smallest prediction error. This simple approach satisfies the rather subjective need to account for parsimony, and we have used it in our recent studies (Tsanas et al., 2010c; Tsanas et al., 2011a; Tsanas et al., 2011b).

4.4.3 Statistical hypothesis and surrogate tests to validate models

In § 4.1.3 we briefly introduced statistical hypothesis tests, where the aim was to establish whether a relationship between two random variables is statistically significant. Statistical hypothesis tests are also used to demonstrate that the observed accuracy in a model cannot be attributed to chance alone. The objective is usually to *reject* the null hypothesis, which in this instance aims to compare the errors computed using the model, and the errors computed using a simple benchmark (one example of a simple benchmark is to consider the predicted response equal to the mean or the median value of the response, and compute the errors). We describe two general, non-parametric statistical hypothesis tests, which make no assumptions about the underlying distributions of the data, and hence are very generally applicable. We remark that in special cases where the data follow the Gaussian distribution, it may be preferable to select appropriate tests which are more sensitive.

The first statistical hypothesis test is the two sample Kolmogorov-Smirnov (KS) test which compares the distributions of two random variables. For the purposes of validating a model, we can compare the error distribution computed when applying the model ($p1$), and the error distribution computed using a benchmark ($p2$). The null hypothesis is that the two distributions are drawn from the same continuous distribution, and the alternative hypothesis is the two distributions are not drawn from the same distribution. Specifically, the KS test works as follows: we compute the empirical cumulative density functions cdf_1 and cdf_2 from the samples of the examined random variables X, Y at some pre-defined grid of values (we represent with c each evaluation point). Then, the KS test finds the maximum distance between the empirical cumulative density functions:

$$cdf_1(c) = \frac{\#i{:}\, x_i \leq c}{N} \qquad cdf_2(c) = \frac{\#i{:}\, y_i \leq c}{N} \tag{4.39}$$

$$\text{KS}_{statistic} = \max(|cdf_1(c) - cdf_2(c)|). \tag{4.40}$$

Finally, the p-*value* in the KS test is computed as a function of the $\text{KS}_{statistic}$ and N, and the distributions are considered to be statistically significantly different if the computed p-*value* is below the user-specified significance level.

Another statistical hypothesis test we use is the Wilcoxon rank sum test (also known as the Mann-Whitney U test). The null hypothesis for this test is that the distributions $p1$ and $p2$ are independent samples from continuous distributions with equal medians. The alternative hypothesis is that the distributions $p1$ and $p2$ do not have equal medians. The computation of the significance value for the Wilcoxon test is similar to the computation we have described in establishing whether correlations are statistically significant (§ 4.1.3). First the $\{x_i, y_i\}_{i=1}^{N}$ realizations of the random variables are transformed to their rank scores. Then, we form sets redefining the pairs of rank scores using all possible combinations of each x_i sample with each y_j sample. If X and Y have equal medians, then each x_i sample can be larger or smaller than each y_j sample with $P = 0.5$. Then, we compute the number of times $U_1 = x_i > y_j$ and the number of times $U_2 = x_i < y_j$. The p-*value* for the Wilcoxon test is then defined from contingency tables as a function of $2 \cdot \min(U_1, U_2)$.

In addition to the two statistical hypothesis tests presented above (two sample KS test, and Wilcoxon rank sum test), we propose using a simple *surrogate test*. Surrogate tests provide a convenient means of null hypothesis testing, and hence are complementary to the statistical hypothesis tests discussed so far. They aim to obtain data which are similar to the original dataset, and ultimately test how the design matrix obtained using a surrogate approach can be used to predict the response. If the original features contain information which is predictive of the response, then it should be expected that the error with the surrogate data (which is purely random and so non-predictive by design) would be considerably larger. As with the statistical hypothesis tests, the objective is to reject the null hypothesis that the random features do not contain predictive information (we accept the null hypothesis as true if the error computed with the surrogates is not statistically significantly larger than the error computed using the features). We propose using the simplest approach to obtain data that have similar properties to the original features, which is to use a *shuffled* version (random permutation) of each feature. This approach preserves the empirical probability density of the data, and as we shall see later, may be quite challenging for some FS algorithms.

In summary, we propose using three hypothesis tests to verify the validity of the model: a) the two sample KS test, b) the Wilcoxon rank sum test, and c) the surrogate testing with the randomly permuted features. The first two tests aim to validate the model against a naïve benchmark, whereas the last test verifies that the computed features contain predictive information about the response.

4.5 Summary of the proposed methodology for analysing data

Here we summarize the methodology which was described in some detail in this chapter. Specifically, we propose the following steps in supervised learning setups:

1) Plot the data using density plots and scatter plots. This step can indicate whether there are any obvious relationships between the features and the response variable, and suggest possible transformations of features (for example log-transforming a feature).

2) Compute correlation coefficients and more robust metrics (such as MI) which can indicate a relationship between each feature and the response, and compute the statistical significance of the relationships.

3) Apply standard classification methods, for example the kNN classifier. Similarly, problems where the response variable spans a continuous range of values should be tackled using standard regression algorithms, for example the ordinary least squares approach (Hastie et al., 2009). The computed accuracy of such a simple learner sets a useful benchmark, which we aim to beat using more complicated prediction algorithms. Use more complicated learners such as SVM and ensembles of decision trees. Use all features to predict the response, to set a benchmark against which the performance of a dataset with fewer features will be compared.

4) Select features using FS algorithms. Then, determine the optimal feature subset for each FS algorithm and compare the performance of the learner(s) using the selected feature subsets.

5) Potentially, in some datasets reducing the number of features can reduce the error metric (due to the curse of dimensionality); while in other cases the use of a larger number of features might offer *insignificant* performance improvement in the error metric. This undesirable trait may be biased, making the resulting model computationally expensive, obscuring its interpretability, and potentially failing to generalize well on a novel dataset. In that case, we suggest using the "one standard error rule" to compromise between model complexity and performance in order to obtain a parsimonious dataset.

6) Use a new dataset, or when not available, K-fold cross-validation with at least 100-1,000 repetitions to ensure that the *out-of-sample* prediction error results are robust.
7) Use statistical hypothesis and surrogate data tests to ensure the model developed is practically useful, in that it can significantly outperform some simple benchmarks.

The list can easily be modified and is purposefully general, so that it is applicable to similar signal processing applications. The field of data analysis and knowledge discovery cannot be possibly covered exhaustively here; we refer to the survey of Kurgan and Musilek (2006) for a relatively recent authoritative overview. A less detailed overview of this topic of data analysis for a non-mathematically oriented audience is presented in Tsanas et al. (2012a).

Chapter 5

Applying the signal processing and machine learning tools to data

In Chapters 3 and 4 we introduced tools for speech signal processing and statistical machine learning, respectively. In this chapter we conduct empirical studies to evaluate the performance of those tools, and determine the techniques that may be widely applicable in practical circumstances. Specifically, in § 3.2.1 we have described ten PDAs, and introduced novel approaches for combining the individual F_0 estimators to obtain a superior ensemble PDA. Here, we use sustained vowel /a/ phonations generated by a physiological model of speech production, to examine rigorously the performance of the F_0 estimation algorithms.

In the remainder of this chapter we compare the statistical machine learning techniques described in Chapter 4. First we compare FS algorithms on a variety of artificial and real datasets, aiming to (a) determine which FS algorithms identify the true features (that is, discard artificial features which do not contribute towards predicting the response, and are commonly referred to as *probes*) in datasets, and (b) minimize the loss function metric (that is, minimize the number of misclassified samples) as a result of identifying a parsimonious, information-rich subset.

Finally, we evaluate empirically the performance of two widely used and powerful nonlinear classifiers: SVM and ensembles of decision trees. Here, we look at various publicly available datasets which are diverse both in terms of the application and in terms of the dataset type, to test how accurately SVM and ensembles of decision trees predict the response.

5.1 Comparing fundamental frequency estimation algorithms

Although accurate F_0 estimation is desirable, there may be no single best PDA which is applicable in all applications (Talkin, 1995). We have already mentioned that different applications may have different requirements, and intuitively we expect that some PDAs may be better suited to particular applications depending on the type of speech signals (e.g. conversational signals, singing, sustained vowels and running speech); computational considerations may also be an issue (for example in embedded speech coding applications).

In Chapter 2 we have seen that sustained vowels are often used in voice quality assessment. In particular, since we will focus exclusively on analysing the sustained vowel /a/, we will evaluate the performance of the PDAs only for those types of signals. Using the sustained vowel /a/ alleviates some of the difficulties in F_0 assessment: (a) the need to characterize frames as voiced or unvoiced, (b) reduces the range of possible F_0 values, and (c) minimizes the masking effects formants may have on F_0 (for example when the formants of a word complicate the identification of F_0).

There are three approaches to validate the accuracy of PDAs: (a) comparing F_0 estimates against benchmark values which have been provided by expert speech scientists following manual inspection of the glottal cycles from plots of the signal, (b) using electroglottographs (EGG) which provide the glottal closure instances (so that we can infer F_0), and (c) using synthetic signals where the ground truth F_0 values are known by means of knowing the values inserted in the model used to generate the data. Although all three approaches are not without limitations, the first two may fail to yield practically accurate ground truth F_0 for validating PDAs. This is because speech experts observing signal plots often do not agree on the exact length of each vocal period (Talkin, 1995), and hence it is not clear how to define the ground truth unambiguously in this context. Similarly, EGGs often provide faulty estimates which are corrected manually by speech experts, casting doubt on the validity of this approach (Colton et al., 1990; Henrich et al., 2004). Therefore, we argue that the third approach, using synthetic signals where the ground truth is known in advance, may be the most appropriate method for validating PDAs. This implicitly assumes that signals closely resembling actual speech signals can be generated. The ability to accurately replicate disordered voice signals is related to the nature of the model used to synthesize the signals, and its capacity to mimic the origin and effects of different speech pathologies.

Here, we used a speech database that was developed by Matias Zañartu specifically for the needs of this study. In short, the sustained vowel /a/ signals were generated using a physiological model of speech production where the F_0 values are known in the form of glottal closure instants, i.e. vocal fold collision instants (or minimum glottis area, when there is no collision for pathological voices). The model is described in detail in Zañartu (2010), and is capable of mimicking various normal, hyper-functional (inappropriate patterns of vocal behaviour that are likely to result in organic voice disorders), and pathological voices, where the exact system fluctuations were known.

Using this physiological model, 100 sustained vowel /a/ phonations each of one second duration were generated. Following manual inspection, eight phonations were discarded because they were unnatural-sounding. Thus, we processed 92 signals to evaluate the performance of the PDAs. The distributions of the ground truth F_0 values for all signals are summarized graphically in Fig. 5.1, depicting the median and the interquartile range values for each phonation. The generated speech signals have a relatively wide range of possible F_0 values, with variable F_0 fluctuations (jitter), which gives some confidence that we are covering a broad type of signals that might occur in practice.

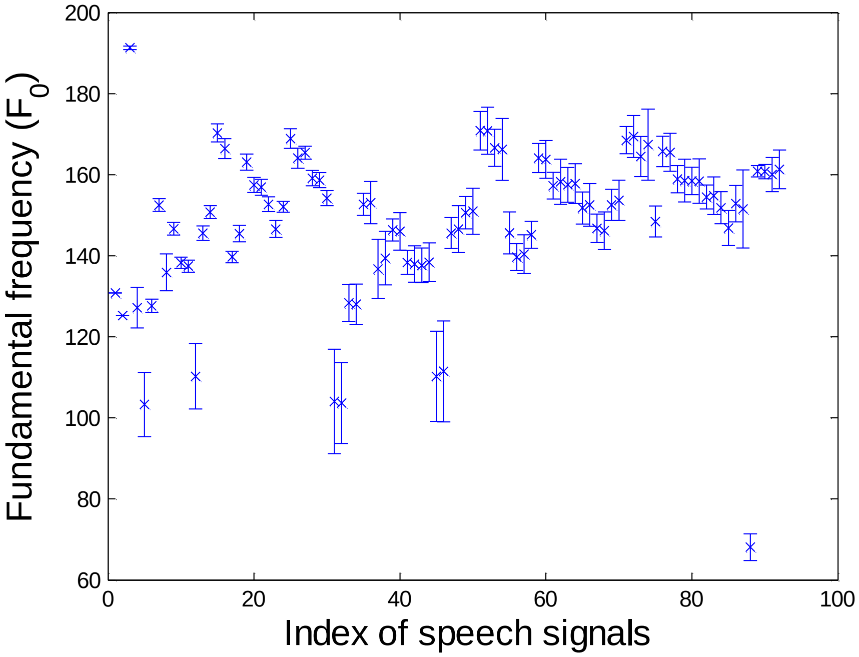

Fig. 5.1: Summary of ground truth F_0 values for the 92 speech signals used in this study. The middle point represents the median and the bars represent the interquartile range. The index refers to the speech signal used in the study.

Most PDAs provide F_0 estimates at specific time intervals (typically at successive instances using a fixed time window of a few milli-seconds). Here, wherever possible, we obtained F_0 estimates from the PDAs every 10 ms, at the reference time instances [100, 110, 120 …950] ms (thus, we have 86 F_0 values for each synthetic phonation signal and each PDA or PDA ensemble). Given that the generated speech signals exhibit inherent instabilities because the physiological model requires some 4-5 vocal cycles to fall into stable oscillation, and that many PDAs provide reliable estimates only some milli-seconds into the speech signal, we discarded the F_0 estimates prior to 100 ms. A few PDAs do not provide F_0 estimates at pre-specified time intervals, but at intervals which are identified as part of the algorithm (this is the case with RAPT, for example). Other PDAs, such as NDF, provide high-frequency F_0 estimates (every millisecond). In those cases where the PDAs do not provide F_0 estimates at the exact time instances described above, we used piecewise linear interpolation between the two closest time intervals of the PDA to obtain the F_0 estimate at the reference time instances. The time instances where F_0 was estimated in RAPT did not differ considerably from the reference time instances, and thus piecewise linear interpolation should not markedly affect its performance.

The ground truth F_0 time series from the physiological model is given in the form of glottal closure time instances, which are directly translated to F_0 estimates in Hertz. However, we need to obtain ground truth F_0 values at the reference time instances. Hence, piecewise linear interpolation was used to obtain the ground truth at the reference instances. Similarly, we used piecewise linear interpolation to obtain F_0 estimates from DYPSA at the reference time instances (DYPSA is the only PDA in this study that aims to identify glottal closure instances, instead of using time windows).

Summarizing, each PDA or PDA ensemble provides 86 F_0 estimates for every synthetic speech signal. These estimates for every speech signal are compared against the 86 ground truth F_0 values at the reference time instances. In total, we processed 92 speech signals which provide $N = 92 \times 86 = 7912$ values over which we compare the performance of the PDAs and ensembles. In a few cases, the algorithms PRAAT2 and TEMPO failed to provide outputs (towards the beginning or end of the signal). Those instances were substituted with the median estimate from the other PDAs. Overall, the F_0 outputs from the ten PDAs for all 7912 cases were concatenated into a matrix **X** with 7912×10 elements. The PDA ensembles are directly computed using this matrix. The ground truth was stored in a vector **y** which comprised $N = 7912$ elements. There were no missing or invalid entries in the matrix **X** or the ground truth vector **y**.

The deviation from the ground truth for each signal and each PDA is computed as $e_i = \hat{y}_i - y_i$, where $\hat{y}_i$ is the ith F_0 estimate ($i \in 1 \dots 86$), and y_i is the ith ground truth F_0 value. We use three metrics to evaluate the performance of the PDAs using MAE, MRE, and RMSE, since there is no universal agreement amongst researchers about which metric to use when evaluating F_0 estimation accuracy.

The weighted ensemble PDAs use 91 training signals to obtain the weights, and are tested on the signal left out of the training process (92nd signal); this process of training the weights and testing on the signal left out of the training is repeated for all 92 signals. This leave-one-signal-out validation is done to provide an estimate of the out-of-sample performance of the ensemble. A stable weighting ensemble scheme would be expected to exhibit very similar weights across all leave-one-signal-out computations for all contributing PDAs in the ensemble.

Table 5.1 compares the average performance of the ten individual PDAs in accurately tracking the F_0 contour of the 92 signals. Fig. 5.2 presents an illustrative example of errors in F_0 estimates over time, for one randomly selected signal, for each of the individual PDAs.

Table 5.1 Performance of the F_0 estimation algorithms

Algorithm	MAE (Hz)	MRE (%)	RMSE (Hz)
DYPSA	3.39 ± 5.34	2.60 ± 5.09	7.35 ± 16.04
PRAAT1	10.79 ± 22.16	7.14 ± 14.44	12.70 ± 22.37
PRAAT2	5.88 ± 14.44	3.90 ± 9.23	8.13 ± 16.34
RAPT	12.42 ±7.81	8.27 ± 4.72	24.97 ± 11.22
SHRP	3.28 ± 3.47	2.27 ± 2.38	7.84 ± 9.43
SWIPE	1.90 ± 1.14	1.37 ± 1.07	2.46 ± 1.56
YIN	21.06 ± 16.23	13.93 ± 10.23	37.41 ± 21.28
NDF	**1.39 ± 0.73**	**0.98 ± 0.66**	**1.80 ± 1.01**
TEMPO	1.71 ± 0.89	1.20 ± .078	2.21 ± 1.23
XSX	2.10 ± 1.12	1.47 ± 0.91	2.74 ± 1.49

The evaluation of the F_0 estimation algorithms uses all the 92 speech signals, where for each signal we use 86 F_0 estimates (thus $N = 92 \times 86 = 7912$). The results are in the form mean ± standard deviation. The best individual pitch detection algorithm (PDA) is highlighted in bold.

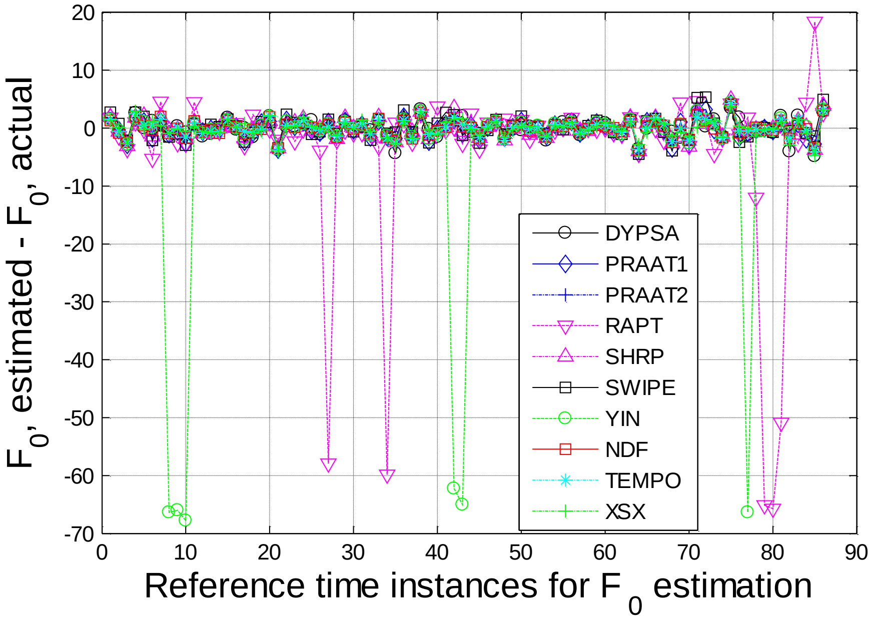

Fig. 5.2: Overview of the tracking errors of the F_0 contour (86 reference time instances of F_0 evaluation) for a randomly selected signal (number 33) using the 10 individual PDAs. Some of the PDAs have large spikes, temporarily deviating considerably from the ground truth F_0.

Overall, almost all the PDAs can estimate the F_0 contour fairly accurately. The best individual PDA is NDF, closely followed by TEMPO and SWIPE. YIN and RAPT exhibit spiky behaviour where, in a few instances, large outlying deviations from the ground truth are observed, which strongly suggests that a post-processing filter may improve the final estimates. We refer to Bagshaw (1994) for further details on post-processing ideas.

Next, we investigate the 12 ensemble PDA schemes defined in § 3.2.1.10 which rely on selecting and/or weighting the outputs of the 10 individual PDAs. The average performance of the ensemble PDAs is presented in Table 5.2. We remark that the naïve ensemble approaches (using the mean and the median F_0 from all PDAs) do not outperform the single best PDA. We decided on the optimal number of PDAs K for each ensemble based on the out of sample MAE performance.

Table 5.2 Performance of the ensemble F_0 estimation algorithms

Algorithm	MAE (Hz)	MRE (%)	RMSE (Hz)
Ensemble 1	5.03 ± 4.51	3.38 ± 2.92	6.78 ± 4.68
Ensemble 2	1.58 ± 0.89	1.12 ± 0.77	2.14 ± 1.35
Ensemble 3	1.52 ± 1.72	1.09 ± 1.50	2.30 ± 4.27
Ensemble 4	1.51 ± 1.69	1.07 ± 1.41	2.27 ± 4.17
Ensemble 5	1.53 ± 1.72	1.09 ± 1.50	2.30 ± 4.27
Ensemble 6	1.49 ± 1.67	1.06 ± 1.40	2.23 ± 4.06
Ensemble 7	1.25 ± 0.70	0.89 ± 0.66	1.65 ± 1.04
Ensemble 8	1.25 ± 0.71	0.89 ± 0.66	1.65 ± 1.04
Ensemble 9	1.26 ± 0.68	0.89 ± 0.63	1.65 ± 0.97
Ensemble 10	1.25 ± 0.70	0.89 ± 0.66	1.65 ± 1.04
Ensemble 11	1.26 ± 0.68	0.89 ± 0.63	1.65 ± 0.97
Ensemble 12	**1.25 ± 0.68**	**0.89 ± 0.63**	**1.65 ± 0.97**

The evaluation of the F_0 estimation algorithms uses all 92 synthetic speech signals, where for each signal we use 86 F_0 values at the reference time instances (giving $N = 92 \times 86 = 7912$). The ensemble weights for the ensembles are computed using a leave-one-signal-out scheme and validated on the out of sample signal. The results are in the form mean ± standard deviation. See § 3.2.1.10 for the definition of all ensemble PDAs.

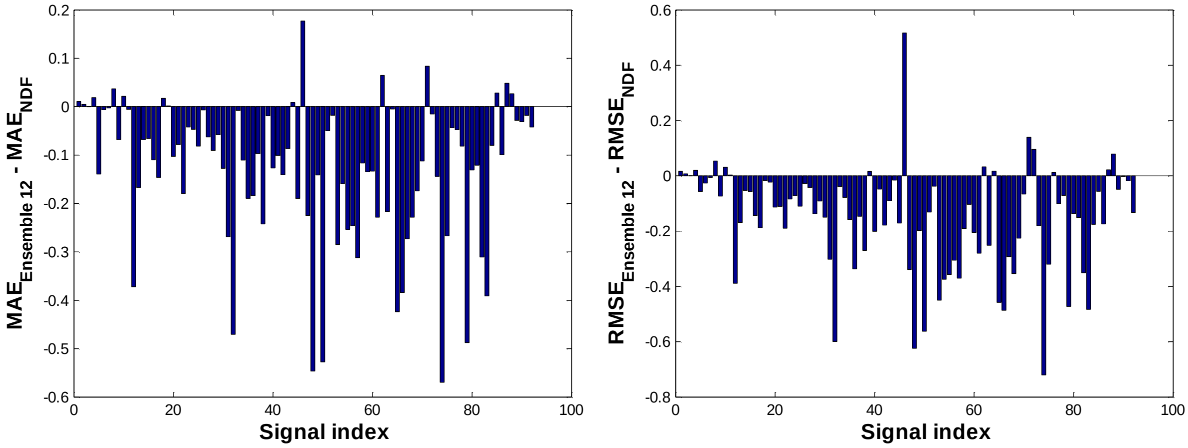

Fig. 5.3: Comparing the performance of Ensemble 12 with the best individual F_0 estimation algorithm, NDF, using (a) MAE, and (b) RMSE. Very similar results are obtained for MRE. For the vast majority of signals used in this study, Ensemble 12 is better than NDF.

Overall, the lowest MAE was with Ensemble 12, which uses IRLS comprising five individual PDAs: NDF, SWIPE, SHRP, TEMPO, and DYPSA. Ensemble 12 was chosen over the other ensembles with similar performance because it consisted of a lower number of PDAs. In Fig. 5.3 we present graphically the performance of this ensemble PDA (Ensemble 12) versus NDF in the average F_0 for all 92 signals, which demonstrates that in the vast majority of signals this ensemble scheme is more accurate than NDF.

By design, all the weighted ensemble schemes were obtained by leaving one signal out of the training dataset also used to compute the PDA weights, and then testing the performance of the final model using the signal left out of the training process. This validation process was repeated for all 92 signals, leaving one signal out each time. Specifically, the final ensemble (Ensemble 12) F_0 estimate we propose is computed as follows:

$$\begin{aligned} F_0 = 2.189 \cdot \mathrm{NDF} - 0.069 \cdot \mathrm{SWIPE} + 0.004 \cdot \mathrm{SHRP} - 1.125 \cdot \mathrm{TEMPO} \\ + 0.002 \cdot \mathrm{DYPSA} \end{aligned} \tag{5.1}$$

where NDF, SWIPE, SHRP, TEMPO, and DYPSA are the F_0 estimates from the corresponding PDAs. The IRLS weights of the individual PDAs which form Ensemble 12 appear in Eq. (4). These weights were very stable for all 92 training cases, specifically the standard deviation of the weights were: 0.006 for NDF, 0.0015 for SWIPE, 0.002 for SHRP, 0.00057 for TEMPO, and 0.0001 for DYPSA.

To demonstrate that Ensemble 12 is a genuine improvement over NDF, we compared the errors for the 92 signals (7912 elements) obtained using NDF and Ensemble 12 using the rank-sum test. The null hypothesis (which we want to reject) is that the errors have the same median, against the alternative hypothesis that the medians are different. The rank-sum test rejected the null hypothesis ($p < 0.01$), which coupled with the results in Tables 5.1 and 5.2 suggests that Ensemble 12 is statistically significantly better than NDF.

Comparing the results in Tables 5.1 and 5.2 we observe an almost 10% improvement of Ensemble 12 over NDF: MAE 1.25 versus 1.39, MRE 0.89 versus 0.98, and RMSE 1.65 versus 1.80. Moreover, the standard deviation of Ensemble 12 is lower than the standard deviation of NDF in all cases. Overall, these findings suggest that weighted ensembles have promising potential in accurate F_0 estimation. Future work could investigate more sophisticated combinations of PDAs to build on those promising results. Moreover, the ensemble investigations in this study did not try to leverage the F_0 estimates considered as a time series, i.e. the temporal variation of the PDA estimates. For example, it is possible that an ensemble constructed of current and some previous F_0 estimates could lead to more accurate F_0 estimation. In other words, it would be worth experimenting with a temporally local ensemble which would combine temporal smoothing with integration over the output of each contributing PDA.

In Eq. 5.1 we notice that SHRP and DYPSA contribute with minimal weights towards Ensemble 12. We remark that although these weights are statistically significant, we have found that discarding those PDAs and using an ensemble with the current weights using only NDF, SWIPE and TEMPO gives rise to only very slightly worse results. Invoking the principle of parsimony, and to account for PDA complexity we could use an ensemble with only those three PDAs. The sampling frequency of the experimental setup of this study was 44.1 KHz, with 16 bits of resolution. Since the speech signals used later in this study (see Chapter 6) use a sampling frequency of 24 KHz, we repeated the analysis resampling the original signals at this frequency. We have found that the performance degradation as a result of down-sampling from 44.1 KHz to 24 KHz was less than 0.01 Hz for both the individual PDAs and for the ensemble PDAs.

We emphasize that the current investigation considered only the sustained vowel /a/, and so the current findings cannot be generalized to all speech signals solely on the evidence presented here. More comprehensive studies are required to investigate PDA performance in other scenarios, e.g. with the other corner vowels (see § 2.2.4), which are also sometimes used in voice quality assessment (Titze, 2000) or with running speech.

5.2 Machine learning datasets

The remaining of this chapter focuses on FS: this section introduces the datasets used in this study, and the following sections evaluate the FS algorithms on these datasets. Table 5.3 summarizes these datasets: all datasets used here are publicly available, and most have been previously used in the FS literature. In cases of missing entries in a dataset, the corresponding sample in the data matrix was discarded. To test the FS algorithms we used three artificial datasets (where the true features and the probes are known in advance), and ten real datasets. Here, we provide a very brief description of each dataset and refer to the original studies and the publicly available repositories cited in Table 5.3 for further details about each dataset.

The first artificial dataset we use is the well-known MONK-1. It consists of 124 samples and six features: three features are predictive of the response, and three features are probes. The relationship of the features to the response is based on logical operators, a setting which is difficult to handle for some FS algorithms. We also used Isabelle Guyon's artificial dataset generator to obtain two datasets, which we call *Artificial1* and *Artificial2*. The Artificial1 dataset consists of 500 samples and 150 features: there are 50 independent, 50 dependent, and 50 repeated features, and only 10 are "true" features. Moreover, there is a 10% fraction of flipped responses in this binary classification problem. The Artificial2 dataset consists of 1000 samples and 100 features: there are 50 independent, 25 dependent, and 25 repeated features, and 20 are the “true” features. There are 10 classes in a nicely balanced dataset where we have 100 data samples for each class.

One real dataset widely used in FS algorithm comparisons is the hepatitis dataset (Diaconis and Efron, 1983). It includes 155 patients and the binary outcome (healthy control subject versus subject with hepatitis disease) depends on 19 features. This dataset has been studied in detail by Breiman (2001), who concluded that features X_{17} and X_{12} were highly predictive of the response (and highly correlated with each other). Breiman suggested that either of those two features individually carries almost as much information as the entire feature set. The features X_{19} and X_{11} were also identified as conveying some additional information for predicting the response. More recently Tuv et al. (2009) identified the following feature subset using a scheme based on random forests: $\{\mathrm{X}_6, \mathrm{X}_{17}, \mathrm{X}_{14}, \mathrm{X}_{19}, \mathrm{X}_{11}\}$. That study contrasted their proposed FS algorithm with three alternative FS algorithms, which unanimously selected variable X_2.

The Parkinson’s dataset (Little et al., 2009) uses 22 dysphonia measures obtained from 195 sustained vowel phonations. In the original study, the optimal feature subset was selected by using a two-step approach: first a simple filter approach eliminated one of the pair of highly correlated features (absolute correlation coefficients larger than 0.95). In the second step, a brute force search determined the best feature subset out of the remaining 10 features using a wrapper approach with SVM. The optimal feature subset according to Little et al. (2009) is $\{\mathrm{X}_{16}, \mathrm{X}_{17}, \mathrm{X}_{18}, \mathrm{X}_{22}\}$.

The Sonar dataset (Gorman and Sejnowski, 1988) is from the application of sonar signals (frequency-modulated chirp rising in frequency) aiming to predict whether the targeted object is a mine or a rock. Each of the 60 features represents the energy within a particular frequency band integrated over a period of time.

The wine dataset comes from the chemical analysis of wines grown in the same region, where 13 features (such as alcohol, magnesium and colour intensity) are used to differentiate the three cultivars.

The image segmentation dataset uses 19 features from images in order to identify whether the investigated region of the image (each sample) belongs to the following seven classes: brickface, sky, foliage, cement, window, path, grass. Its developers have split it into two subsets: a training set with 210 samples and a testing set with 2100 samples. We use the training set to select the features, and then use 10-fold cross validation with 100 repetitions to evaluate the performance of the learners on the testing set.

The cardiotocography dataset (Ayres-de Campos et al., 2000) has 2129 fetal cardiotocograms which were processed and classified by three expert obstetricians. The class label refers to a morphological pattern and was assigned 10 possible classes obtained by consensus from the three experts.

Finally, we use two datasets where the number of features is larger than the number of samples (also known as 'fat' datasets): these problems are inherently difficult for many machine learning algorithms, and have attracted the dedicated attention of researchers (Hastie et al., 2009). The ovarian cancer dataset consists of 72 samples and 592 features. There are many ovarian cancer datasets in the machine learning literature; here we use the dataset from Guan et al. (2009). In that study, the focus was on developing a biomarker of ovarian cancer based on metabolic changes in biological systems in order to differentiate subjects into the binary classes "healthy" versus "cancer". We used leave-one-sample-out to obtain 72 candidate feature subsets for each FS algorithm; then we used the voting scheme described in Table 4.4 to select the final features for each FS algorithm. The performance of the FS algorithms was assessed using leave-one-sample-out validation with RF.

The Small Round Blue-Cell Tumors (SRBCT) dataset (Khan et al., 2001) has 88 samples and 2308 features, which in this application are expressions profiles of genes, and is one of the most widely used datasets for validating FS algorithms in the domain of bioinformatics. The four-class response denotes the type of the tumor. We used 63 samples for selecting the features and training the classifier, and tested the performance of RF using the selected feature subsets on the remaining 25 samples. We used the partitioning of the samples into training and testing sets suggested by Hastie et al. (2009).

Table 5.3: Summary of the datasets

Dataset	**Design matrix**	**Associated task**	**Type**
MONK1[35]	124×6	Classification (2 classes)	D (6)
Artificial 1	500×150	Classification (2 classes)	C(150)
Artificial 2	1000×100	Classification (10 classes)	C(100)
Hepatitis[35]	155×19	Classification (2 classes)	C (17), D (2)
Parkinson's[35]	195×22	Classification (2 classes)	C (22)
Sonar[35]	208×60	Classification (2 classes)	C (60)
Wine[35]	178×13	Classification (3 classes)	C (13)
Image segmentation[35]	2310×19	Classification (7 classes)	C (16), D (3)
Cardiotocography[35]	2129×21	Classification (10 classes)	C (14), D (7)
Ovarian cancer[36]	72×592	Classification (2 classes)	C (592)
SRBCT[37]	88×2308	Classification (4 classes)	C (2308)

The size of each design matrix is $N \times M$, where N denotes the number of instances (samples), and M denotes the number of features. The last column denotes the type of the design matrices' variables: continuous (C) or discrete (D). In cases of missing entries, the entire row in the design matrix was deleted.

[35] Downloaded from the *UCI Machine Learning Repository*: http://archive.ics.uci.edu/ml/datasets.html
[36] Downloaded from http://www.biomedcentral.com/1471-2105/10/259/additional
[37] Download from http://www-stat.stanford.edu/~tibs/ElemStatLearn/

5.3 Feature selection results

In general, there are two approaches to evaluate FS algorithms: (1) assessing whether the “optimal” feature subset was selected (“optimal” being the combination of features maximally associated with the response), i.e. no probes or redundant features were selected, and (2) assessing a performance metric of the subsequent learning phase where the selected feature subsets are fed into the learner. The latter is a surrogate approach to evaluate the performance of FS algorithms (Tuv et al., 2009) by introducing an additional layer into the FS problem, and does not necessarily correspond to selecting the true feature subset. In practice, some weakly relevant or redundant features could improve the learners’ performance; conversely, the benefit of discarding relevant features may outweigh loss in information content (Guyon et al., 2007). Moreover, it is possible that using different learners might lead to different conclusions regarding the superiority of the FS algorithms (Hilario and Kalousis, 2008). Nevertheless, both approaches are commonly used in the FS literature (Guyon et al., 2006; Tuv et al., 2009; Sun et al., 2010) and will be used in this study as well.

In the subsequent analysis for the real datasets, the features are chosen following the methodology outlined in § 4.2.4 and in particular the steps in Table 4.2 unless the authors have provided separate sample sets for training and testing: in those cases we use the training set to obtain the features and test the performance of the FS algorithms on the testing set.

5.3.1 Validating feature selection algorithms based on false discovery rate

Determining whether the true set of features has been selected is only possible in artificial datasets, where the true features and the probes are known in advance. This aim is infeasible for real world datasets, because we do not know *a priori* the best feature subset which necessitates using classifier performance to act as a surrogate to evaluate the FS algorithms. Assuming the number of true and false (collectively referring to redundant, irrelevant and noisy) features in a dataset is known (ground truth), we define the False Discovery Rate (FDR) as the number of false features erroneously identified by the FS algorithm as true. For artificial datasets the optimal feature subset is known *a priori*, and therefore it is easy to quantify the performance of each FS algorithm for selecting the optimal feature subset.

We use the MONK dataset, setting $m = 3$ (we check for three true features), and examine the ability of the FS algorithms to detect the true features at each of the iterative $1 \dots m$ steps. In this simple benchmark problem, we have found that all FS algorithms correctly select the first two features in the first two steps. However, mRMR, mRMR$_{\text{Spearman}}$, RRCT$_0$, mRMR$_{\text{ITL}}$ failed to correctly select the true third feature in the third iterative step. It is interesting that RRCT, which is an extension of mRMR$_{\text{Spearman}}$ (in terms of introducing (a) an information theoretic transformation of the correlation coefficient and, (b) a complementarity term), is able to recover correctly the third feature, whereas the FS algorithms using only one of the two extensions introduced in RRCT (i.e. RRCT$_0$, mRMR$_{\text{ITL}}$) fail.

Next, we use the Artificial1 and Artificial2 datasets and report the FDR results in Fig. 5.4. These results should be interpreted sequentially: each step on the x-axis denotes the iterative step in the FS algorithms, and the values in the y-axis denote whether each FS algorithm's choice identified true features in the subset (or whether it selected a probe). For example, a value of 1 in the y-axis for the first iterative step for one of the FS algorithms would denote that the first feature that is selected for the given FS algorithm is a probe. Similarly, for the second iterative step a value of 1 in the y-axis would denote that one out of the first two selected features is a probe, and so on. In Artificial1 we set $m = 10$, i.e. to check how accurately the FS algorithms recover the 10 true features in the first 10 steps, and similarly for

Artificial2 we set $m = 20$ (i.e. we evaluate iteratively how often, and at which steps the FS algorithms fail to correctly detect the true features in the datasets).

The results in Fig. 5.4 illustrate that RRCT works well to discard probes, demonstrating that the additional complementarity term over $\text{mRMR}_{\text{Spearman}}$, is beneficial. LASSO and GSO also exhibit competitive performance in terms of accurately recovering most of the true features. On the contrary, more sophisticated schemes such as mRMR, RELIEF, and LLBFS appear to select many probes amongst their top selected features. In both Artificial1 and Artificial2, RRCT_0 and mRMR_{ITL} select considerably more probes compared to RRCT (results not shown here).

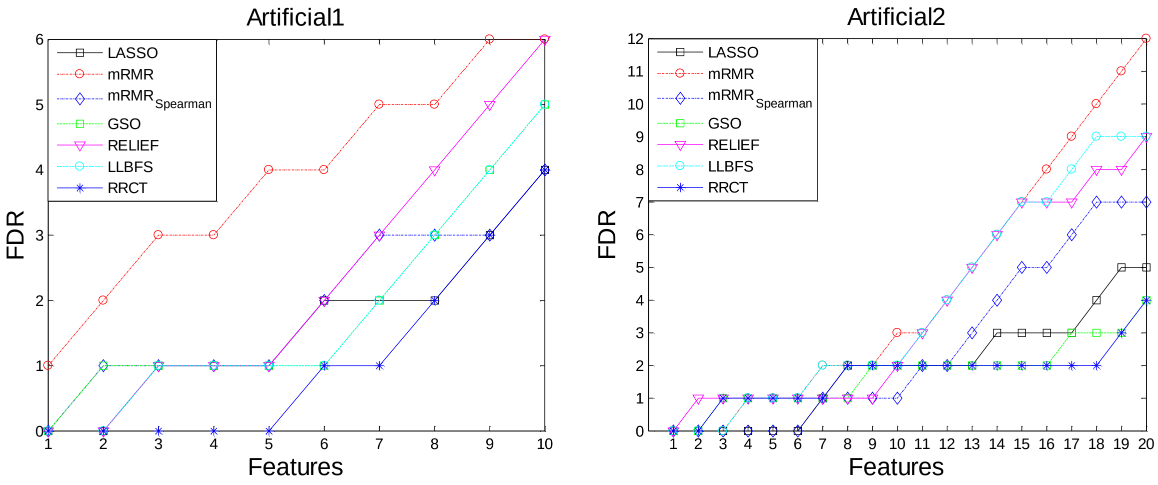

Fig. 5.4: Comparison of the feature selection algorithms in terms of true feature set recovery. The lower the False Discovery Rate (FDR), the better the feature selection algorithm. The horizontal axis denotes the number of features selected during the incremental process.

5.3.2 Validating feature selection algorithms based on learner performance

We have already referred to prediction performance of a learner as a proxy for the accuracy of the FS algorithms, and have highlighted the pitfalls in that approach.

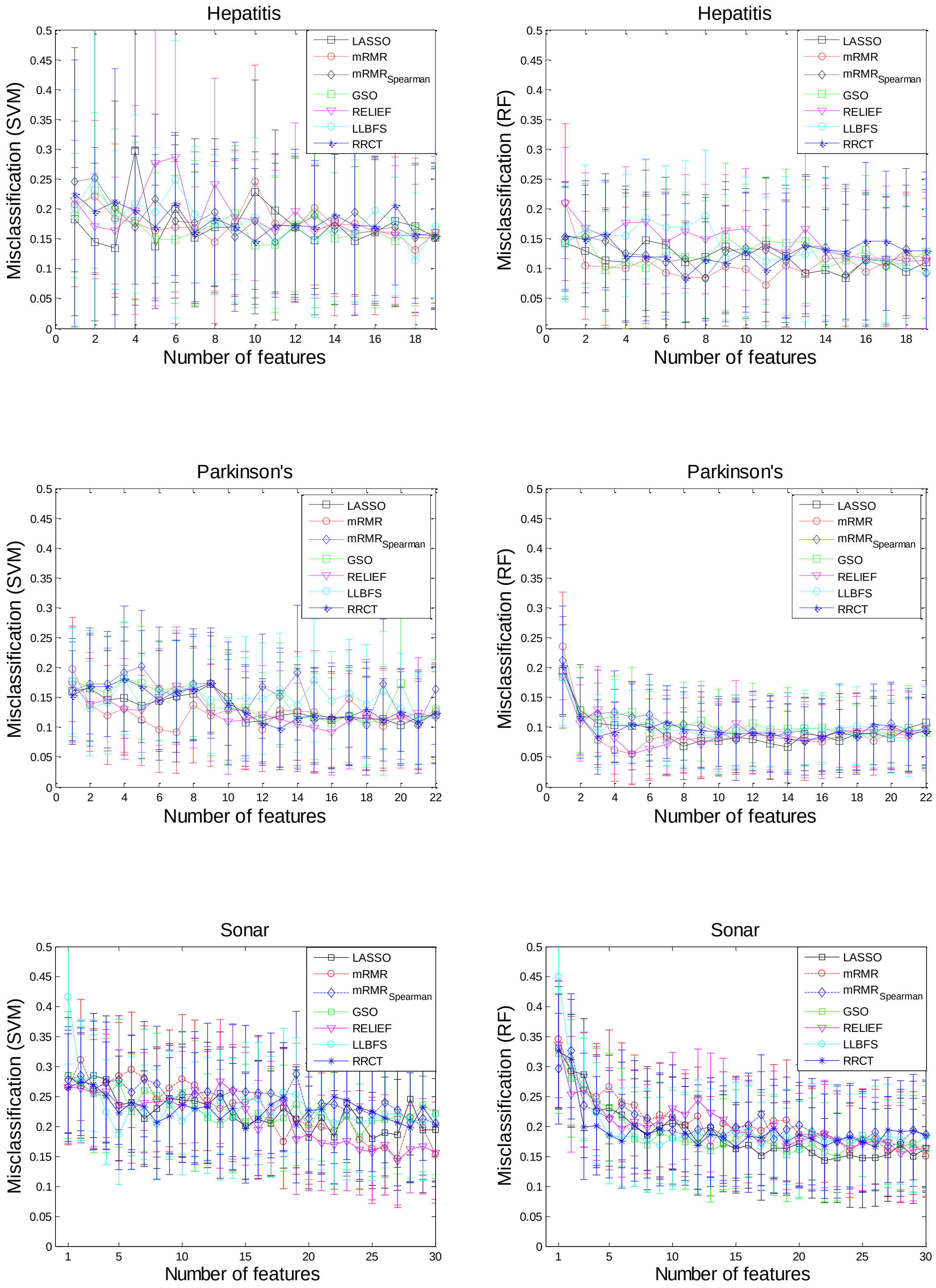

Fig. 5.5: Comparison of the feature selection algorithms based on learner performance (binary classification datasets) using SVM (left) and RF (right). The horizontal axis denotes the number of features selected in the greedy feature selection process.

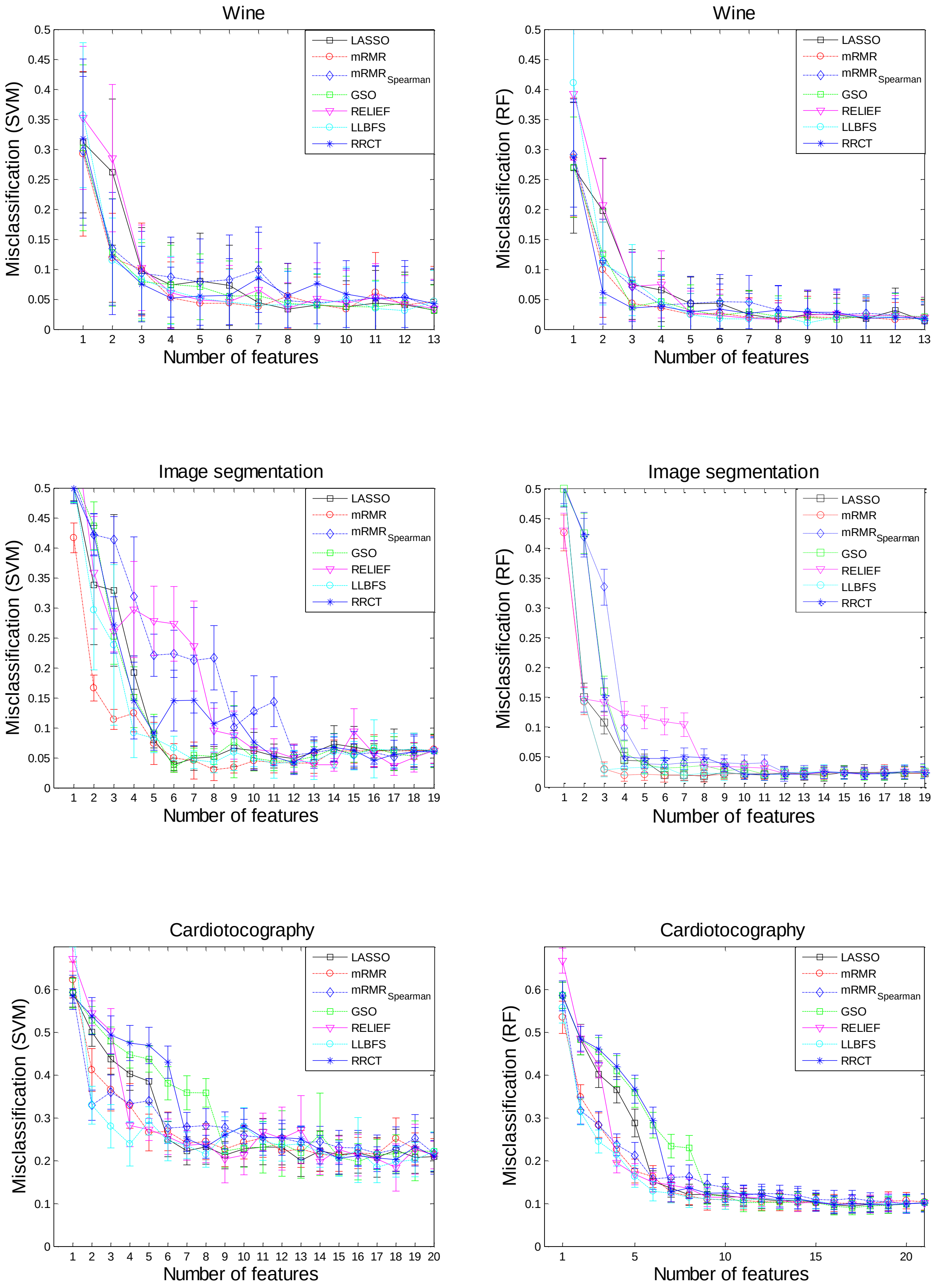

Fig. 5.6: Comparison of the feature selection algorithms based on learner performance (multi-class classification datasets) using SVM (left) and RF (right). The horizontal axis denotes the number of features selected in the greedy feature selection process.

In this respect, Tuv et al. (2009) argued that FS algorithms should be tested mainly to demonstrate how accurately they recover true features and discard probes, and not use the learner performance as an indicator to demonstrate which FS algorithm is superior. Nonetheless, to conform with the literature on FS, we also evaluate the performance of two learners (SVM and RF) as a function of the features selected by each of the FS algorithms. Typically, at least two learners should be used to draw conclusions when learner performance is used to evaluate FS algorithms: it is possible that different learners could promote different FS algorithms, whereas ideally the selected feature set should generalize well to all learners.

The feature subsets selected by each FS algorithm are included in Appendix I to enable other researchers to directly compare their findings against the investigated FS algorithms. Here, we present only the classification performance as a function of the features presented to the two learners and refer to Appendix I for the actual features selected using each FS algorithm. Specifically, Fig. 5.5 presents the results for binary classification settings, Fig. 5.6 for multi-class classification settings, and Fig. 5.7 for 'fat' datasets. Overall, there is no clear winner amongst the competing FS algorithms in terms of performance, but RRCT works very well generally, particularly for the fat datasets.

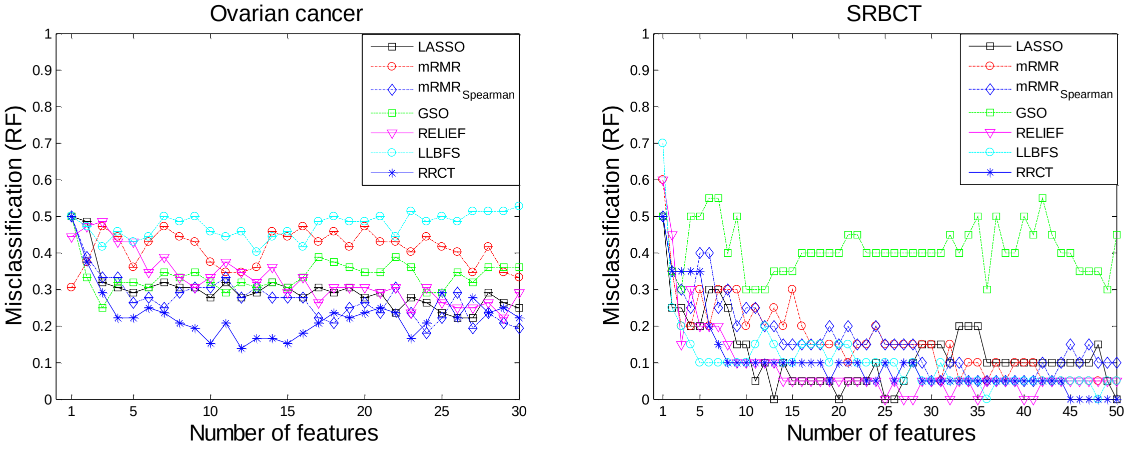

Fig. 5.7: Comparison of the feature selection algorithms based on learner performance ('fat' datasets, where the number of features is larger than the number of samples).

5.3.3 Summarizing the feature selection results

We used two powerful nonlinear classifiers, SVM and RF, to evaluate the FS algorithms as a result of the predictive strength of the selected features. In most cases we have found that the ranking of the FS algorithms is similar for both SVM and RF. Still, for example in the Sonar dataset after the 20th step, RELIEF leads to the best classification performance when using SVM, whereas RF has the best performance when using the features selected by LASSO. RF appear to be slightly better than SVM for the datasets used here. For the fat datasets RF was considerably better; hence the results with SVM are not shown.

We have found that mRMR$_{\text{Spearman}}$ works adequately well for many datasets, avoiding the computational complexity of mRMR which relies on MI. On the contrary, the mRMR-type FS algorithms using the divergence metrics of Table 4.1 led to selected feature subsets with considerably worse classification performance (results not shown). This finding could suggest that these divergence metrics may be overly sensitive or most probably indicate the problems associated with density estimation (the MI in the mRMR scheme was evaluated directly using entropy estimates, thus circumventing this step). The mRMR$_{\text{Spearman}}$ is also the predecessor to the new FS algorithm RRCT, which appears to be very competitive against popular schemes.

The RRCT algorithm appears well suited to discarding probes in the artificial datasets, often outperforming competing FS algorithms. In the datasets examined here, RRCT selected features that led to good learner performance, often outperforming the competing FS algorithms. One particularly attractive characteristic of RRCT is that it may not demonstrate an edge over other specific FS algorithms in all cases, but it appears to be fairly robust and is typically amongst the FS algorithms selecting a feature subset that leads to very good classifier performance. Moreover, although not tested in this study, RRCT has the additional advantage that it can be readily used both for classification and regression applications, whereas some of the investigated FS algorithms cannot be readily generalized to such settings. Perhaps surprisingly, in some cases the Box-Cox transformation leads to degraded performance. This is an inherent problem of the maximum likelihood approach used for the computation of λ in Eq. (4.25) in the presence of outliers; better approaches to power transformations might be more suitable in these cases (Marazzi and Yohai, 2006). Other density normalization techniques might be more appropriate pre-processing steps prior to FS with RRCT (and possibly other FS algorithms) but we do not pursue these here.

We remark that there is no clearly superior FS algorithm for all datasets, which can be seen as one manifestation of the *no free lunch theorem*[38]. For example, mRMR is very good in Image Segmentation, but rather poor in the Sonar dataset. A very large empirical study using diverse datasets to identify the settings where particular FS algorithms excel and fail would be very useful in this regard. Here, we can propose a tentative explanation for why particular FS algorithms might be suited to specific datasets (or domains), which might give a good indication of their performance in similar settings, e.g. in terms of the number of classes, the number features and the number of samples (or possibly their ratio), and the correlation matrix. RRCT is well suited to datasets where feature complementarity is prominent, such as in micro-array datasets (Ovarian cancer and SRBCT). Since it is a correlation-based filter, its weakness is in datasets where the relationship between features and the response can only be captured by higher order moments (e.g. the cardiotocography dataset). LASSO is particularly suitable in minimally correlated datasets, but its performance degrades in highly correlated datasets; similarly mRMR underperforms in settings where complementarity is essential (e.g. the micro-array datasets). RELIEF is most efficient in settings where the redundancy amongst the most relevant features is minimal.

[38] Informally, the *no free lunch* theorem states that there is no machine learning approach which will be *universally* optimal, and finds application in various sub-fields in machine learning such as classification and regression, in addition to FS.

Chapter 6

Data acquisition: speech and Parkinson’s disease

This chapter introduces the speech collection protocol and the speech-PD databases used in Chapter 7. In addition, it describes possible confounding factors when inferring PD symptom severity from speech recordings.

6.1 Speech data collection protocol

Speech processing is an important discipline in its own right and is the focus of dedicated study in phonetics and linguistics. A major workshop sponsored by the US National Center for Voice and Speech (NCVS) was held in Colorado in 1994 “*to reach better agreement on [the] purpose and methods of acoustic analysis of voice signals*” (Titze, 1994). The approved data acquisition recommendations and conclusions that are directly relevant to this study are summarised below *in the wording of the workshop contributors*:

1) Sustained vowels should continue to be used for voice perturbation analysis because they elicit a stationary process in vocal fold vibration[39]
2) A professional-grade condenser microphone (omnidirectional or cardioid)[40] with a minimum sensitivity of -60 dB should be used

[39] We remark that this assertion is not strictly speaking correct, particularly for pathological voices.

[40] These terms appear in various engineering contexts, for example in telecommunications (e.g. for antennas). Omnidirectional refers to a microphone being equally sensitive in a 2 dimensional plane; cardioid refers to sensitivity in the epicycloid with the cusp at the central point of the microphone.

3) For steady vowel phonations, the mouth to microphone distance can be held constant and less than 10 cm (preferably 3-4 cm)
4) A 16-bit ADC is recommended, but this must be accompanied by conditioning electronics (amplifiers, filters) that have signal to noise ratio (SNR) in the 85-95 dB range
5) Sampling frequencies of 20-100 KHz should be used

Dissecting the above points, one can briefly reflect on the most appropriate data acquisition methods regarding speech signals, which potentially offer a valuable monitoring tool. As we have seen in § 2.2.4, there are many compelling reasons for using sustained vowels. High-quality electronics promise high-quality data, capturing acoustic information which may otherwise have been lost. The quality of the signal decays proportionally with the mouth-microphone distance, and using a sampling frequency larger than 20 KHz captures the higher speech harmonics whose properties may have clinical value.

6.2 Speech database used for the discrimination of healthy controls from people with Parkinson’s disease

The National Center for Voice and Speech (NCVS) database comprises 263 phonations from 43 subjects (17 females and 26 males, 10 healthy controls and 33 PWP). It is an extension of the database used originally in Little et al. (2009), and the extended database includes all the voice recordings from the earlier study. The 10 healthy controls (4 males and 6 females), had an age range of 46 to 72 years with (mean ± standard deviation) 61 ± 8.6 years, and we processed 61 healthy phonations. The 33 PWP (22 males and 11 females), had an age range of 48 to 85 (67.2 ± 9.3), time since diagnosis 0 to 28 years (5.8 ± 6.3); we processed 202 PD phonations. This database contains six or seven sustained vowel "ahh…" phonations from each speaker, recorded at a comfortable frequency and amplitude.

The phonations were recorded in an IAC sound-treated booth with a head mounted microphone (AKG C420), which was placed at 8 cm distance from the subject's mouth. The voice signals were sampled at 44.1 kHz with 16 bits resolution, and were recorded directly to computer using CSL 4300B hardware (Kay Elemetrics).

6.3 Speech database used for estimating the Unified Parkinson's Disease Rating Scale (UPDRS)

The At-Home Testing Device (AHTD) database was described in Goetz et al. (2009): it is a novel telemonitoring device built by Intel Corporation for collecting data from PWP. Originally, 52 subjects with idiopathic PD with diagnosis within the previous five years at trial onset were recruited. A PD diagnosis was given if the subject had at least two of the following: rest tremor, bradykinesia (slow movement) or rigidity, without evidence of other forms of Parkinsonism. The study was supervised by six US medical centers: Georgia Institute of Technology (7 subjects), National Institutes of Health (10 subjects), Oregon Health and Science University (14 subjects), Rush University Medical Center (11 subjects), Southern Illinois University (6 subjects) and University of California Los Angeles (4 subjects). All patients gave written informed consent, and remained un-medicated for the six-month duration of the study. We disregarded data from 10 recruits – two that dropped out the study early, and a further eight that provided insufficient test data. The 42 PWP used in this study had at least 20 valid study sessions during the trial period.

The 28 male participants had an age range (mean ± standard deviation) 64.8 ± 8.1, min. 49, max. 78, median 65 years, with 63.0 ± 61.9, min. 1, max. 260, median 48 weeks since diagnosis. UPDRS was assessed at baseline (onset of trial), after three, and after six months: (20.3 ± 8.5, 21.9 ± 8.7, 22.0 ± 9.2), min. (6, 6, 5), max. (36, 38, 41), median (21, 22, 20) points for motor UPDRS, and (27.5 ± 11.6, 30.4 ± 11.8, 31.0 ± 12.4), min. (8, 7, 7), max. (54, 55, 54), median (27, 28.5, 26.5) points for total UPDRS.

The 14 female participants had an age range 63.6 ± 11.6, min. 36, max. 85, median 64 years with 89.7 ± 81.2, min. 4, max. 252, median 60 weeks since diagnosis. Their UPDRS at baseline, after three and after six months was: (17.6 ± 7.4, 21.2 ± 10.5, 20.1 ± 9.4), min. (6, 6, 8), max. (32, 38, 38), median (18, 18.5, 19.5) points for motor UPDRS, and (24.2 ± 9.1, 27.4 ± 12.1, 26.8 ± 10.8), min. (10, 7, 10), max. (42, 46, 49), median (25, 28, 24.5) points for total UPDRS. At baseline, the combined (male and female) scores were 19.42 ± 8.12, min. 6, max. 36, median 18 points for motor UPDRS, and 26.39 ± 10.80, min. 8, max. 54, median 25.5 points for total UPDRS. After three months: 21.69 ± 9.18, min. 6, max. 38, median 21 points for motor UPDRS, and 29.36 ± 11.82, min. 7, max. 55, median 28 points for total UPDRS, and after six months: 29.57 ± 9.17, min. 5, max. 41, median 20 points for motor UPDRS, and 29.57 ± 11.92, min. 7, max. 54, median 26 points for total UPDRS.

Fig. 6.1 displays graphically the data acquisition and UPDRS estimation procedure. The data is collected at the subject's home, transmitted over a dedicated, purpose-built secure server, and processed in the clinic to predict the UPDRS score. The AHTD is designed to facilitate remote, Internet-enabled measurement of a variety of PD-related motor impairment symptoms: it contains a docking station for measuring tremor, paddles and pegboards for assessing upper body dexterity, a high-quality microphone headset for recording patient voice signals and a USB data stick to store test data. A Liquid Crystal Display (LCD) guides the subject in taking the tests.

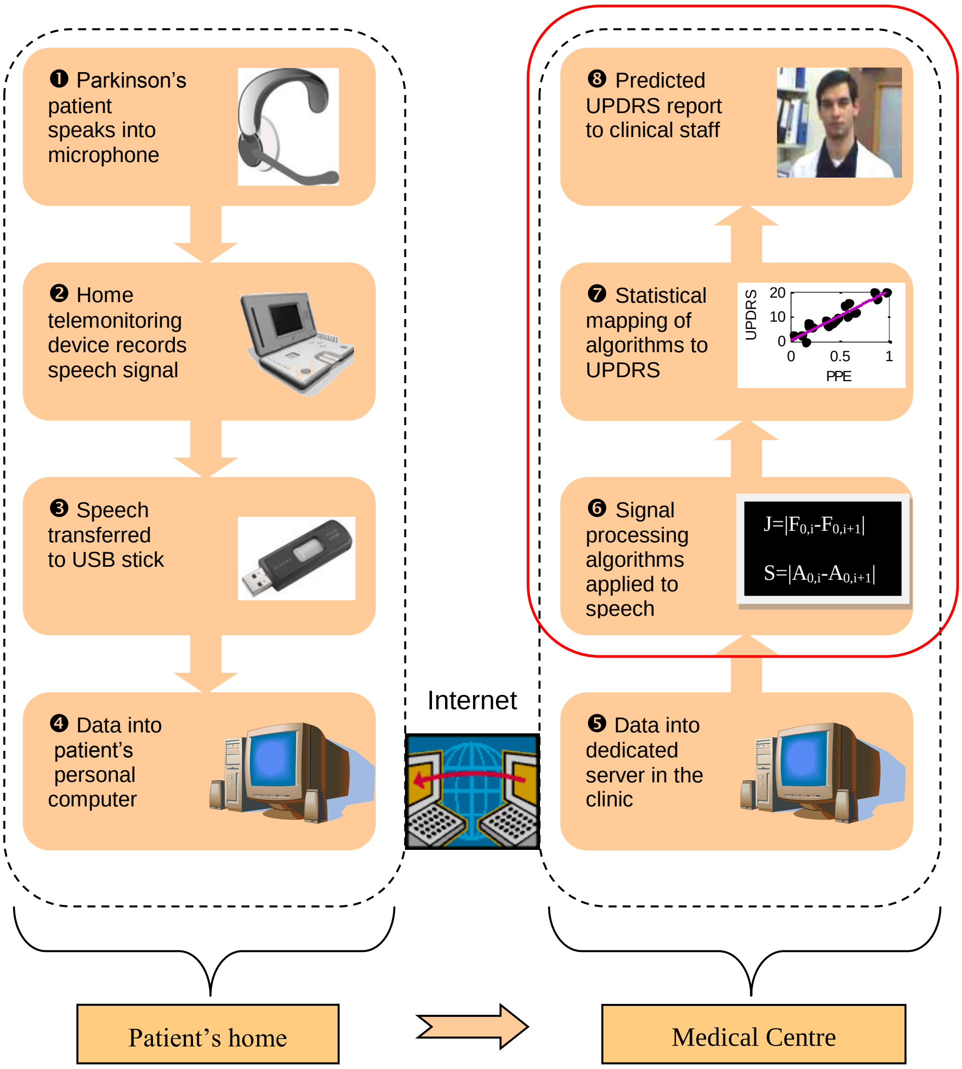
❶ Parkinson’s patient speaks into microphone
❷ Home telemonitoring device records speech signal
❸ Speech transferred to USB stick
❹ Data into patient’s personal computer
Internet
❺ Data into dedicated server in the clinic
❻ Signal processing algorithms applied to speech
J=|F0,i-F0,i+1|
S=|A0,i-A0,i+1|
❼ Statistical mapping of algorithms to UPDRS
UPDRS
20
10
0
0
0.5
1
PPE
❽ Predicted UPDRS report to clinical staff
Patient’s home
Medical Centre

Fig. 6.1: Schematic diagram depicting the data acquisition procedure and the methodology to estimate the average Parkinson’s disease symptom severity expressed using the Unified Parkinson's Disease Rating Scale (UPDRS). The device that collects the data is known as the At-Home-Testing-Device (AHTD). The red box (steps 6-8) is the focus of this study.

Audible prompts instruct the subject to undertake tasks to measure tremor, bradykinesia, complex co-ordinated motor function, speech and voice. As part of a trial to test the effectiveness of the AHTD system in practice, PWP were recruited and trained to use the device. Subsequently, an AHTD was installed in their home and they performed tests on a weekly basis. Each patient specified a day and time of the week during which they had to complete the test protocol, prompted with an automatic alarm reminder on the device. The collected data was encrypted and transmitted to a dedicated server automatically when the USB stick was inserted into a computer with internet connection. Further details of the AHTD apparatus and trial protocol can be found in Goetz et al. (2009). Henceforth, we refer to this trial as the *AHTD trial*.

The audio recordings are of two types: sustained phonations and running speech tests. In the sustained vowel phonations, the subjects were instructed to say "ahh…" and keep the pitch as steady as possible, for as long as possible. In the running speech tests the subject was instructed to describe photographs displayed on the AHTD's screen, specifically chosen to elicit emotional responses. The voice samples were recorded using a head-mounted microphone placed approximately 5 cm from the patient's lips. The AHTD uses a spoken instruction followed by a "beep" prompting the subject to begin phonation; an audio amplitude threshold detector triggered the capture of audio, and subsequently the capture was stopped one second after the detected signal amplitude dropped below that threshold, or 30 seconds of audio had been captured (whichever occurred sooner). The voice signals were recorded directly to the AHTD USB stick sampled at 24 KHz with 16 bit resolution.

Following initial screening to remove faulty recordings (for example failure to record a phonation, subject coughing, or initialization and very early termination of the phonation), 5875 sustained vowel /a/ phonations were digitally processed using algorithms implemented in the Matlab software package. Six phonations were recorded each day on which the test was performed: four at comfortable pitch and loudness and two at twice the initial loudness (but without shouting).

6.4 Confounding factors when inferring Parkinson’s disease symptom severity from speech recordings

For the sake of completeness, the ambiguities and confounding factors of the current study resulting from processing the speech recordings to infer PD symptom severity are explicitly stated:

1. *Distance of microphone from the subject’s mouth*

It is straightforward to relate the measured speech signal **s** with the power emitted by the vocalist’s mouth. As the acoustic wave energy is radiated, it spreads spherically around the vocalist’s head (Flanagan, 1972) giving rise to:

$$I = \frac{W_{out}}{4\pi r^2} \tag{6.1}$$

where W_{out} is the power emitted by the source (the vocalist’s mouth), measured in Watts, r is the distance of the microphone from the source, measured in meters, and I is the intensity, which is a measure of power per unit area (Watts/m^2).

Then, the relationship between **s** and W_{out} is (Titze 2000):

$$W_{out} = 4\pi r^2 \cdot 10^{(\mathbf{s}-120)/10}. \tag{6.2}$$

Eq. (6.2) shows the relationship between the actual measured speech signal and the density of the emitted power from the speaker’s mouth. As can be seen from Eq. (6.1), the distance from microphone scales the recorded amplitude. "Wind noise" can also arise if the subject keeps the microphone very close to the mouth. In the NCVS database, the data was collected under controlled conditions, so we can be fairly confident that the protocol was followed precisely. In the AHTD trial, the PWP were trained for a week on how to use the device, and subsequently they performed the tests at their homes without supervision. We assume that the microphone was placed properly 5 cm away from the vocalist's mouth (the AHTD uses a headset microphone which 'automatically' gets the spacing). Nevertheless, it cannot be verified that the subjects did not interfere with the placing of the microphone. This has important consequences because, as we have already seen in Chapters 2 and 3, vocal intensity is important and is affected in PWP.

2. *The peak amplitude and sustained phonation time depends on lung efficiency*

Given that no medical records are available, it could be that a subject fails to sustain his phonation or has a problem in generating peak amplitude voice signals due to lung inefficiency or other health problem, and not because of PD. We assume that the speech signals are affected only by PD, and not any other underlying pathologies.

3. *Not all demographic data are available*

Some published studies indicate that height, weight and often the profession of the subjects affect phonations. For example, someone who is taller is expected to have a larger larynx and thus lower F_0 (Titze, 2000). The chronological age of the subjects is known, but studies have shown that the vocal performance depends more on *physiological age* rather than *chronological age*[41] (Titze, 2000). We are using the subjects' chronological age to estimate expected F_0, based on Fig. 2.5. Similarly, some additional parameters could affect the subjects' phonations, for example if they are smokers, or whether women subjects are past menopause, but these details are not available to this study.

4. Natural production of phonation

The subjects are expected to produce phonations as naturally as possible. Any phonation which is articulated slightly differently results in different kinds of sound. For example, closing the lips interferes with the generation and transmission of the harmonic content of the speech signal.

5. Time of recordings

In the AHTD study the subjects were instructed to record their voices during a specific time interval in the morning. This is because the voice undergoes constant changes during the day. However, mitigating this potential confound is the fact that data on the time the recordings took place is available. For the NCVS database this information is not available.

[41] Physiological age is the age as perceived by listeners. Chronological age is the actual age in years.

Chapter 7

Parkinson's disease classification using speech signals

In Chapter 4, we outlined a general methodology for analysing high-dimensional data, which is now applied to investigate (a) how accurately we can discriminate healthy controls from PWP, and (b) the relationship of speech and average PD symptom severity, when symptom severity is quantified using the standard clinical metric UPDRS. The aim is to develop a functional mapping of dysphonia measures extracted from the speech signals to (a) a binary classification response for the NCVS database which was described in § 6.2, or (b) motor UPDRS and total UPDRS, using the AHTD database which was described in § 6.3.

7.1 Using speech signals to discriminate healthy controls from people with Parkinson's disease

In this section we work with the NCVS database where we have 263 sustained vowel phonations, and the corresponding response values indicating whether the subject belongs to a binary class: PWP (denoted by '1') versus healthy controls (denoted by '0'). We aim to characterize the speech signals extracting features, and using those features to determine whether we can automatically classify subjects into the two classes. The sustained vowel phonations were analyzed using the dysphonia measures outlined in Chapter 3 and summarized in Table 3.2. Thus, each sustained vowel phonation was characterized by 318 dysphonia measures. In Tsanas et al. (2012b), a subset of the dysphonia measures used in this study was investigated (the 132 dysphonia measures described in Tsanas et al. (2011a)); here we include also the 180 wavelet dysphonia measures used previously in Tsanas et al. (2010c) to investigate whether we can further improve on those findings. Moreover, building on the results of § 5.1, we also use the F_0 contour estimated using two additional F_0 estimation algorithms (NDF, and the ensemble F_0 scheme in Eq. 5.1) to compute six additional dysphonia measures which fall under the label “F_0 related measures” in Table 3.2. In summary, the 318 dysphonia measures in this study are the 132 dysphonia measures from Tsanas et al. (2011a), 180 dysphonia measures from Tsanas et al. (2010c), and six additional dysphonia measures (three measures for each F_0 estimation algorithm).

Now, we follow the methodology outlined in Chapter 4 (see in particular § 4.5 which summarizes the proposed methodology), to investigate how accurately we can differentiate PWP from healthy controls. In order to gain a preliminary understanding of the statistical properties of the features for this application, we computed the Pearson correlation coefficient between each feature and the response variable. The Pearson correlation coefficient is indicated in this application because the response variable is binary.

Table 7.1 presents the most strongly associated dysphonia measure from each dysphonia measure family with the response. These results provide a general overview of the association strength of algorithmically related features with the response. We remark that the dysphonia measures proposed in Tsanas et al. (2010c; 2011a) and in Little et al. (2007) along with the MFCCs appear to be statistically strongly associated with the response. The relatively high absolute correlation coefficient values ($R > 0.3$) provide an initial indication that the binary classification task of differentiating PWP from healthy controls might be successful. Table 7.2 summarizes classification results reported in the literature for the discrimination of PWP versus healthy controls, when sustained vowels are used.

Table 7.1: Statistical associations of indicative dysphonia measures with the response variable to differentiate people with Parkinson’s disease and healthy controls

Dysphonia measure	Description	Correlation coefficient
10th level detail wavelet coef$_{\mathrm{TKEO,std}}$	Wavelet coefficient at the 10th decomposition level summarized using the standard deviation of the TKEO values of the coefficients	**0.399**
VFER$_{\mathrm{entropy}}$	Extent of noise in the speech signal using entropy	**-0.388**
11th MFCC coef	11th Mel Frequency Cepstral Coefficient	**0.369**
4th delta MFCC	1st derivative of the 4th MFCC	**-0.363**
$F_{0,NDF} - F_{0,exp}$	Mean difference of the cycle-to-cycle F_0 estimate and the average expected F_0 in age- and gender-matched healthy controls	**-0.357**
RPDE	Quantify the stochastic component of the deviation of vocal fold periodicity	**0.292**
DFA	Characterizes the extent of turbulent noise, quantifying its stochastic self-similarity	**0.287**
Shimmer$_{\mathrm{PQ11}}$	Amplitude differences using an 11 sample window of F_0 estimates	**0.285**
HNR$_{\mathrm{mean}}$	Signal to noise ratio measure	**-0.285**
Jitter$_{\mathrm{F0,TKEO,std}}$	Standard deviation of the TKEO of the fundamental frequency perturbations quantified with jitter	**-0.268**
GQ$_{\mathrm{std,open}}$	Standard deviation of the glottal quotient for the duration where vocal folds are apart	**0.237**
GNE$_{\mathrm{std}}$	Standard deviation of the glottal to noise excitation	**0.231**

For illustration, one dysphonia measure from each algorithmic family is presented and the results are sorted using the absolute value of the Pearson correlation coefficient. All reported correlations were statistically significant ($p < 0.001$). If there was no measure from an algorithmic family that was statistically significant, that algorithmic family was not represented in the table. In addition, the Mann Whitney statistical test suggests that the differences in the distributions of the features belonging to the two different classes are statistically significant ($p < 0.001$). The response was defined as ‘0’ for healthy controls and ‘1’ for people with Parkinson’s disease; therefore dysphonia measures associated with positive correlation coefficient sign indicate that those dysphonia measures have on average larger values for Parkinson’s disease phonations compared to healthy control phonations.

Those studies used the exact design matrix originally computed in Little et al. (2009) which comprised 31 subjects (195 phonations) and 22 features (some jitter variants, some shimmer variants, HNR, DFA, RPDE and PPE). The results in the present study are obtained using a considerably larger database with 43 subjects (263 phonations) and 318 features. To facilitate comparison with the original study of Little et al. (2009), Table 7.2 reports also the classification accuracy obtained with the classification algorithms used in the current study, when the optimal feature subset computed by Little et al. (2009) is fed into the classifiers.

Table 7.2: Summary of classification results reported in the literature for the application of discriminating people with Parkinson's and healthy controls using sustained vowels.

Study	**Learning and validation scheme**	**Reported accuracy (%)**
Guo et al. (2010)	GP-EM, 10-fold cross-validation	93.1 ± 2.9
Das (2010)	Neural network, 35% of the data used for testing following random initial partitioning	92.9
Sakar and Kursun (2010)	SVM, bootstrap with 50 replicates	92.8 ± 1.2
Little et al. (2009)	SVM, bootstrap with 50 replicates	91.4 ± 4.4
Psorakis et al. (2010)	Non-sparse E-M, 10-fold cross-validation with 10 repetitions	89.5 ± 6.6
Shahbaba and Neal (2009)	dpMNL, 5-fold cross-validation	87.7 ± 3.3
*Optimal feature subset from Little et al. (2009)	SVM methodology in this study, 10-fold cross-validation with 100 repetitions, features recalculated	89.3 ± 6.9
*Optimal feature subset from Little et al. (2009)	RF methodology in this study, 10-fold cross-validation with 100 repetitions, features recalculated	89.3 ± 7.2
*All 318 features	SVM, 10-fold cross-validation with 100 repetitions	97.7 ± 2.8
*All 318 features	kNN, 10-fold cross-validation with 100 repetitions	93.1 ± 5.2
*All 318 features	RF, 10-fold cross-validation with 100 repetitions	90.2 ± 5.9

The results are presented in the form mean ± standard deviation where appropriate. The asterisk (*) indicates new results of the present study. SVM stands for *support vector machine*, dpMNL for *Dirichlet process multinomial logit*, GP-EM for *genetic programming and the expectation maximization algorithm*, E-M for *expectation maximization* algorithm, kNN for *k-nearest neighbours*, and RF for *random forests*. All cited studies used the features derived in Little et al. (2009) with 31 subjects; the results in the present study are from an expanded database with 43 subjects, with all features recalculated.

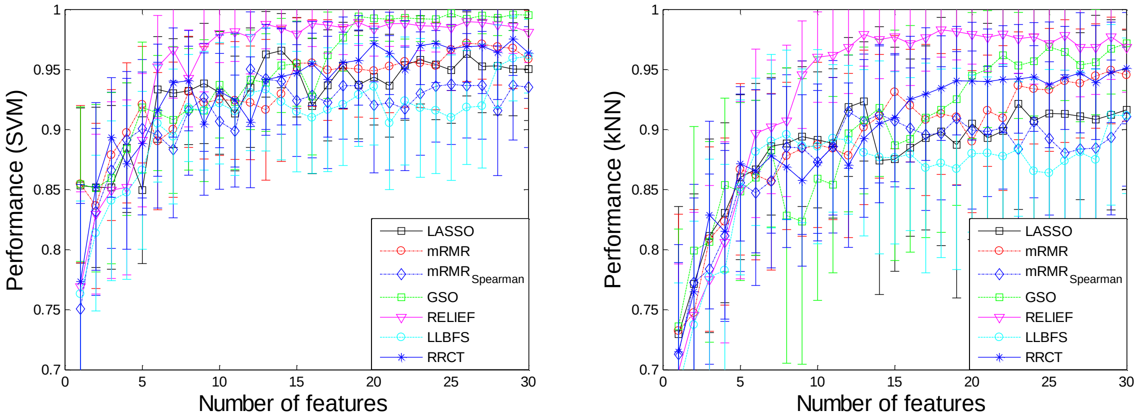

Fig. 7.1 Comparison of out of sample performance results with confidence intervals (one standard deviation around the quoted mean performance) using the features selected by each of the seven feature selection algorithms (for clarity, only the first 30 steps are presented). These results are computed using 10-fold cross validation with 100 repetitions.

The best classification accuracy according to the reports in the literature was about 93%. Using all the 318 features leads to a noticeable improvement in accuracy over the results reported in the literature: 97.7 ± 2.8% using SVM. Interestingly, using a simple classifier (kNN) we get 93.1 ± 5.2% accuracy. Although we have used cross-validation to provide an estimate of the generalization error, it is possible those results might not generalize well on a novel dataset. Following the principle of parsimony, it would be desirable to reduce the dimensionality of the feature set, which might also potentially lead to further improvement in classification accuracy. For this reason, we computed the performance of the learners using the features selected by seven FS algorithms (see Table 7.3) as a function of the number of features fed into the learner (see Fig. 7.1). We report our findings using SVM and kNN ($k = 1$), which in this application gave superior results to RF. SVM consistently outperforms kNN, but it is interesting that a simple classifier such as kNN can lead to 98% accuracy when presented with a parsimonious, information-rich feature subset.

Table 7.3: Selected feature subsets and classification performance differentiating people with Parkinson’s disease and healthy controls

LASSO	mRMR	mRMR Spearman	GSO	RELIEF	LLBFS	RRCT
10th level detail wavelet coef$_{TKEO,std}$	1st level approximation wavelet coef$_{TKEO,mean}$	Shimmer$_{PQ11}$	10th level detail wavelet coef$_{TKEO,std}$	VFER$_{NSR,SEO}$	VFER$_{NSR,SEO}$	Shimmer$_{PQ11}$
VFER$_{NSR,TKEO}$	4th level detail coef log entropy log-F0	VFER$_{NSR,SEO}$	VFER$_{NSR,SEO}$	2nd MFCC coef	11th MFCC coef	VFER$_{NSR,SEO}$
4th delta MFCC	VFER$_{NSR,SEO}$	4th delta MFCC	4th level detail coef log entropy log-F0	VFER$_{NSR,TKEO}$	VFER$_{NSR,TKEO}$	2nd MFCC coef
VFER$_{NSR,SEO}$	VFER$_{NSR,TKEO}$	GNE$_{std}$	HNR$_{mean}$	0th MFCC coef	4th level detail coef log entropy log-F0	GNE$_{std}$
11th MFCC coef	2nd MFCC coef	11th MFCC coef	VFER$_{SNR,SEO}$	11th MFCC coef	10th level detail wavelet coef$_{TKEO,std}$	11th MFCC coef
HNR$_{mean}$	GNE$_{TKEO}$	9th level wavelet coef wavelet energy	GNE$_{std}$	1st MFCC coef	2nd MFCC coef	13th delta MFCC
GNE$_{std}$	5th delta delta MFCC coef	12th MFCC coef	12th MFCC coef	3rd MFCC coef	9th level detail wavelet coef$_{TKEO,std}$	4th delta MFCC
Shimmer$_{PQ11}$	11th MFCC coef	4th MFCC coef	Jitter$_{PQ5}$	VFER$_{NSR,SEO}$	Entropy of F0 1st level approx. coef	9th level detail wavelet coef$_{TKEO,mean}$ log-F0
VFER$_{entropy}$	8th level detail coef log entropy log-F0	Shimmer$_{CV}$	VFER$_{SNR,SEO}$	9th MFCC coef	Entropy of F0 2nd level approx. coef	GQ$_{std,open}$
Shimmer %	4th delta MFCC	GQ$_{std,open}$	6th MFCC coef	Log energy	Entropy of F0 3rd level approx. coef	VFER$_{NSR,TKEO}$
94.7 ± 4.6 TP: 97.9±3.2 TN: 85.8±14.3	94.1 ± 3.9 TP: 97.6±3.3 TN: 84.3±13.2	93.8 ± 4.4 TP: 97.3±3.3 TN: 82.7±16.3	96.3 ± 3.4 TP: 99.4±2.1 TN: 88.3±12.2	98.5 ± 2.3 TP: 99.4±1.3 TN: 94.0 ± 9.4	96.9 ± 3.9 TP: 98.2±1.4 TN: 88.1±14.2	96.1 ± 3.2 TP: 99.5±1.6 TN: 85.1±12.7

The last row presents the % accuracy when the selected features from each algorithm are fed into the SVM classification algorithm. The results are given in the form mean ± standard deviation and are out of sample computed using 10-fold cross validation with 100 repetitions. In the last row, TP stands for *true positive* (true assessment of PD) and TN for *true negative* (true assessment of healthy controls).

7.2 UPDRS statistics and structure

We start by examining the statistical properties of the UPDRS using the AHTD data. First, we plot the motor-UPDRS and total UPDRS densities in Fig. 7.1, in order to get an intuitive understanding of the spread of these metrics. The probability densities have their peak approximately in the middle of the recorded range of values, although motor-UPDRS appears to have two modes. In Appendix II we present more thorough analysis of the UPDRS metric, investigating grouping of sections, and relationships between the UPDRS components. Amongst the key findings is that we verified that motor-UPDRS is very strongly correlated to the total UPDRS (Table II.1). In fact, motor-UPDRS (third component of the UPDRS metric) is practically a reflection of the total-UPDRS score (Spearman $R = 0.95$).

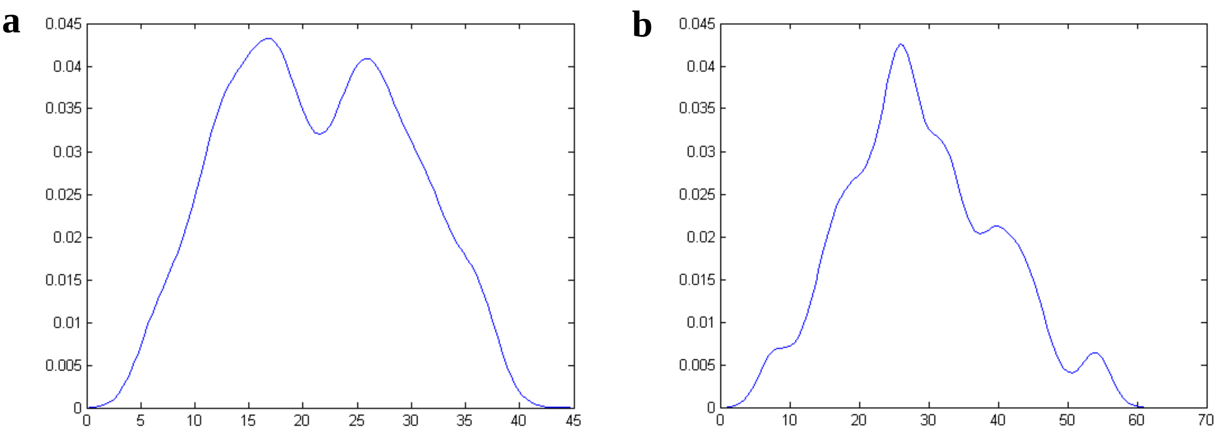

Fig. 7.2 Probability densities of the a) motor-UPDRS, b) total-UPDRS. The probability densities were estimated using kernel density estimation with Gaussian kernels.

This large association strength expressed using the Spearman correlation coefficient could be expected: the motor-UPDRS contributes 108 points out of the 176 points of the total-UPDRS, and the motor component of the UPDRS quantifies the hallmark symptoms of PD. Therefore, this finding justifies the widespread use in clinical practice of motor-UPDRS as a general indicator of general PD symptom severity (we have already mentioned in Chapter 2 that many studies focus solely on motor-UPDRS). The second component of UPDRS (part 2 – ADL) is also strongly correlated to total-UPDRS (Spearman $R \cong 0.7$). Interestingly, the components two (ALD) and three (motor) are also statistically significantly correlated with association strength of about 0.5. The correlation strength indicated by the Spearman correlation coefficient of the first UPDRS component (MBM) with components two and three is markedly lower. In Tables II.2 and II.3 in Appendix II we get an overview of the statistical correlation strengths between sections of the motor-UPDRS.

Stebbins et al. (1999) have used factor analysis (for a brief introduction see § 4.3.1) to determine the motor-UPDRS structure, by grouping UPDRS sections and identifying a number of common factors. They reported *motor-UPDRS* can be assessed on six distinct and clinically useful factors: speech, facial expression, balance and gait (factor I), rest tremor (factor II), rigidity (factor IV), right and left bradykinesia (factors III and V), and postural tremor (factor VI). They found relatively low correlations between the six factors, suggesting all contribute to accurate UPDRS estimation by capturing different aspects of PD symptoms. We have used factor analysis on the AHTD data and observed there is generally good agreement with the findings in Stebbins et al. (1999). However, we will not go into deeper detail here because factor analysis does not offer a unique representation of the data, and therefore most statisticians are very cautious in interpreting the results inferred by applying this method (Hastie et al., 2009).

This study builds on the premise that sustained vowel phonations can capture average PD symptom progression expressed by UPDRS. There is a strong relationship between speech and UPDRS, and this can be shown in statistical correlations. *Speech* appears explicitly twice in the UPDRS metric (see Appendix III): in section 5 (part of the ADL component) and in section 18 (part of motor-UPDRS). These two sections, taken together, are strongly correlated to motor-UPDRS ($p < 0.001$, Spearman $R = 0.44$) and total-UPDRS ($p < 0.001, R = 0.51$) indicating strong association between speech and UPDRS. These statistically significant findings intuitively suggest that the extraction of subtle features from speech signals could accentuate this concealed relationship.

The exploratory statistical analysis in this section was deemed necessary in order to try and understand the AHTD trial data. We have tentatively determined the internal structure of UPDRS, and computed correlation coefficients between the UPDRS sections (correlation matrix). We now proceed to study mapping speech dysphonias to UPDRS.

7.3 Functional mapping of dysphonia measures to UPDRS

The aim of this study is to characterise the speech signals with signal processing algorithms (dysphonia measures), select the most parsimonious set of the dysphonia measures (features), and map the selected feature subset to UPDRS. Ultimately, we want to replicate the clinicians' motor-UPDRS and total-UPDRS assessments as accurately as possible, using only the speech signals. The actual UPDRS values were obtained at three month intervals (baseline, three-month and six-month into the AHTD trial), whereas the voice recordings were obtained *weekly*; therefore *weekly* UPDRS estimates need to be derived to associate with each phonation.

The simplest approach to obtain those weekly estimates is to use nearest neighbour interpolated UPDRS scores. However, this would imply a sudden sharp UPDRS change mid-way between assessments and physiologically this is unlikely. Instead, in our studies we have used a straightforward piecewise linear interpolation, with the interpolation going exactly through the measured UPDRS scores (Tsanas et al., 2010a; 2011b; 2011a; 2012e). We interpolated both motor UPDRS and total UPDRS to assess the efficacy of the dysphonia measures for predicting both scores. The tacit assumption is that symptom severity did not fluctuate wildly within the three-month intervals over which the UPDRS were obtained. The assumption of average *linear* PD progression is the most *parsimonious* interpolation when lacking frequent UPDRS assessments, and has been verified in a number of previous studies many of which are reviewed in Chan and Holford (2001), and Maetzler et al., (2009). Particularly important for the argument of linear PD progression is a recent study by Schüpbach et al. (2010), where non-medicated PD subjects diagnosed within less than 5 years at trial onset were followed for 12 months: they showed that linear UPDRS progression is a very reasonable assumption on average. We have found that in this application it is better to discretize the interpolated UPDRS scores and work with *classifiers* instead of *regressors*; hence both motor-UPDRS and total-UPDRS were rounded to the closest integer value, giving rise to a multi-class classification setting. Discretising the (real-valued) response variable to transform a regression problem into a classification problem is well known in the machine learning literature, and often this step can enhance the prediction performance of the learner. For another application of this problem transformation from regression to classification in a different domain see for example Tsanas and Xifara (2012d).

Similarly to § 7.1, we extracted 318 dysphonia measures which will be used to estimate the two response variables: motor-UPDRS and total-UPDRS. Therefore, we have a design matrix 5875×318, which contained no invalid or missing entries. In many practical applications, partitioning the data may often provide improved classification or regression accuracy. We follow our previous studies where we partitioned the data according to gender, to investigate whether PD progression can be captured more accurately (Tsanas et al., 2011a; 2011b; 2012e). That is, instead of using the original design matrix with all the data (5875×M), we used a design matrix of size 4010×M for male PWP and 1865×M for female PWP.

Table 7.4: Statistical associations of indicative dysphonia measures with motor-UPDRS and total-UPDRS for the male subset

Dysphonia measure	**Description**	**Motor-UPDRS relevance and correlation**		**Total-UPDRS relevance and correlation**	
		MI	Spearman R	MI	Spearman R
DFA	Characterizes the extent of turbulent noise, quantifying its stochastic self-similarity	0.126	-0.16	**0.147**	-0.204
Log energy	Estimate of the log-energy	0.142	0.148	**0.145**	0.168
9th level approximation wavelet coef$_{\text{log-entropy}}$	Wavelet coefficient at the 9th decomposition level summarized using the log-entropy values of the coefficients	0.135	-0.104	**0.142**	-0.052
0th MFCC	0th Mel Frequency Cepstral Coefficient	0.132	0.171	**0.141**	0.196
VFER-NSR$_{\text{TKEO}}$	Ratio of the sum of the log-transformed mean TKEO of the band-pass signals for frequencies >2.5 kHz to the sum of the mean TKEO of the band-pass signals for frequencies <2.5 kHz	0.116	0.157	**0.113**	0.186
IMF-SNR$_{\text{entropy}}$	Signal to noise ratio using EMD-based entropy of energy	0.084	-0.138	**0.084**	-0.179
HNR$_{\text{std}}$	Standard deviation of the signal to noise ratio quantified using auto-correlation concepts	0.067	0.058	**0.074**	0.134
RPDE	Quantifies the stochastic component of the deviation of vocal fold periodicity	0.071	0.003	**0.072**	0.065
GQ$_{\text{std, cycle closed}}$	Standard deviation of the vocal fold collision time	0.072	-0.1	**0.067**	-0.097
Shimmer$_{\text{PQ3}}$	Amplitude differences using a 3 sample window of F_0 estimates	0.064	-0.068	**0.066**	-0.115
GNE$_{\text{NSR,SEO}}$	Extent of noise focusing on different frequency bands	0.062	0.097	**0.059**	0.11
Jitter$_{\text{TKEO,std}}$	Standard deviation of the TKEO of the vocal fold duration differences	0.06	-0.136	**0.055**	-0.068
Std $F_{0,PRAAT}$	Standard deviation of the F_0 contour estimated using PRAAT	0.036	0.171	**0.039**	0.145

For illustration, one dysphonia measure from each algorithmic family is presented and the results are sorted using the mutual information (MI) value. All reported correlations were statistically significant ($p < 0.001$). If there was no measure from an algorithmic family that was statistically significant, that algorithmic family was not represented in the table. The reported MI is normalized (i.e. MI lies between 0-1, where 0 denotes that UPDRS is independent on the dysphonia measure, and 1 indicates that UPDRS is completely determined by the dysphonia measure - see Section 3.2 for details). All speech signals from the male PWP were used to generate these results (N = 4010 phonations). The F_0 subscript text refers to the algorithm used to extract F_0.

Table 7.5: Statistical associations of indicative dysphonia measures with motor-UPDRS and total-UPDRS for the female subset

Measure	**Description**	**Motor-UPDRS relevance and correlation**		**Total-UPDRS relevance and correlation**	
		MI	Spearman R	MI	Spearman R
0th MFCC	0th Mel Frequency Cepstral Coefficient	0.221	-0.327	**0.225**	-0.344
Log energy	Estimate of the log-energy	0.21	-0.457	**0.204**	-0.488
9th level approximation wavelet coef$_{\text{log-entropy}}$	Wavelet coefficient at the 9th decomposition level summarized using the log-entropy values of the coefficients	0.153	-0.048	**0.139**	-0.106
PPE	PPE quantifies the impaired control of stabilised pitch	0.14	0.435	**0.133**	0.397
RPDE	Quantifies the stochastic component of the deviation of vocal fold periodicity	0.131	0.299	**0.126**	0.318
Std $F_{0,Rapt}$	Standard deviation of the extracted $F_{0,Rapt}$	0.11	0.473	**0.117**	0.47
Jitter$_{\text{pitch}}$ %	Percentage difference in pitch estimates	0.117	0.433	**0.105**	0.406
VFER-NSR$_{\text{TKEO}}$	Ratio of the sum of the log-transformed mean TKEO of the band-pass signals for frequencies >2.5 kHz to the sum of the mean TKEO of the band-pass signals for frequencies <2.5 kHz	0.104	-0.06	**0.105**	-0.092
HNR$_{\text{mean}}$	Mean of the signal to noise ratio quantified using auto-correlation concepts	0.086	-0.418	**0.099**	-0.436
Shimmer$_{\text{PQ11}}$	Amplitude differences using an 11 sample window of F_0 estimates	0.085	0.362	**0.091**	0.357
GQ$_{\text{std, cycle closed}}$	Standard deviation of the vocal fold collision time	0.084	0.235	**0.079**	0.25

For illustration, one dysphonia measure from each algorithmic family is presented and the results are sorted using the mutual information (MI) value. All reported correlations were statistically significant ($p < 0.001$). If there was no measure from an algorithmic family that was statistically significant, that algorithmic family was not represented in the table. The reported MI is normalized (i.e. MI lies between 0-1, where 0 denotes that UPDRS is independent on the dysphonia measure, and 1 indicates that UPDRS is completely determined by the dysphonia measure - see Section 3.2 for details). All speech signals from the female PWP were used to generate these results (N = 1865 phonations). The F_0 subscript text refers to the algorithm used to extract F_0.

Prior to feature selection, we have all the 318 dysphonia measures (i.e. initially, M = 318). Now, we follow the methodology outlined in Chapter 4: we identify statistical associations, select a robust parsimonious feature subset using different FS algorithms, and map the feature subsets to the response variables (motor-UPDRS and total-UPDRS). The statistical associations appear in Table 7.4 for males, and in Table 7.5 for females. Similarly to Table 7.1, we report the most strongly associated dysphonia measure from each algorithmic dysphonia measure family with motor-UPDRS and total-UPDRS. These results provide a general overview of the association strength of algorithmically conceptually related features with UPDRS. We observe that the recently proposed nonlinear dysphonia measures exhibit statistically stronger association with UPDRS compared to the classical dysphonia measures, results which are in broad agreement with the findings in Table 7.1.

Following this initial statistical analysis, we use FS algorithms to determine parsimonious, information-rich feature subsets for males and females. The feature subsets selected using the seven FS algorithms are summarized in Table 7.6 for males and Table 7.7 for females. The number of the features was decided using the "one-standard-error" rule (Hastie et al. 2009): we pick the most parsimonious subset in which the MAE is no more than one standard deviation above the MAE of the best subset of the best performing feature subset. For fair comparison of the FS algorithms, we use the same number of features. We used the standard 10-fold cross-validation approach to evaluate the *generalization performance* of the classifiers (kNN, SVM, RF). Specifically, the initial dataset consisting of N data samples (4010 for males and 1865 for females) was split into a training subset of $0.9 \cdot N$ (3609 for males and 1679 for females) phonations and a testing (*out of sample*) subset of $0.1 \cdot N$ (401 for males and 186 for females) phonations. The process was repeated a total of 100 times, each time randomly permuting the data before splitting into training and testing subsets. Overall, we can estimate motor-UPDRS within approximately 1.5 UPDRS points, and total-UPDRS within approximately 2 UPDRS points from the clinicians' estimates. The RF appear to consistently outperform SVM and kNN in this application.

Table 7.6: Selected dysphonia measures using seven feature selection algorithms and classification performance for motor-UPDRS and total-UPDRS for males.

LASSO	**mRMR**	**mRMR** **Spearman**	**GSO**	**RELIEF**	**LLBFS**	**RRCT**
6th MFCC	VFER-NSR$_{TKEO}$	6th MFCC	6th MFCC	9th level app.coef$_{entropy}$ of the log-F0	9th level app.coef$_{entropy}$ of the log-F0	6th MFCC
8th MFCC	6th MFCC	2nd MFCC	VFER-SNR$_{TKEO}$	DFA	DFA	2nd MFCC
8th delta MFCC	7th MFCC	8th delta MFCC	8th MFCC	6th MFCC	8th MFCC	8th delta MFCC
VFER-SNR$_{TKEO}$	8th MFCC	12th delta MFCC	8th delta MFCC	3rd MFCC	7th MFCC	Std $F_{0,NDF} - F_{0,exp}$
0th MFCC	3rd level detail wav.coef$_{entropy}$	8th MFCC	HNR$_{std}$	5th MFCC	6th MFCC	10th level detail wav.coef$_{entropy}$
2nd MFCC	Log energy	10th level detail wav.coef$_{entropy}$	3rd MFCC	8th MFCC	VFER-NSR$_{TKEO}$	3rd delta MFCC
IMF-SNR$_{TKEO}$	8th delta MFCC	Shimmer$_{TKEO,std}$	2nd MFCC	7th MFCC	5th MFCC	12th delta MFCC
12th delta MFCC	3rd MFCC	3rd delta MFCC	12th delta MFCC	4th MFCC	1st MFCC	3rd MFCC
HNR$_{std}$	1st MFCC	Std $(F_{0,NDF} - F_{0,exp})$	11th MFCC	10th MFCC	3rd MFCC	8th delta MFCC
Shimmer$_{PQ5}$	VFER-SNR$_{TKEO}$	11th delta MFCC	VFER-NSR$_{TKEO}$	11th MFCC	4th MFCC	Std $F_{0,SHRP} - F_{0,exp}$
3rd MFCC	HNR$_{std}$	0th MFCC	RPDE	9th MFCC	9th MFCC	12th MFCC
4th level detail wav.coef$_{entropy}$	5th MFCC	12th MFCC	Shimmer$_{PQ5}$	Log energy	Log energy	11th delta MFCC
10th level detail wav.coef$_{entropy}$	4th MFCC	GNE-NSR$_{TKEO}$	10th level detail wav.coef$_{entropy}$	2nd MFCC	12th MFCC	Log energy
GNE-SNR$_{SEO}$	8th level detail wav.coef$_{entropy}$	Std $F_{0,SHRP}$	2nd delta MFCC	VFER-NSR$_{TKEO}$	10th MFCC	GNE-SNR$_{SEO}$
2nd delta MFCC	11th MFCC	VFER-NSR$_{TKEO}$	9th delta MFCC	12th MFCC	$(F_{0,NDF} - F_{0,exp})_{prc5-95}$	GNE-SNR$_{TKEO}$
2.25 ± 0.19	1.60 ± 0.17	2.32 ± 0.19	2.21 ± 0.20	1.33 ± 0.14	1.31 ± 0.13	2.15 ± 0.18
2.84 ± 0.24	2.02 ± 0.19	2.95 ± 0.28	2.85 ± 0.27	1.59 ± 0.17	1.70 ± 0.18	2.79 ± 0.25

The two entries in the last row for each column denote the motor-UPDRS and total-UPDRS mean absolute error (MAE) results computed using the selected dysphonia measure subsets, and are computed using random forests and 10-fold cross validation with 100 repetitions. They are presented in the form mean ± standard deviation.

Table 7.7: Selected dysphonia measures using seven feature selection algorithms and classification performance for motor-UPDRS and total-UPDRS for females.

LASSO	mRMR	mRMR Spearman	GSO	RELIEF	LLBFS	RRCT
Log energy	Std $F_{0,RAPT}$	Log energy	Log energy	4th MFCC	Log energy	Log energy
Std $F_{0,RAPT}$	Log energy	NHR$_{\text{std}}$	10th MFCC	Log energy	Std $F_{0,RAPT}$	NHR$_{\text{mean}}$
HNR$_{\text{mean}}$	$(F_{0,ensemble} - F_{0,exp})_{\text{mean}}$	PPE	PPE	2nd MFCC	4th MFCC	$(F_{0,ensemble} - F_{0,exp})_{\text{mean}}$
4th level detail wav.coef$_{\text{energy}}$ log-F0	10th MFCC	4th level detail wav.coef$_{\text{TKEO, mean}}$ log-F0	0th MFCC	0th MFCC	11th MFCC	12th MFCC
10th MFCC	4th level detail wav.coef$_{\text{energy}}$	12th MFCC	8th MFCC	1st MFCC	1st MFCC	10th MFCC
PPE	VFER-SNR$_{\text{TKEO}}$	0th MFCC	12th MFCC	5th MFCC	2nd MFCC	4th level app. wav.coef$_{\text{energy}}$
12th MFCC	12th MFCC	3rd level detail wav.coef$_{\text{TKEO, mean}}$ log-F0	HNR$_{\text{std}}$	Std $F_{0,RAPT}$	10th MFCC	3rd delta MFCC
8th MFCC	4th level detail wav.coef$_{\text{TKEO,std}}$ log-F0	6th MFCC	11th MFCC	6th MFCC	PPE	PPE
11th MFCC	1st MFCC	2nd delta MFCC	Jitter$_{\text{F0,TKEO,mean}}$	PPE	0th MFCC	4th MFCC
5th level detail wav.coef$_{\text{energy}}$ log-F0	3rd MFCC	Jitter$_{\text{F0,TKEO,mean}}$	4th MFCC	8th MFCC	3rd MFCC	3rd level detail wav.coef$_{\text{TKEO,std}}$ log-F0
5th level detail wav.coef$_{\text{TKEO, mean}}$ log-F0	0th MFCC	5th level detail wav.coef$_{\text{TKEO, mean}}$ log-F0	3rd delta MFCC	3rd MFCC	4th level detail wav.coef$_{\text{entropy}}$	12th delta MFCC
HNR$_{\text{std}}$	3rd level detail wav.coef$_{\text{entropy}}$	10th MFCC	7th MFCC	7th MFCC	5th MFCC	3rd level detail wav.coef$_{\text{entropy}}$ log-F0
4th level detail wav.coef$_{\text{TKEO,std}}$ log-F0	4th level detail wav.coef$_{\text{entropy}}$ log-F0	4th level detail wav.coef$_{\text{energy}}$ log-F0	10th delta MFCC	11th MFCC	6th level detail wav.coef$_{\text{TKEO,std}}$	5th level detail wav.coef$_{\text{TKEO,std}}$ log-F0
2nd delta MFCC	6th level detail wav.coef$_{\text{TKEO, mean}}$ log-F0	HNR$_{\text{mean}}$	4th level detail wav.coef$_{\text{energy}}$ log-F0	10th MFCC	3rd level detail wav.coef$_{\text{entropy}}$	Shimmer$_{\text{AM}}$
3rd level detail wav.coef$_{\text{entropy}}$ log-F0	11th MFCC	12th delta MFCC	$(F_{0,ensemble} - F_{0,exp})_{\text{mean}}$	3rd level detail wav.coef$_{\text{entropy}}$ log-F0	HNR$_{\text{mean}}$	2nd delta MFCC
2.83 ± 0.29	1.99 ± 0.20	2.42 ± 0.25	2.02 ± 0.23	1.58 ± 0.19	1.72 ± 0.18	2.37 ± 0.22
3.26 ± 0.31	2.42 ± 0.24	2.99 ± 0.30	2.43 ± 0.20	1.95 ± 0.19	2.13 ± 0.20	2.91 ± 0.26

The two entries in the last row for each column denote the motor-UPDRS and total-UPDRS mean absolute error (MAE) results computed using the selected dysphonia measure subsets, and are computed using random forests and 10-fold cross validation with 100 repetitions. They are presented in the form mean ± standard deviation.

The preceding results were obtained using the average out of sample MAE from 100 random selections of phonations from the *entire* dataset, where we try to approximate the clinicians' UPDRS evaluation. Here, we aim to demonstrate that it is possible to estimate the UPDRS progression of *specific* individuals for duration of the AHTD trial, that is, *UPDRS tracking* (weekly UPDRS estimation of an individual for the six month duration of the trial using the speech recordings). The simplest UPDRS tracking scheme would be to train the classifier using the entire dataset with the exception of the dysphonia measures from the specific subject under investigation. However, this is a very unstable scheme due to the low number of subjects in the AHTD trial (see Chapter 8 for a detailed discussion). For that reason, we have introduced a proxy UPDRS tracking approach (Tsanas et al., 2011a).

The PWP in the AHTD trial under investigation completed 20-30 weekly tests during the 6-month period. As part of the data acquisition process, six sustained vowel phonations were recorded on each of those days, and therefore we have approximately 150 data samples from each subject. Now, instead of leaving all the data samples from a single subject to test the performance of the developed methodology on individual subjects, we proposed leaving out of the training process the data samples derived from one of the weekly tests, and evaluating the out of sample performance of the classifier on those samples. That is, we train the algorithm using all the data samples with the exception of those data samples derived from the first of each of the weekly phonations (about 20-25) for the individual under investigation, which are used for testing. The methodology is successively repeated leaving out the second, third, fourth, fifth, and sixth of the weekly phonations of the individual under investigation. Finally, the six weekly out-of-sample UPDRS estimates are averaged, resulting in a single UPDRS estimate. We have found that this setting where the average UPDRS estimates from the dysphonia measures of the six weekly phonations are averaged is a more robust method compared to randomly selecting a UPDRS estimate from one of the six weekly phonations.

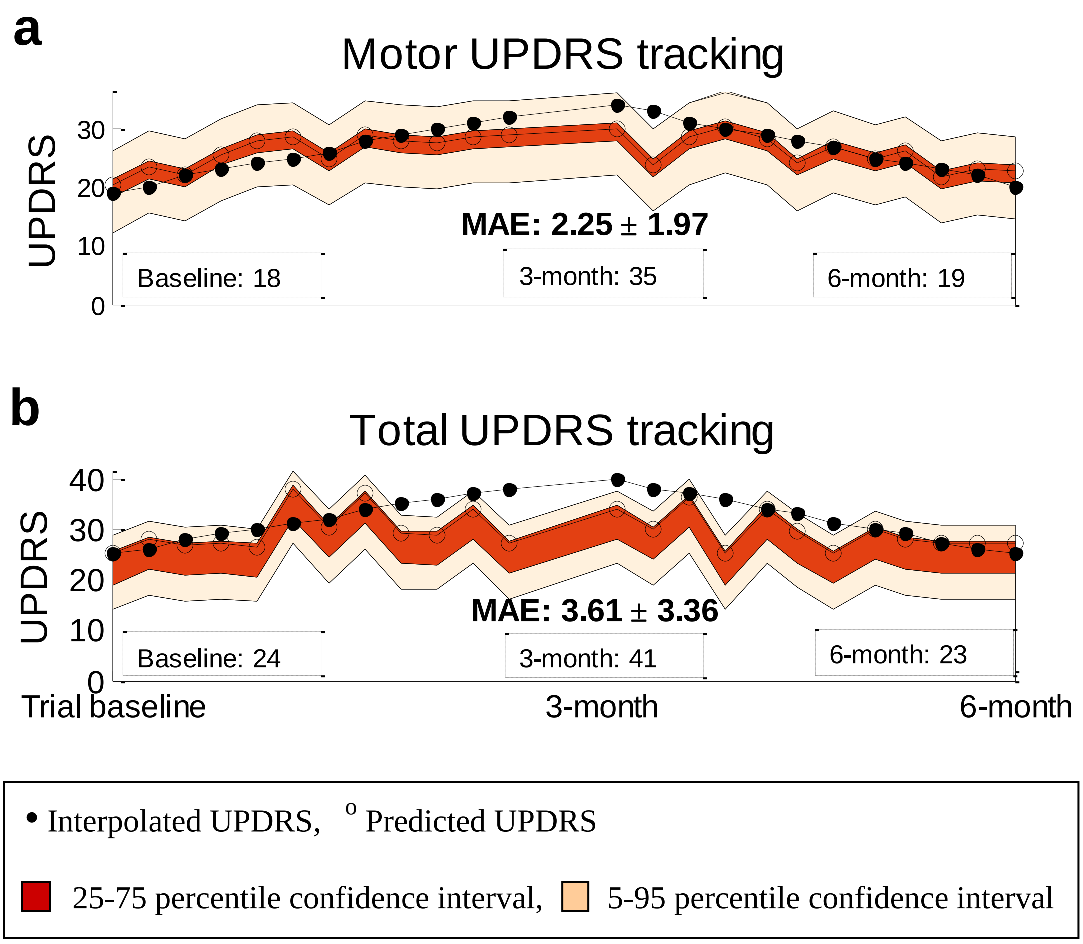

Fig. 7.3 Motor-UPDRS and total-UPDRS tracking over the six month period of the AHTD trial for the subject with the largest and most uncharacteristic UPDRS progression (sharp UPDRS increase three months into the trial, and sharp UPDRS decrease six months into the trial). The computation of the out-of-sample MAE and the confidence intervals were estimated from the average MAE of the six weekly error estimates throughout the six month duration of the trial for the specific individual. Slight deviations from a straight interpolation line are observed because of the subsequent rounding of the interpolated UPDRS values.

Figure 7.3 presents the UPDRS tracking of the subject with the most uncharacteristic UPDRS progression in the AHTD trial (sharp UPDRS increase three months into the trial, and subsequent sharp UPDRS decline six months into the trial) using RF and the selected feature subset using RELIEF in Table 7.7 (this subject is a female). The choice of a subject with a non-typical UPDRS pattern (PD is a progressive disorder which is reflected in typically increasing UPDRS scores, although it is possible to have UPDRS decline in the short term) serves to illustrate that the proposed methodology is applicable and yields satisfactory results even in such scenarios. We remark that in most PWP UPDRS increases monotonically, and the estimated UPDRS tracking is even more precise than the results in Fig. 7.3.

7.4 Validating the results using statistical hypothesis tests and surrogate hypothesis tests

In this section some formal statistical tests are used to reinforce the validity of the results reported in the previous section. So far, we focused on quantifying the ability of the machine learning approach to replicate the clinicians' estimates in order to obtain the response variable (UPDRS). It is often useful in practical applications to have a point of reference (*benchmark*) against which to compare the results of the machine learning algorithm, in order to demonstrate whether (and the extent) the proposed method outperforms a typically *naïve* approach. The benchmark chosen depends on the fine points of the application, and is typically the mean of the response variable. In this study, we used the *normalized MAE* (Eq. 7.1) as a performance metric to illustrate that the findings reported in the previous section outperform the mean response variable benchmark.

$$MAE_{normalized} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}|} \tag{7.1}$$

where $\bar{y}$ is the mean of the response variable (UPDRS) and N is the number of samples used in each examined subset (four sets: motor UPDRS total UPDRS for males and females). If

$MAE_{normalized} = 1$, the model predicts the mean of the response; the lower that value is, the better the quality of the prediction. In all four cases we have found that $MAE_{normalized} < 0.25$, which indicates that the dysphonia measures can outperform a naïve benchmark such as the mean of the response variable.

We test whether these findings are statistically significant using two formal statistical tests. First, we use the two-sample Kolmogorov-Smirnov test to check whether we get significant differences between the distributions of the MAE using the dysphonia measures and the MAE using the mean of the response variable. The *null hypothesis* is that the MAE using the dysphonia measures and the MAE using the mean of the response variable to predict the subjects' UPDRS are from the *same continuous distribution*. The alternative hypothesis is that they are from different continuous distributions. We check for significance at the $p = 0.01$ level. The Kolmogorov – Smirnov test rejects the null hypothesis that the distributions of the MAE using the dysphonia measures and the MAE derived using the mean as forecast are from the same distribution. We also use the Wilcoxon (Mann-Whitney) test to perform a two-sided rank sum test. The *null hypothesis* is that the MAE with the dysphonia measures and the MAE using the mean of the response variable to obtain response estimates are independent samples from identical continuous distributions with *equal medians*. We check for significance at the $p = 0.01$ level. The Wilcoxon test rejects the null hypothesis that the samples of the MAE with the dysphonia measures and the MAE using the mean of the response variable stem from distributions with equal medians.

As a last check, we plot the empirical cumulative distribution functions of the MAE, to directly compare the performance of the dysphonia measures and the mean of the response variable (the benchmark). The red and green lines are the upper and lower confidence bounds. The aim is to demonstrate that the MAE curve from the dysphonia measures is on the left (i.e. lower) compared to the MAE derived from the benchmark. The MAE are as before in the out-of-sample case with the 100-run 10-fold cross validation. Fig. 7.4 clearly illustrates that the dysphonia measures provide a very competitive MAE compared to issuing naively the mean as a future UPDRS estimate in the 100 run 10-fold cross validation setting.

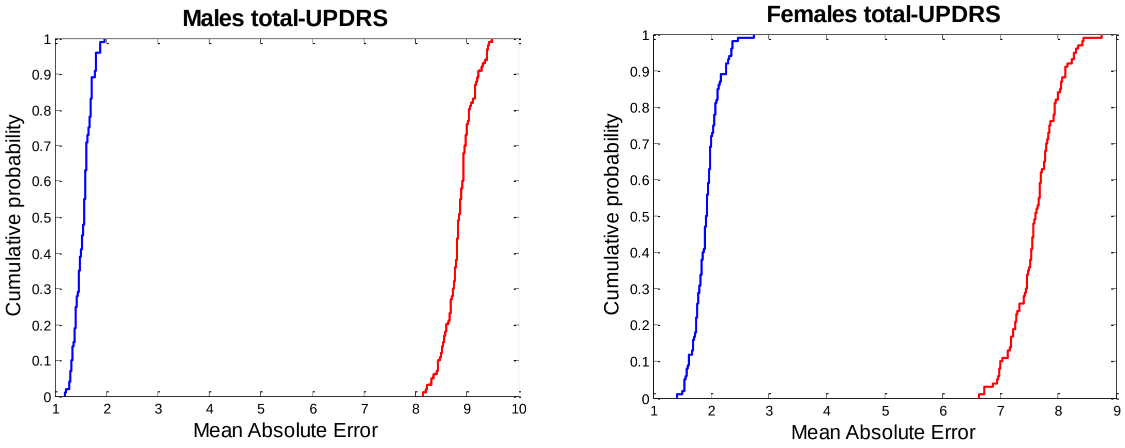

Fig. 7.4: Empirical cumulative distribution functions of the model (denoted by blue) versus a naïve benchmark (denoted by red) for estimating the total-UPDRS. Similar results are observed for estimating the motor-UPDRS.

In addition to the statistical hypothesis tests described so far, we use a simple surrogate analysis test (see § 4.4.3) to illustrate that the dysphonia measures carry useful information towards predicting UPDRS. Specifically, we randomly permute the data samples: if the dysphonia measures convey information towards predicting UPDRS, then the randomly shuffled dataset should lead to considerably worse results. The null hypothesis is that the dysphonia measures do not carry information towards predicting UPDRS. For statistical confidence, we have repeated the process a total of 10 times, each time randomly permuting the features in the dataset and associating it with the (non-permuted) response (UPDRS). In all cases $MAE > 7$ UPDRS points, which leads to the rejection of the null hypothesis that the dysphonia measures do not carry information to predict UPDRS.

Chapter 8

Conclusions and future work

This study investigated the potential of using speech signal analysis to (a) differentiate people with Parkinson’s disease (PD) from healthy controls, and (b) replicate PD symptom severity as defined by the standard reference clinical metric Unified Parkinson’s Disease Rating Scale (UPDRS). Towards this aim, many novel speech signal processing algorithms were developed, which extract clinically useful information in this context (Chapter 3). Moreover, we tackled the complex problem of identifying the ‘true’ fundamental frequency (F_0) in speech signals (Chapters 3 and 5). We systematically studied ten F_0 estimation algorithms and proposed a new ensemble approach that appears to be particularly promising, outperforming by more than 10% the single best individual F_0 estimator. In biomedical datasets such as those used in this study (Chapter 7), researchers often collect a large number of (possibly highly correlated) features. For this reason, we studied feature selection (FS) algorithms, in order to tentatively understand the most important properties of the data when inferred from the underlying physiological meaning of the selected features. We have extended some FS algorithms, and introduced a new algorithm which we call Relevance, Redundancy and Complementarity Trade-off (RRCT) that is very competitive with widely used contemporary schemes. Finally, we compared the performance of two powerful state of the art classifiers, Support Vector Machines (SVM) and ensembles of decision trees (random forests – RF), in a variety of general settings, both in terms of applications, and also in terms of dataset complexity.

The extent of speech disorders was quantified using a wide range of speech signal processing algorithms known as *dysphonia measures*. We demonstrated that we can differentiate PD subjects and healthy controls with almost 99% accuracy, and also that motor-UPDRS can be estimated within approximately 1.6 UPDRS points (out of 108) and total-UPDRS within 2 UPDRS points (out of 176). These UPDRS predictions are lower than the *inter-rater variability*, which is about 4-5 UPDRS points (Post et al., 2005). These results were obtained using the 100 runs 10-fold cross-validation scheme, and reflect our best estimate of the *asymptotic* out-of-sample prediction error as argued in Tsanas et al. (2010a; 2011a; 2012b). Since there are samples from the same subject used both in the training and the testing subsets, one could argue that this might affect the reliability of the cross-validation generalization error estimate (cross-validation implicitly assumes statistical independence between samples). This fact would indicate that a different validation scheme might be more appropriate, such as leaving one individual out (training the system with the phonations from $L-1$ subjects, and testing it on the phonations of the remaining subject, for all subjects). However, this scheme contradicts an important general consideration of model validation: there must be sufficient hold-out data to form a reliable estimate of the *asymptotic* out-of-sample prediction error. With the current number of subjects, any subject-specific cross-validation is not really reliable: there is not enough hold-out data and in our own experimental computations the variance of the mean absolute errors was too large. Therefore, simple leave one individual out is too unstable to form a reliable estimate of the asymptotic out-of-sample prediction error (Tsanas et al., 2010a; Tsanas et al., 2011a, Tsanas et al., 2012b).

This study also demonstrated the feasibility of tracking UPDRS changes in time with clinically useful accuracy (see § 7.3 and in particular Fig. 7.3). From a practical perspective, the satisfactory reception of the patients themselves towards the At-Home testing device (AHTD) and speech tests (Goetz et al., 2009) suggests this field may be promising for further experimentation. The 42 people with Parkinson's (PWP) in the AHTD trial were diagnosed with PD within the previous five years at trial onset. As a result, the range of UPDRS investigated here did not span the full range (max motor-UPDRS 41, max total-UPDRS 55); hence more extensive studies are required to test the generalization of the current findings outside the examined range. We believe that the promising performance of the developed methodology to accurately replicate UPDRS in PWP who do not exhibit profound symptoms (as a result of being recently diagnosed) may be indicative of the feasibility of successful UPDRS tracking in more severely affected patients.

This study looked into various FS algorithms, to address the curse of dimensionality. One attractive aspect of FS is the insight offered into the most important aspects into the examined problem, tentatively inferred from the selected features. FS algorithms can be generally categorized in terms of the compromise amongst three main terms: *relevance*, *redundancy*, and *complementarity*. In Chapter 5 we have verified a well-established truth: that there is no universally best FS algorithm. We reported that the new FS algorithm proposed in this study, RRCT, worked well in many datasets and was shown to be fairly robust. In Chapter 7 we found RELIEF clearly outperformed the competing FS algorithms both in discriminating PWP from healthy controls (§ 7.1), and also in estimating UPDRS using the dysphonia measures (§ 7.3). It is worthwhile reflecting on this finding, particularly since RELIEF is inherently an algorithm which does not account for redundancy (there are abundant studies in the research literature that demonstrate this aspect is critical in selecting a parsimonious information-rich feature subset, e.g. Peng et al. (2005), and Guyon et al. (2006)). We attribute the success of RELIEF in the datasets analysed in § 7.1 and § 7.3 to two factors: (a) it can identify features which are highly nonlinearly related with the response, and (b) the information from dysphonia measures is often complementary. Hence, FS algorithms which rely on simple metrics such as LASSO, RRCT and GSO, may fail to determine the most predictive features. Similarly, feature complementarity may be crucial in this domain where the combination of features may indicate degree of voice pathology, and therefore FS algorithms such as mRMR fail to account for this critical aspect of the dataset.

In Chapter 7 the FS algorithms have consistently selected the non-classical dysphonia measures over the classical schemes such as jitter and shimmer. This is compelling evidence that these new measures quantify clinically useful information in PD voices which may not be captured by the classical dysphonia measures. The results in § 7.3 suggest that gender differentiation in PD is useful, supporting recent findings in speech signal analysis (Cnockaert et al., 2008; Fraile et al., 2009), and possibly general PD symptom severity (Tsanas et al., 2012c). The MFCCs and some of the novel dysphonia measures appear to convey substantially useful information in both genders (Tables 7.6 and 7.7). The success of these dysphonia measures may be at least partly attributed to the fact they circumvent F_0 estimation (Godino-Llorente et al., 2006; Fraile et al., 2009; Tsanas et al., 2011a). Furthermore, these findings accentuate the imperious need to target the vocal tract to obtain clinically useful information in addition to the vocal folds, which hitherto have been the main focus in speech PD research. Interestingly, the MFCCs and the VFER family of dysphonia measures dominate in the features selected in Table 7.6 (males), whereas in Table 7.7 (females) dysphonia measures which focus on the fundamental frequency appear to be most predictive of UPDRS. This finding may suggest there is a distinct voice pattern in female PWP which may be masked in male PWP due to the physiology of natural male voice production (Tsanas et al., 2011a). Since higher fundamental frequencies are reported to be more stable (i.e. have lower perturbations) (Baken and Orlikoff, 2000), and women have higher F_0 in general (Titze, 2000), we argue that slight distortions in vocal performance (e.g. irregular vocal fold vibration pattern) could likely reflect voice pathology in females, whereas similar distortions in males' vocal performance could be attributed (at least partly) to normal vibrato. Thus, voice degradation quantified using some of the dysphonia measures which inherently rely on F_0 (e.g. the F0-related measures, jitter, and PPE) may represent general symptom degradation in females, whereas similar quantification of the voice perturbations in males could be part of the physiological variability in normal male voice production mechanisms.

Speech appears explicitly in two UPDRS sections: once in the activities of daily living component (part II, section 5) and once in the motor component (part III, section 18) – for details of all UPDRS sections see Appendix III. Whereas the link between speech and general motor function may be intuitively easy to grasp, this study has provided compelling evidence to suggest that speech can help quantify not only motor symptoms (as part of the motor component in UPDRS), but generalized diverse symptoms in PD.

With the exception of our previous studies on replicating a PD symptom severity metric (Tsanas et al., 2010a; 2010b; 2011a) we are not aware of any related studies in the research literature that have focused on this topic. In a recent study, Patel et al. (2009) have studied segmented aspects of the UPDRS metric (tremor, bradykinesia, and dyskinesia), using accelerometers. Unsurprisingly, accelerometers can replicate the clinicians' evaluation of PD symptom severity in tremor, bradykinesia, and dyskinesia more accurately than speech signals. Nevertheless, notwithstanding the importance of those three UPDRS elements, they do not encompass the breadth of PD symptoms which is reflected in the comprehensive UPDRS metric that offers a concise quantification of PD symptom severity.

We believe the current findings provide compelling evidence for further research on telemonitoring PD using speech signals. We envisage this technology finding use in future clinical trials, offering clinical staff the prospect of frequent, remote, and accurate UPDRS monitoring, particularly in those cases where PWP are reluctant or unable to make frequent physical visits to the clinic. Finally, this technology could be invaluable in future clinical trials of novel PD treatments requiring very large study populations who can be monitored frequently, and cost-effectively.

This study has touched upon many topics and has raised a number of important questions. Some major points are highlighted below which may be the starting points of further research:

- Development of new dysphonia measures, which capture speech signal characteristics that convey clinically useful information both in this application (speech and PD) and other related applications. These dysphonia measures could be extensions of the already developed algorithms, application of known algorithms from other disciplines, or completely new concepts. On the basis of our findings, we tentatively suggest that signal to noise measures appear very promising. Also, dysphonia measures targeting the vocal tract have not been particularly popular amongst researchers working in clinical applications of speech; the results in some recent studies and this study suggest that there is additional clinically useful information which can be extracted from resonances in the 'filter' part of the source-filter vocal production mechanism.
- It would be interesting to combine information using various primary signals in addition to speech in this application. It is highly likely that combining information from speech and information extracted from dexterity and pegboard tests, which are also collected with the AHTD, might further reduce the UPDRS prediction error and enhance the clinical value of such multimodal testing in telemedicine applications.
- Although the dysphonia measures have tentative physiological interpretations, it is difficult to link them with the underlying physiology. Establishing a physiologically-based model which would explain the data-driven findings in this study in terms of the relevant physiological changes that occur in PD would be particularly useful. In this direction, Gomez-Vilda et al. (2007) have evaluated voice pathologies by interpreting the

vocal fold biomechanical parameters of a lumped mechanical equivalent model. Similarly, researchers have developed plausible basal ganglia models explaining PD behavior (Gurney et al., 2001a; Gurney et al., 2001b; Humphries et al., 2006). Combining neuronal input models to speech production mechanisms, e.g. through simple electrical circuit equivalents (Tsanas et al., 2009), could help in the understanding of the physiological mechanisms and pose new research questions.

- Speech is particularly suitable for telemonitoring applications, because the required equipment to capture these signals is readily available to the vast majority of people in the form of mobile phones. The potential of using the standard cellular mobile network for PD telemonitoring was recently featured in Tsanas et al. (2012e). We demonstrated that UPDRS tracking is adequately accurate and provides clinically useful information, whilst being an attractive option in those cases where access to expensive equipment such as the AHTD is not possible. That study focused on a simulation environment; we intend to apply the proposed methodology in a practical setting studying various realistic settings with different mobile phones, and under different recording conditions.
- The UPDRS assessment is subjective and the clinical raters' scoring often varies (Post et al., 2005), that is there is no truly *objective* definition of PD symptom severity. Moreover, in practice, UPDRS is only recorded every 3-6 months. This study builds on the assumption that the UPDRS scores used in the AHTD trial were accurate, and offers a machine learning approach to replicate those evaluations. Nonetheless, the most relevant aspect of any disease progression and treatment is *the patient's perception* of symptoms, that is, the patient's own self-rating. It would be interesting to have a metric similar to UPDRS which is based on the patients' assessments. One interesting application would be to use smart phones and record a large number of signals from PWP *on a daily basis* which we could somehow project to a new universal clinical PD severity metric.

- The machine learning research community has progressed tremendously in the last 10-15 years, offering valuable tools such as ensemble learning methods (Kuncheva, 2004; Polikar, 2006; Bishop, 2007; Hastie et al., 2009). In this study, we proposed an ensemble F_0 estimation algorithm which appears to consistently outperform the best single F_0 estimation approach. The applicability of ensembles, which was originally proposed to fuse learners, can be extended to other domains in principle, such as FS. For example, it would be interesting to use a (possibly weighted) voting scheme from different FS algorithms to determine the most predictive feature subset. The challenge in this case is to determine which FS algorithms are most applicable in particular settings (for example in large scale problems, in tall and fat datasets, and in datasets with varying degrees of feature correlations).
- The F_0 estimation algorithms in this study were only validated for the case of the sustained vowel /a/, because the physiological model used can only generate this type of signals. Future work could look into extending these results to other sounds, if accurate ground truth F_0 values can be obtained (either from a physiological model, or from very accurate detection of the glottal cycles by direct measurements while a subject speaks).

The topic of statistical machine learning in general, and in the context of this study in particular, leave abundant scope for creativity with new notions and ideas constantly emerging. Having spent many long hours studying this challenging and fascinating topic, I have only one thing to regret: I did not spend longer. During the following years I hope that I will be able to contribute new ideas and concepts both in biomedical applications, and also in more wide impact, wide applicability statistical machine learning algorithms.

References

M.C.P. Apps, P.C. Sheaff, D.A. Ingram, C. Kennard, D.W. Empey, “Respiration and sleep in Parkinson's disease,” *Journal of Neurology, Neurosurgery, and Psychiatry*, Vol. 48, pp. 1240-1245, 1985

D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, L. Pereira-Leite, “SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms”, *Journal of Maternal Fetal Medicine*, Vol. 5, pp. 311-318, 2000

T. Backstrom, P. Alku, and E. Vilkman, “Time domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range,” *IEEE Transactions on Speech and Audio Processing*, Vol. 10, pp. 186-192, 2002

R.J. Baken and R.F. Orlikoff, *Clinical measurement of speech and voice*, San Diego: Singular Thomson Learning, 2nd ed., 2000

M. Baldereschi, A. De Carlo, W.A. Rocca et al., “Parkinson’s disease and parkinsonism in a longitudinal study. Two fold higher incidence in men,” *Neurology*, Vol. 55, pp. 1358-1363, 2000

R.E. Bellman, *Adaptive control processes: a guided tour*, Princeton University Press, 1961.

A.L. Benabid, S. Chabardes, J. Mitrofanis and P. Pollak, “Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson’s disease,” *Lancet Neurology*, Vol. 8, pp. 67-81, 2009

H. Bernheimer, W. Birkmeyer, O. Hornykiewicz, K. Jellinger and F. Seitenberger, “Brain dopamine and the syndromes of Parkinson and Huntington,” *Journal of Neurological Sciences*, Vol. 20, pp. 425-255, 1973

W.R. Berry, *Clinical dysarthria*, San Diego, College-Hill, 1983

C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2007

P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *IFA Proceedings* 17, pp. 97-110, 1993

P. Boersma, “Should jitter be measures by peak picking of by waveform matching?,” *Folia Phoniatrica et Logopaedica*, Vol. 61, pp. 305-308, 2009

J.H. Bower, D.M. Maraganore, S.K. McDonnell and W.A. Rocca, “Incidence and distribution of parkinsonism in Olmsted County, Minnesota,” 1976-1990, *Neurology*, Vol. 52, pp. 1214-1220, 1999

L. Breiman, “Statistical modelling: the two cultures,” *Statistical Science*, Vol. 16, No. 3, pp. 199-231 (with comments and discussion), 2001a

L. Breiman, “Random forests,” *Machine learning*, Vol. 45, pp. 5-32, 2001b

D.J. Brooks, “Assessment of Parkinson’s disease with imaging,” *Parkinsonism and Related Disorders*, Vol. 13, pp. 268-275, 2007

G. Brown, A. Pocock, M. Zhao, M. Lujan, “Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection,” *Journal of Machine Learning Research*, Vol. 13, pp. 27-66, 2012

C.J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, Vol. 2, pp. 121–167, 1998

A. Camacho, J.G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” *Journal of the Acoustical Society of America*, Vol. 124, pp. 1638-1652, 2008

S. von Campenhausen, B. Bornschein, R. Wick, K. Bötzel, C. Sampaio, W. Poewe, W. Oertel, U. Siebert, K. Berger and R. Dodel, “Prevalence and incidence of Parkinson's disease in Europe,” *European Neuropsychopharmacology*, Vol. 15, No. 4, pp. 473-490, 2005

E.J. Candes, M.B. Wakin and S.P. Boyd, “Enhancing sparsity by reweighted L_1 minimization,” *The Journal of Fourier analysis and applications*, Vol. 14, pp. 877-905, 2008

P.L.S. Chan and N.H.G. Holford, “Drug treatment effects on disease progression,” *Annual Review of Pharmacology and Toxicology*, Vol. 41, pp. 625-659, 2001

C-C. Chang and C-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27:1--27:27, 2011

Z. Chen, P.C. Ivanov, K. Hu, and H. E. Stanley, “Effect of nonstationarities on detrended fluctuation analysis,” *Physical Review E*, Vol. 65(4), 041107, 2002

L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebve and F. Grenez, “Low frequency vocal modulations in vowels produced by Parkinsonian subjects,” *Speech Communication*, Vol. 50, pp. 288-300, 2008

J. Cohen, P. Cohen, S.G. West, L.S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Routledge Academic, 3rd ed., 2002

R.H. Colton, E.G. Conture, “Problems and pitfalls of electroglottography,” *Journal of Voice*, Vol. 4 (1), pp. 10-24, 1990

G.C. Cotzias, “L-dopa for Parkinsonism,” *New England Journal of Medicine*, Vol. 278, pp. 630, 1968

T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, Vol. IT-13, pp. 21-27, 1967

T.M. Cover and J.A. Thomas, *Elements of information theory*, Wiley-interscience, 2nd edition, 2006

M.G. Christensen and A. Jakobsson, *Multi-pitch estimation*, Synthesis lectures on speech and audio processing, Morgan & Claypool Publishers, 2009

A. de Cheveigne and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, Vol. 111 (4), pp. 1917-1930, 2002

C. Cortes, V. Vapnik, “Support vector networks,” *Machine Learning*, Vol. 20, pp. 273–297, 1995

S.B. Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustic and Speech Signal Processing*, Vol. 28, No. 4, pp. 357-366, 1980

F.L. Darley, A. E. Aronson and J. R. Brown, “Differential diagnostic patterns of dysarthria,” *Journal of Speech and Hearing Research*, Vol. 12, pp. 246-269, 1969

F.L. Darley, A. E. Aronson and J. R. Brown, “Clusters of deviant speech dimensions in the dysarthrias,” *Journal of Speech and Hearing Research*, Vol. 12, pp. 462-469, 1969

F.L. Darley, A. E. Aronson and J. R. Brown, *Motor Speech Disorders*, Philadelphia: WB Saunders, 1975

R. Das, “Classification of Parkinson's disease by using voice measurements,” *Expert systems with applications*, Vol. 37, pp. 1568–1572, 2010

M.C. de Rijk, W.A. Rocca, D.W. Anderson, M.O. Melcon, M.M.B. Breteler and D.M. Maraganore, “A population perspective on diagnostic criteria for Parkinson’s disease,” *Neurology*, Vol. 48, pp. 1277-1281, 1997

M.C. de Rijk, L.J. Launer, K. Berger, M.M. Breteler, J.F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder and A. Hofman, “Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts,” *Neurology*, Vol. 54, pp. 21–23, 2000

P. Diaconis and B. Efron, “Computer intensive methods in statics”, *Scientific American*, Vol. 248, pp.116-131, 1983

D. Donoho, “For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution,” *Communications on Pure and Applied Mathematics*, Vol. 59 (7), 904-934, 2006

B. Duchenne, De l'électrisation localisée et de son application à la physiologie, à la pathologie et à la thérapeutique, published by Chez J.-B. Baillière, Paris, 1855 (available online on Google books, accessed on 12 Nov. 08)

J.R. Duffy, *Motor Speech Disorders: substrates, differential diagnosis and management,* New York: Mosby, 2nd ed., 2005

A. Elbaz, J.H. Bower, D.M. Maraganore, et al., “Risk tables for parkinsonism and Parkinson’s disease,” *Journal of Clinical Epidemiology*, Vol. 55, pp. 25-31, 2002

S. Fahn, R. Elton, Members of the UPDRS Development Committee, in: S. Fahn, C.D. Marsden, D.B. Calne, M. Goldstein, eds. *Recent developments in Parkinson's disease*, Vol. 2, Florham Park, NJ, Macmillan Health Care Information pp. 153-163, 1987

G. Fant, *Accoustic theory of speech production*, The Hague: Mouton, 1960

J.M. Fearnley and A.J. Lees, “Ageing and Parkinson’s disease: substantia nigra regional selectivity,” *Brain*, Vol. 114, pp. 2283-2301, 1991

C.A. Ferrer, E. Gonzalez, M.E. Hernandez-Diaz, *Evaluation of time and frequenc domain-based methods for the estimation of harmonics-to-noise-ratios in voice signals*, In Progress in pattern recognition, image analysis and applications, J.F. Martinzed-Trinidad, J.A. C. Ochoa, J. Kittler (Eds.), LNCS 4225, pp. 406-415, 2006

J. Flanagan, *Speech analysis, synthesis and perception*, New York: Springer-Verlag, 1972

C. Fox and L. Ramig, “Vocal sound pressure level and self-perception of speech and voice in men and women with idiopathic Parkinson disease,” *American Journal of Speech and Language Pathology*, Vol. 2, pp. 29-42, 1997

R. Fraile, N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, C. Fredouille, “Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex,” *Folia Phoniatrica et Logopaedica*, Vol. 61, pp. 146-152, 2009

Y. Freund and R.E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, Vol. 55 (1), pp. 119–139, 1997

J. Gamboa, F.J. Jimenez-Jimenez, A. Nieto, J. Montojo, M. Orti-Pareja, J.A. Molina, E. Garcia-Albea, I. Cobeta, “Acoustic voice analysis in patients with Parkinson’s disease treated with dopaminergic drugs,” *Journal of Voice*, Vol. 11, pp. 314-320, 1997

D. Gerhard, *Pitch extraction and fundamental frequency: history and current techniques*, Technical report, TR-CS 2003-06, Department of Computer Science, University of Regina, Canada, 2003

R. Gilad-Bachrach, A. Navot, and N. Tishby, “Margin based feature selection - theory and algorithms,” *21st International Conference on Machine learning* (ICML), pp. 43-50, 2004

A. Giovanni, M. Ouaknine, J.L. Triglia, “Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis,” *Journal of Voice*, Vol. 13 (3), pp. 341-354, 1999

B.R. Glasberg, B.C.J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, Vol. 47, pp. 103-108, 1990

J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, “Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters”, *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 1943-1953, 2006

J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gomez-Vilda, M. Blanco-Velasco, F. Cruz-Roldán, “The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders,” *Journal of Voice*, Vol. 24, pp. 47-56, 2010

C.G. Goetz et al., “Movement Disorder Society Task Force on Rating Scales for Parkinson’s disease, UPDRS: status and recommendations,” *Movement Disorders*, Vol. 18, pp. 738-750, 2003

C.G. Goetz et al., “Movement Disorder Society-sponsored revision of the Unified Parkinson’s disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results,” *Movement Disorders*, Vol. 23, No 15, pp. 2129-2170, 2008

C.G. Goetz, G.T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A.D. Wu, P.H. Kraus, L.M. Blasucci, E.A. Shamim, K.D. Sethi, J. Spielman, K. Kubota, A.S. Grove, E. Dishman, C.B. Taylor, "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, Vol. 24 (4), pp. 551-556, 2009

P. Gomez-Vilda, R. Fernandez-Baillo, A. Nieto, F. Diaz, F.J. Fernandez-Camacho, V. Rodellar, A. Alvarez, R. Martinez, "Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters," *Journal of Voice*, Vol. 21 (4), pp. 450-476, 2007

R.P. Gorman, T.J. Sejnowski, "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets," *Neural Networks*, Vol. 1, pp. 75-89, 1988

A. Gray, A. Moore, "Rapid evaluation of multiple density models", *Artificial Intelligence in Statistics* (AISTATS), 2003

W. Guan, M. Zhou, C.Y. Hampton, B.B. Benigno, L.D. Walker, A. Gray, J.F. McDonald, F.M. Fernández, "Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines," *BMC Bioinformatics*, 10:259, 2009

P-F. Guo, P. Bhattacharya and N. Kharma, *Advances in detecting Parkinson's disease*, Medical Biometrics, Lecture Notes in Computer Science, Vol. 6165/2010, pp. 306-314, 2010

K. Gurney, P. Redgrave & T.J. Prescott, "A computational model of action selection in the basal ganglia I. A new functional anatomy," *Biological Cybernetics*, Vol. 84, pp. 401-410, 2001a

K. Gurney, P. Redgrave & T.J. Prescott, "A computational model of action selection in the basal ganglia II. Analysis and simulation of behaviour," *Biological Cybernetics*, Vol. 84, pp. 411-423, 2001b

I. Guyon and A. Eliseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182, 2003

I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications*, Springer, 2006

I. Guyon, J. Li, T. Mader, P. A. Pletscher, G. Schneider, M. Uhr, "Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark," *Pattern Recognition Letters*, 28:1438-1444, 2007

I. Guyon, *Practical feature selection: from correlation to causality*. In Mining Massive Data Sets for Security. IOS Press, 2008. Available online at http://eprints.pascal-network.org/archive/00004038/01/PracticalFS.pdf

I. Guyon, A. Saffari, G. Dror, G. Cawley, "Model Selection: Beyond the Bayesian/Frequentist Divide," *Journal of Machine Learning Research*, Vol. 11 pp. 61-87, 2010

A.C. Guyton and J.E. Hall, *Textbook of Medical Physiology*, 11th edition, Elsevier Saunders, 2006

C.A. Haaxma, B.R. Bloem, G.F. Borm, et al., "Gender differences in Parkinson's disease, Journal of Neurology," *Neurosurgery and Psychiatry*, Vol. 78, pp. 819-824, 2007

D. Hand, "Classifier technology and the illusion of progress," *Statistical Science*, Vol. 21, pp. 1-15, 2006

D. Hanson, B. Gerratt and P. Ward, "Cinegraphic observations of laryngeal function in Parkinson's disease," *Laryngoscope*, Vol 94, pp. 348-353, 1984

L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis: A survey," *Folia Phoniatr. Logop*. Vol. 46, pp. 9-17, 1994

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*, vol. 56, pp. 24–29, 2004

T. Hastie, and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, pp. 607-616, 1996

T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2nd ed., 2009

N. Henrich, C. d'Alessandro, B. Doval, M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *Journal of Acoustical Society of America*, Vol. 115 (3), pp. 1321-1332, 2004

W.J. Hess, Pitch and Voicing Determination, in *Advances in Speech Signal Processing*, S. F. a. M.M. Sondhi, Ed. New York, NY: Marcel Dekker, Inc., pp. 3-48, 1991

M. Hilario and A. Kalousis, "Approaches to Dimensionality Reduction in Proteomic Biomarker Studies", *Briefings in Bioinformatics*, Vol. 9, pp. 102-118, 2008

A. Ho, R. Iansek, C. Marigliani, J. Bradshaw, S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioral Neurology,* Vol. 11, pp. 131-37, 1998

A.K. Ho, R. Iansek and J. L. Bradshaw, "Motor instability in parkinsonian speech intensity," *Neuropsychiatry Neuropsychology and Behavioral Neurology*, Vol. 14, pp. 109-116, 2001

A. Ho, J. L. Bradshaw and R. Iansek, "For better or for worse: the effect of Levodopa on Speech in Parkinson's disease," *Movement Disorders*, Vol. 23, No. 4, pp. 574-580, 2008

M.M. Hoehn, M.D. Yahr, "Parkinsonism: onset, progression, and mortality," *Neurology*, Vol. 17, pp. 427–42, 1967

C.C. Holmes and N.M. Adams, "A probabilistic nearest neighbour method for statistical pattern recognition," *Journal Royal Statistical Society Ser. B*, Vol. 64 (2), pp. 295–306, 2002

R.J. Homes, J.M. Oates, D.J. Phyland, A.J. Hughes, "Voice characteristics in the progression of Parkinson's disease", *International Journal of Language and Communication Disorders*, vol. 35, pp. 407-418, 2000

F.C. Hoppensteadt and C.S. Peskin, *Modelling and Simulation in Medicine and the Life Sciences*, Springer, 2nd ed., 2002

S. Howison, *Practical applied mathematics: modelling, analysis, approximation*, Cambridge University Press, New York, 2005

C-W. Hsu, C-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on neural networks,* Vol. 13, No. 2, pp. 415-425, March 2002

C-W. Hsu, C-C. Chang, C-J. Lin, *A practical guide to support vector classification*, Technical report, National Taiwan University, 2010

A. J. Hughes, S. E. Daniel, S. Blankson and A. J. Lees, "A clinicopathologic study of 100 cases of Parkinson's disease", *Archives of Neurology*, vol. 50, pp. 140–148, 1993

M.D. Humphries, R.D. Stewart, K. Gurney, "A physiologically plausible model of action selection and oscillatory activity in the basal ganglia," *Journal of Neuroscience*, Vol. 26 (50), pp. 12921-42, 2006

E.J. Hunter, "A comparison of a child's fundamental frequencies in structured elicited vocalizations versus unstructured natural vocalizations: A case study," *International Journal of Pediatric Otorhinolaryngology*, Vol. 73 (4), pp. 561-571, 2009

J. Kaiser, “On a simple algorithm to calculate the 'energy' of a signal,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, pp. 381-384, Albuquerque, NM, USA, April 1990

H. Kantz, T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press, 2nd edition, 2004

H. Kawahara, H. Katayose, A. de Cheveigne, R.D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” *Eurospeech*, pp. 2781-2784, Budapest, Hungary, September 1999

H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi, T. Irino, “Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT,” *Interspeech*, pp. 537-540, Lisbon, Portugal, September 2005

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, “Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” *ICASSP 2008*, Las Vegas, pp. 3933-3936, 2008

J. Keener, J. Sneyd, *Mathematical Physiology*, Springer, 1998

J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature Medicine*, Vol. 7 (6), pp. 673-679, 2001

J. King, L. Ramig, J.H. Lemke, et al., “Parkinson’s disease: longitudinal changes in acoustic parameters of phonation,” *Journal of Medical Speech and Language Pathology*, Vol. 2, pp. 29-42, 1994

K. Kira, L. Rendell, “A practical approach to feature selection,” In Sleeman and P. Edwards (Eds.) *Proceedings of the Ninth International Conference on Machine Learning* (ICML-92), pp. 249-256, Morgan Kaufmann, 1992

R. Kohavi, G.H. John, “Wrappers for subset feature selection”, *Artificial Intelligence*, Vol. 97, pp. 273-324, 1997

I. Kononenko, “Estimating attributes: Analysis and extension of RELIEF,” In F. Bergadano and L. De Raedt (Eds.), *Proceedings of the European Conference on Machine Learning*, April 6-8, pp. 171-182, Catania, Italy, 1994

A. Kounoudes, P.A. Naylor, M. Brookes, “The DYPSA algorithm for estimation of glottal closure instants in voices speech,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, (ICASSP), pp. 349-352, Orlando, FL, 2002

A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical Review E*, Vol. 69 (6), 2004

L.A. Kurgan and K.J. Cios, “CAIM Discretization Algorithm,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 2, pp. 145-153, 2004

L.A. Kurgan and P. Musilek, “A survey of Knowledge Discovery and Data Mining process models,” *The Knowledge Engineering Review*, Vol. 21(1), pp. 1-24, 2006

L.I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley, 2004

N. Kwak, C-H. Hoi, “Input feature selection by mutual information based on Parzen window,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 12, pp. 1667-1771, 2002

A.E. Lang and A.M. Lozano, “Parkinson's disease - First of two parts,” *New England Journal Medicine*, vol. 339, pp. 1044-1053, 1998

K. Larson, L.O. Ramig and R.C. Scherer, “Acoustic and glottographic voice analysis during drug-related fluctuations in Parkinson’s disease,” *Journal of Medical Speech and Language Pathology*, Vol. 2, pp. 211-226, 1994

C.A. Laughton, M. Slavin, K. Katdare, L. Nolan, J.F. Bean, D.C. Kerrigan, E. Phillips, L.A. Lipsitz and J.J. Collins, “Aging, muscle activity, and balance control: physiologic changes associated with balance impairment,” *Gait and Posture*, Vol. 18, pp. 101-108, 2003

G.R. Lewis, “Ageing and the accuracy of jaw muscle control,” *Gerodontology*, Vol. 7, pp. 139-144, 2006

M.A. Little, *Biomechanically Informed Nonlinear Speech Signal Processing*, D.Phil. Thesis, University of Oxford, 2006

M.A. Little, P. E. McSharry, S. J. Roberts, D. Costello and I. M. Moroz, “Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection,” *Biomedical Engineering Online*, vol. 6 (23), 2007

M.A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,” *IEEE Transactions Biomedical Engineering*, Vol. 56 (4), pp. 1015-1022, 2009

M.A. Little, “Mathematical foundations of nonlinear, non-Gaussian, and time-varying digital speech signal processing,” *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Vol. 7015, pp. 9-16, 2011

H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu and G. Forman, “Evolving Feature Selection,” *IEEE Intelligent Systems*, Vol. 20 (6), pp. 64-76, November 2005

J.A. Logemann, H.B. Fisher, B. Boshes and E.R. Blonsky, “Frequency and coocurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients,” *Journal of Speech and Hearing Disorders*, Vol. 43, pp. 47-57, 1978

W. Maetzler, I. Liepelt, D. Berg, “Progression of Parkinson's disease in the clinical phase: potential markers,” *Lancet Neurology*, Vol. 8, pp. 1158-1171, 2009

S. Mallat, *A wavelet tour of signal processing*, Academic press, 3rd edition, 2009

C.L. Mallows, “Some comments on CP,” *Technometrics*, Vol. 42 (1), pp. 87-94, 2000

P. Maragos, J.F. Kaiser and T.F. Quatieri, “On amplitude and frequency demodulation using energy operators,” *IEEE Transactions on Signal Processing*, Vol. 41 (4), pp. 1532-1550, 1993

E. Marchiori, “Class conditional nearest neighbour for large margin instance selection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32 (2), pp. 364-370, 2010

M.R. McNeil, *Clinical Management of Sensorimotor Speech Disorders*, Thieme, New York, 1997

N. Meinshausen and B. Yu, “Lasso-type recovery of sparse representations for high-dimensional data,” *Annals of Statistics*, Vol. 37(1), pp.246-270, 2009

P. Mermelstein, *Distance measures for speech recognition, psychological and instrumental*, in Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., Academic, New York, pp. 374–388, 1976

J. Metter, and W. Hanson, “Clinical and acoustical variability in hypokinetic dysarthria,” *Journal of Communication Disorders*, vol. 19, pp. 347-366, 1986

P.E. Meyer, C. Schretter and G. Bontempi, “Information-theoretic feature selection in microarray data using variable complementarity”, *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Genomic and Proteomic Signal Processing*, Vol. 2, pp. 261-274, 2008

D. Michaelis, M. Frohlich and H.W. Strube, “Glottal to noise excitation ratio - a new measure for describing pathological voices”, *Acustica/acta acustica*, Vol. 83, pp. 700–706, 1997

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*, vol. 103(3), pp. 1628-1639, 1998

D. Michie, D.J. Spiegelhalter, C.C. Taylor (Eds.), *Machine learning, neural and statistical classification*, Ellis Horwood, 1994 (The book is freely available on the web: http://www1.maths.leeds.ac.uk/~charles/statlog/)

K.S.R. Murty and B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition,” *IEEE Signal Processing Letters*, Vol. 13, No. 1, pp. 52-55, 2006

P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, “Estimation of glottal closure instants in voices speech using the DYPSA algorithm,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, pp. 34-43, 2007

H. Ney, “Dynamic programming algorithm for optimal estimation of speech parameter contours,” *IEEE Transactions on Systems, Man, Cybernetics*, Vol. 13, pp. 208-214, 1983

M. Nishio and S. Niimi, “Changes in speaking fundamental frequency characteristics with aging,” *Folia phoniatrica et logopaedica*, Vol. 60, pp. 120-127, 2008

J.T. Ottesen, M.S. Olufsen and J.K. Larsen, *Applied Mathematical Models in Human Physiology*, SIAM (Monographs of Mathematical Modelling and Computation series), 2004

R. Pahwa and K. E. Lyons (Eds.), *Handbook of Parkinson’s Disease*, 4th edition, Informa Healthcare, USA, 2007

R. Paredes and E. Vidal, “Learning weighted metrics to minimize nearest-neighbor classification error,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28 (7), pp. 1100-1110, 2006

J. Parkinson, *Essay on the shaking palsy*, Whittingham and Rowland, London, 1817

E. Parzen, “On estimation of a probability density function and mode,” *Annals of Math. Statistics*, Vol. 33, pp. 1065-1076, 1962

S. Patel, K. Lorincz, R. Hughes, N. Huggins, J. Growdon, D. Standaert, M. Akay, J. Dy, M. Welsh, P. Bonato, “Monitoring Motor Fluctuations in Patients With Parkinson’s

Disease Using Wearable Sensors", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13 (6), pp. 864-873, 2009

H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226-1238, 2005

K. Perez, L. Ramig, M. Smith, C. Fromey, "The Parkinson larynx: tremor and video-stroboscopic findings," *Journal of Voice*, Vol. 10, pp. 354-361, 1996

G.E. Peterson and H.L. Barney, "Control methods used in a study of vowels," *Journal of the Acoustical Society of America*, Vol. 24, pp. 175-184, 1952

W. Pirker, S. Djamshidian, S. Asenbaum, et al., "Progression of dopaminergic degeneration in Parkinson's disease and atypical parkinsonism: a longitudinal β-CIT SPECT study," *Movement Disorders*, Vol. 17, pp. 45-53, 2002

R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, Vol. 6, No. 3, pp. 21-45, 2006

B. Post, M.P. Merkus, R.M.A. de Bie, R.J. de Haan, J.D. Speelman, "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?," *Movement Disorders*, Vol. 20, No. 12, pp. 1577-1584, 2005

Praat: doing phonetics by computer (Version 5.1.15) [Computer program], by P. Boersma and D. Weenink. Retrieved from http://www.praat.org/, 2009

J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: principles, algorithms and applications*, Prentice Hall, 4th edition, 2006

I. Psorakis, T. Damoulas, M.A. Girolami, "Multiclass relevance vector machines: sparsity and accuracy", *IEEE Transactions on Neural Networks*, Vol. 21, pp. 1588-1598, 2010

L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 25, pp. 24-33, 1977

A.H. Rajput, K.P. Offord, C.M. Beard, L.T. Kurland, "Epidemiology of Parkinsonism: incidence, classification and mortality," *Annals of Neurology*, Vol. 16, pp. 278-282, 1984

A.H. Rajput, B. Rozdilsky, and A. Rajput, "Accuracy of clinical diagnosis in parkinsonism – a prospective study," *Canadian Journal of Neurological Sciences*, vol 18 (3), pp. 275-278, 1991

A.H. Rajput, B. Rozdilsky, L. Ang and A. Rajput, "A significance of Parkinsonian manifestations in essential tremor," *Canadian Journal Neurological Science*, Vol. 20, pp. 114-117, 1993

M. Rajput, A. Rajput and A.H. Rajput, Epidemiology (Chapter 2). In *Handbook of Parkinson's disease*, R. Pahwa and K. E. Lyons (Eds), 4th edition, Informa Healthcare, USA, 2007

C. Ramaker, J. Marinus, A.M. Stiggelbout and B.J. van Hilten, "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease," *Movement Disorders*, Vol. 17, pp. 867-876, 2002

B. Ravina, D. Eidelberg, J.E. Ahlskog, R.L. Albin, D.J. Brooks, et al., "The role of radiotracer imaging in Parkinson's disease," *Neurology*, Vol. 64, pp. 208-215, 2005

B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, 1996

R.M. Roark, "Frequency and Voice: perspectives in the time domain," *Journal of Voice*, Vol. 20, No. 3, pp. 325-354, 2006

M. Robnik-Sikonja, I. Kononenko, "An adaptation of relief for attribute estimation in regression", *Proceedings of the 14th International Conference in Machine Learning* (ICML), pp. 296-304, 1997

M. Robnik-Sikonja, I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF", *Machine Learning*, Vol. 53, pp. 23-69, 2003

K.M. Rosen, R.D. Kent and J.R. Duffy, "Task-based profile of vocal intensity decline in Parkinson's disease," *Folia Phoniatr. Logop*, Vol. 57, pp. 28-37, 2005

S. Ross, *Probability and statistics for engineers and scientists*, 4th ed., Elsevier Academic press, 2009

C.O. Sakar and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia," *Journal of medical systems*, Vol. 34, pp. 591-599, 2010

S. Sapir, J. Spielman, L. Ramig, B. Story and C. Fox, "Effects of intensive voice treatment (LSVT) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings," *Journal of Speech, Language and Hearing Research*, 2006

S. Sapir, L.O. Ramig and C. Fox, "Voice, speech and swallowing disorders", In *Handbook of Parkinson's disease*, R. Pahwa and K. E. Lyons (Eds), 4th edition, Informa Healthcare, USA, 2007

S. Sapir, L. Ramig, J. Spielman, C. Fox, "Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech", *Journal of Speech Language and Hearing Research*, Vol. 53, pp. 114-25, 2010

J. Schoentgen and R. de Guchteneere, "Time series analysis of jitter," *Journal of Phonetics*, Vol. 23, pp. 189-201, 1995

A. Schrag, Y. Ben-Schlomo and N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?," *Journal of Neurology, Neurosurgery Pshychiatry*, Vol. 73, pp. 529-535, 2002

M.W.M. Schüpbach, J-C. Corvol, V. Czernecki, M.B. Djebara, J-L. Golmard, Y. Agid and A. Hartmann, "The segmental progression of early untreated Parkinson disease: a novel approach to clinical rating," *Journal of Neurology, Neurosurgery and Psychiatry*, Vol. 81(1), pp. 20-25, 2010

B. Shahbaba, R. Neal, "Nonlinear models using Dirichlet process mixtures," *Journal of Machine Learning Research*, Vol. 10, pp. 1829-1850, 2009

C.E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948

H. Shimazaki and S. Shinomoto, "Kernel bandwidth optimization in spike rate estimation," *Journal of Computational Neuroscience*, Vol. 29, pp. 171-182, 2010

B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, 1986

D.J. Sinder, Speech synthesis using an aeroacoustic fricative model, Ph.D. dissertation, department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 1999

N. Singh, V. Pillay, and Y.E. Choonara, "Advances in the treatment of Parkinson's disease," *Progress in Neurobiology*, vol. 81, pp. 29-44, 2007

S. Skodda, H. Rinsche, U. Schlegel, "Progression of dysprosody in Parkinson's disease over time – A longitudinal study," *Movement Disorders*, Vol. 24 (5), pp. 716-722, 2009

M. Small, *Applied Nonlinear Time Series Analysis*, World Scientific Publishing, 2005

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. van der Linde, "Bayesian measures of complexity and fit," *J. R. Statist. Soc. B*, Vol. 64, pp. 583-639, 2002

J. Stark and K. Hardy, "Chaos: Useful at last?," *Science*, Vol. 301, pp. 1192-1193, 2003

N. Stergiopulos, B.E. Westerhof, N. Westerhof, "Total Arterial Inertance as the fourth element of the windkessel model," *Am. J. Physiol. Heart Circ. Physiol*, Vol. 276, pp. 81-88, 1999

G.T. Stebbins, C.G. Goetz, A.E. Lang, E. Cubo, "Factor analysis of the motor section of the Unified Parkinson's Disease Rating Scale during the off-state," *Movement Disorders*, Vol. 14 (4), pp. 585-589, 1999

D. Stirzaker, *Elementary probability*, Cambridge University press, 2nd edition, 2003

H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of Machine Learning Research*, Vol. 3, pp. 1399–1414, 2003

X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *IEEE International Conference on Acoustics, Speech and Signal Processing*, (ICASSP), Orlando, Florida, 2002

Y. Sun, S. Todorovic and S. Goodison, "Local learning based feature selection for high dimensional data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, pp. 1610-1626, 2010

D. Talkin, *A robust algorithm for pitch tracking*, in Speech coding and synthesis (Eds. W.B. Kleijn and K.K. Paliwal), Elsevier, 1995

H.M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28 (5), pp. 599-601, 1980

R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *J. R. Statist. Soc. B* 58, 267-288, 1996

I.R. Titze, Workshop on Acoustic Voice Analysis, Summary statement by I. R. Titze, Denver, Colorado, Feb. 1994 (available online at: http://www.ncvs.org/museum-archive/sumstat.pdf - last time accessed: 2 January 2010)

I.R. Titze, *Principles of Voice Production*. National Center for Voice and Speech, Iowa City, US, 2nd printing, 2000

I.R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *Journal of Acoustical Society of America*, Vol. 123 (5), pp. 2733-2749, 2008

I.R. Titze and H. Liang, "Comparison of F0 extraction methods for high precision perturbation measurements," *Journal of Speech, Language, and Hearing Research*, Vol. 36, pp. 1120-133, 1993

K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, Vol. 3, pp. 1415-1438, 2003

C.-J. Tsai, C.-I. Lee, W.-P. Yang, "A discretization algorithm based on class-attribute contingency coefficient," *Information Sciences*, Vol. 178 (3), pp. 714-731, 2008

A. Tsanas, J.Y. Goulermas, V. Vartela, D. Tsiapras, G. Theodorakis, A.C. Fisher and P. Sfirakis, “The Windkessel model revisited: a qualitative analysis of the circulatory system,” *Medical Engineering and Physics*, Vol. 31, pp. 581-588, 2009

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, “Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests,” *IEEE Transactions Biomedical Engineering*, Vol. 57, pp. 884-893, 2010a

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, “Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, (ICASSP), Dallas, Texas, US, pp. 594-597, 14-19 March 2010b

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, “New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson’s disease symptom severity,” *International Symposium on Nonlinear Theory and its Applications* (NOLTA), Krakow, Poland, 5-8 September 2010c

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity”, *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011a

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, “Robust parsimonious selection of dysphonia measures for telemonitoring of Parkinson's disease symptom severity”, *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (MAVEBA), Florence, Italy, pp. 169-172, 25-27 August 2011b

A. Tsanas, M.A. Little, P.E. McSharry, *A methodology for the analysis of medical data*, In Handbook of Systems and Complexity in Health, Wiley, (in press), 2012a

A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, “Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease”, *IEEE Transactions on Biomedical Engineering*, Vol. 59, pp. 1264-1271, 2012b

A. Tsanas, M.A. Little, P.E. McSharry, B.K. Scanlon, S. Papapetropoulos, “Statistical analysis and mapping of the Unified Parkinson’s Disease Rating Scale to Hoehn and Yahr staging”, *Parkinsonism and Related Disorders*, Vol. 18 (5), pp. 697-699, 2012c

A. Tsanas, A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools", *Energy and Buildings*, Vol. 49C, pp. 560-567, 2012d

A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, "Using the cellular mobile telephone network to remotely monitor Parkinson's disease symptom severity", *IEEE Transactions on Biomedical Engineering*, (under review), 2012e

A. Tsanas, M. Zañartu, M.A. Little, P.E. McSharry, "Robust fundamental frequency estimation in sustained vowels using ensembles", *IEEE Transactions on Audio, Speech, and Language Processing* (under review), 2012f

E. Tuv, A. Borisov, G. Runger and K. Torkkola, "Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination", *Journal of Machine Learning Research*, Vol. 10, pp. 1341-1366, 2009

V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995

L.T. Vinh, N.D. Thang, Y-K. Lee, "An improved maximum relevance and minimum redundancy feature selection algorithms based on normalized mutual information," *10th Annual international symposium on applications and the internet*, pp. 395-398, Seoul, Korea, July 2010

L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, *Dimensionality Reduction: A Comparative Review*, Tilburg University Technical Report, 2009

A. Webb, *Statistical Pattern Recognition*, John Wiley and Sons Ltd, 2002

T.T. Wu and K. Lange, "Coordinate descent algorithms for LASSO penalised regression," *Annals of Applied Statistics*, Vol. 2 (1), pp. 224-244, 2008

M. Zañartu, *Acoustic coupling in phonation and its effect on inverse filtering of oral airflow and neck surface acceleration*, Ph.D. dissertation, School of Electrical and Computer Engineering, Purdue University, 2010

Y. Zhang, J.J. Jiang, L. Biazzo, M. Jorgessen, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral larungeal paralysis," *Journal of Voice*, Vol. 19 (4), pp. 519-528, 2005

H. Zou, "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, Vol. 101, pp. 1418-1429, 2006

Glossary of Key Terms

A discipline is only mastered when someone truly understands the abbreviations and concepts lying herein. This short glossary is an initial attempt to collect various key terms useful to this project in the wider sense.

- ☞ **Abduction:** movement of the vocal folds apart, “opening”

- ☞ **Action Potential**: The ‘spike’ or pulse of a neuron. An action potential is fired from a neuron when its membrane has been sufficiently depolarized. The normal resting value of the neuron membrane is -70 mV and an action potential occurs if the voltage increases at about 15 mV (depolarization), reaching -55 mV. The action potentials are the result of ion changes (particularly sodium (Na^+) and potassium (K^+), also known as the Na^+/K^+ pump) across the semi-permeable membrane. It is often abbreviated as AP.

- ☞ **Adduction:** movement of the vocal folds towards each other, “closing”

- ☞ **AMPA receptors**: The α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor is credited with fast EPSPs (*Excitatory* Post-Synaptic Potentials)

- ☞ **Bifurcation**: qualitative change of a nonlinear dynamical system when one of its parameters changes.

- ☞ **Blood-brain barrier** (BBB): Cellular property of the central nervous system that restricts the passage of various chemical substances and microscopic objects between the bloodstream and the neural tissue itself, while still allowing the passage of substances essential to metabolic function.

- ☞ **Breathiness**: varying degrees of noise, usually increased in people with Parkinson’s disease

- ☞ **CNS:** Central nervous system

- **DFA**: Detrended Fluctuation Analysis, is a technique for identifying the extent of *fractal self-similarity* in a signal

- **EPSP**: Excitatory Post-Synaptic Potentials

- **Fluctuation**: Backward and forward irregular movement, usually indicating instability. See also *perturbation*.

- **Fundamental frequency**: The largest frequency value after which the signal repeats.

- **GABA**: gamma (γ)-aminobutyric acid. It is the main *inhibitory* neurotransmitter (as opposed to glutamate, which is an *excitatory* neurotransmitter) of the nervous system. It is also responsible for muscle control. The term is often met as 'GABAergic', and refers to neurons which produce GABA at their output.

- **Glottis**: The airspace between the vocal folds.

- **Glutamate**: The main *excitatory* neurotransmitter

- **i.i.d:** Independent, identically distributed

- **IPD or PD**: Idiopathic Parkinson's disease or Parkinson's disease. It differs from parkinsonism, which refers to any symptom profile similar to that of PD but with a known etiology

- **Jitter:** Fluctuations in the variation in frequency of phonatory signals

- **LDA**: Linear Discriminant Analysis, is a simple technique that allows *linear separation* by modelling the data conditional upon each class using joint Gaussian probability densities

- **LSVT**: Lee Silverman Voice Treatment – see Sapir et al. (2006)

- **MS**: Multivariate surrogates: surrogates which retain both the linear temporal and linear spatial cross-correlations. These are further broken down into Global multivariate (GM) or Block multivariate (BM) surrogates.

- **NMDA receptor**: The NMDA (N-methyl D-aspartate) receptor contributes to *excitatory* synaptic transmission by depolarizing the postsynaptic cell membrane. It is credited with slow EPSPs (Excitatory Post-Synaptic Potentials).

- **Oscillation**: Repeated backwards and forwards movement (as in pendulum). When it continues without an applied external force, it is *self-sustained*.

- **Perturbation**: A slight change in a cyclic variable of the system (can be the amplitude, or the frequency). The system is slightly disturbed but is stable. See also *fluctuation*.

- **Pitch Period:** the inverse of the fundamental frequency

- **Principle of parsimony**: the more general model explains more phenomena with a smaller number of assumptions. Also known as *Occam's razor*.

- **Quality of Life**: Important factor in general clinical practice. It refers to how severely detrimental a pathology/disease is, and how this prevents a patient from leading a 'normal' life. It is usually weighed in the decision for the most appropriate treatment for each patient.

- **Shimmer:** Fluctuations of the amplitude of speech signals

- **Sodium-potassium pump**: Also known as the Na$^+$/K$^+$ pump, this membrane protein is responsible for the movement of sodium and potassium ions across the semi-permeable cell membrane. The pump transfers two K$^+$ ions in the cell, in exchange for pumping out three Na$^+$ ions. Functionally, it is necessary to maintain the membrane resting potential and regulate cell volume. The neurons rely on it to evoke action potentials responding to external stimulation.

- ☞ **Sparsity (sparse problem):** A term which appears often in high-dimensional data applications and refers to the fact that the function depends on only a few of the features present or data collected. Practically speaking, the desired quantity y depends on a small amount of the measured variables $\mathbf{x}$. The LASSO algorithm is particularly good at detecting *sparsity* and eliminating the superfluous variables. The number of features associated with non-zero coefficients in a regression setting for feature selection (such as the LASSO) is known as *sparsity level*.

Appendix I: Feature selection results

In this Appendix, we first present results which generalize LLBFS to multi-class classification problems in order to justify the selected approach. This generalization approach is similar in spirit to Kononenko (1994) who extended RELIEF to multi-class classification problems, and Hsu and Lin (2002) who investigated different generalizations of SVM.

In the next subsection of this Appendix, we succinctly present the features selected by each of the feature selection (FS) algorithms. These results should be read in conjunction with section 5.3. The aim is to allow other researchers to directly compare their findings with the FS algorithms used in this study. For simplicity, we refer to the features using simply their index number in each dataset instead of the actual feature name.

1. Six approaches to generalize LLBFS to multi-class classification problems

The six approaches tested to generalize LLBFS to multi-class classification problems were:

(a) **LLBFS1**: Normalized weights for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use One-Against-One (OAO) for the construction of the binary sub-problems.

(b) **LLBFS2**: Original (un-normalized) weights for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use OAO for the construction of the binary sub-problems.

(c) **LLBFS3**: Original weights also taking into account the number of data samples for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use OAO for the construction of the binary sub-problems.

(d) **LLBFS4**: Normalized weights for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use One-Against-All (OAA) for the construction of the binary sub-problems.

(e) **LLBFS5**: Original (un-normalized) weights for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use OAA for the construction of the binary sub-problems.

(f) **LLBFS6**: Original weights also taking into account the number of data samples for each binary sub-problem, and then taking the mean of the feature weights across the sub-problems. Use OAA for the construction of the binary sub-problems.

We use the three multi-class classification datasets presented in § 5.2 to tentatively draw conclusions. First, we use the Artificial2 dataset to investigate which of the LLBFS approaches leads to the lowest FDR, and present the results in Fig. I.1.

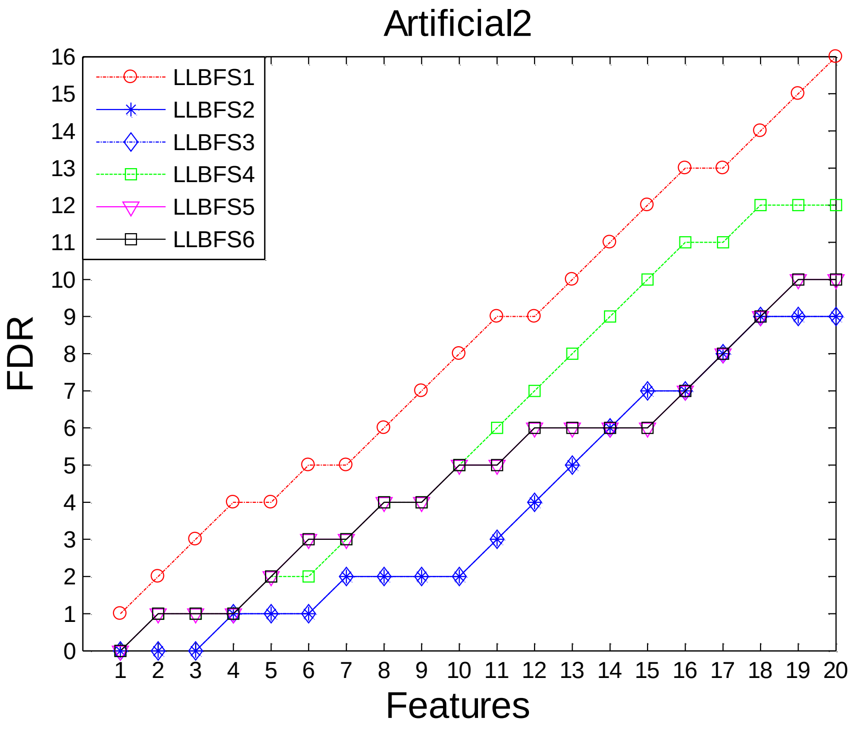

Fig. I.1. Comparison of the feature selection algorithms in terms of true feature set recovery.

Next, we feed the features selected as a result of applying these six different types of LLBFS to a RF for the multi-class classification datasets in § 5.2. We present the results for the Image segmentation and the Cardiotocography dataset; there was almost complete overlap in performance for the LLBFS variants in the Wine dataset.

The findings in Figures I.1 and I.2 suggest that LLBFS3 selects datasets which lead to somewhat better performance compared to the alternatives for multi-class classification. Therefore, this is the version of LLBFS we will be using for multi-class classification problems in this thesis. Moreover, using the OAA approach turns out to be computationally more demanding. It is interesting that Hsu and Lin (2002), who studied OAO and OAA for the generalization of SVM applied to multi-class classification problems, reached similar conclusions and advocate using OAO as the default method for SVM multi-class classification.

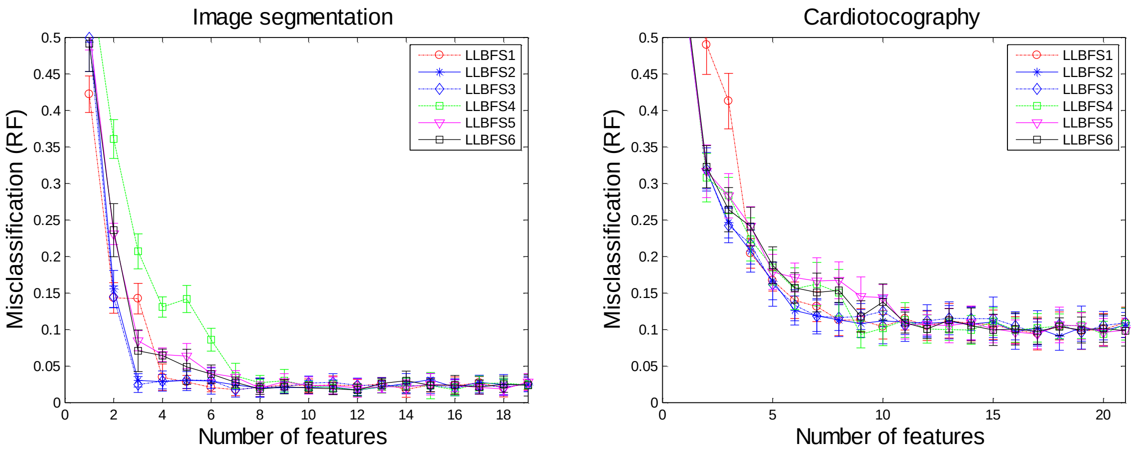

Fig. I.2. Comparison of the LLBFS approaches in terms of learner performance.

2. Features selected by the feature selection algorithms

We present the results for two binary classification datasets, two multi-class classification datasets, and a fat dataset from the datasets introduced in § 5.2.

Table I.1: Selected features for the Hepatitis dataset.

LASSO	mRMR	mRMR$_{Spearman}$	GSO	RELIEF	LLBFS	RRCT
12	18	12	17	19	12	12
17	19	7	19	12	13	7
19	8	18	18	11	15	19
18	2	19	7	6	19	18
14	12	8	6	13	11	17
7	7	17	14	9	6	13
6	15	2	16	5	14	6
8	13	13	1	3	10	1
16	3	1	12	10	18	8
2	6	6	2	17	3	2

Entries refer to the indices of the features in the original dataset. Only the top 10 features are reported here.

Table I.2: Selected features for the Parkinson’s dataset.

LASSO	mRMR	mRMR$_{Spearman}$	GSO	RELIEF	LLBFS	RRCT
19	1	22	19	19	19	22
3	19	2	3	1	1	2
20	17	19	18	22	3	20
1	11	3	21	11	2	3
21	18	13	7	14	20	13
18	20	20	1	18	22	18
2	3	18	4	9	4	21
15	2	21	20	12	5	5
11	8	8	17	17	6	17
7	21	1	6	20	7	19

Entries refer to the indices of the features in the original dataset. Only the top 10 features are reported here.

Table I.3: Selected features for the Image Segmentation dataset.

LASSO	mRMR	$\text{mRMR}_{\text{Spearman}}$	GSO	RELIEF	LLBFS	RRCT
2	17	2	2	19	2	2
19	2	1	1	12	19	1
1	19	9	14	17	12	14
14	1	19	15	10	14	16
16	18	16	10	13	17	9
18	14	6	18	11	18	4
5	16	4	12	16	1	5
11	11	8	5	2	10	6
7	6	15	16	15	13	19
4	15	7	17	14	11	11

Entries refer to the indices of the features in the original dataset. Only the top 10 features are reported here.

Table I.4: Selected features for the Cardiotocography dataset.

LASSO	mRMR	$\text{mRMR}_{\text{Spearman}}$	GSO	RELIEF	LLBFS	RRCT
5	20	5	5	8	2	5
10	2	2	10	5	5	8
8	18	11	7	13	8	7
7	9	7	4	2	10	4
20	5	3	1	10	13	11
2	12	8	20	12	18	1
4	8	6	8	21	11	2
1	11	16	21	18	7	6
21	14	21	2	1	12	3
12	13	20	14	9	17	16

Entries refer to the indices of the features in the original dataset. Only the top 10 features are reported here.

Table I.5: Selected features for the SRBCT dataset.

LASSO	mRMR	mRMR$_{\text{Spearman}}$	GSO	RELIEF	LLBFS	RRCT
1194	1389	1194	1194	1389	1955	1194
187	2	1301	1301	1955	842	1301
1207	1888	2247	2247	246	1389	1937
1003	1980	187	419	1319	2022	1207
1105	545	1634	1826	107	1954	1002
1634	174	1207	265	1954	2162	509
188	1784	867	2053	545	187	1954
1536	867	2046	682	187	1066	1723
251	1634	1536	434	1645	246	188
867	1194	1003	1888	1708	545	2146
849	277	970	149	509	1915	1706
1955	566	509	179	2162	174	1888
335	71	1112	718	867	1427	187
758	1158	335	148	1003	1601	2046
123	823	1105	1678	1194	1093	1920
2046	2162	1888	569	2050	107	251
1964	1954	941	1495	1980	819	1112
558	1009	1760	2133	566	1319	1093
2081	246	761	1138	1353	742	1105
1301	1645	2146	1667	129	788	123
850	1884	1723	51	2046	2198	558
970	1708	1093	1877	153	867	867
808	2144	910	150	607	1980	1647
1896	2047	1706	1944	1066	509	2230
1914	2258	469	1432	2022	1353	1536

Entries refer to the indices of the features in the original dataset. The features which were selected and subsequently removed by LASSO in the first 25 steps are not reported. Only the top 25 features are reported here.

Appendix II: Correlations, statistics, and errors

1. UPDRS correlations between components

We examine the correlation between the three UPDRS components, and between each component and the total UPDRS score. All correlations and *p-values* were determined using the Spearman correlation test. We use the data from the 42 patients of the AHTD study taking the UPDRS scores at all three distinct times (baseline, 3-month and 6-month intervals).

Table II.1: Statistical association and statistical significance between UPDRS components. Bold italics indicate statistically significant correlation (at the 95% level).

	UPDRS component 1 (MBM) *16 points*	**UPDRS component 2** (ADL) *52 points*	**UPDRS component 3** (Motor) *108 points*
UPDRS component 2 (ADL) *52 points*	*p-val:* 0.060 *R:* 0.168		
UPDRS component 3 (Motor) *108 points*	*p-val*: 0.068 *R:* 0.163	*p-val:* **<0.0001** *R:* 0.486	
UPDRS total (Sum all parts) *176 points*	*p-val*: **0.002** *R:* 0.278	*p-val:* **<0.0001** *R:* 0.695	*p-val:* **<0.0001** *R:* 0.951

MBM stands for ‘Mentation, Behaviour and Mood’, and ADL stands for ‘Activities of Daily Living’, which are the first and second components of the UPDRS metric.

We remark that motor-UPDRS (UPDRS component 3) is the most significantly correlated component to the total UPDRS. In fact, *motor-UPDRS is practically a reflection of the total-UPDRS score* (Spearman $R = 0.95$). Component 2 is also strongly correlated to total-UPDRS (Spearman $R \approx 0.7$). Interestingly, components 2 and 3 are also statistically significantly correlated with association strength of about 0.486. The correlation strength denoted by the nonparametric Spearman R of component 1 with components 2 and 3 is markedly lower.

2. UPDRS correlations between sections

Table II.2: p-values of the motor-UPDRS sections.

	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
19	0																									
20	0.776	0.775																								
21	0.118	0.124	0.40																							
22	0.703	0.766	0.76	0																						
23	0.457	0.654	0.087	0.0660	0.736																					
24	0.2700	0.537	0.822	0.020	0	0.438																				
25	0.987	0.694	0.899	0	0.484	0.184	0.720																			
26	0.880	0.760	0.407	0.001	0	0.052	0.008	0.002																		
27	0	0.015	0.196	0.0710	0.709	0.038	0.712	0.627	0.157																	
28	0.029	0	0.05	0	0.003	0.991	0.214	0.006	0.975	0																
29	0.001	0	0.35	0.007	0	0.134	0.005	0.619	0	0.063	0.587															
30	0.123	0.267	0.118	0.928	0.843	0.003	0.116	0.124	0.009	0	0	0.037														
31	0.003	0.026	0.675	0.389	0.001	0.22	0.007	0.584	0	0	0.084	0	0													
32	0.003	0.006	0.664	0	0.031	0.086	0.008	0	0.33	0.001	0	0.975	0.045	0.208												
33	0.001	0	0.916	0.171	0.001	0.039	0.59	0.915	0	0.345	0.515	0	0.235	0.032	0.085											
34	0.39	0.298	0.132	0.001	0.387	0.366	0.094	0	0.1	0	0	0.217	0.004	0.682	0	0.35										
35	0.02	0	0.842	0.012	0	0.304	0.005	0.493	0	0.249	0.259	0	0.843	0.004	0.363	0	0.263									
36	0.078	0.267	0.195	0	0.446	0.171	0.036	0	0.087	0	0	0.089	0.003	0.255	0	0.505	0	0.562								
37	0.007	0.002	0.758	0.053	0	0.417	0.012	0.697	0	0.014	0.988	0	0.182	0	0.324	0	0.145	0	0.006							
38	0.036	0	0.868	0	0.095	0.336	0.174	0.029	0.948	0	0	0.783	0	0.032	0	0.923	0	0.625	0	0.581						
39	0.001	0.001	0.301	0.682	0.004	0.588	0.038	0.54	0.111	0.012	0.068	0	0.017	0	0.174	0	0.379	0	0.442	0	0					
40	0.934	0.712	0.046	0.242	0.079	0.091	0.802	0.149	0.049	0.003	0.154	0.511	0.001	0.02	0.844	0.557	0.004	0.234	0.045	0.032	0.01	0.005				
41	0.002	0	0.502	0.689	0.008	0.525	0.11	0.581	0.016	0.009	0.162	0.021	0.104	0.02	0.004	0.017	0.3	0	0.015	0	0.003	0.001	0.702			
42	0.231	0.167	0.412	0.965	0.369	0.245	0.934	0.228	0.212	0.342	0.23	0.535	0.038	0.16	0.297	0.858	0.018	0.119	0.055	0.178	0.003	0.25	0.241	0		
43	0.796	0.393	0.75	0.14	0.579	0.978	0.833	0.978	0.04	0.582	0.945	0.012	0.962	0.596	0.749	0.11	0.727	0.074	0.766	0.396	0.061	0.694	0.372	0.19	0.07	
44	0	0	0.894	0.522	0.337	0.715	0.503	0.08	0.172	0	0.007	0.08	0	0.002	0.005	0.579	0.016	0.037	0.025	0.042	0.002	0.013	0.781	0	0	0.366

Table II.3: Spearman R nonparametric correlation coefficients of the motor-UPDRS sections. Bold indicates strength of correlation >0.5.

	18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**
19	**0.57**																									
20	0.03	-0.03																								
21	0.14	0.14	0.08																							
22	-0.04	0.03	-0.03	-0.34																						
23	-0.07	-0.04	0.16	0.17	-0.03																					
24	-0.1	0.06	-0.02	-0.21	0.36	0.07																				
25	0	0.04	0.01	0.37	-0.07	0.12	-0.03																			
26	-0.01	0.03	0.08	-0.3	0.44	0.18	0.24	0.29																		
27	0.38	0.22	0.12	0.17	-0.04	0.19	0.03	0.05	0.13																	
28	0.2	0.34	0.18	0.37	-0.27	0	-0.12	0.25	0	0.38																
29	0.3	0.33	-0.09	-0.25	0.42	-0.14	0.26	-0.05	0.41	0.17	0.05															
30	0.14	0.1	0.14	0.01	-0.02	0.27	0.15	0.14	0.24	0.68	0.42	0.19														
31	0.27	0.21	0.04	-0.08	0.29	0.11	0.25	0.05	0.34	**0.51**	0.16	**0.5**4	**0.57**													
32	0.27	0.25	0.04	0.45	-0.2	0.16	-0.24	0.38	0.09	0.32	**0.57**	0	0.19	0.12												
33	0.31	0.36	0.01	-0.13	0.3	-0.19	0.05	0.01	0.35	0.09	0.06	**0.5**	-0.11	0.2	0.16											
34	0.08	0.1	0.14	0.29	-0.08	0.08	-0.16	0.42	0.15	0.33	**0.5**	-0.12	0.26	0.04	**0.59**	0.09										
35	0.21	0.32	-0.02	-0.23	0.37	-0.1	0.26	0.06	**0.53**	0.11	-0.11	**0.52**	-0.02	0.26	0.09	0.73	0.1									
36	0.16	0.1	0.12	0.49	-0.07	0.13	-0.19	0.38	0.16	0.45	**0.53**	-0.16	0.27	0.11	0.71	0.06	0.72	0.05								
37	0.25	0.28	0.03	-0.18	**0.55**	-0.08	0.23	0.04	0.6	0.23	0	0.44	0.12	0.32	0.09	0.6	0.14	0.65	0.25							
38	0.19	0.35	-0.02	0.37	-0.15	0.09	-0.13	0.2	-0.01	0.38	**0.57**	-0.03	0.37	0.2	0.61	-0.01	**0.51**	-0.05	**0.5**	0.05						
39	0.29	0.31	-0.1	-0.04	0.26	-0.05	0.19	-0.06	0.15	0.23	0.17	0.33	0.22	0.36	0.13	0.36	0.08	0.34	0.07	0.39	0.41					
40	0.01	0.03	0.18	-0.11	0.16	0.16	0.02	-0.13	0.18	0.27	0.13	0.06	0.31	0.21	-0.02	0.06	0.26	0.11	0.19	0.2	0.24	0.26				
41	0.29	0.39	-0.06	-0.04	0.25	-0.06	0.15	-0.05	0.22	0.24	0.13	0.21	0.15	0.21	0.26	0.22	0.1	0.36	0.22	0.47	0.28	0.32	0.04			
42	0.11	0.13	-0.08	0	0.08	-0.11	-0.01	0.11	0.12	0.09	0.11	0.06	0.19	0.13	0.1	0.02	0.22	0.14	0.18	0.13	0.27	0.11	0.11	0.35		
43	-0.02	-0.08	-0.03	0.14	-0.05	0	0.02	0	-0.19	0.05	-0.01	-0.23	0	-0.05	-0.03	-0.15	0.03	-0.17	0.03	-0.08	0.17	0.04	0.08	0.12	0.17	
44	0.34	0.45	0.01	0.06	0.09	0.03	0.06	0.16	0.13	0.37	0.25	0.16	0.32	0.28	0.26	0.05	0.22	0.19	0.21	0.19	0.28	0.23	0.03	0.48	0.38	-0.08

Appendix III: UPDRS guide

The following document presents the UPDRS guide in marking PD symptoms and was obtained from http://www.mdvu.org/library/ratingscales/pd/updrs.pdf. The original UPDRS was introduced by Fahn et al. (1989). The author of this report does not claim any originality, and the document is only included for completeness to illustrate the format and guidelines of the actual UPDRS collection process. Only the first three components of the UPDRS are included in this appendix; the fourth is about the effect of treatment, but in the AHTD trial all patients remained untreated. Also, note that the UPDRS dataform is more detailed in the motor-UPDRS, e.g. with details of left and right hand/feet whereas here they appear united. For more details about the metrics, the dataforms and the PD marking please refer to: http://www.parkinson.org/ and to http://www.mdvu.org/library/ratingscales/pd/.

The following is reprinted with approval from WE MOVE, New York, NY 2012.

I. MENTATION, BEHAVIOR AND MOOD

1. Intellectual Impairment
0 = None.
1 = Mild. Consistent forgetfulness with partial recollection of events and no other difficulties.
2 = Moderate memory loss, with disorientation and moderate difficulty handling complex problems. Mild but definite impairment of function at home with need of occasional prompting.
3 = Severe memory loss with disorientation for time and often to place. Severe impairment in handling problems.
4 = Severe memory loss with orientation preserved to person only. Unable to make judgements or solve problems. Requires much help with personal care. Cannot be left alone at all.

2. Thought Disorder (Due to dementia or drug intoxication)
0 = None.
1 = Vivid dreaming.
2 = "Benign" hallucinations with insight retained.
3 = Occasional to frequent hallucinations or delusions; without insight; could interfere with daily activities.
4 = Persistent hallucinations, delusions, or florrid psychosis. Not able to care for self.

3. Depression
1 = Periods of sadness or guilt greater than normal, never sustained for days or weeks.
2 = Sustained depression (1 week or more).
3 = Sustained depression with vegetative symptoms (insomnia, anorexia, weight loss, loss of interest).
4 = Sustained depression with vegetative symptoms and suicidal thoughts or intent.

4. Motivation/Initiative
0 = Normal.
1 = Less assertive than usual; more passive.
2 = Loss of initiative or disinterest in elective (nonroutine) activities.
3 = Loss of initiative or disinterest in day to day (routine) activities.
4 = Withdrawn, complete loss of motivation.

II. ACTIVITIES OF DAILY LIVING (for both "on" and "off")

5. Speech
0 = Normal.
1 = Mildly affected. No difficulty being understood.
2 = Moderately affected. Sometimes asked to repeat statements.
3 = Severely affected. Frequently asked to repeat statements.
4 = Unintelligible most of the time.

6. Salivation
0 = Normal.
1 = Slight but definite excess of saliva in mouth; may have nighttime drooling.
2 = Moderately excessive saliva; may have minimal drooling.
3 = Marked excess of saliva with some drooling.
4 = Marked drooling, requires constant tissue or handkerchief.

7. Swallowing
0 = Normal.
1 = Rare choking.
2 = Occasional choking.
3 = Requires soft food.
4 = Requires NG tube or gastrotomy feeding.

8. Handwriting
0 = Normal.
1 = Slightly slow or small.
2 = Moderately slow or small; all words are legible.
3 = Severely affected; not all words are legible.
4 = The majority of words are not legible.

9. Cutting food and handling utensils
0 = Normal.
1 = Somewhat slow and clumsy, but no help needed.
2 = Can cut most foods, although clumsy and slow; some help needed.
3 = Food must be cut by someone, but can still feed slowly.
4 = Needs to be fed.

10. Dressing
0 = Normal.
1 = Somewhat slow, but no help needed.
2 = Occasional assistance with buttoning, getting arms in sleeves.
3 = Considerable help required, but can do some things alone.
4 = Helpless.

11. Hygiene
0 = Normal.
1 = Somewhat slow, but no help needed.
2 = Needs help to shower or bathe; or very slow in hygienic care.
3 = Requires assistance for washing, brushing teeth, combing hair, going to bathroom.
4 = Foley catheter or other mechanical aids.

12. Turning in bed and adjusting bed clothes
0 = Normal.
1 = Somewhat slow and clumsy, but no help needed.
2 = Can turn alone or adjust sheets, but with great difficulty.
3 = Can initiate, but not turn or adjust sheets alone.
4 = Helpless.

13. Falling (unrelated to freezing)
0 = None.
1 = Rare falling.
2 = Occasionally falls, less than once per day.
3 = Falls an average of once daily.
4 = Falls more than once daily.

14. Freezing when walking
0 = None.
1 = Rare freezing when walking; may have starthesitation.
2 = Occasional freezing when walking.
3 = Frequent freezing. Occasionally falls from freezing.
4 = Frequent falls from freezing.

15. Walking
0 = Normal.
1 = Mild difficulty. May not swing arms or may tend to drag leg.
2 = Moderate difficulty, but requires little or no assistance.
3 = Severe disturbance of walking, requiring assistance.
4 = Cannot walk at all, even with assistance.

16. Tremor (Symptomatic complaint of tremor in any part of body.)
0 = Absent.
1 = Slight and infrequently present.
2 = Moderate; bothersome to patient.
3 = Severe; interferes with many activities.
4 = Marked; interferes with most activities.

17. Sensory complaints related to parkinsonism
0 = None.
1 = Occasionally has numbness, tingling, or mild aching.
2 = Frequently has numbness, tingling, or aching; not distressing.
3 = Frequent painful sensations.
4 = Excruciating pain.

III. MOTOR EXAMINATION

18. Speech
0 = Normal.
1 = Slight loss of expression, diction and/or volume.
2 = Monotone, slurred but understandable; moderately impaired.
3 = Marked impairment, difficult to understand.
4 = Unintelligible.

19. Facial Expression
0 = Normal.
1 = Minimal hypomimia, could be normal "Poker Face".
2 = Slight but definitely abnormal diminution of facial expression
3 = Moderate hypomimia; lips parted some of the time.
4 = Masked or fixed facies with severe or complete loss of facial expression; lips parted 1/4 inch or more.

20. Tremor at rest (head, upper and lower extremities)
0 = Absent.
1 = Slight and infrequently present.
2 = Mild in amplitude and persistent. Or moderate in amplitude, but only intermittently present.
3 = Moderate in amplitude and present most of the time.
4 = Marked in amplitude and present most of the time.

21. Action or Postural Tremor of hands
0 = Absent.
1 = Slight; present with action.
2 = Moderate in amplitude, present with action.
3 = Moderate in amplitude with posture holding as well as action.
4 = Marked in amplitude; interferes with feeding.

22. Rigidity (Judged on passive movement of major joints with patient relaxed in sitting position. Cogwheeling to be ignored.)
0 = Absent.
1 = Slight or detectable only when activated by mirror or other movements.
2 = Mild to moderate.
3 = Marked, but full range of motion easily achieved.
4 = Severe, range of motion achieved with difficulty.

23. Finger Taps (Patient taps thumb with index finger in rapid succession.)
0 = Normal.
1 = Mild slowing and/or reduction in amplitude.
2 = Moderately impaired. Definite and early fatiguing. May have occasional arrests in movement.
3 = Severely impaired. Frequent hesitation in initiating movements or arrests in ongoing movement.
4 = Can barely perform the task.

24. Hand Movements (Patient opens and closes hands in rapid succesion.)
0 = Normal.
1 = Mild slowing and/or reduction in amplitude.
2 = Moderately impaired. Definite and early fatiguing. May have occasional arrests in movement.
3 = Severely impaired. Frequent hesitation in initiating movements or arrests in ongoing movement.
4 = Can barely perform the task.

25. Rapid Alternating Movements of Hands (Pronation-supination movements of hands, vertically and horizontally, with as large an amplitude as possible, both hands simultaneously.)
0 = Normal.
1 = Mild slowing and/or reduction in amplitude.
2 = Moderately impaired. Definite and early fatiguing. May have occasional arrests in movement.
3 = Severely impaired. Frequent hesitation in initiating movements or arrests in ongoing movement.
4 = Can barely perform the task.

26. Leg Agility (Patient taps heel on the ground in rapid succession picking up entire leg. Amplitude should be at least 3 inches.)
0 = Normal.
1 = Mild slowing and/or reduction in amplitude.
2 = Moderately impaired. Definite and early fatiguing. May have occasional arrests in movement.
3 = Severely impaired. Frequent hesitation in initiating movements or arrests in ongoing movement.
4 = Can barely perform the task.

27. Arising from Chair (Patient attempts to rise from a straightbacked chair, with arms folded across chest.)
0 = Normal.
1 = Slow; or may need more than one attempt.
2 = Pushes self up from arms of seat.
3 = Tends to fall back and may have to try more than one time, but can get up without help.
4 = Unable to arise without help.

28. Posture
0 = Normal erect.
1 = Not quite erect, slightly stooped posture; could be normal for older person.
2 = Moderately stooped posture, definitely abnormal; can be slightly leaning to one side.
3 = Severely stooped posture with kyphosis; can be moderately leaning to one side.
4 = Marked flexion with extreme abnormality of posture.

29. Gait
0 = Normal.
1 = Walks slowly, may shuffle with short steps, but no festination (hastening steps) or propulsion.
2 = Walks with difficulty, but requires little or no assistance; may have some festination, short steps, or propulsion.
3 = Severe disturbance of gait, requiring assistance.
4 = Cannot walk at all, even with assistance.

30. Postural Stability (Response to sudden, strong posterior displacement produced by pull on shoulders while patient erect with eyes open and feet slightly apart. Patient is prepared.)
0 = Normal.
1 = Retropulsion, but recovers unaided.
2 = Absence of postural response; would fall if not caught by examiner.
3 = Very unstable, tends to lose balance spontaneously.
4 = Unable to stand without assistance.

31. Body Bradykinesia and Hypokinesia (Combining slowness, hesitancy, decreased armswing, small amplitude, and poverty of movement in general.)
0 = None.
1 = Minimal slowness, giving movement a deliberate character; could be normal for some persons. Possibly reduced amplitude.
2 = Mild degree of slowness and poverty of movement which is definitely abnormal. Alternatively, some reduced amplitude.
3 = Moderate slowness, poverty or small amplitude of movement.
4 = Marked slowness, poverty or small amplitude of movement.

Appendix IV: Courses attended

Date	Name of Course/Seminar	Speaker	Skills developed/purpose
5 November 2008 14.30-15.30	**Jorge Cham Lecture** *Dept. Physics, Clarendon Lab*	Jorge Cham	Introduction to graduate student life
Nov-Feb 2008 (8x) 17.30-18.30	**Building a Business** *Said Business School*	Various	Business practices, negotiation skills
24 November 2008 9.00-18.30	**KTN Seminar** *St. Catherine's college*	KTN projects Alfio Quarteroni	More open-minded, various project ideas
25 November 2008 10.30-14.30	**Basic Presentation Skills** *Chemistry Research Lab (13)*	Alison Trinder	Preparing and giving a successful presentation
9 December 2008 11.30-14.30	**Good practice in research** *Chemistry Research Lab (13)*	Karen Melham, Barbara Gabrys	Ethos in research
12 January 2009 9.00-17.00	**Communication Skills** *Department of Zoology*	Various	Conveying science to different groups of people
14 January 2009 9.00-17.00	**Inverse Problems Workshop** *St. Anne's College*	Various (KTN)	Mathematical aspects of inverse problems
21 January 2009	**Deep Brain Stimulation Surgery** (J.R. Hospital)	Tipu Aziz	Witnessed brain surgery to treat PD

3 February 2009 9.30-13.00	**Managing your DPhil** *Centenary room, Career Serv.*	Louise Baron	Time allocation, personal relationships
3 March 2009 12.00-14.00	**Networking** *Centenary room, Career Serv.*	Various	Meeting important people and conveying research
4 March 2009	**PUMMA group** *IBME, Oxford*	I gave a talk	Feedback on my cardiovascular modeling project
Jan – Mar 2009	**Introduction to statistical machine learning (×5)**	Max Little	Statistical machine learning ideas
Sep 2008 – Dec 2009	**Telephone conferences (Intel)**	M. Little, M. Deisher, B. Deleeuw, S. Sharma	Updates on the project, setting plans
Dec 2008 – Jun 2009	**Meetings with medical and nursing staff** (J.R. Hospital)	Ralph Gregory Tipu Aziz	Practical aspects of neurological disease
Oct 2008 – Jun 2009	**OCIAM/JAMS Seminars (and 1 Biomedical Engineering Seminar)**	Various	Interesting ideas, could be applicable to my project
4 November 2009	**Online training course**: Protecting Human Research Participants (certification number: 333277)	National Institutes of Health	Ethics in research and learning involving human participants
30 November 2009 9.00-18.30	**KTN Seminar** *St. Catherine's college*	Various * I gave a talk	More open-minded, various project ideas
1 February	**Bibliometrics - the black**	Various	Impact factors,

2010 **12.30-13.30**	**art of citation rankings** *Oxford, OUCS*		alternative methods to measure journal importance
24 February 2010 **14.00-15.30**	**Transfer Viva** *Oxford, OCIAM*	Examiners: Prof. Stephen Roberts, Dr. Irene Moroz	Feedback on my work, ideas to pursue next in the project
14-19 March 2010	**ICASSP conference** *Dallas, Texas, US*	Various * I presented a poster	Ideas on signal processing, possible collaborations
10 June 2010 **14.00-16.00**	**Developing your professional network** *Oxford, Career service*	Natalie Lundsteen and Claire Conway	Ideas on meeting people and keeping contacts
5 September 2010	**NOLTA conference** *Krakow, Poland*	Various * Oral presentation	Ideas on time series analysis
2 November 2010	**JAMS** *OCIAM, Oxford*	I gave a talk	Feedback on my work (PD project)
3 November 2010 **10.00-17.00**	**Feature selection for the sciences seminar** *Department of Physics, Oxford*	Various *I gave a talk *Co-organised the seminar	Feedback on my work (feature selection), alternative feature selection concepts
22 November 2010	**KTN Seminar** *St. Catherine's college, Oxford*	Various *I presented a poster	Feedback on my work (PD project)
January-February/11 (6x)	**MPLS reading group** *Oxford, Career service*	Catherine Baillie, Barbara Gabrys	Ideas about teaching in higher education, reflecting on students' needs

24 February 2011	**IBME Seminar** *IBME, Oxford*	I gave a talk	Feedback on my work (PD project)
1 March 2011	**Dynamical systems group seminar** *OCIAM, Oxford*	I gave a talk	Feedback on my work (feature selection)
16 March 2011	**Oxford Robotics Research Group Seminars** *Department of Engineering Science, Oxford*	I gave a talk	Feedback on my work (PD project)
5 full days March 2011	**Genetic Algorithms course** *IBME, Oxford*	Andrew Kramer	Learned the basic ideas of Genetic Algorithms
4 May 2011	**Statistics for the analysis of medical data** *Attikon hospital, Athens, Greece*	I gave a talk	Presenting work to a non-mathematically oriented audience (clinicians)
9 May 2011	**Oxford Robotics Research Group Seminars** *Department of Engineering Science, Oxford*	Arthur Gretton	Good ideas on kernels, we could use them for the feature selection project
18 May 2011	**Developing Learning and Teaching: Portfolio workshop**	Ian Finlay	Portfolio writing for obtaining the teaching qualification for higher education
24 May 2011	**IBME Seminar** *IBME, Oxford*	I gave a talk	Feedback on my work (feature selection)
26 May 2011	**Time series seminar** *Balliol college, Oxford*	I gave a talk	Feedback on my work (PD project)

May 2011 **(3x)**	**Meetings with Clinicians** *Birmingham, UK*	Declan Costello, Caren Morrison	Clinical insight and possibilities for extending our work with new collaborations
16 June 2011	**Oxbridge Wooly Owl competition in Applied Maths**	Various	Interesting ideas on various projects
23 June 2011	**OCIAM Differential equations and applications seminar series**	Prof. Q-C. Zhong *I arranged the talk and hosted the event	Talk on time-delay systems, some concepts could be applied to my work
15 July 2011 **10.00-13.30**	**Confirmation Viva** *Oxford, OCIAM*	Examiners: Prof. Philip Maini, Dr. Gari Clifford	Feedback on my work, ideas to pursue next in the project
21-30 August 2011	**MAVEBA conference** *Florence, Italy*	Various * Oral presentation	Networking with experts in speech + PD
23 September 2011	**Meeting with clinicians and phoneticians** *Oxford, UK*	Declan Costello, Elinor Payne, Ladan	Investigating extensions of my work on various other vocal pathologies
10 January 2012	**Invited talk – Harvard Medical School** *Boston, US*	I gave a talk	Feedback from clinicians – possible collaboration on related projects
23 February 2012	**IBME Seminar** *IBME, Oxford*	I gave a talk	Feedback on my work (feature selection)
26-27 March 2012	**Time series symposium** *London, UK*	Various	Interesting concepts and applications

8 May 2012	**IBME Seminar** *IBME, Oxford*	I gave a talk	Feedback on my work (F0 estimation)
17 May 2012	**OCIAM Differential equations and applications seminar series**	Dr. Gavin Brown *I arranged the talk and hosted the event	Talk on feature selection
14 June 2012	**Numeric Algorithms Group** *Oxford*	I gave a talk	Possibility for future collaboration

Appendix V: List of software tools used

All programming was completed in Matlab (MATLAB®, version 2010b, The MathWorks). For my experiments I have made use of the following packages and toolboxes, for which I am indebted to their developers for making them freely available or providing me access to their source code:

- **CLOP Toolbox** (Matlab toolbox developed on top of Spider) by I. Guyon http://clopinet.com/isabelle/Projects/ETH/Feature_Selection_w_CLOP.html
- **Dimensionality reduction toolbox** (version 0.7.2), by L.J.P. van der Maaten http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
- **Empirical mode decomposition (Hilbert-Huang transform)**, (Matlab function) by G. Rilling and P. Flandrin, http://perso.ens-lyon.fr/patrick.flandrin/emd.html
- **F0 estimation algorithms (TEMPO, NDF)**, by H. Kawahara (the algorithms are not publicly available – obtained by contacting the developer)
- **KDE Toolbox**, by A. Ihler and M. Mandel, (Matlab and C files) http://www.ics.uci.edu/~ihler/code/kde.html
- **LASSO path** determination (Matlab function) by K. Skoglund, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897
- **LIBSVM Toolbox** (version 2.9.1), by C-C. Chang and C-J. Len, *C++ code accessed through a Matlab interface*, property of the LIBSVM developers
- **PRAAT software package (Praat: doing phonetics by computer)** version 5.2.26, *C++ code accessed through a Matlab interface developed by M. Little*, by P. Boersma and W. Weenink, http://www.fon.hum.uva.nl/praat/
- **R software package** (The R project for statistical computing) *accessed through a Matlab interface*, http://www.r-project.org/
- **SHRP (pitch determination algorithm)**, (Matlab function) by X. Sun, http://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm
- **Spider Toolbox** (version 1.71), by J. Weston, A. Elisseeff, G. Bakir and F. Sinz http://people.kyb.tuebingen.mpg.de/spider/main.html

- **SWIPE (pitch determination algorithm)**, by A. Camacho http://www.cise.ufl.edu/~acamacho/publications/swipep.m
- **Voicebox Toolbox** by M. Brookes, (Speech processing toolbox for Matlab), http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- **Weka software package** (version 3.6.4), Machine Learning group at the University of Waikato, *originally programmed in Java and accessed through a Matlab interface* http://www.cs.waikato.ac.nz/ml/weka/
- **YIN** algorithm for estimating the fundamental frequency by A. de Cheveigne http://audition.ens.fr/adc/

This study led to the development of two Matlab toolboxes, which we may make freely available in the future:

- **Speech disorders toolbox**: This is mainly a collection of heavily annotated *.m files and a few *.mex files which compute the dysphonia measures explained in detail in Chapter 3 of the thesis. All the functions have typical default values which may work well for most problems, but we suggest possible ranges of values for specific parameters over which experienced users may want to experiment.

- **Statistical machine learning toolbox**: This is also a collection of heavily annotated *.m files, including references to the prototype algorithms used therein. These functions are on topics such as data exploration and statistical analysis, supervised feature selection, and statistical mapping of the feature matrix **X** to the response **y**. All the functions have typical default values which should work well for most problems, but we suggest possible ranges of values for specific parameters over which experienced users may want to experiment. Most processes are automated, so that reasonable outputs could be obtained in most cases simply providing the feature matrix **X**, and the response **y**. The algorithmic details for many of these functions are described in Chapter 4 of the thesis. For completeness, some additional statistical tests are included which are not directly relevant to this work.