Automated decision support for intrapartum fetal surveillance

Philip A. Warrick

Doctor of Philosophy

Department of Biomedical Engineering

McGill University

Montreal, Quebec

2010-01-07

A thesis submitted to McGill University in partial fulfilment
of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-66503-9
Our file Notre référence
ISBN: 978-0-494-66503-9

DEDICATION

To the late Dr. Robert Usher (1929-2006), a mentor who introduced me to the field of obstetrical and newborn care.

ACKNOWLEDGEMENTS

I wish to thank my co-supervisors, Robert Kearney and Doina Precup, for their support and encouragement. Their careful editing and promotion of lively discussion was a motivation to strive for excellence. I thank Emily Hamilton for her unfailing enthusiasm for the project and for her clinical guidance from inception to completion. I would also like to thank my daughter Nolwenn for her insatiable and inspiring curiosity. Finally I thank my partner Livia for her love and support, and her encouragement to integrate all of my life and be fully human.

ABSTRACT

Recording of maternal uterine pressure (UP) and fetal heart rate (FHR) during labour and delivery is a procedure referred to as cardiotocography. Delay or failure to recognize abnormal patterns in these recordings can result in a failure to prevent fetal injury. Clinical interpretation has been predominantly visual, creating significant problems of intra- and inter-subject variability, as well as significant debate about its utility due to the low specificity of visual interpretation that contributes to unnecessary interventions (Cesarian sections). Taking a more automated and objective approach, we modelled the UP-FHR signal pair, for the first time, as an input-output system using a system identification approach to estimate their dynamic relation in terms of an impulse response function. We also modelled FHR baseline with a linear fit and FHR variability unrelated to UP using the power spectral density computed from an auto-regressive model. Using a perinatal database of normal and pathological cases, we trained support-vector-machine classifiers with feature sets from these models. We used the classification in a detection process. We obtained the best results with a detector that combined the decisions of classifiers using both feature sets. It detected half of the pathological cases, with very few false positives (7.5%), one hour and forty minutes before delivery. This would leave sufficient time for an appropriate clinical response. These results clearly demonstrate the utility of our method for the early detection of cases needing clinical intervention.

ABRÉGÉ

L'enregistrement de la pression utérine (PU) et de la fréquence cardiaque fœtale (FCF) pendant le travail et l'accouchement est une procédure appelée cardiotocographie (CTG). Une reconnaissance tardive ou erronée des anomalies dans ces données peut augmenter les risques de lésions au fœtus. L'analyse de ces données est principalement pratiquée visuellement; la difficulté de cette méthode, les problèmes liés aux importantes variations intra- et inter-sujets et le risque de faux positifs entrainant des interventions inutiles (césariennes) entretiennent un important débat sur l'utilité de cette technique. Par une approche plus objective et automatique, nous avons pour la première fois modélisée la pression utérine et la fréquence cardiaque en tant que système entrée-sortie utilisant une approche d'identification de système pour estimer leur relation dynamique en termes d'une fonction de réponse impulsionnelle. Nous avons également modélisé le rythme de base de la FCF à l'aide d'une régression linéaire, ainsi que la variabilité de la FCF sans lien avec la PU à l'aide de densités spectrales de puissance issues d'un modèle autorégressif. Ces modèles, appliqués à une base de données périnatale de cas normaux et pathologiques ont servi à l'apprentissage de machines à vecteurs de support (séparateurs à vaste marge) utilisées dans un processus de détection pour la classification du FCF. Les meilleurs résultats sont obtenus avec un détecteur combinant les décisions de classificateurs utilisant les deux ensembles de caractéristiques. Ils ont détectés la moitié des cas pathologiques, avec très peu de faux positifs (7.5%), une heure et quarante minutes avant l'accouchement. Cela laisserait suffisamment de temps pour une réponse médicale appropriée. Ces résultats démontrent clairement l'utilité de notre méthode pour la détection prématurée des cas nécessitant une intervention clinique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Motivation

Although childbirth is a natural process and outcomes are generally good, approximately 1-7 in 1000 babies experience sufficient oxygen deprivation during labor to cause death or brain injury [2,3,15]. During labor the fetus is relatively inaccessible and the clinician must rely upon indirect measures of fetal condition to assess its tolerance to labor. The objective of this monitoring is to detect the fetus at substantial risk of hypoxic injury so that intervention can prevent its occurrence. Today over 90% of labors are monitored electronically with sensors that measure and record fetal heart rate (FHR) and maternal uterine pressure (UP). Delay or failure to recognize abnormal patterns in these recordings can lead to fetal injury.

In fact, multiple reviews of cases with birth-related brain injury suggest that around 50% of such injuries are related to preventable medical errors, most often centering on incorrect analysis of the FHR recording [15,59,67,75]. The financial burden is massive and rising, reflecting the 4.5 million annual births in North America, the frequency of errors and the cost of an individual settlement for a baby with permanent birth related brain injury. The median jury award in single cases involving childbirth reached US$1,000,000 in 2000, and continues to climb [1]. Thus it is not surprising that childbirth healthcare services continue to generate the most frequent malpractice claims and lawsuits as well as the greatest liability exposure and cost of all medical specialties [5]. Furthermore, the debate over obstetrical litigation costs is not merely an American one; even in Canada the

fact that malpractice insurance fees for obstetricians in some provinces are ten-times the average of all health-care disciplines has been called the 'current crisis in obstetrics' [36].

Assessing fetal condition from cardiotocography (CTG), also known as electronic fetal monitoring (EFM), poses many challenges to obstetrical care. Visual interpretation of UP and FHR data has been common since the 1960's when Hon and Lee first demonstrated clinically that heart-rate variability correlated to fetal morbidity [28], but there is much debate surrounding its use due to significant inter- and intra-subject variability in the interpretations. Visual assessment by obstetrical staff has not lived up to expectations: while it has provided benefits to monitoring of labour, especially in reducing fetal mortality, studies show that its use has been associated with unnecessary intervention and increased Cesarean-section rates [37]. Fundamentally, the FHR captures only a portion of the relevant fetal-health information, so incorporating it into decision-making is an inherently probabilistic endeavour.

Given the paucity of data regarding the precise correlation between the degree and duration of abnormal FHR patterns and hypoxic injury, it is not surprising that clinicians are uncertain about when to intervene. Faced with the extremely debilitating effects of delayed intervention and against the backdrop of potential litigation, there is a tendency towards early intervention, and consequently, a corresponding escalation of the Cesarean section rate.

1.2 Problem Areas

The application of modern computational techniques and hardware to this problem opens the possibility of automatic and therefore consistent real-time assessment of the fetal state. This assessment could be used as an adjunct to the information already available to care-givers. This is an inherently multi-disciplinary problem which draws from, in addition

to obstetrics, the combined technical expertise of signal estimation, system identification, machine learning and statistics.

Signal estimation addresses the problem of describing signals that exhibit some degree of random behaviour. It can also be posed as forming an opinion about the numerical value of a quantity that cannot be observed directly. Both aspects are relevant to our problem in that compact mathematical representations (models) of the measureable signals are required to capture succinctly their information content and interactions while suppressing noise and other disturbances. On the other hand, the fetal state during labour can be considered a quantity that is not directly observable but must be inferred from available measurements. Given the range of possible fetal conditions during labour and delivery and the indirect nature of CTG as a measure of fetal state, it is reasonable to pose fetal-state assessment as an estimation problem. With these estimates, one can hope to form a system model that approximates the behaviour of the underlying system. If this modelling phase is successful, it is reasonable to expect that inferences about the unobserved (or 'hidden') fetal state may be made by tracking how the system model changes with time.

If the signal interactions mentioned above can be interpreted as stimulus-response phenomena, it is possible to model them as an input-output system. The input and output in the context of fetal monitoring is UP and FHR, respectively. Such input-output modeling is referred to as 'system identification' and there are a number of approaches that aim to construct both linear and non-linear models. Most real-world problems are non-linear in nature although linear models can be adequately descriptive for certain phenomena.

In fact, this thesis focuses on linear system models, since they are able to represent the signals of interest with sufficient fidelity. To our knowledge, the system-identification approach applied to modelling this UP-FHR interaction has not been described in the literature. When we complement the system identification model with model parameters

of the FHR baseline and variability, we obtain a novel overall FHR model that is a good representation of the measured FHR.

This modelling approach is to be distinguished from hypothesis-driven (or 'deductive') modeling where accepted physiological models drive the search for features correlating to those models. Many previous works aiming to describe UP and FHR have adopted such a feature-detection-based paradigm. On the other hand, the 'data-driven' (or 'inductive') approach of this thesis directly models the signals and their coupling and in so doing introduces fewer *a priori* assumptions on the information content of the data. The two approaches should be considered as two poles of a continuous spectrum since hypotheses can provide reasonable starting points for the inductive analysis (e.g. an appropriate timescale for local analysis in a non-stationary environment). In turn, the data-driven and relatively unbiased knowledge can be used to inform hypothesis-driven research [49].

The research area of machine learning concerns itself with the construction of computer programs that automatically improve with experience. This field is currently undergoing a significant surge in multi-disciplinary research and industrial interest due to the advancement of powerful algorithms and hardware in recent years. Many open research avenues remain that build on the foundation now in place. For our problem, drawing from a population of healthy and pathological fetuses, the concise models of CTG described above provide the 'experience' while machine learning provides the statistical and probabilistic tools to assess their discriminating power. The main machine learning tool used in this thesis is a support-vector machine classifier.

1.3 Thesis Objectives

CTG interpretation is a problem where consistently applied policies that lead to improvements of a single percentage point in sensitivity and specificity can have significant impact on the quality of obstetrical care. With this in mind, the goals of this thesis are:

1. to produce models of UP and FHR, accounting for their interactions as much as possible with system-identification techniques. This is analysis that is informed by physiology but primarily driven by data, exploiting the database of the industrial partner of the research, LMS Medical Systems, Ltd.
2. to produce automated, intelligent and timely assessments of the risk of fetal oxygen deprivation from these models that have accuracy and confidence levels as high as the data permits and that compare favourably to and integrate with current clinical assessment policies.

1.4 Statement of originality for co-authored works

This thesis follows the manuscript model and therefore is structured around three journal articles:

- **Journal Article I: Feature Extraction**
  - Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of singular spectrum analysis. *Methods of Information in Medicine*, 46(2):196–201, 2007
- **Journal Article II: System Identification**
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Identification of the dynamic relationship between intra-partum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *IEEE Transactions on Biomedical Engineering*, 56(6):1587–1597, June 2009
- **Journal Article III: Fetal-State Classification**
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Classification of normal and hypoxic fetuses from systems modelling of

intra-partum cardiotocography. *IEEE Transactions on Biomedical Engineering*, October 2009. accepted

As first author of these papers, I did all the data analysis and the writing of the manuscripts. These works were multidisciplinary, however, with thesis co-supervisors Robert Kearney and Doina Precup providing guidance and direction on the data analysis (especially signal processing, modelling and machine learning) and editing contributions. Guidance for the clinical obstetrics sections was provided by Emily Hamilton.

1.5 Overview

This thesis is structured as follows. Chapter 2 contains a literature review of the most relevant physiological and clinical information related to fetal surveillance, followed by a selected review of automated fetal-state assessment techniques. A description of the background methodology to the journal papers is given in chapter 3. The journal papers are given in chapters 4, 5 and 6. The conclusions drawn from the work appear in chapter 8.

CHAPTER 2
Literature review

2.1 Physiological-clinical literature

2.1.1 Neonatal neuropathology

The two main adverse outcomes attributable to oxygen deprivation ('asphyxia') during labour are hypoxic ischemic encephalopathy (HIE) and cerebral palsy (CP). The medical literature includes several similar and interrelated conditions which are described below.

Pediatric neurological mortality/morbidity refers to any childhood neurological condition and is relatively common with an incidence of 50-70 per 1000 live births (see Table 2–1 for a summary of all incidences in this section). Encephalopathy is a collection of signs indicating cerebral injury. A clinical diagnosis requires the presence of at least two of the following 1) abnormal conscious state; 2) feeding or respiratory difficulty of central origin; or 3) abnormal tone often accompanied by seizures. Neonatal encephalopathy (NE) refers to any encephalopathy from any cause occurring in the first 7 days of life, and has an incidence of 1.8-7.7 per 1000 births. Reports attribute about 10 to 30% of NE to birth asphyxia involving hypoxic ischemic encephalopathy [13, 25, 53]. HIE refers to a subset of NE where there is also evidence of birth asphyxia (most reliably defined as the presence of an elevated base deficit in an umbilical artery blood sample) and no other obvious cause such as infection.

Intrapartum asphyxia (described as the presence of an elevated base deficit) occurs in approximately 20 per 1000 births. It is generally asymptomatic. However, when followed by signs of moderate or severe encephalopathy, permanent handicap or death are relatively

common; 20-30% of fetuses with moderate HIE and 72-80% of fetuses with severe HIE experience serious handicap, as shown in Table 2–2. It is generally thought that episodes of hypoxia in excess of one hour are required before such neuropathologic damage responsible for motor and cognitive deficits develop [34].

CP is a group of conditions associated with developmental disorder, and has an incidence of 1.5-2.5 per 1000 births. It results from brain lesions or abnormalities occurring early in development and can be traced to HIE in approximately 10% of the CP population. These insults cause motor impairment that neither resolves nor progresses, characterized by muscle attributes of spasticity, increased tone and reflexes, and altered resistance to stretch. CP can result in delayed acquisition of motor skills, clumsiness, gait/ambulatory difficulties and is often co-occurring with epilepsy and cognitive deficiency. It is usually diagnosed between the ages of 2 and 5 years [68].

Condition	Incidence (/1000 live births)
Intrapartum Asphyxia (BD > 12)	20
Neonatal encephalopathy	1.8 - 7.7 (1.0 due to HIE)
Cerebral Palsy	1.5 - 2.5 (0.15 - 0.5 due to HIE)
Overall neurological mortality/morbidity	50 - 70

Table 2–1: Incidence of neuropathology. BD refers to post-natal umbilical cord arterial base deficit. [63, 68]

HIE Severity	Mortality	Serious handicap
Mild	Near 0%	Near 0%
Moderate	5.6-14%	20-30%
Severe	61%	72-80%

Table 2–2: HIE Outcomes [20, 35, 57]

The costs of neurological impairment of newborns are clearly significant for the individual, their families and for society as a whole. As shown in Fig. 2–1, potentially avoidable events are associated with the injury pathway *asphyxia-HIE-neonatal encephalopathy*. Also

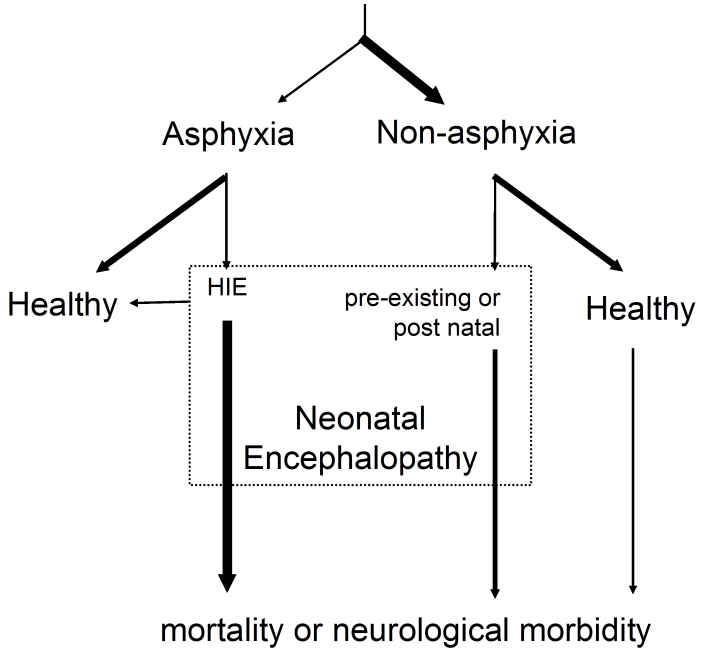


Figure 2–1: Pathways to neuropathology from the fetus in labour and delivery to neonate. Arrow widths indicate relative incidence. The Asphyxia-HIE-Neonatal Encephalopathy injury pathway occurs during labour and is potentially preventable.

note that due to variability in 1) the degree and duration of insult; and 2) the robustness of the fetus, not all episodes of perinatal asphyxia are severe enough to cause injury, as shown by the pathway *asphyxia-healthy*. The pathways to neurological pathology are not restricted to labour and delivery events. Chronic, or underlying pre-labour conditions may cause neurological impairment themselves or render the fetus more susceptible to experience intrapartum oxygen deprivation. The other injury pathway *non-asphyxia-preexisting or post-natal-neonatal encephalopathy* involves conditions that precede or follow labour and delivery (i.e. not intrapartum asphyxia).

However, autopsy and neurological imaging in the early neonatal period indicates that the great majority ($> 80\%$) of neonates with HIE have only acute injury or intrapartum

lesions without evidence of chronic lesions [13]. Therefore, improving policies of care where there are potentially preventable adverse outcomes is clearly a worthwhile effort.

2.1.2 Intrapartum asphyxia

Biochemical mechanism

Biochemically, intrapartum asphyxia can be seen as a gradual diminishing of the buffering response to rising acidity, due to impaired gas exchange. Fig. 2–2 shows the equilibrium equation of the main buffering action involving bicarbonate ($HCO3^{-}$), and how it can be disrupted. As tissue and blood CO_2 levels rise (i.e., a rise in CO_2 concentration on the left side of the equation), the primary buffering agent bicarbonate is consumed, in the presence of the catalyst enzyme Carbonic anhydrase, to maintain pH levels (i.e., to reduce H^+ concentration on the right side of the equation). In this context and under conditions of aerobic metabolism, the buffering action may begin to lag behind rising CO_2 levels, and mild respiratory acidosis can occur. But under increased fatigue and lowered O_2 supply, anaerobic metabolism develops, adding lactic acid to the buffering requirements. Under these more extreme conditions, termed metabolic acidosis, the buffering capacity can become overwhelmed. Consequently, the pH falls and the bicarbonate buffer stores become depleted. Thus if fetal respiration is malfunctioning, persistent respiratory acidosis will lead to metabolic acidosis.

$$CO_2 + H_2O \underset{\text{anhydrase}}{\overset{\text{Carbonic}}{\rightleftharpoons}} H^+ + HCO_3^- \qquad \text{(bicarbonate buffering)}$$

$$\twoheadleftarrow\twoheadleftarrow \text{Lactic acid}$$

Figure 2–2: Disruption of buffering during metabolic acidosis.

Etiology

The causes for intrapartum asphyxia fall into three categories: maternal, fetal and placental. All relate to effective O_2 and CO_2 transport; this can be compromised by the maternal or fetal circulatory systems or in the ability of these systems to exchange materials. Maternal factors are outlined in Table 2–3. These can involve the integrity of the circulatory system due to disease or trauma, disorders of the blood, inability of the heart to maintain adequate cardiac output, or cardio-respiratory disease. Factors related to the fetus are shown in Table 2–4. The effectiveness of the umbilical cord is of paramount importance. The cord may be compressed due to 1) insufficient amniotic fluid (oligohydramnios) which normally surrounds and protects the fetus, to 2) cord loops (especially around the neck, termed nuchal loops), or to 3) prolapse, where the cord falls into the birth canal ahead of the fetus. Feto-maternal hemorrhage, fetal blood and fetal heart disorders can also play a role. Finally the placenta, the site of material exchange, can be compromised by separation from the uterus (abruption), insufficient size, lack of vascularization in development, arterio-venal shunting, and obstructions (infarcts).

Category	Condition	Source
Blood circulation	hypotension, hypovolemia	hemorrhage, vasovagal attack, drugs (eg. epidural), vasoconstriction
Blood disorders	O_2 transport compromise	anemia, hemoglobinopathy
Heart	heart failure	coronary artery disease, valvular dysfunction, malformations
Lung	O_2 uptake compromise	asthma, pneumonia, pulmonary embolus, pulmonary edema

Table 2–3: Maternal factors contributing to intrapartum asphyxia

Risk factors and markers

During the prenatal period, there are numerous risk factors associated with intrapartum asphyxia. On the maternal side, bleeding, lowered O_2 or blood-pressure levels

Category	Condition	Source
Blood circulation	cord compression	oligohydramnios, knots, prolapse, nuchal loops
	feto-maternal hemorrhage	vascular malformation, micro-vascular leakage across the placental villus, trauma
Blood disorders	O_2 transport compromise	anemia, hemoglobinopathy
Heart	heart failure	malformation

Table 2–4: Fetal factors contributing to intrapartum asphyxia

can indicate inadequate maternal blood supply to the placenta. Fetal indicators can include growth restrictions (measured by ultrasound), poor response to tactile stimulation or vibro-acoustic stimulation, and meconium passages. Measurement of FHR is normally non-invasive except in the cases when it is derived from a fetal-scalp electrode. Measured ECG (especially the ST segment [12, 54, 62]) and direct measurement of oxygen saturation via pulse oximetry [6, 22] require intra-uterine access to the fetus and provide diagnostic information, but do not have widespread acceptance in clinical practice. The post-natal infant can undergo more direct examination. Apgar scores are routinely calculated from qualitative assessments (presence of breathing, muscle tone, reflexes, etc.) to give an initial overall indication of the neonatal vigour.

Markers of neonatal encephalopathy and measurement of arterial umbilical cord pH and base deficit give a more specific indication of neuropathology and a more quantitative measure of exposure to hypoxia. The presence of neonatal encephalopathy and umbilical cord base deficit will be used as markers of outcome in this thesis. Bicarbonate buffer concentration is readily calculated using the bicarbonate buffering equation of Fig. 2–2 using measurement of the blood H^+ concentration (1/pH) and the partial pressure of CO_2 (H_2O concentration is considered constant). 'Base deficit' refers to the reduction in bicarbonate concentration compared to normal conditions. There is a growing body

of evidence that the threshold of metabolic acidosis beyond which newborn morbidity or mortality may occur is a base deficit of 12-16 mmol/L [37].

2.1.3 Overall circulatory responses to hypoxia

In the laboratory animal studies, fetal hypoxia can be induced by either reducing the maternal blood oxygenation (hypoxemia) or by reducing the umbilical cord blood flow by maternal vascular occlusion (e.g. of the uterine artery) or by occlusion of the fetal vessels in the cord itself. All have the effect of reducing O_2 delivery to the fetus, but there are different redistribution patterns of the fetal blood to maintain or increase O_2 delivery to the vital central organs (the heart, brain and adrenals) in response to the stress. In the case of maternal hypoxemia, fetal blood flow to the central organs is increased to the degree that O_2 delivery is also markedly increased, while for other organs and the periphery (muscle, skin, bone), blood flow and O_2 delivery is diminished. In the case of vascular occlusion, the fetus attempts to maintain central organ O_2 delivery with modest increases in blood flow to the central organs while elsewhere flow is maintained (apart from the lungs) but with reduced O_2 delivery [64].

Extreme conditions of sustained acute asphyxia (e.g. by complete cord compression) can only be tolerated for about four minutes, after which severe acidosis leads to the demise of fetal sheep [30]. This estimate of the survival horizon should be seen in the light of typical clinical conditions; cord occlusion of the human fetus during uterine contraction is intermittent and most often partial rather than complete.

These changing fetal circulation patterns are accompanied by other circulatory control mechanisms. While adults respond to acute hypoxia with transient tachycardia, the fetus responds with deceleration. This is followed by hypertension and under very severe conditions of prolonged hypoxia, subsequent hypotension. As hypoxia progresses to acidosis, the intrinsic myocardial contraction rate is reduced while at the same time cardiac efficiency

is impaired; together these compound the impact of the other deceleration mechanisms on cardiac output. These neural mechanisms will be discussed in detail in the next section.

2.1.4 Fetal neural response to hypoxia

Four main neuro-mechanisms govern the fetal response to acute hypoxia. The most important effect of these control mechanisms is transient bradycardia, clinically known as deceleration, which can last from about 15s to several minutes. Vagal (parasympathetic) mechanisms are primarily responsible for the earliest onset of heart rate decrease. The two main afferent pathways are arterial chemo-receptors and baro-receptors which respond to the decreased partial pressure of oxygen in the blood and increased arterial blood pressure, respectively. Both transmit through the brainstem vagal center to efferent nerves acting on the sinoatrial node to depress heart rate, and both responses are enhanced by increased degree and duration of hypoxia. When severe hypoxic conditions exist, direct myocardial depression begins, reducing the intrinsic myocardial cell contraction rate in addition to the effects of extrinsic control mentioned previously. Finally, sympathetic mechanisms play a role in increasing heart rate via chemo-receptors, both at the onset of an episode of hypoxia, as well as during recovery. Their effects are enhanced by increasing hypoxia [51,94]. Figure 2–3 illustrates the major pathways.

A spectrum of deceleration morphologies, therefore, are possible from these combined mechanisms. Short and steep decelerations (referred to as 'variable' by clinicians) are presumably primarily baro- and chemo-receptor mediated while 'late' decelerations are associated slower responses, with myocardial depression causing the most delay. Decelerations accompanied by acceleration at recovery (known clinically as 'overshoot') may be indicative of central nervous system dysfunction [21].

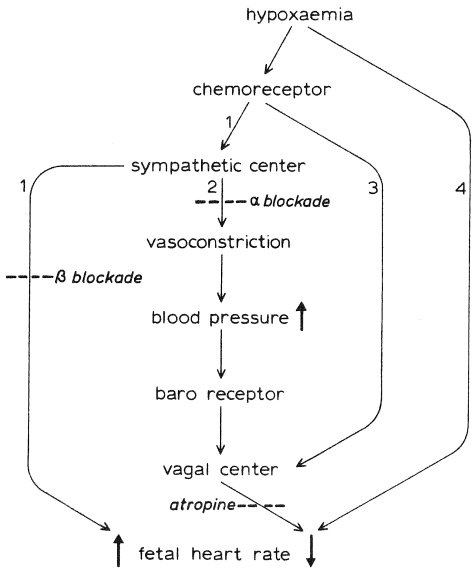


Figure 2–3: Neural responses to hypoxia are chemo- (path 1 and 3) and baro- (path 2) reflex responses and direct myocardial depression (path 4) (from [51]).

2.1.5 Clinical relevance of FHR features

The clinical relevance of some FHR signal characteristics is well understood. The average FHR level, or baseline, reflects the cardiac output; a value within the range 120-160 beats per minute (bpm) indicates that blood delivery is adequate. Small, random fluctuations around baseline (normally 5-15 bpm), known as FHR variability (HRV), indicate that the central nervous system is intact and providing a healthy modulating influence [2, 21].

Temporary decreases in FHR (from 15 s to several minutes in duration and >15 bpm in amplitude) are known as decelerations and reflect events such as compression of the umbilical cord by uterine contractions, malfunction of the fetal heart muscle, or premature separation of the placenta. Generally, larger insults are indicated by recurring episodes of deep, long decelerations whose onsets occur late with respect to the uterine contractions. Temporary increases in FHR (>15 s, >15 bpm), called accelerations, accompany fetal movement and are generally thought to indicate a healthy state.

FHR interpretation in this framework suffers numerous limitations, most notably inter- and intra-subject variability. In addition it only partially reflects the fetal-state and is thus a non-specific descriptor. False positive interpretations result in unnecessary intervention (emergency Cesarean section) and interpretation delays can result in hypoxic injury [15]. Some have questioned its effectiveness compared to intermittent auscultation via stethoscope [76]. Nevertheless its non-invasiveness and perceived sensitivity have made it the standard of care in North-American and European hospitals.

2.2 Scientific-engineering literature

2.2.1 Feature detection

There have been numerous studies in the literature that describe fetal-state assessment based on computerized interpretation of the CTG signal. By far the majority of these have been based on a paradigm of detection and estimation of the clinical CTG features described above. Feature attributes are selected to mirror visual interpretation of the obstetrician and assumed physiological events.

Such CTG analysis is hampered by the fact that the UP and FHR signals are very noisy. This is especially true when the signals are collected under clinical conditions, as is the case for our data. Because the sensors are attached to the maternal abdomen, there is often a problem of sensor contact or missing data when the mother wishes to be more mobile and is temporarily detached from monitoring. These sensor disturbances result in frequent artifacts where the signal drops to a lower value in both the UP and FHR. As well, UP is acquired with a pressure sensor and so is especially sensitive to contact variations. The FHR can also include interference from the maternal heart rate causing the signal to drop to a much lower value. These disturbances pose significant challenges to CTG analysis.

A simple moving-average of the FHR is used most often as a simple estimate of the FHR baseline. The more elaborate rules of Dawes [14] and Mantel [47] are also common since they tend to be less biased by asymmetry of the excursions of accelerations and deceleration (decelerations tend to be deeper). I described another rule-based approach developed at LMS Medical Systems Ltd. which uses a measure of the local variability to set adaptively the allowable excursion about an assumed baseline level. This baseline estimate compared favourably to clinical annotation of the baseline signal [80].

Accelerations and decelerations are frequently detected and measured as is the temporal relationship of decelerations to uterine contractions. Again the rules of Mantel [48] were one of the first standards in the field. They used fixed thresholds of depth and duration of excursions around their baseline estimate as decision criteria. The superiority of non-linear classifiers to thresholding approaches has been shown with several neural-network-based systems. Rosen [61] used a 10min window of raw FHR (300 samples) as a feature vector and saw improved performance of a multi-layer-perceptron (MLP) neural network over a decision-tree classifier. Using raw FHR and measured deceleration features with a recurrent neural-network classifier, Ulbricht [78] reported even better results. I have reported three classifiers that use a bank of band-pass filters to collect candidate decelerations. From these candidates, clinically informed features [80] or discrete-cosine-transform (DCT) features modelling the spectrum [91] were calculated and used to train an MLP neural-network. These results (receiver-operator characteristic area under curve 0.935) were similar to the best results of Ulbricht's recurrent network, except that they were based on a larger dataset (161 cases).

The standard clinical measure of variability–the peak-to-peak deviation about baseline–often forms a simple estimate of the FHR variability in computerized systems. The FHR standard deviation is also used widely. A host of other heart-rate-variability estimates have

been proposed to obtain more information than these simple measurements describe. Spectral analysis is often used, with autoregression (AR) modelling, based on optimal linear Wiener filtering, being most common [9, 72].

It is common in heart-rate variability analysis to use the raw beat-to-beat (or 'RR') intervals rather than the derived heart-rate which is a zero-order sampling of the RR interval. Some samples are delayed and some are repeated by this sampling, adding jitter to the FHR measurement and having the effect of low-pass filtering in the spectrum. Numerous researchers have pointed out the superiority of spectral estimates from the RR interval in terms of information content due to the low-pass filtering effect [11, 33].

Variability has also been estimated by wavelet spectral analysis in order to obtain the frequency response at non-uniform resolution [32, 66, 96], appropriate for the multi-resolution nature of the FHR frequency content. The response of heart-rate variability to contractions has also been investigated [55, 60]). Of particular interest to this thesis, Romano found that there were FHR responses to contractions at frequencies quite different from the contractions themselves, and with frequency-dependent lags; in other words, a non-linear input-output response.

More recently, drawing from the well-developed literature in chaos and adult heart-rate variability, various measures of entropy have appeared which attempt to measure the divergence from regularity or determinism of the FHR signal. Pincus first pointed out the potential for distinguishing acidotic from non-acidotic fetuses from the approximate entropy measure [58]. Since then it has been used in many studies to distinguish normal and pathological conditions (e.g. [50, 97], with researchers at the Politecnico di Milano being most active (Signorini 1989-2006 [4, 9, 18, 19, 38, 41–46, 69–73]).

2.2.2 Models

Apart from the prevalent use of AR modeling for FHR, other related models have also been proposed. Cazares used a 10-th order AR model of UP [8] to improve contraction detection by obtaining better estimates of the signal energy in the known contraction frequency band. An earlier, oft-cited study [29] proposed a multi-stage model of FHR variability including a 2-pole AR model, a second-order non-linearity, and a first-order difference noise term to simulate the jitter of FHR resampling from RR. Then they used an extended Kalman filter to obtain maximum likelihood estimates of the 7 model parameters from 1 min of data. This study is significant because it demonstrated 1) the usefulness of non-linear autoregressive models of variability; and 2) a model that was flexible enough to account for the distortion of the RR to FHR resampling. See Fig. 2–4 for a schematic of this arrangement.

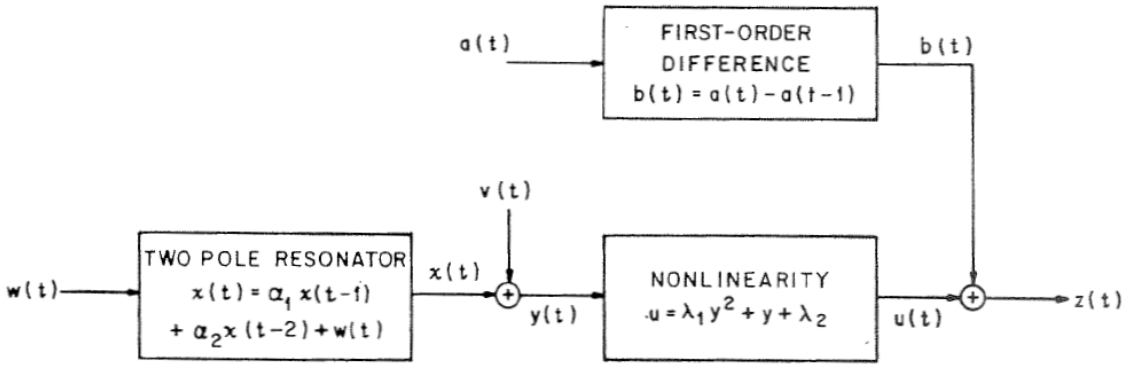


Figure 2–4: A non-linear model of heart-rate variability consisting of a two-pole AR model, a second-order static non-linearity, and noise sources, included FHR resampling jitter noise, modelled as a first-order difference. The models are driven by independent Gaussian white noises $a(t)$, $w(t)$ and $v(t)$. From Jarisch et. al 1980 [29].

2.2.3 Classification

Attempts to classify the fetal state based on feature detection have used domain-specific rules [10, 16], standard linear methods such as linear and logistic regression [7],

and linear-discriminant analysis [26]. However, linear or rectangular partitioning of a multidimensional feature space is often insufficient when applied to the complex boundaries of most classification problems. Non-linear methods are therefore more common: multi-layer perceptron neural networks, fuzzy-logic systems and support-vector machines have all been applied to this problem. However, these methods are usually static; they do not include a notion of how recent events in the past affect the present (i.e. state dynamics), which is very likely important to model in an evolving non-stationary environment such as labour. Accordingly, hidden-Markov models, which have seen widespread use in speech recognition because of their ability to infer state (words or phonemes) from observations (acoustical measurements), have also been to applied to dynamic fetal-state estimation [23].

Table 2–5 summarizes selected automated fetal-surveillance research that addresses both the feature and classification problems as described in previous paragraphs. Several issues are apparent upon examination of these efforts. One is that performance measures often do not meet the minimum requirements for clear conclusions to be made since results are reported without explicit use of independent test data [7, 10, 65]. Secondly, the criteria for outcome class labeling can be quite subjective (e.g. clinician assessment of the signal [7,31,74]) or far from a gold standard (e.g. fixed rules [40]). In other cases, representative examples from entire classes of the population are removed making the performance measure of questionable relevance (e.g. [38] removes all pathological cases except for intrauterine growth retardation (IUGR) pregnancies and discriminates these from normal). In the majority of these efforts, a paucity of data is apparent (and generally recognized by authors); sample sizes of less than 100 cases are common. The larger databases in the studies of Elliott [17] and Jezewski [31] are exceptions in that regard; consequently their results should be more accurate. Given the difficulty of the problem of CTG interpretation

for clinicians, it seems probable that the 'experience' of larger databases will be required for machines to distinguish between the patterns of normal and abnormal fetal labour.

2.2.4 Thesis approach and rationale

Drawing from this rich backdrop of accumulated research, this thesis will:

- use outcome-class labelling that is as objective as possible. The CTG database provided by LMS Medical Systems, Inc. satisfied this objective since it includes as fetal attributes the base deficit and signs of neonatal neuropathology. These are some of the best early neonatal indicators of fetal hypoxic distress.
- be *data-driven*: model the data as much as possible rather than choosing features *a-priori* and in so doing extract as much of the information content as possible.
- obtain the most informative *input-output* models from UP and FHR. While coarse measures of UP-FHR interaction have been described, this thesis is the first to propose a system-identification approach to the problem.
- report performance using an independent test set, using multiple k-fold cross-validation simulations whenever possible so that all performance estimates also include confidence levels.

Reference	Data						Classifier	Results			
	Selection	Normal	Abnormal		Total	Indep. Test?		Sens.	Spec.	Accuracy	ROC AUC
		#	#	Criteria							
Elliott 2008 [17]	intrapartum 3 hrs	2132	60	BD≥12mmol/l, neonatal encephalopathy	2182	✓	Clinical Rules				0.83
Jezewski 2007 [31]	intrapartum >20min	434	251	ObGyn experts	685	✓	NN	84%	84%		
Tsoulos 2006 [77]	intrapartum >30min	130	30	$pH < 7.1$	160	✓	NN	63%	87%		
Salamalekis 2006 [65]	intrapartum, 2nd stage	24	18	Non-reassuring FHR, $pH < 7.2$	42		Normal densities	72%	59%		0.66
Cao 2006 [7]	intrapartum >30min	104	44	ObGyn experts	148	?	LogReg				0.84
Signorini 2005 [38]	antepartum wk30-35	25	35	IUGR	60	✓	SVM	-	-	78%	
Georgoulas 2004 [23]	intrapartum 20min	20	16	$pH < 7.05$	36	✓	HMM	85%	81%		
Skinner 2000 [74]	intrapartum 15min	?	?	ObGyn experts	95	✓	Fuzzy	86%	31%		
Maeda 1998 [40]	intrapartum 50min	7	42	Manual Rules	49	✓	NN			86%	
Chung 1995 [10]	High-risk, >3hr	65	8	$pH < 7.15$ or $BD > 8$	73	?	Clinical Rules	76%	82%		

Table 2–5: Selected automated fetal-state estimate systems: classification from feature detection. Abbreviations: LogReg = logistic regression, NN = neural network, Fuzzy = ANFIS Fuzzy Logic, SVM = Support-vector machine, HMM = Hidden-markov model, BD = base deficit, IUGR = intra-uterine growth retardation

CHAPTER 3
Methodological background

3.1 Database of CTG signals

This thesis used a database of intrapartum CTG recordings provided by LMS Medical Systems, Inc. The cases in this database were screened to include only pregnancies having a birth gestational age greater than 36 weeks and having no known genetic malformations [27]. Only records with at least 3 hours of recording were considered. All data was provided in compliance with institutional regulations.

We labelled each recording by outcome according to its arterial umbilical-cord base deficit and neonatal indications of neurological impairment. An elevated base deficit measurement is an important indicator of metabolic acidosis large enough to cause neurological injury [37, 39, 56]. The majority of the recordings were from normal fetuses (base deficit $< 8\,\text{mmol/L}$); the rest were severely pathological (base deficit $\geq 12\,\text{mmol/L}$, death or evidence of hypoxic ischemic encephalopathy). The proportion of pathological cases was much higher than their natural incidence [2, 3, 15].

Data collection was performed by clinicians using standard clinical fetal monitors to acquire the CTG. The monitors reported at uniform sampling rates of 4 Hz for FHR (measured in beats per minute (bpm)) and 1 Hz for UP (measured in mmHg), which we up-sampled to 4 Hz by zero-insertion and low-pass filtering. In the majority of cases, the UP or FHR sensors were attached to the maternal abdomen; the FHR was acquired from an ultrasound probe and the UP was acquired by tocography. In a few exceptional cases, they were acquired internally via an intra-uterine (IU) probe and/or a fetal-scalp electrode.

UP acquired by an IU probe is considered the gold-standard measurement. Tocography, on the other hand, is not precisely calibrated; its amplitude depends on the pressure sensor contact and abdominal tissue thickness [79]. Additionally, tocography underestimates the duration of contractions [52]. As a result, the uterine contraction onset in the UP signal is delayed with respect to that acquired by an IU probe.

3.2 Introduction to system identification

It is well known that the primary physiological mechanisms for FHR decelerations are 1) contraction-induced umbilical-cord compression and 2) contraction-related decreases in oxygen delivery through an impaired utero-placental unit. Furthermore there is a general consensus that deceleration frequency and timing with respect to contractions can be an indicator of the ability of the fetus to withstand these types of insults. Hypothesis-driven modeling from these facts would focus on contraction-deceleration detection and gross estimates of timing between these events. However, it is also possible to redirect attention to the interactions between the UP and FHR signal pair, which can be viewed as the maternal stimulus and fetal response. It would be natural to model this signal arrangement as an input-output system where the fetus is the 'system' that senses the UP and reacts with changes to the FHR. In this way, contractions, decelerations and their temporal relationships (and possibly other phenomena) are implicitly rather than explicitly modeled.

This investigation studied the feasibility of generating input-output models of UP and FHR, as shown in Fig. 3–1, using linear system-identification techniques. We used the CTG database to generate models of the system impulse response over several hours of data collection. We analyzed the evolution of these models over time and compared models of normal and pathological cases. The ultimate purpose of the model is to describe

more fully the maternal-fetal interaction available via CTG. The input-output modelling procedure will be described below.

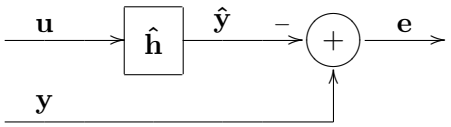


Figure 3–1: General linear input-output model

In the general depiction of system identification in Fig. 3–1, the estimated system $\hat{\mathbf{h}}$ responds to the input UP signal $\mathbf{u}$ and forms an estimate of the output FHR $\hat{\mathbf{y}}$. When the estimate is imperfect, the residue signal $\mathbf{e}$ is generated, where $\mathbf{e}$ is defined as the difference between the true and estimated FHR signals $\mathbf{y} - \hat{\mathbf{y}}$. A quality measure of the estimated model is the *percent variance accounted for* (%VAF) defined as

$$\%\text{VAF} = 100 \times \left(1 - \frac{\sigma_e^2}{\sigma_y^2}\right) \tag{3.1}$$

where σ_e^2 and σ_y^2 are the variances of the residue and desired signals, respectively. Lower residual energy thus corresponds to higher %VAF values.

The fact that the variables are shown above in bold indicates that they are vector quantities. Indeed, a linear *dynamic* model of the finite-impulse-response (FIR) type is one in which the output estimate $\hat{y}_n$ at sample n, is a function of M previous inputs $\mathbf{u}_n^T = [u_n\, u_{n-1}\, \ldots\, u_{n-M+1}]$. The model is dynamic in the sense that it depends on the recent history of inputs. Ideally the desired output y_n can be written as a linear convolution of the inputs with some unknown vector $\mathbf{h} = [h_0\, h_1\, \ldots\, h_{M-1}]^T$:

$$\begin{aligned} y_n &= \mathbf{h} * \mathbf{u}_n \\ &= h_0 u_n + h_1 u_{n-1} + \cdots + h_{M-1} u_{n-M+1} \end{aligned} \tag{3.2}$$

If it is assumed more realistically that the output is corrupted by the measurement-noise signal $\mathbf{v}$, the measured output can be written as $\mathbf{z} = \mathbf{y} + \mathbf{v}$. Better estimates of $\mathbf{h}$ can be calculated over a time period with multiple output samples, written succinctly in matrix form for an analysis of N input and output samples as $\mathbf{z} = \mathbf{U}\mathbf{h} + \mathbf{v}$ where

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} u_1 & 0 & \dots & 0 \\ u_2 & u_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_N & u_{N-1} & \dots & u_{N-M-1} \end{pmatrix} \tag{3.3}$$

The least-squares estimate of $\mathbf{h}$ is then given by: [95]

$$\begin{aligned} \hat{\mathbf{h}} &= (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{z} \\ &\approx \boldsymbol{\Phi}_{\boldsymbol{uu}}^{-1}\boldsymbol{\phi}_{\boldsymbol{uz}} \end{aligned} \tag{3.4}$$

where, for $N \gg M$, $\mathbf{U}^T\mathbf{U}$ (the Hessian $\mathbf{H}$) and $\mathbf{U}^T\mathbf{z}$ are approximated by the input autocorrelation matrix $\boldsymbol{\Phi}_{\boldsymbol{uu}}$ and the input-output cross-correlation $\boldsymbol{\phi}_{\boldsymbol{uz}}$, which are readily calculated.

3.3 Introduction to classification with support-vector machines

Consider a set of labelled data with a number of attributes (the *training* set) and another set of data with the same attributes but unknown labels (the *test* set). In machine learning, *classification* refers to the problem of inferring the unknown labels of the test set, given the training set. Fig. 3–2 illustrates this for our problem domain of inferring fetal state from CTG signals. The training database consists of n UP-FHR signal pairs, which are described by the attributes of some modelling process. In our case, these attributes are derived from the system identification model $\hat{\mathbf{h}}$ parameters described above. With these attributes and known fetal outcomes, the learning procedure determines a classifier that

maps attributes to decisions. This classifier can then be used to form decisions about the test data for which the label is unknown (for example, in an intra-partum clinical context).

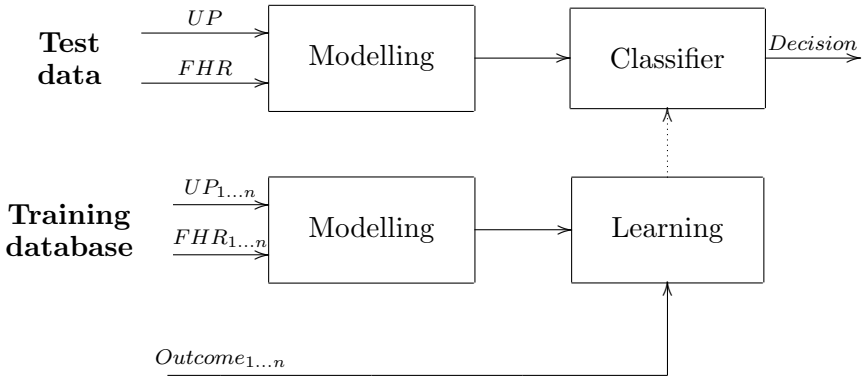


Figure 3–2: Classification applied to inferring fetal state from CTG signals

We chose support-vector machine (SVM) classification because it gives state-of-the-art performance, is robust to overfitting, and computationally efficient implementations are available. SVM learning algorithms process labelled data to determine an optimal decision boundary. This boundary is represented using a subset of the data called the support vectors. The decision boundary can be linear or nonlinear, depending on the kernel chosen. We used a Gaussian kernel in this thesis to allow a nonlinear boundary. SVMs solve the non-linear optimization problem of finding a hyperspace boundary between classes in a given dataset such that the minimum distance of any class example from the boundary is maximized. Fig. 3–3 shows an example of SVM classification.

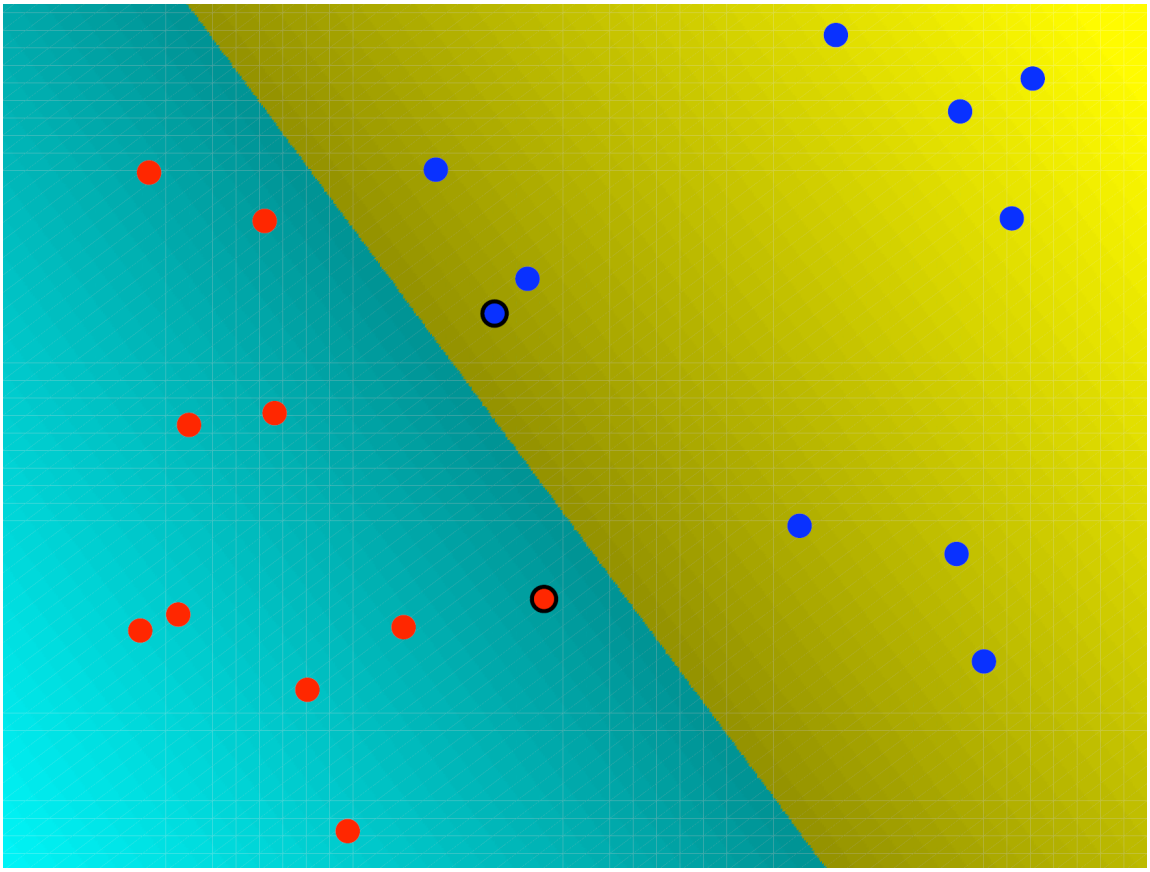


Figure 3–3: Example of SVM classification of two classes (red and blue) using a linear kernel. The gold and turquoise decision regions (corresponding to the red and blue classes, respectively) are determined during learning. The two support vectors closest to the boundary are outlined in black.

CHAPTER 4
Journal article I: Feature extraction

Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of singular spectrum analysis. *Methods of Information in Medicine*, 46(2):196–201, 2007

4.1 Introduction

The focus of this paper is detection of FHR decelerations using a feature extraction approach called singular spectrum analysis (SSA). SSA decomposes the FHR signal by the Karhunen-Loève transformation to obtain a signal representation containing the main eigenvectors with the majority of the signal energy. When this representation changes, a *change-point* is considered to have occurred (in our case, a deceleration). We modified the SSA algorithm to use the discrete cosine transform to approximate the KL transform in far less computational time. As well, to adapt the algorithm to episodes of repeating decelerations, we retained the signal representation that existed before the episode to detect the full time extent of the decelerations. We referred to this modification of the SSA algorithm as *base-hold* SSA and show that it improved deceleration detection.

This paper exemplifies approaches where the focus is on detection of features considered clinically important. By modelling the signal by SSA, however, this approach does not simply mimic the signal analysis of the human expert, as many studies have done, but represents the key characteristics of the signal in a more objective and orthogonal way.

Fetal Heart Rate Deceleration Detection using a Discrete Cosine Transform Implementation of Singular Spectrum Analysis

Philip A. Warrick[1,4], Doina Precup[2], Emily F. Hamilton[3,4], and Robert E. Kearney[1]

Abstract

We present a method for decomposing a signal into near-orthogonal components via the discrete cosine transform (DCT) and apply this in a novel way to change-point detection based on singular-spectrum analysis (SSA). The algorithm is applied to fetal heart rate (FHR) monitoring to improve the detection of deceleration events.

I. INTRODUCTION

FHR monitoring provides an important diagnostic signal for the assessment of intra-partum fetal condition. The FHR manifests several patterns that have postulated physiological significance (e.g. variability, accelerations and decelerations) and visual recognition of these patterns is an important skill of an obstetrician. Signal analysis of the FHR can be used to automate detection of these patterns although the presence of noise and artifact pose significant challenges to success. This study will focus on the deceleration event, a momentary decrease in the FHR that may last from 15s to 5 min. Detection of these events is complicated by their varying length, morphology, and degrees of background variability as well as the distracting presence of artifactual events.

Signal decomposition into orthogonal components is a standard approach for generating compact representations of non-stationary signals. Singular-spectrum analysis (SSA) is one such method that has been known since the original paper of Broomhead and King [1], but which has only recently been

Manuscript submitted May 23, 2005. Corresponding author: Philip Warrick (e-mail: philip.warrick@mcgill.ca). The authors acknowledge the financial support of this work by LMS Medical Systems, Inc.

[1]Biomedical Engineering Department, McGill University, Montreal, Quebec, Canada. [2]School of Computer Science, McGill University, Montreal, Quebec, Canada. [3]Department of Obstetrics and Gynecology, McGill University, Montreal Quebec, Canada. [4]LMS Medical Systems, Inc., Montreal, Quebec, Canada.

applied in a sequential manner suitable for on-line analysis [2]. The SSA technique uses the Karhunen-Loève transform (KLT) to determine the main structure of the underlying signal. By restricting the signal representation to the eigenvectors containing the majority of the signal energy, a compact signal approximation can be extracted.

As their name implies, change-point detection algorithms attempt to detect the location of changes in the characteristics of time-series data. Many such algorithms exist, but they have often been applied to simple parametric models [3]. In [2] an on-line form of the non-parametric SSA change-point detection algorithm is presented. The SSA produces an L-dimensional hyperplane from M-dimensional time-series data ($L \leq M$); the distance of the M-dimensional vectors generating the model (and which form the correlation matrix $\mathbf{R}$) projected onto the L-dimensional hyperplane should be small. The change-point detection is obtained by projecting recent data (i.e. data not used to generate the model) and applying a decision rule to the distance from the hyperplane.

Electro-cardiogram (ECG) ST-segment analysis has used KLT in numerous settings to assess cardiac condition. In [4] the KLT was used to characterize the ST segment of 2-channel ECG signal to detect ischemic episodes. They use the first five KL coefficients as a feature vector whose trajectory over time is monitored with reference to a mean feature vector. Stepwise transitions in the trajectory are considered normal while smooth changes indicate a possible ischemic ST episode. In [5] the T-wave is characterized over time by the first KL coefficient to detect sleep apnea.

This paper will briefly present the key concepts underlying KL decomposition. An efficient discrete cosine transform (DCT) implementation of the on-line SSA algorithm will be described and applied to the classification of deceleration candidate events from a database of FHR tracings (see Figure 1). To the knowledge of the authors, no such DCT-based SSA implementation has been reported in the literature.

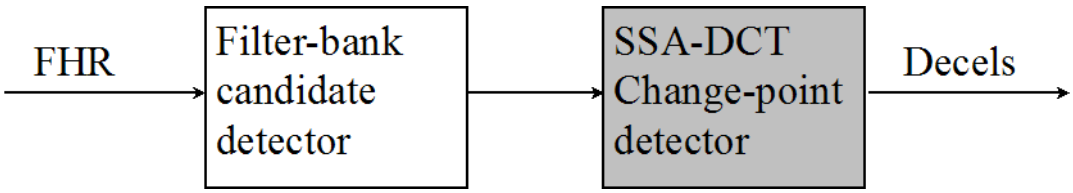


Fig. 1. Block diagram of overall processing. The second shaded block is the focus of this paper.

II. BACKGROUND

A. KL decomposition

In general, given any set of orthonormal basis functions ϕ_n that span a linear vector space of dimension N, any function $f(k)$ in that space can be represented as [6]:

$$f(k) = \sum_{n=0}^{N-1} \theta_n \phi_n(k), \qquad 0 \leq k \leq N-1 \tag{1}$$

where the spectral coefficients θ_j are given by the inner product

$$\theta_j = \sum_{k=0}^{N-1} f(k)\theta_j^*(k), \qquad 0 \leq j \leq N-1 \tag{2}$$

It is a well-known result that the KLT finds the set of basis functions that best represent the signal in a mean-squared error sense. In addition, the decomposition of the signal with these basis functions is such that 1) the mean-square error of a truncated representation is also minimized, 2) it has optimal energy compaction: it contains the most variance (energy) in the fewest number of transform coefficients and 3) it minimizes the total representation entropy of the sequence. The result of the minimization criterion is that:

$$(\mathbf{R} - \lambda_i \mathbf{I}_N)\phi_i = 0, \quad 0 \leq i \leq N-1 \tag{3}$$

where $\mathbf{R} = \mathbf{E}[\mathbf{f}\mathbf{f}^{\mathbf{T}}]$ is the covariance matrix of the input signal f. This is called the eigenvalue problem where λ_i and ϕ_i are the eigenvalues and the eigenvectors of $\mathbf{R}$, respectively. The truncation error of reconstruction is minimized when the eigenvalues are ranked in descending order.

The drawback of the KLT is that the basis functions are signal dependent and therefore cannot be predetermined. In addition, finding the eigenvalues of $\mathbf{R}$ is a computationally intensive task, especially as N becomes large. To overcome these impracticalities, the discrete cosine transform (DCT) is often a viable alternative as it provides a good approximation to the KLT for many signals, while using a fixed set of basis vectors. The DCT transform basis vectors are defined as:

$$\begin{aligned} \theta_k &= \tfrac{1}{c_r} \cos \tfrac{(2k+1)r\pi}{2N}, \quad 0 \leq k, r \leq N-1, \\ \text{where} \quad c_r &= \begin{cases} \sqrt{N} & r = 0 \\ \sqrt{\frac{N}{2}} & r \neq 0 \end{cases} \end{aligned} \tag{4}$$

The eigenvalue approximation from the calculation of DCT coefficients over n signal realizations of length N is given by the average of the squared DCT coefficients [7]:

$$\lambda_j(n) = \frac{1}{n} \sum_{i=1}^{n} C_j^2(i), \quad j = 0, 1, \ldots, N-1 \tag{5}$$

where C_j is the j-th DCT coefficient. It has been shown that for a stationary zero-mean, first-order Markov process the DCT is asymptotically equivalent to the KLT as the sequence length increases and as the adjacent correlation coefficient ρ tends to unity [7]. The DCT is therefore near optimal for many correlated signals encountered in practice and fast algorithms exist for its calculation; indeed it is the industry-standard in image and speech transform coding [6].

B. On-line SSA change-point detection

For a time series $x_1, x_2, \ldots, x_N$, let M be an integer called the lag, where $M \leq \frac{N}{2}$ and let $K = N - M + 1$. Define the $M \times K$ Hankel matrix

$$\begin{aligned} \mathbf{X} &= [X_1 X_2 \ldots X_K] \\ &= \begin{pmatrix} x_1 & x_2 & \ldots & x_K \\ x_2 & x_3 & \ldots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & \ldots & x_N \end{pmatrix} \end{aligned} \tag{6}$$

where $x_{ij} = x_{i+j-1}$. $\mathbf{X}$ can be considered as multivariate data with the M rows as characteristics and the K column vectors X_j as observations. Defining the $M \times M$ lag-covariance matrix $\mathbf{R} = \mathbf{X}\mathbf{X}^T$ and calculating its singular value decomposition (SVD), we determine the orthonormal eigenvectors U_i and eigenvalues λ_i, $i = 1, 2, \ldots, M$ such that $\mathbf{RU} = \mathbf{UD}$, where $\mathbf{D}$ is a diagonal matrix with λ_i along the diagonal and $\mathbf{U} = [U_1 U_2 \ldots U_M]$.

The magnitude of the eigenvalue λ_i is equal to the variance in $\mathbf{R}$ spanned by (or energy content of) the eigenvector U_i[6]. In our case, given the orthonormal set of eigenvectors U_i, reconstruction of X_i is possible by summing the product of the spectral coefficients (the inner products of X_i and U_k) and the eigenvectors:

$$X_i' = \sum_{k=1}^{L} (X_i \cdot X_k) U_k, \quad 0 \leq L \leq M \tag{7}$$

If $L = M$ we have perfect reconstruction while for $L < M$ approximations to X_i are formed using the L eigenvectors having the highest eigenvalues. The assumption made by the SSA algorithm is that given an appropriate choice of L, the reconstruction will tend to correspond to signal while the residue energy (for the unused terms $L + 1, \ldots, M$) will correspond to noise. The residual energy e_r is calculated as the difference in Euclidean norms:

$$e_r(X_i) = \|X_i\|^2 - \|X_i'\|^2 \tag{8}$$

The SSA algorithm for change-point detection calculates the basis functions over a base interval and evaluates the residual energy of the reconstruction of another interval called the test interval. The intervals are shifted in time across available samples and the residual energy is calculated for each shift. Specifically, the base and test intervals are used to construct corresponding base and test matrices as follows. For time shift n, the $\mathbf{X}$ matrices have the following form:

$$\underbrace{\begin{pmatrix} x_{n+1} & x_{n+2} & \dots & x_{n+K} \\ x_{n+2} & x_{n+3} & \dots & x_{n+K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & \dots & x_{n+N} \end{pmatrix}}_{K\text{-vector base matrix}} \tag{9}$$

$$\underbrace{\begin{pmatrix} x_{n+P+1} & x_{n+P+2} & \dots & x_{n+P+Q} \\ x_{n+P+2} & x_{n+P+3} & \dots & x_{n+P+Q+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n+P+M} & x_{n+P+M+1} & \dots & x_{n+P+Q+N} \end{pmatrix}}_{Q\text{-vector test matrix}} \tag{10}$$

where P is the position of the test interval relative to the beginning of the base interval and Q is the number of observations taken from the test interval.

The difference metric at sample n is given by the summation of the residual energies of the test matrix vectors

$$D_n = \sum_{i=n+P+1}^{n+P+Q} e_r(X_i) \tag{11}$$

where the basis functions U_i for the reconstruction are calculated from the base matrix. This difference metric is then adjusted by an estimate of the underlying noise energy, taken as the difference metric of the base matrix where the last null hypothesis $\mathbf{H_0}$ (no change) occurred: i.e. m is chosen as the largest value of $m \leq n$ under $\mathbf{H_0}$:

$$S_n = \frac{D_n}{v_n} \quad \text{where} \quad v_n = \sum_{i=m+1}^{m+K} e_r(X_i) \tag{12}$$

If v_n is considered an estimate of the variance of the noise at sample m (where it is assumed that there is very little signal energy), S_n normalizes the distance metric such that it is not depend on this (assumed unknown and possibly non-stationary) quantity.

It is difficult to select a threshold for the change statistic that could apply to an arbitrary sequence. Nonetheless, under some simplifying assumptions, it is shown in [2] that the following broad rule applies:

when the distribution of the change statistic D_n is assumed to be Gaussian, a threshold can be chosen such that the probability of an event under $\mathbf{H_0}$ is approximately α, where $\alpha > 0$ is a significance level. In [2] the correspondence of α to a threshold h of the S statistic is elaborated, and we use this h calculation in the subsequent discussion.

C. FHR database and candidate detector

The database for this study consisted of 15 FHR tracings containing 535 decelerations as marked by an expert obstetrician. We have described in [8] an approach that uses filtering to detect deceleration candidates. The filtering stage consists of a bank of band-pass filters responsive to decelerations over various overlapping frequency ranges. Candidate events, whose extent corresponds to the zero-crossings of a particular band-pass filter, are placed in a competition with overlapping events generated by the other band-pass filters, and only the highest amplitude events survive. This detector detects 528 of the 535 decelerations of the study database, although it also detects 1285 false positives. One of the goals of this study is to improve on the discrimination of the candidate detector.

III. METHODS

A. SSA Change Point Detection: Block DCT implementation

While the DCT implementation provides some reduction of the processing load compared to the KLT, further improvements were made by using a block implementation of the DCT version of the algorithm. Because the DCT eigenvalue approximations are summations of the DCT coefficients of the observation vectors (eqn. 5), a recursive calculation at each time sample is possible [7]:

$$\lambda_j(n) = \lambda_j(n-1) + \tfrac{1}{N}(C_j^2(n) - C_j^2(n-N)), \quad j = 0, 1, \ldots, N-1 \tag{13}$$

Performing the DCT calculations in block fashion produced a more efficient SSA implementation. For one typical FHR segment of approximately 2400s duration, a 40-fold decrease in the SSA processing time was observed with the block implementation (from 1000s to 25s) compared to repeated and highly redundant DCT calculations at one-sample time shifts of the analysis block.

Using several calibration FHR tracings, we observed good agreement between the KL and DCT implementations of the SSA algorithm (see one example in Figure 2). Some higher frequency spikes are better resolved by the KLT, but the responses are otherwise similar.

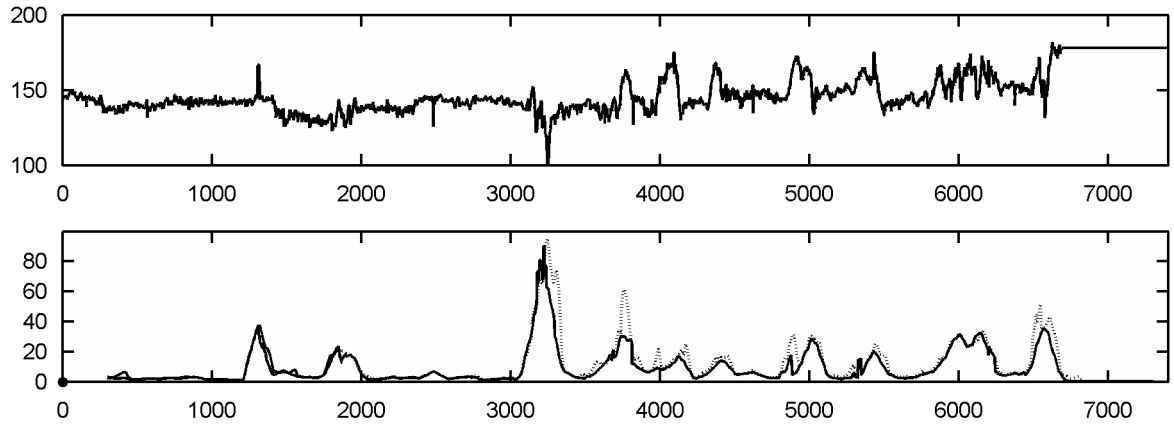


Fig. 2. The lower plot compares the KLT (dashed) and DCT (solid) implementations of the SSA statistic S. The input FHR appears above.

B. SSA Parameter selection

The SSA algorithm is quite parameter-sensitive. Perhaps the single most important parameter for selection is the observation length M. Our preliminary investigations suggested that M should be chosen to be of similar duration to the event of interest. The significance level of the threshold h for the SSA S statistic was set to values close to $\alpha = 0.1$ (as suggested in [2]). The number of observations K needs to be large enough to create a relatively stable base interval representation but not so long as to smooth out distinct regions of quasi-stationarity [2]. The other main choice is the P value: whether to place the test interval within the base interval or following it. Embedding the test interval in the base interval has the advantage of including the before- and after-event signal in the determination of the base matrix, which may be appropriate for the typical deceleration event that is preceeded and followed by some common steady state. On the other hand, placing the test interval after the event should accentuate the difference between the base and test intervals since no test signal is included in the base matrix calculation. We used the latter approach in the experiments that follow. The length of the test interval Q was chosen to be somewhat smaller than M (e.g. $Q = \frac{M}{4}$) and the number of eigenvectors L was kept as small as possible (e.g. $L = 5$) since often the most important changes occur in the highest energy components.

C. SSA on candidate detected FHR

Using the location of previously detected deceleration candidates focused the analysis by matching the M value to the event duration and reducing the search space significantly. Given the set of candidates, local SSA was done in the vicinity of each candidate, using the candidate length for the SSA observation

length M in each case (i.e. as an estimate of the true event length). For larger values of M, the processing burden was excessive and for this reason the data was decimated. The decimation factor was determined by the ratio of the event length L_C to the minimum event length L_M of 15s, i.e. the decimation factor is defined as $D = \text{round}(2L_C/L_M)$, where round$(\cdot)$ denotes rounding to the nearest integer. With the minimum candidate length restricted to 15s, the number of decimated samples $L_D = \text{round}(L_C/D)$ falls into the range of 25 to 37 samples. Thus the results of the analysis are relatively scale-invariant over a broad range of event durations. The SSA analysis begins and ends at fixed factors of M before and after the event, chosen in a way to give sufficient context to the analysis of the event. By experimentation, we found that factors of 16 and 8, respectively, provided enough context without incurring excessive decision delay.

D. Base-hold SSA

The standard SSA algorithm encounters difficulty in intervals where repeated decelerations occur in close proximity. Under these conditions the first events of the series may be recognized, but subsequent events are not since the base and test intervals tend to contain similar energy (i.e. of morphologically similar decelerations) as analysis progresses forward in time. To address this, we modified the algorithm to hold the base interval constant when changes are detected until the next null hypothesis $\mathbf{H_0}$ is encountered.

For these conditions, Figure 3 illustrates how holding the SSA base interval to the last $\mathbf{H_0}$ interval can improve detection. The standard SSA base interval B_S contains an event, reducing its distinction from the test interval T. The base-hold interval B_H, having less low-frequency energy, is retained as the base interval for the successive event candidates, and the energy difference between B_H and T is sufficient to signal a change-point.

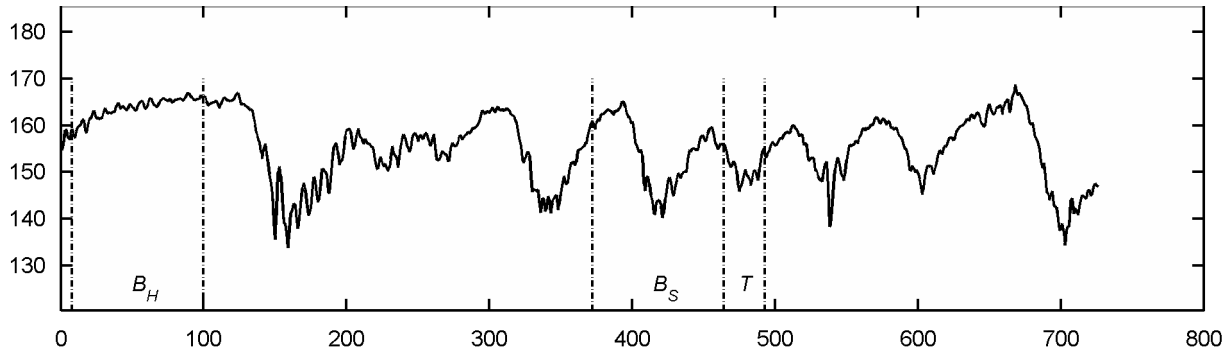


Fig. 3. Example base-hold. For the sixth event of Figure 4, the locations of the base intervals for the standard and base-hold SSA are shown between dotted vertical lines as B_S and B_H respectively. The test interval is indicated by T.

	TP In	FP In	TP Missed	FP Reduction
Standard SSA	528	1285	64	279
Base-hold SSA	528	1285	49	278

TABLE I

SSA-DCT DECELERATION DETECTION RESULTS

IV. RESULTS

The results of the detection performance of the standard and base-hold SSA are shown in Table I. The true and false positives (TP, FP) of the candidate detector are the input to the two variants of the change-point detector. Both detectors reduced the number of false positives by approximately 21.6% (279 or 278 of the original 1285). However, the base-hold approach had 23% fewer misses (49 vs. 64). Thus for equivalent false-positive rates, the base-hold approach shows improved sensitivity over that of standard SSA.

Figure 4 shows an example of repetitive deceleration patterns, which are problematic for standard SSA. The base-hold SSA algorithm detects two decelerations that are missed by standard SSA. Both approaches have the same false positives.

V. DISCUSSION

The standard SSA algorithm has been used successfully in many contexts where changes are infrequent [2]. Our efficient DCT implementation of SSA is of general applicability to these problems, especially where the processing data volume is large and the input signal is sufficiently correlated. In FHR analysis, the assumption of infrequent changes does not always apply and our modifications improve the resolution of single events within a series of similar events. While this has shown to improve overall results on our dataset, it is also conceivable that a series of similar non-events can be wrongly classified as a group as a result of this modification.

Rather than using this change-point detector in isolation in our FHR analysis, we are investigating the incorporation of the SSA statistic into an ensemble of other features within a more comprehensive classifier.

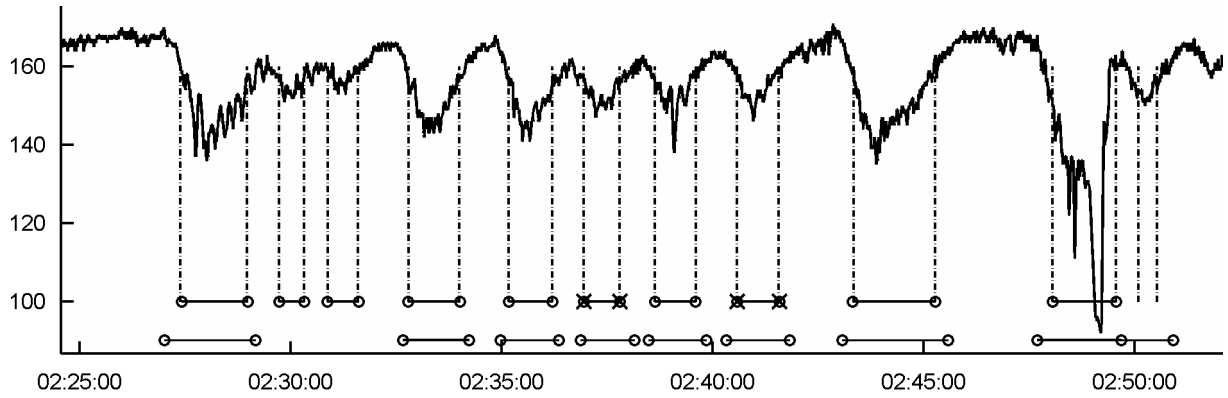


Fig. 4. Example of a region of repetitive decelerations. The dotted vertical lines indicate the extents of candidate detector markings. The lower and upper horizontal bars indicate the extents of the expert markings and SSA markings, respectively. The standard SSA algorithm misses the sixth and eighth decelerations (marked by $\times$) whereas the base-hold approach detects them. Both approaches have the same false positives (the second and third events).

REFERENCES

[1] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D: Nonlinear Phenomena*, vol. 20, no. 2-3, pp. 217–236, 1986.

[2] V. Moskvina, "Application of the singular spectrum analysis for change-point detection in time series," Ph.D. dissertation, Cardiff School of Mathematics, Cardiff University, 2001.

[3] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes: Theory and Applications*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[4] F. Jager, R. Mark, G. Moody, and S. Divjak., "Analysis of transient ST segment changes during ambulatory monitoring using the Karhunen-Loève transform," in *Computers in Cardiology*, 1992, pp. 691–694.

[5] C. Maier, H. Dickhaus, M. Bauch, and T. Penzel, "Comparison of heart rhythm and morphological ECG features in recognition of sleep apnea from the ECG," in *Computers in Cardiology*, A. Murray, Ed., 2003, pp. 311–314.

[6] A. Akansu and R. Haddad, *Multiresolution Signal Decomposition:Transforms, Subbands, Wavelets*, 2nd ed. Englewood Cliffs, New Jersey: Academic Press, 2001.

[7] S. Haykin, *Adaptive Filtering*, 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 2002.

[8] P. Warrick, E. Hamilton, and M. Macieszczak, "Neural network based detection of fetal heart rate patterns," in *Neural Networks, 2005. Proceedings. 2005 IEEE International Joint Conference on*, vol. 4, 2005, pp. 2400–2405.

CHAPTER 5
Journal article II: System identification

Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Identification of the dynamic relationship between intra-partum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *IEEE Transactions on Biomedical Engineering*, 56(6):1587–1597, June 2009

5.1 Introduction

Hypothesis-driven modelling of CTG, based on an understanding of physiology and clinical interpretation, has focused on contraction and deceleration detection in the face of noisy signals, as in the previous paper.

This paper, which represents the core of the thesis, is a novel departure from these feature detection approaches; it focuses on the dynamic relationship between uterine pressure (as an input) and FHR (as an output). We use a system-identification approach to estimate system dynamics in terms of an impulse response function (IRF). Our system-identification approach also incorporates noise suppression to address the challenge of analyzing very noisy CTG signals. Such an approach has not appeared in the literature to date. This resulting model represents very-low-frequency FHR energy (i.e., <0.03 Hz) linearly related to UP and therefore is complementary to other components of the FHR normally studied. We show that the models for normal and pathological cases is our database were quite different, with pathological cases having stronger, more delayed and more predictable responses. While we used a novel modelling approach in this study, these results remained consistent with clinical expectation.

Identification of the Dynamic Relationship Between Intra-Partum Uterine Pressure and Fetal Heart Rate for Normal And Hypoxic Fetuses

Philip A. Warrick[1,4], *Student Member, IEEE*, Emily F. Hamilton[3,4], Doina Precup[2] and Robert E. Kearney[1], *Fellow, IEEE*

Abstract

Labour and delivery is routinely monitored electronically with sensors that measure and record maternal uterine pressure (UP) and fetal heart rate (FHR), a procedure referred to as cardiotocography (CTG). Delay or failure to recognize abnormal patterns in these recordings can result in a failure to prevent fetal injury. We address the challenging problem of interpreting intra-partum CTG in a novel way by modelling the dynamic relationship between UP (as an input) and FHR (as an output). We use a non-parametric approach to estimate the dynamics in terms of an impulse response function (IRF). We apply singular value decomposition to suppress noise, IRF delay and memory estimation to identify the temporal extent of the response, and surrogate testing to assess model significance. We construct models for a database of CTG recordings labelled by outcome, and compare the models during the last three hours of labour as well as across outcome classes. The results demonstrate that the UP-FHR dynamics can be successfully modelled as an input-output system. Models for pathological cases had stronger, more delayed, and more predictable responses than those for normal cases. In addition, the models evolved in time, reflecting a clinically plausible evolution of the fetal state due to the stress of labour.

I. Introduction

Manuscript received June 27, 2008; revised November 27, 2008 and January 19, 2009. First published February 20, 2009; current version published June 10, 2009. This work was supported by the LMS Medical Systems, Inc., and by the Natural Sciences and Engineering Research Council of Canada (NSERC). Corresponding author: Philip Warrick (e-mail: philip.warrick@mcgill.ca).

[1]Biomedical Engineering Department, McGill University, Montreal, Quebec, Canada. [2]School of Computer Science, McGill University, Montreal, Quebec, Canada. [3]Department of Obstetrics and Gynecology, McGill University, Montreal Quebec, Canada. [4]LMS Medical Systems, Inc., Montreal, Quebec, Canada.

THE lifelong disability that can result from oxygen deprivation during childbirth is rare but devastating for families, clinicians and the health-care system. The hallmark indications that significant fetal cerebral hypoxia has occurred during labour are metabolic acidosis and neurological signs such as altered levels of consciousness or seizures. Between 1 and 7 in 1000 fetuses experience oxygen deprivation during labour that is severe enough to cause fetal death or brain injury [2]–[4]; the range of this estimate reflects considerable regional variation and some clinical debate on the definition of brain injury. Unfortunately, non-invasive methods to measure directly the fetal acid-base status and cerebral oxygenation do not exist. Consequently, clinicians must rely upon indirect measures of oxygen delivery and neurological function.

A standard approach is cardiotocography (CTG), which measures maternal uterine pressure (UP) and fetal heart rate (FHR); these signals are the cumulative result of many concomitant physiologies. Visual pattern recognition and inference are the basis of clinical interpretation. However, these are inconsistently applied [5]. Furthermore, classical patterns have low specificity. Because significant hypoxia is rare, false alarms are common, leading physicians to disregard truly abnormal signals. Indeed, approximately 50 % of birth-related brain injuries are deemed preventable, with incorrect CTG interpretation leading the list of causes [4], [6]–[8]. The social costs of such errors are massive: intra-partum care generates the most frequent malpractice claims and the greatest liability costs of all medical specialties [9]. Thus, there is great motivation to find better methods to discriminate between healthy and hypoxic conditions.

Clinicians' interpretation of intra-partum CTG signals relies on the temporary decreases in FHR (FHR *decelerations*) in response to uterine contractions. FHR decelerations are due mainly to two contraction-induced events: 1) umbilical-cord compression and 2) a decrease in oxygen delivery through an impaired utero-placental unit. There is general consensus that deceleration depth, frequency and timing with respect to contractions are indicators of both the insult and the ability of the fetus to withstand it. Fig. 1 shows an example CTG demonstrating this response, where an FHR deceleration follows shortly after the onset of each of the four successive uterine contractions. Hypothesis-driven modeling based on this understanding of physiology and clinical interpretation has focused on contraction and deceleration detection [10]–[15].

In this paper, we propose a new approach that focuses on the dynamic relationship between UP (as an input) and FHR (as an output). Although the FHR is subject to numerous influences (i.e., it is the result of a multiple-input system), UP is the only input that is accessible by external monitoring and routinely used in clinical practice; indeed, clinicians already interpret certain UP-FHR relationships as indications of pathology. We use a system-identification approach to estimate system dynamics in terms of an impulse response function (IRF). This model represents very-low-frequency FHR energy (i.e., <0.03 Hz) linearly

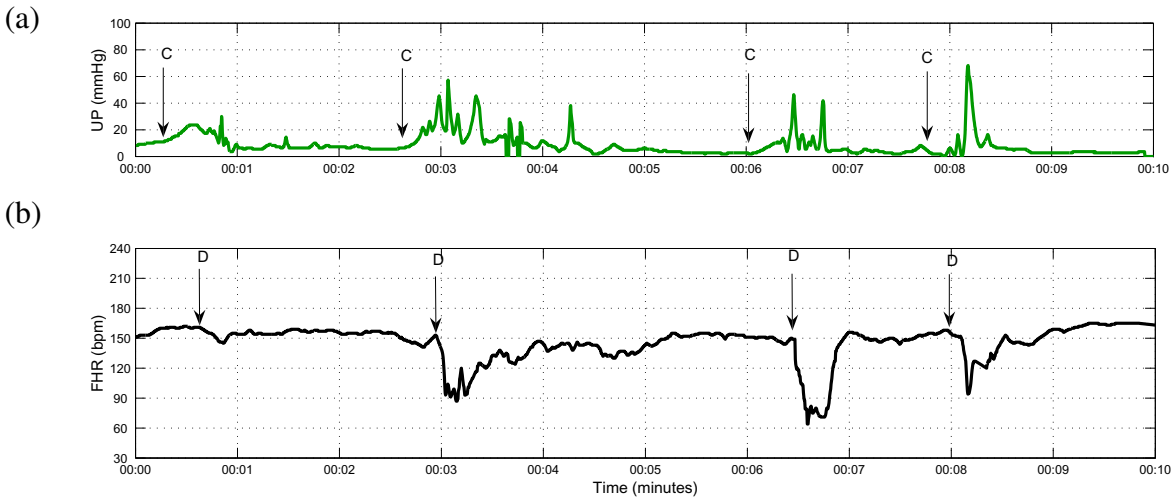


Fig. 1. CTG stimulus-response from 4 contraction-deceleration pairs over 10 min. (a) UP signal with contraction onsets ('C') indicated. (b) FHR signal with deceleration onsets ('D') indicated.

related to UP and therefore is complementary to other components of the FHR normally studied. System identification from CTG is a challenge because 1) measurement disturbances are prevalent; and 2) the UP input has narrow bandwidth, is not calibrated, and is subject to unknown measurement delay.

Using a database of CTG recordings labelled by outcome data available after birth, we compared IRFs over time (during the last three hours of labour) and with respect to their outcome classes (either normal or pathological). We found that our method successfully modelled CTG, and that the models were better for the pathological cases; these models consistently explained more of the output variance. Model parameters were significantly different between the normal and pathological cases, with the latter having IRFs with larger gain and longer delay (indicating a stronger and more delayed FHR response to changes in UP). The models also evolved over time toward stronger and later responses; this progression is consistent with other studies regarding the timing of acute intra-partum hypoxic injury [16]. These results suggest that the model parameters could be useful for the automatic classification of the fetal state (as we have done in other studies [17], [18] and detail in the discussion).

II. Methods

A. Data

We used a database consisting of 264 intrapartum CTG recordings for pregnancies having a birth gestational age greater than 36 weeks and having no known genetic malformations [19]. We only

considered records with at least 3 hours of recording.

We labelled each recording by outcome according to its arterial umbilical-cord base deficit and neonatal indications of neurological impairment. An elevated base deficit measurement is an important indicator of metabolic acidosis large enough to cause neurological injury [5], [20], [21]. The majority of the recordings were from normal fetuses (221 cases: base deficit < 8 mmol/L); the rest were severely pathological (43 cases: base deficit ≥ 12 mmol/L, death or evidence of hypoxic ischemic encephalopathy). This proportion of pathological cases was much higher than their natural incidence [2]–[4].

Data collection was performed by clinicians using standard clinical fetal monitors to acquire the CTG. The monitors reported at uniform sampling rates of 4 Hz for FHR (measured in beats per minute (bpm)) and 1 Hz for UP (measured in mmHg), which we up-sampled to 4 Hz by zero-insertion and low-pass filtering. In the majority of cases, the UP or FHR sensors were attached to the maternal abdomen; the FHR was acquired from an ultrasound probe and the UP was acquired by tocography. In a few exceptional cases, they were acquired internally via an intra-uterine (IU) probe and/or a fetal scalp electrode. UP acquired by an IU probe is considered the gold-standard measurement. Tocography, on the other hand, is not precisely calibrated; its amplitude depends on the pressure sensor contact and abdominal tissue thickness [22]. Additionally, tocography underestimates the duration of contractions [23]. As a result, the uterine contraction onset in the UP signal is delayed with respect to that acquired by an IU probe. Jezewski et. al. [24] compared tocography to uterine electromyography, a measure that is consistent temporally with the IU measurement; they found that in a set of 108 recordings taken within 24 hours of labour, the contraction onset, when measured by tocography, occurred later (mean, standard deviation and maximum values of 14 , 18 , and 80 s, respectively).

B. Overall processing

We modelled UP-FHR system dynamics by linear system identification. The processing steps are shown schematically in Fig. 2. A preprocessing step cleaned and segmented the UP and FHR into 20-min epochs of input and output data ($\mathbf{U}$ and $\mathbf{f}$). Next, non-parametric system-identification (SI) methods were used to estimate the IRF ($\hat{\mathbf{h}}$) and the best values for the IRF delay d and memory M. Finally, the models were validated using a model significance filter. Each step is detailed below.

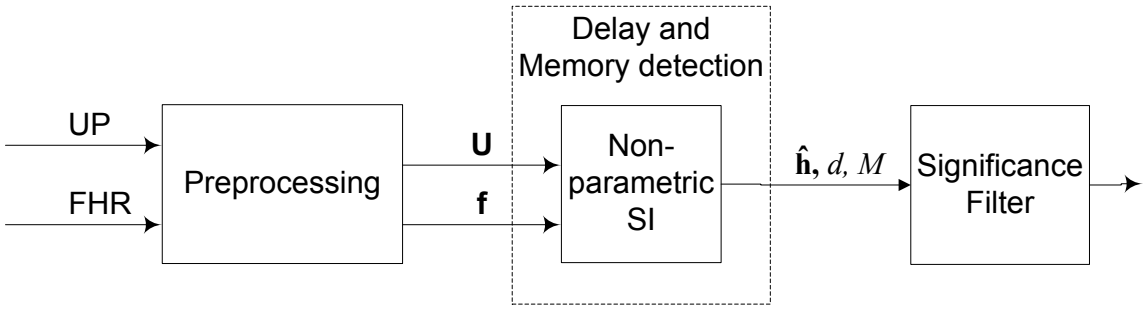


Fig. 2. Data processing for estimation of UP-FHR dynamics. Preprocessing cleans and segments the UP and FHR into 20-min epochs of input **U** and output **f**. Non-parametric system identification (SI) estimates the impulse response function $\hat{\mathbf{h}}$, delay d, and memory M. The significance filter validates the resulting models.

C. Preprocessing

Because it is recorded in a clinical setting, CTG is subject to specific types of noise. The loss of sensor contact can temporarily interrupt the UP or FHR signals, and interference from the (much lower) maternal heart rate can corrupt the FHR. Interruptions appear in the signal as a sharp drop to much lower amplitude followed by a sharp signal restoration. A simple threshold detector for such segments would be prone to oscillate for noisy signals near the threshold; to avoid this, we used a Schmitt trigger, which defines separate detection thresholds for down-going and up-going transitions. We chose thresholds of $\max(\mu - \alpha_i\sigma, 0)$ where μ and σ are the mean and standard deviation, respectively, over the entire signal ; we used $\alpha_1 = 2$, $\alpha_2 = 1$ for the down-going and up-going transitions, respectively. Some UP dropout durations were underestimated by the Schmitt trigger; therefore, we extended dropout segments forward and/or backwards in time if their overall standard deviation remained below a threshold (1 mmHg). We also removed segments that had passed the Schmitt trigger but had negligible slope over a period of more than 15 s. We merged sequential dropout segments in close proximity (≤4 samples). We bridged dropouts lasting less than 15 s by linear interpolation; we removed longer dropouts from consideration. We chose all thresholds empirically to obtain a qualitatively acceptable segmentation of signal and artifact. More objective segmentation methods were outside the scope of this study, but are an area of active research [25], [26].

Besides UP, FHR is influenced by a number of other physiological factors that are unobserved and intrinsic to the fetus (reflected in FHR baseline and central-nervous-system-regulated variability); we did not consider these factors in our input-output model. Also, the UP baseline can fluctuate considerably

depending on the sensor contact. To reduce the impact of these factors, we applied a high-pass filter to remove the baseline, and a low-pass filter to reduce the variability. We detrended the signals by a high-pass filter selected to pass a long contraction or deceleration without incurring excessive filter delay. We chose a filter with a cutoff frequency of $\frac{1}{220s} = 4.5 \times 10^{-3}$ Hz as a compromise between these competing demands. We selected the parameters for low-pass filtering based on two considerations. First, the narrow-band UP energy (i.e., near the contraction rate and $\ll 4$ Hz) determines the bandwidth of the UP-FHR linear interaction. However, we also wanted to retain sufficient time resolution for the delay and memory estimation step (i.e., on the order of seconds). For these reasons, we decimated the signals by an anti-aliasing filter and sub-sampling to 0.5 Hz .

Model quality generally increases with data length. However, we expected UP-FHR dynamics to be non-stationary since the state of the fetus tends to degrade with time. To resolve this trade off, we extracted 20-min epochs with 10-min overlap between successive epochs; this epoch length is much longer than the typical FHR deceleration response to a contraction (i.e., 1-2 min). We extracted as many such epochs as possible starting from the beginning of a clean (artifact-free) segment; to include any remaining data at the end of the segment (i.e., <10 min), the overlap was increased for the last epoch. We estimated models for each epoch.

D. Non-parametric model

We denote the input UP, and output FHR, at time sample n ($n = 1 \ldots N$), by u_n and f_n, respectively. We modelled the response, f_n, to input u_n as the convolution sum:

$$f_n \approx \sum_{i=0}^{M-1} (h_i \Delta t) u_{n-d-i} = \mathbf{h} * \mathbf{u}_n \tag{1}$$

where Δt is the sampling period, and $\mathbf{h}$ is the impulse response function (IRF) beginning at sample d (called the *delay*), and of length M (called the *memory*). We denote by $\mathbf{u}_n$ the length-M vector of input samples $[u_{n-d-M+1}, \ldots, u_{n-d-1}, u_{n-d}]$ used to estimate f_n at sample n. For causal (physically realizable) systems, $d \geq 0$, but in the presence of an input measurement delay (as in our case), d may be negative [27].

Let $\mathbf{U}$ be an $N \times M$ Toeplitz matrix formed from $\mathbf{u}_n$. The least-squares estimate of $\mathbf{h}$ is given by:

$$\hat{\mathbf{h}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{f} \approx \boldsymbol{\Phi}_{uu}^{-1}\boldsymbol{\phi}_{uf} \tag{2}$$

where, for $N \gg M$, $\mathbf{U}^T\mathbf{U}$ and $\mathbf{U}^T\mathbf{f}$ are estimated by the input autocorrelation matrix $\boldsymbol{\Phi}_{uu}$ and the input-output cross-correlation $\boldsymbol{\phi}_{uf}$, respectively [27]. Generally, increasing N improves the correlation estimates and hence improves $\hat{\mathbf{h}}$.

The residual signal $\mathbf{e} = \mathbf{f} - \mathbf{U}\hat{\mathbf{h}}$ was used to generate a figure of merit for the estimated model, in terms of the *percent variance accounted for* (%VAF):

$$\%\text{VAF} = 100 \times \left(1 - \frac{\sigma_e^2}{\sigma_f^2}\right) \tag{3}$$

where σ_e^2 and σ_f^2 are the variances of the residual and observed signals, respectively; lower residual energy corresponds to higher VAFs and better models. VAF calculations excluded the initial M output samples to avoid filter end effects.

Using VAF to compare normal and pathological models might be misleading since difference in FHR variability might confound the ability of VAF to assess model quality. Therefore, we also calculated model VAF after low-pass filtering the pre-processed FHR. We chose the very-low frequency band below 0.03 Hz (VLF, c.f. [28]) since that frequency band contained practically all (99.97%) of the average spectral power of the models; at 0.03 Hz this power had dropped by 35dB, compared to the peak power. We refer to this VAF measure as VAF_{VLF}.

E. Pseudo-inverse for noise reduction

The input u_n was not controlled and was usually a narrow-band signal dominated by the contraction frequency. On the other hand, the FHR was typically rich in frequency content. In performing linear system identification with such an input, frequencies far from the contraction frequency will be hard to identify. In other words, the least-squares problem of (2) may be ill-conditioned; spurious input-output correlations may result in noisy IRF estimates. In this respect it should be noted that FHR variability at frequencies greater than 0.03 Hz will act as noise for the identification procedure since there is no power in the UP signal above this frequency (i.e., FHR variability is either unrelated or non-linearly related to UP). These higher frequency components of FHR are known to contain information that is potentially discriminating [29], [30], which indicates that they may be complementary to our model. In order to reduce the noise for the linear model, we used SVD and retained only the most significant IRF components, as described in [31] and summarized below.

The noise characteristics of the IRF estimate can be analyzed by examining $(\mathbf{U}^T\mathbf{U})^{-1}$ and $\mathbf{U}^T\mathbf{f}$ in (2) separately [32]. First, the Hessian $\mathbf{G} = \mathbf{U}^T\mathbf{U}$ is Hermitian and positive definite, so its SVD is equivalent to calculating its eigenvectors and eigenvalues as $\mathbf{G} = \mathbf{V}\mathbf{S}\mathbf{V}^T$, where $\mathbf{V}$ contains a set of orthonormal vectors and $\mathbf{S}$ is a diagonal matrix. Secondly, $\mathbf{U}^T\mathbf{f}$ can be rewritten as

$$\mathbf{U}^T\mathbf{f} \quad = \quad \mathbf{U}^T\mathbf{U}\mathbf{h} + \mathbf{U}^T\mathbf{e} \tag{4}$$

Let $\rho = \mathbf{V}^T\mathbf{h}$ and $\eta = \mathbf{V}^T(\mathbf{U}^T\mathbf{e})$ be the projections of the IRF $\mathbf{h}$ and the estimated input-noise cross-correlation $\mathbf{U}^T\mathbf{e}$, respectively, onto the Hessian eigenvectors $\mathbf{V}$. With these transformations, (2) can be rewritten as:

$$\begin{aligned}\hat{\mathbf{h}} &= \mathbf{V}\mathbf{S}^{-1}\mathbf{V}^T(\mathbf{V}\mathbf{S}\rho + \mathbf{V}\eta) \\ &= \sum_{i=0}^{M-1}(\rho_i + \frac{\eta_i}{s_i})\mathbf{v}_i \\ &\triangleq \sum_{i=0}^{M-1}\hat{\mathbf{h}}_i \end{aligned} \tag{5}$$

Thus, the SVD of $\mathbf{G}$ leads to a corresponding decomposition of $\hat{\mathbf{h}}$. Eq. (5) shows that for small eigenvalues s_i, the associated noise terms η_i will be amplified and corrupt the model term ρ_i when projected onto the i-th eigenvector $\mathbf{v}_i$ [32]. Limiting the summation to the most significant eigenvalues should improve the signal-to-noise ratio of the estimate $\hat{\mathbf{h}}$. Note that computing the pseudo-inverse of $\mathbf{G}$ in this way follows the same decomposition procedure as applying principal components analysis (PCA) or the Karhonen-Loève Transformation (KLT) to the input matrix $\mathbf{U}$.

F. Order estimation

We used the minimum description length (MDL) criterion [33] to choose the number of terms in (5) to retain, given by

$$\mathrm{MDL}(m) = \left[1 + \frac{mP\log(N)}{N}\right]\sum_{i=1}^{N}\left[f_i - \hat{f}_i(m)\right]^2 \tag{6}$$

where m is the number of singular values retained, $\hat{f}_i(m)$ is the output estimate at the i^{th} time step with m singular values retained, and P is a penalty factor, usually set to 1. In [32] it is shown that the summation in (6) can be efficiently calculated as $\sigma_f^2 - \sum_{i=1}^{m} s_i\left(\mathbf{v_i}^T\hat{\mathbf{h}}\right)^2$.

We analyzed the effect of the penalty P on the number of terms in (6). Our goal was two-fold: first, to maximize IRF quality, as measured by the output VAF; second, to minimize IRF variability, as measured by the sum of the IRF coefficient absolute differences

$$\mathrm{TV} = \sum_{i=0}^{M}\left|h_i - h_{i-1}\right|. \tag{7}$$

where h_{-1} and h_M are defined to be zero. This measure is called *total variation* [34] and can be interpreted as the total amplitude of IRF oscillations.

To resolve the tradeoff between these measures, we compared the average VAF and average TV over all epochs for different values of P. From these results and empirical analysis of individual cases, we

chose P=4 for our subsequent modelling. Compared to the results with P=1, P=4 reduced average TV by 70.7 % (with a corresponding reduction in TV variance), while only reducing average VAF from 35.9 % to 31.5 % (a relative reduction of 12.3 %). The VAF continued to decrease with higher values of P, with little improvement in the TV.

Fig. 3a shows the effect of changing P on the shape of the IRF for an example epoch from a pathological case. With the default MDL penalty factor P = 1, the IRF included all six components; with P = 4, only the first two components were included. Fig. 3b shows that the final four components $\hat{\mathbf{h}}_{3,4,5,6}$ contributed little to the output VAF (i.e., they were effectively noise); the two-component IRF had very similar VAF (59.1 % vs. 62.9 %), but the TV was substantially reduced (i.e., from 98 % to 22 % of its maximum value).

G. Delay and memory estimation

We initially assumed system causality and estimated the IRFs with no delay (i.e., d=0 in (1)) [35]. However, the input measurement delay associated with the UP sensor may result in a negative IRF delay. In contrast, we expect the physiological response to have a positive delay. The combination of these two delays can produce an FHR response that occurs before or after the measured UP contraction onset (i.e., d could be negative or positive). We found that better estimates of d led to better model prediction; therefore, we developed an algorithm to estimate d for each epoch. Based on [24], we set the lower bound to d_{min} = -80 s. We observed that for some pathological cases, the FHR response occurred as late as 80 s after the UP response, so we set d_{max} to this value.

The IRF length M should be long enough for the IRF to decay to 0, but as short as possible to improve the conditioning associated with estimating IRFs, given the short duration of the epochs (20 min). Furthermore, UP periodicity introduced artifactual IRF periodicity if M was too long. To handle this problem, we performed a search over values of M as well.

The objective of the IRF delay and memory estimation was to identify the temporal extent of the IRF; this meant searching for IRFs that 1) started and ended close to zero amplitude, 2) were as short as possible, and 3) were most predictive of the output. This was not always straightforward because of the IRF periodicity mentioned above. Therefore, we identified candidate intervals that satisfied goal (1), and then chose a best interval using the MDL criterion, which implicitly satisfied goals (2) and (3).

Occasionally, when competing candidates had similar MDL values, the automatically selected delay and/or memory estimates were not consistent with those of neighbouring epochs. Therefore, we performed a post-processing step to choose the most plausible candidate. First, we median filtered (independently)

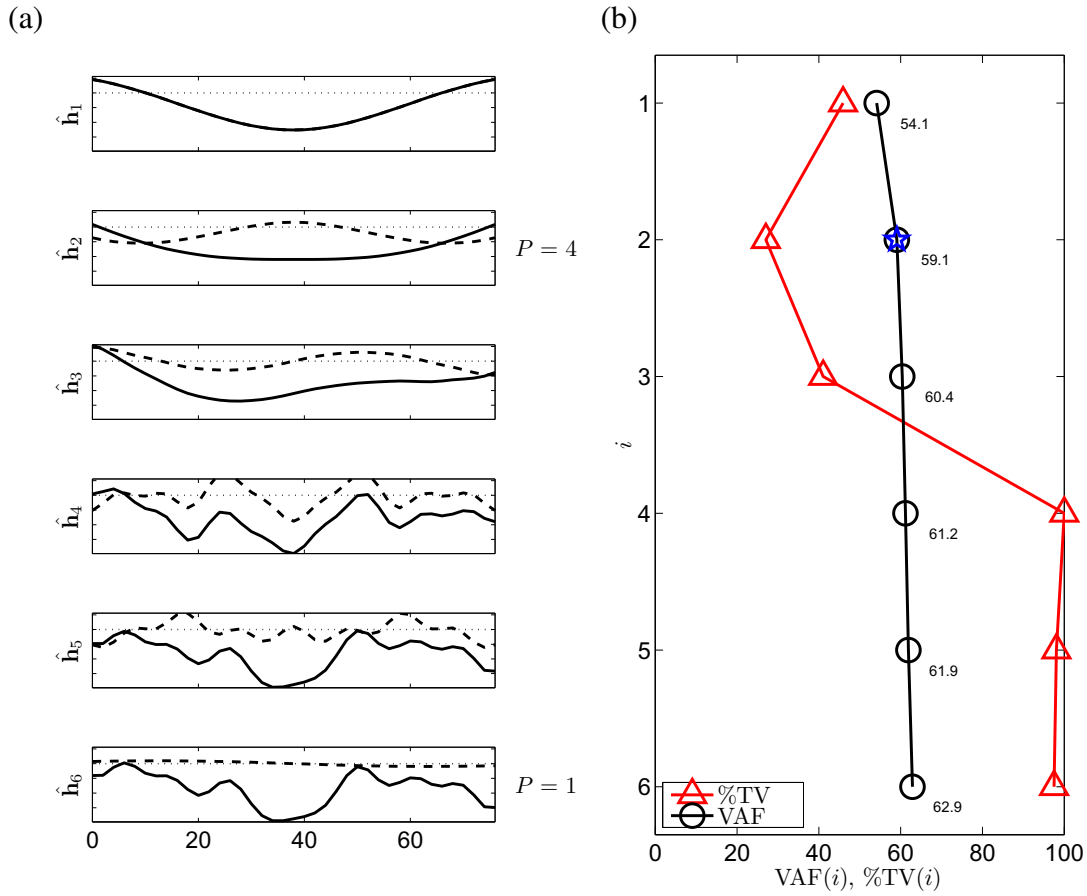


Fig. 3. Order selection for an example epoch of a pathological case. (a) IRF principal components $\hat{\mathbf{h}}_i$ (dashed) and cumulative sum $\sum \hat{\mathbf{h}}_i$ (solid) for orders $i \leq 6$ are shown. The vertical scales on the IRF plots are identical (zero amplitude is shown as a dotted horizontal line) and the time units are seconds. With default MDL penalty factor P = 1, the IRF included all six components; with selected P = 4, only the first two components were included. (b) VAF (black circles) and %TV (red triangles) at each order. Compared to the six-component IRF (VAF=62.9%, TV=98% of its maximum value), the two-component IRF had very similar VAF (~ 59.1 %, indicated by the blue star), but the total variance was reduced to 22% of its maximum value.

the initial selections of delay $\{d_i\}$ and recovery $\{r_i\} = \{d_i + M_i\}$ over all epochs i as $\{\hat{d}_i\}$ and $\{\hat{r}_i\}$. The median filtering included up to a total of 7 neighbouring samples in the median calculation at each sample and no fewer than 4 samples (at the end points). Then, at each epoch i, we chose the candidate model that was the nearest neighbour to $(\hat{d}_i, \hat{r}_i, \mathrm{MDL}_i)$, where MDL_i was the MDL of the initial candidate at epoch i; a Euclidean distance metric was used, with each dimension normalized to units of standard deviation about the mean, to give the three dimensions approximately equal weight. In this way, the refined candidates had IRF start and end time values that were more consistent with their neighbouring epochs. Additionally, their quality remained similar to the initial candidate.

As a further refinement, we performed a local search for IRFs with minimal starting and ending amplitudes, by varying the d and r found in the previous step by ± 10 s. Then, from a final median filtering of d, we removed delays outside the 95^{th} percentile of the deviation about the median-filtered value at each epoch (considering only epochs with models that had passed the significance filter, described below). We selected both the median filter order and the refinement neighbourhood empirically.

To achieve goal (1), we defined 'close' to zero amplitude as $<25\,\%$ of the maximum IRF amplitude. Therefore we removed models where the first and last IRF lags, h_0 and h_{M-1}, did not satisfy this criterion (i.e., remove if $\left|\frac{h_i}{h_{max}}\right| \geq 0.25$, $i \in \{0, M-1\}$, where h_{max} is the IRF maximum amplitude). We considered that delay-memory estimation had not been successful for these models; furthermore, their VAFs were generally very low (the mean, standard deviation, and median values were 11.0, 11.3, and 6.2 %, respectively).

Finally, we removed models having insignificant steady-state gain $G < 10^{-5}$ (which indicated that UP-FHR dynamics were virtually absent); the VAFs of these models were also generally very low (the mean, standard deviation, and median values were 8.7, 11.2, and 4.6 %, respectively). We calculated steady-state gain (henceforth referred to as *gain*) as the sum of the IRF coefficients

$$G = \sum_{i=0}^{M-1} h_i \tag{8}$$

H. Determining model significance using surrogates

Because the model output VAF was often quite low, we wanted to be confident that the IRFs modelled system dynamics rather than noise [36]. To do so, we compared IRFs estimated from the preprocessed FHR and from multiple FHR surrogates; the latter had similar properties to the former, except that their relationship to UP had been destroyed. We generated these surrogates using the amplitude adjusted Fourier transform (AAFT) [37], an algorithm that approximately preserves the power spectrum while maintaining the amplitude probability distribution of the original FHR. However, the process randomizes the phase so any relation to the UP will be removed.

The null hypothesis was that there were no UP-FHR linear dynamics in the epoch. This was rejected when the quality of the model obtained with the original FHR was superior, statistically, to models computed using surrogate FHR. In particular, the model was deemed significant if the VAF with the original FHR was in the 95^{th} percentile of VAFs computed with the original and surrogate FHRs. This is justified in [37], [38]. To limit computational requirements, we generated Q=99 surrogates, and chose the parameters of the test to obtain a significance threshold of 95 %.

Fig. 4 shows the model significance validation rates for normal (Fig. 4a) and pathological (Fig. 4b) cases. The bars for VAF range 0-100 (at right) shows that overall validation rates were higher for pathological cases (66.8 %) than for the normal cases (54.0 %), mainly because pathological cases tended to have higher VAFs. This is seen in the bars to the left that have more precise VAF ranges. In particular, pathological cases had a smaller proportion (82/346=23.7 %) of very low VAF (<10 %) compared to the normal cases (917/2322 = 39.4 %); these models tended to be rejected for both classes. Acceptance rates increased for higher ranges of VAF. For the pathological cases, acceptance was 4.9 % for models with VAF<10 %, 63.4 % for 10 %≤VAF<20 %, 88.4 % for 20 %≤VAF<40 %, and 100 % for models with VAF≥40 %. These proportions were similar for the normal cases.

For comparison, to achieve an overall validation acceptance rate equivalent to that obtained with surrogates (55.7 %), a hard threshold of approximately VAF=13 % would have been required. Although our approach also tends to reject low VAF models and accept high VAF models, it does so more adaptively. We note that retaining the FHR variability component (i.e., FHR power ≥ 0.03 Hz) in the VAF calculation did not affect the relative rates of acceptance across classes. The level of agreement between significance tests that used VAF and those that used VAF_{VLF} was 94.9% for normal cases and 95.9% for pathological cases.

III. Results

A. *Processing summary*

The proportion of epochs eliminated at each processing step is shown in Table I. A significant portion of the data failed to meet minimum criteria for modelling and were discarded. Preprocessing rejected all epochs for 2/221 normal cases and 5/43 pathological cases. Subsequent processing (i.e., removal of models that had low gain, or that failed delay-memory estimation, or that did not pass the significance test using surrogate data) removed all epochs for a further 7 pathological and 20 normal cases. We therefore retained for analysis 189 normal cases with 1195 epochs (out of a possible 3978 epochs, or 30 % of the data) and 31 pathological cases with 219 epochs (out of 774 epochs, or 28 % of the data). The pathological data were significantly noisier than the normal data: a much larger portion of the former was rejected as artifactual during preprocessing (42 % vs. 8 % for the normals). On the other hand, of the data that passed the preprocessing step, a larger proportion of the pathological cases had models that passed validation (28/(100-42) = 48 %) than for the normals (30/(100-8) = 33 %).

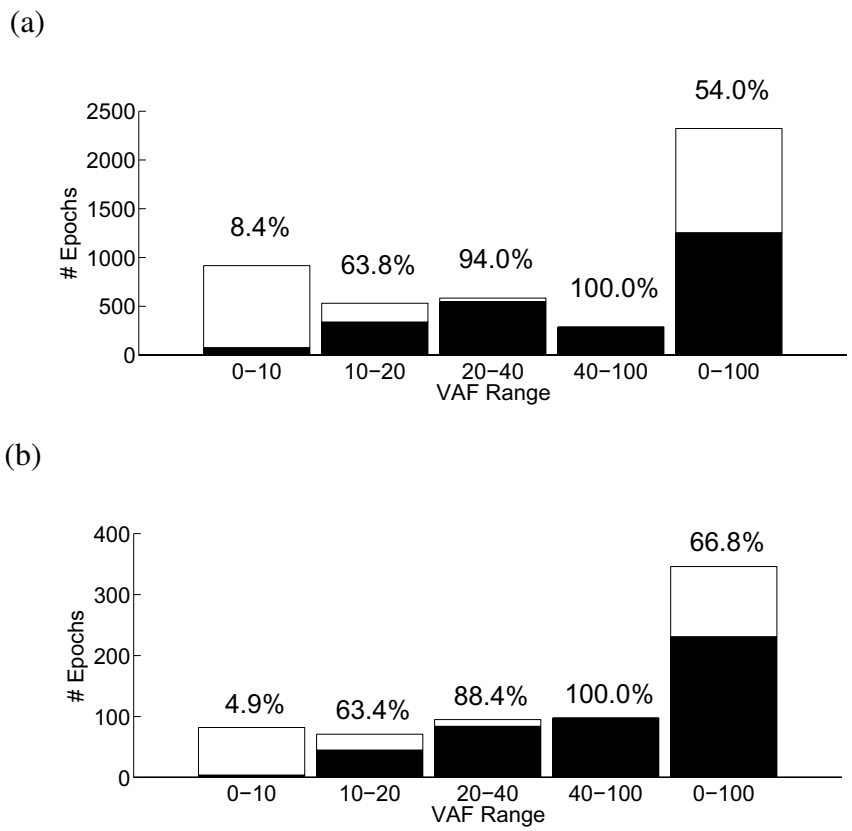


Fig. 4. Acceptance rates of significance validation for a) normal and (b) pathological cases. For each VAF range, the number of models validated (height of black bars) out of the total tested (combined height of black and white bars) is shown and the percentage of validated models is indicated.

	Normal	Pathological
# Epochs (100%)	3978	774
Artifact	8%	42%
Low G	13%	4%
d/M estimation failed	21%	9%
Insignificant	27%	15%
Outlier d	2%	2%
Validated	30%	28%

TABLE I

RELATIVE PROPORTION OF EPOCHS ELIMINATED AT EACH PROCESSING STEP AND FINAL PROPORTION OF VALIDATED EPOCHS FOR NORMAL AND PATHOLOGICAL CASES. '# EPOCHS' IS THE NUMBER OF EPOCHS POSSIBLE IN THREE HOURS OF DATA, CORRESPONDING TO 100%. THE PROCESSING STEPS REMOVED: (1) ARTIFACT DURING PREPROCESSING; (2) LOW GAIN MODELS; (3) MODELS WITH NON-ZERO IRF AT END LAGS DURING DELAY/MEMORY ESTIMATION; (4) INSIGNIFICANT MODELS; (5) MODELS WITH OUTLIER DELAY VALUES.

B. IRFs of typical epochs

Fig. 5 shows the raw and preprocessed input UP, the raw, preprocessed and predicted output FHR, the residual e, and estimated IRF for a typical normal epoch. The model parameters were VAF=44.7 %, delay d=-16 s and gain G=-0.17 bpm/mmHg. This epoch illustrates some of the challenges to successful modelling of the data. First, the background FHR variability was significant relative to the deceleration amplitude. Secondly, our artifact detection procedure bridged an FHR dropout at 1100 s, introducing some high-frequency artifact. Despite these factors, the predicted FHR shows that the model captured the FHR response to UP.

Fig. 6 shows the same signals for a typical pathological epoch. In this case, two FHR dropout artifacts were bridged (at 350 s and 950 s). Model parameters for the pathological case were markedly different from those of the normal case; there was larger VAF (69.6 %), longer delay (d=34 s) and larger gain (G=-0.50 bpm/mmHg) for the pathological case.

The output predictions of Figs. 5b and 6b (in red) were very smooth, suggesting that most prediction occurred in the lower frequency band. A frequency analysis of the same epochs confirmed this. Figs. 7 and 8 show the input UP power spectrum S_{uu}, and the preprocessed and predicted output FHR power spectra S_{ff} and $S_{\hat{f}\hat{f}}$. In both the normal and pathological case, most of the input and output energies were in the lower frequency band (i.e., near the UP contraction rate and far below the half-sampling rate of 0.25 Hz). This was the dominant frequency band of the dynamics and the model accounted for most of the energy here: the preprocessed and predicted FHR power spectra S_{ff} and $S_{\hat{f}\hat{f}}$ are similar over the main spectral peak (i.e., below ~0.01Hz). The figure also shows that most of the model power was contained in the VLF band below 0.03 Hz; at this frequency the response was attenuated by ~ 40 dB with respect to the peak power. It is also apparent that the higher-frequency FHR energy (>0.03 Hz) falls off at later frequencies for the normal compared to the pathological case, indicating that it contained relatively more FHR variability.

C. IRF time progression of typical cases

Fig. 9a shows the progression of the IRFs for a normal case over time. This case showed marked gain and VAF progression with time: it started at low gain and VAF values (G = −0.136 bpm/mmHg, VAF=13.7 %) and steadily increased to its largest values (G = −0.810 bpm/mmHg, VAF=58.6 %) just before delivery. The delay varied between -20 s and -4 s.

Fig. 9b shows the same time progression for a pathological case. Overall, the pathological case had larger VAF, longer delay and larger gain. The gain was lowest for the initial epochs (G = −0.172 bpm/mmHg,

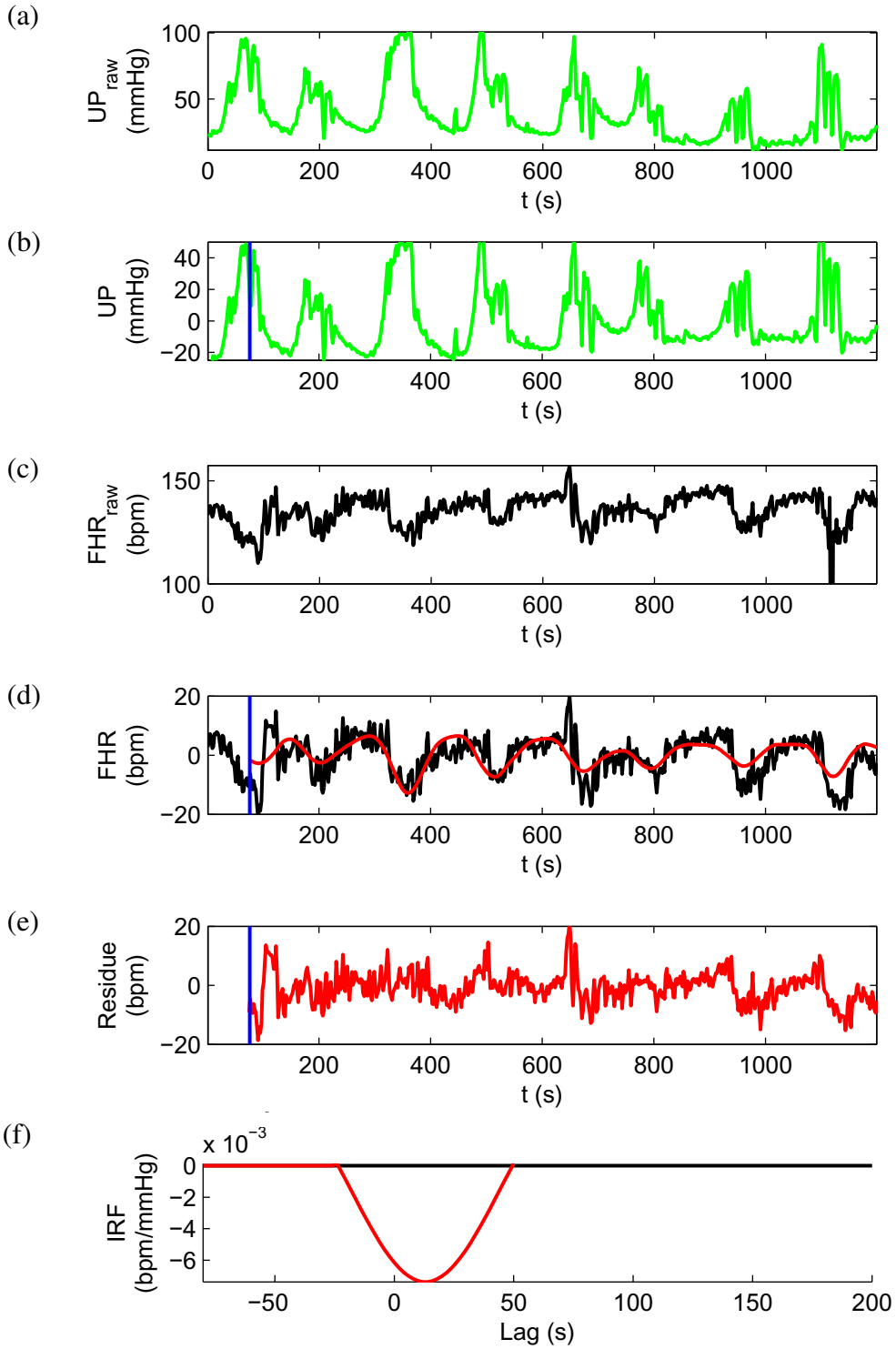


Fig. 5. Modelling results for an example epoch of a normal case. (a) raw input UP (b) pre-processed UP (c) raw output FHR (d) preprocessed (black) and predicted (red) output FHR (e) residual FHR (f) final IRF. Vertical blue bars indicate memory length M=76 s (VAF=44.7 %, $d = -16$ s and $G = -0.17$ bpm/mmHg).

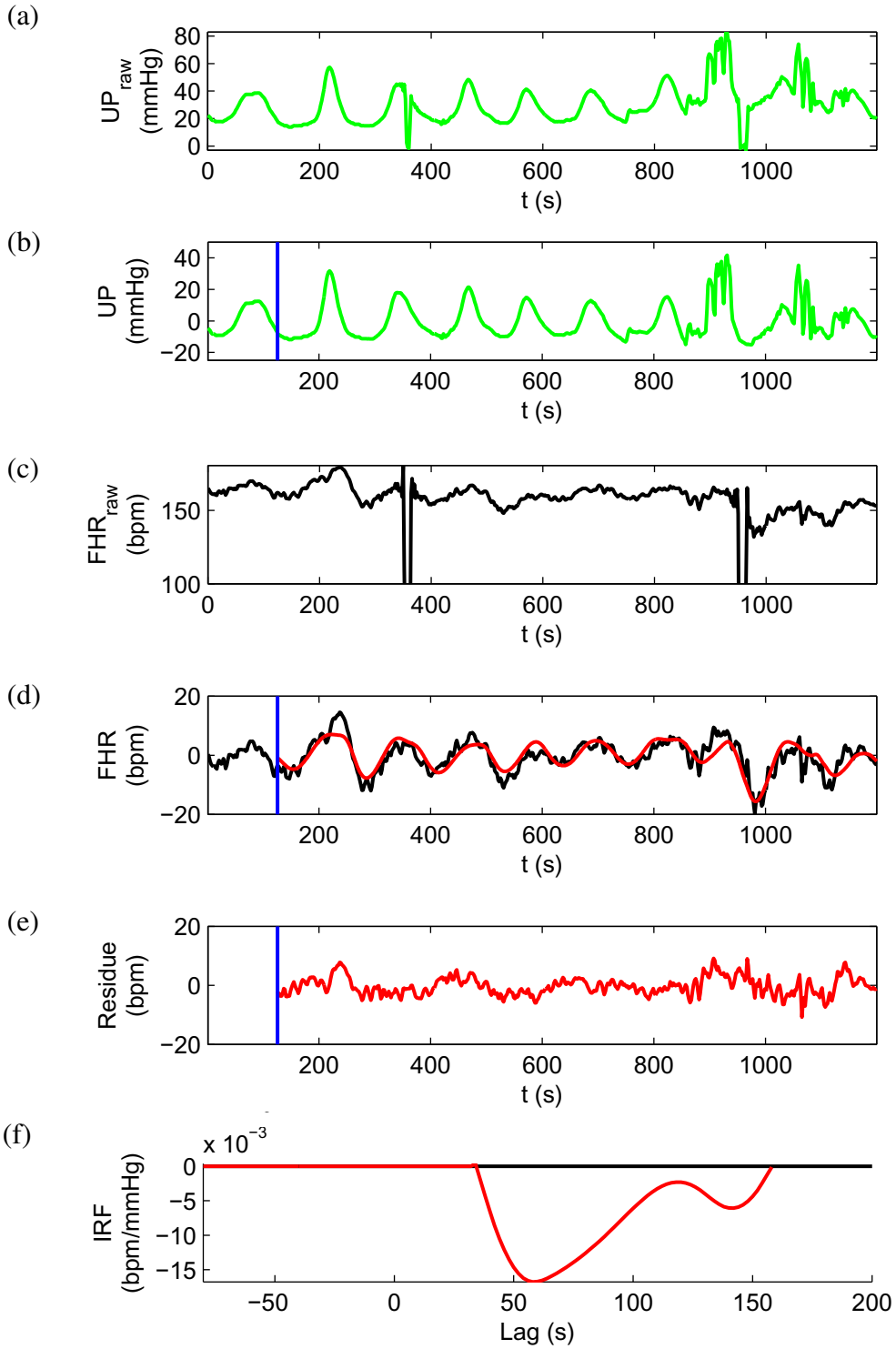


Fig. 6. Modelling results for an example epoch of a pathological case. (a) raw input UP (b) pre-processed UP (c) raw output FHR (d) preprocessed (black) and predicted (red) output FHR (e) residual FHR (f) final IRF. Vertical blue bars indicate memory length M=126 s (VAF=69.6 %, d = 34 s and G = −0.50 bpm/mmHg).

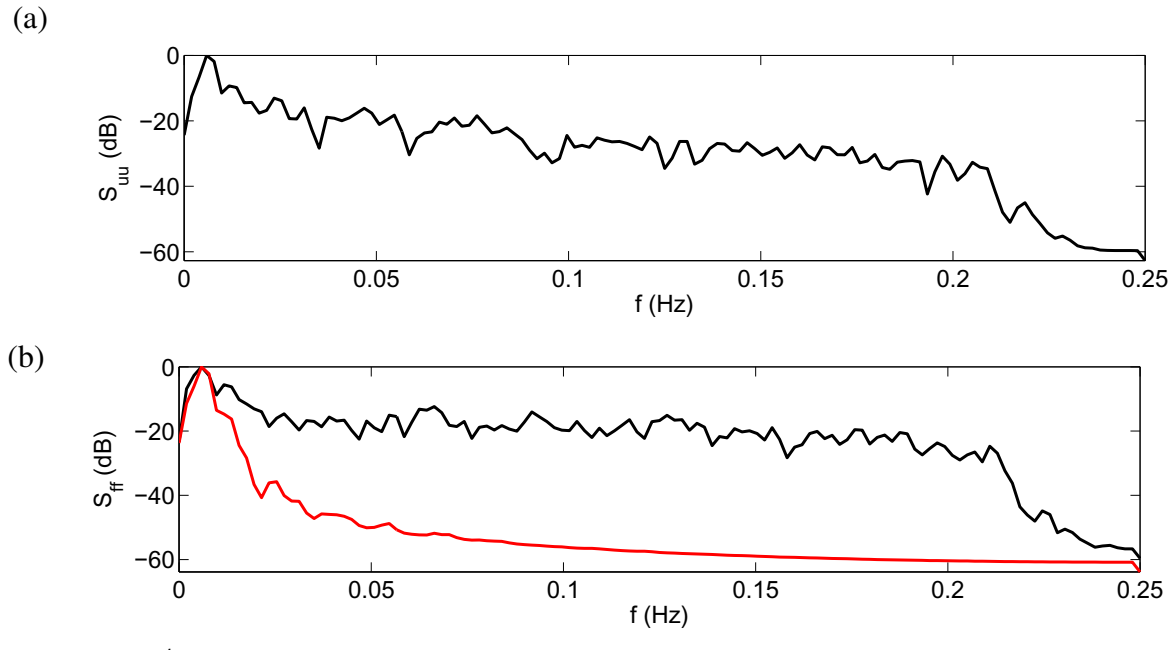


Fig. 7. Frequency analysis for an example epoch for a normal case. (a) input UP spectrum S_{uu} (b) preprocessed (black) and predicted (red) output FHR spectra S_{ff} and $S_{\hat{f}\hat{f}}$.

VAF=37.6 %), followed by a larger and relatively stable gain and VAF ($G = -0.616$ bpm/mmHg, VAF=54.9 %) at epoch -10 until delivery. The delay was longer and steady for the pathological case (~16 s). These results are consistent with the clinical expectation that pathological cases have later and stronger responses.

D. IRF parameters by class

Fig. 10 shows the time progression, by epoch, of the average model parameters. There were 6 epochs per hour. For the normal cases, the delay was relatively steady over time. G and VAF, however, tended to increase over time, especially in the last hour (i.e., the last 6 epochs). Overall, the pathological cases had longer delay d, larger (negative) gain G and higher VAF. For the pathological cases, the delay increased substantially in the final hour; G and VAF increased from the first to second hour and then dropped in the final hour, likely reflecting the influence of greater sensor disturbance near delivery on the smaller pathological population.

To assess whether the class parameter differences were statistically significant, we performed hypothesis tests at each epoch. We compared class distributions with the Kolmogorov-Smirnov (KS) hypothesis test, which uses the maximum difference of the empirical cumulative distribution functions (cdfs) as a test statistic. More precisely, denoting by F_N and F_P the cdfs of the normal and pathological populations,

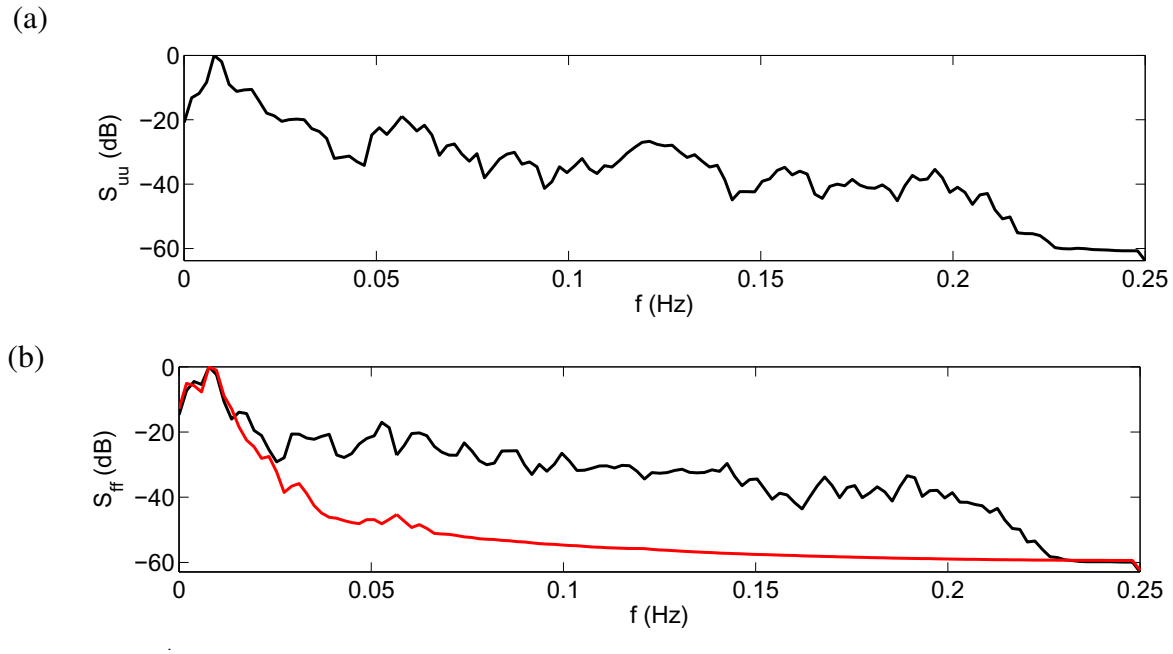
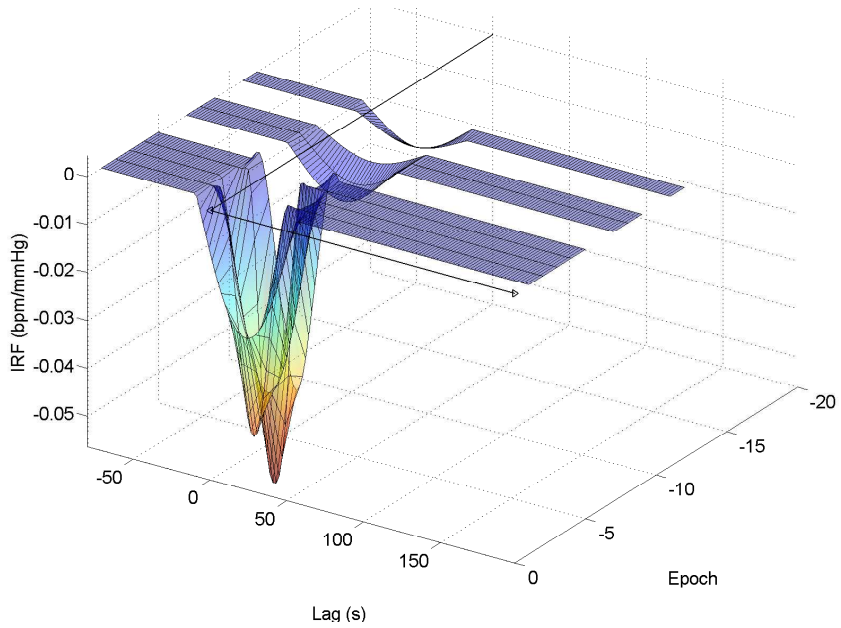


Fig. 8. Frequency analysis for an example epoch for a pathological case. (a) input UP spectrum S_{uu} (b) preprocessed (black) and predicted (red) output FHR spectra S_{ff} and $S_{\hat{f}\hat{f}}$.

respectively, the KS test computes $\max(|F_N(x) - F_P(x)|)$ for all values of the model parameter x. We deemed a class difference significant if the test rejected the null hypothesis at the $p < 0.05$ level (shown in the figure by blue asterisks). The delay difference was consistently significant over the epochs of the final hour. The VAF and gain were significant for most epochs between the end of hour -3 and hour -2. Memory M did not pass the hypothesis tests for significant class differences. We noted that the K-S test was more conservative than simpler mean-based tests such as the t-test, which tended to reject the null hypothesis more often for the VAF, delay and gain parameters.

Using the modified VAF_{VLF} calculation increased the VAF of the normal case of Fig. 5 from 44.7% to 54.2% and the pathological case of Fig. 6 from 69.6% to 74.5%. The time progression of the class-average VAF_{VLF} of Fig. 10(b) had a mean trend and variance that was very similar to the original VAF of Fig. 10(a), except that both the normal and pathological trends were shifted towards higher VAF. There was also a slight increase (from 5 to 8) in the number of epochs that had statistically significant differences. These results confirm that the across-class differences we observed in model quality were due mainly to system dynamics rather than FHR variability.

(a)



(b)

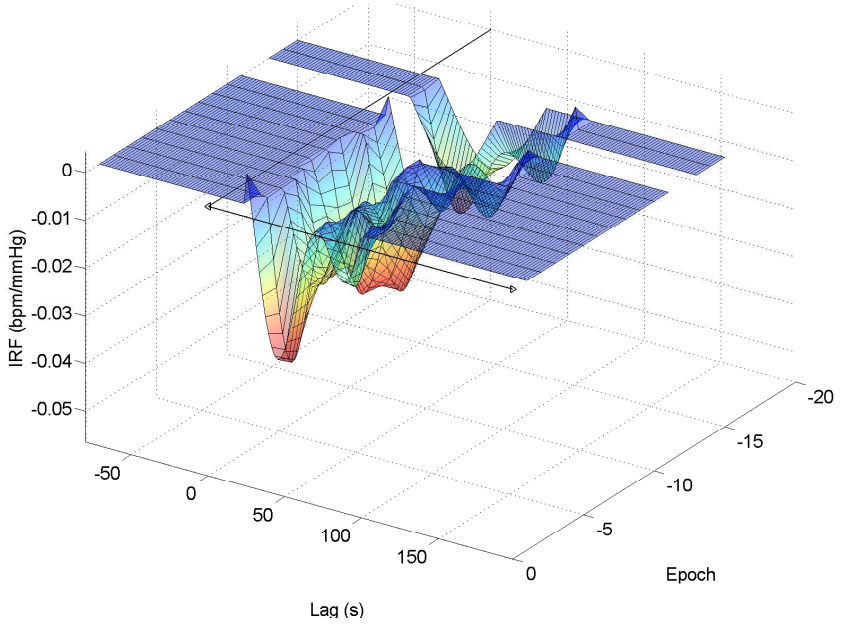


Fig. 9. IRF progression over epochs for (a) normal and (b) pathological cases. The perpendicular lines indicate IRF lag t=0 and the time of delivery (epoch 0, marked with triangles). The IRF amplitudes have the same scale with the mesh colour ranging from blue (no response) to red (strong response) . Overall, the pathological case had longer delay and larger gain. The normal case steadily progressed in time towards its largest gain just before delivery, while the pathological case had lower gain in the first few epochs, but soon increased to a larger gain that remained almost constant until delivery.

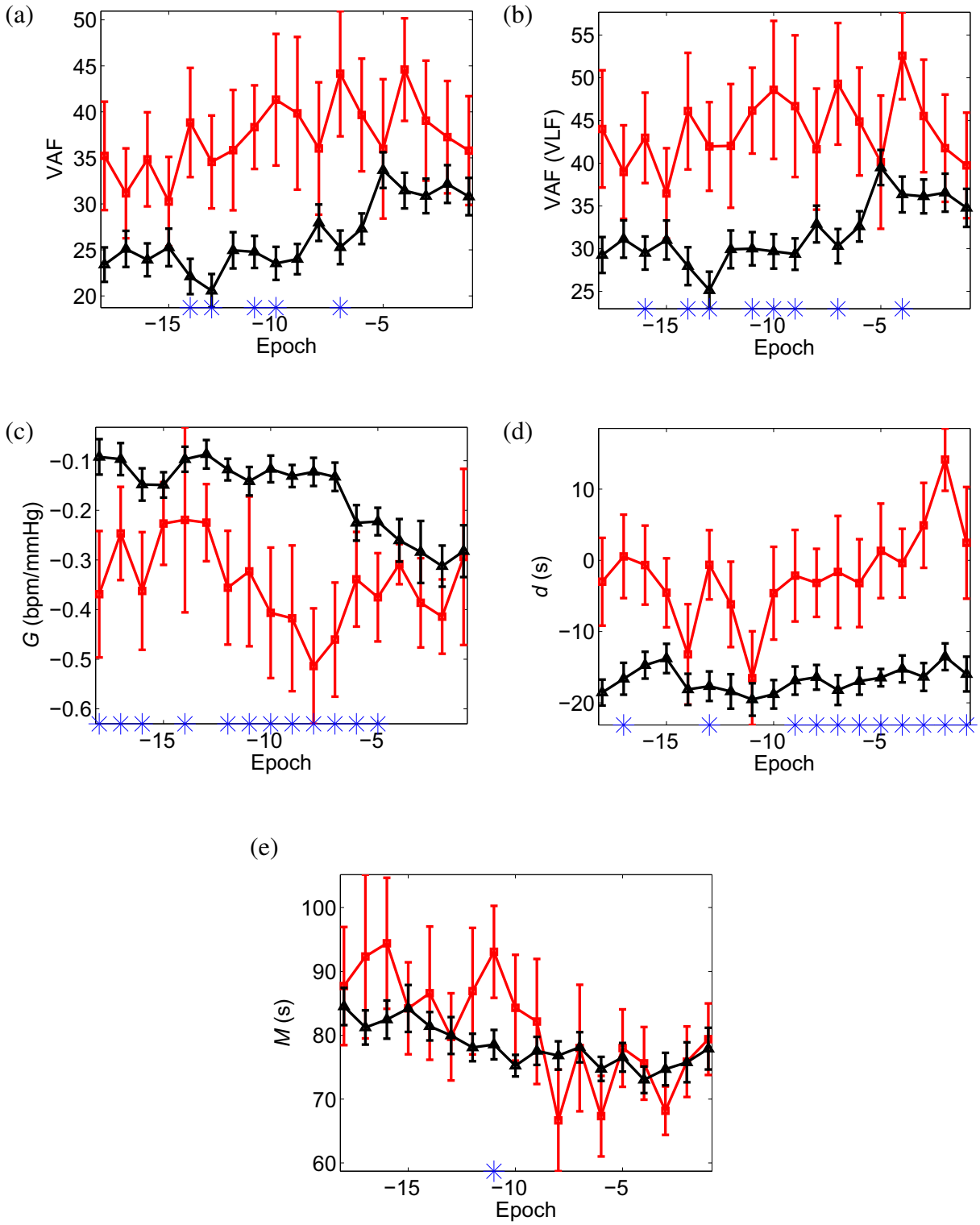


Fig. 10. Time progression of model parameters in the last three hours of labour and delivery. The mean and standard error for (a) VAF, (b) VAF_{VLF} (c) gain G, (d) delay d and (e) memory M are plotted by epoch for normal (black triangles) and pathological (red squares) cases. For each parameter, epochs where hypothesis tests indicated significant differences between classes ($p < 0.05$) are displayed by asterisks (*).

IV. Discussion

A. Technical approach

Our system identification approach is an entirely automated method that extracts information from the two CTG signals in a novel way. The IRF implicitly captures the strength and timing of the FHR response to UP, in contrast to a feature detection approach where these relations must be measured explicitly (e.g., in terms of amplitude ratios and time delays) from noisy signals. To our knowledge, this is the first successful attempt to model UP-FHR dynamics.

The particularities of the (uncontrolled) UP input signal required special treatment to regularize the ill-posed identification problem. We used SVD decomposition to suppress noise as is customary, but instead of using the standard MDL formulation, which tended to generate overly complex IRFs, we used a modified MDL to choose a more appropriate order. The narrow-band UP introduced periodicity in the estimated IRF; this and the unknown UP sensor measurement delay introduced ambiguity about the temporal extent of the IRF response. We solved this by developing a method of estimating the IRF delay and memory.

Modelling such data required extensive validation; the final surrogate testing provided an objective, adaptive assessment of model quality. When our modelling was successful, our techniques generated models that were reliable; they were clean (as evidenced by the reduction in model total variation by the modified MDL criterion for order selection), concise (via the MDL criterion for delay and memory estimation), and predictive (as evidenced by the VAF). Moreover, these models were consistent with clinical expectation, as we discuss in the next section.

B. Scientific and clinical significance

We have shown that our modelling revealed UP-FHR system dynamics that correspond well to the clinical understanding of the UP-FHR interaction. Fetuses in distress tend to have stronger and later FHR responses to UP, described in clinical terms as deeper and later decelerations. Healthy fetuses, on the other hand, tend to respond less overall, and with strong responses only for a short period just prior to delivery. Indeed, our results showed significant differences across outcome classes in the IRF parameters corresponding to response strength and timing (the IRF gain and delay), despite these parameters being subject to the measurement problems of UP tocography.

The differences in delay also align with clinical expectation. Sensor-related delays would be expected to affect both outcome classes equally; hence, in general, the longer delays in pathological cases were

dominated by delays that were physiological in nature. This result is also consistent with animal studies, which have shown that under conditions of mild hypoxia, relatively rapid parasympathetic chemo-receptor mechanisms dominate the fetal response to contractions. However, with increasing hypoxia, slower parasympathetic baro-receptor and direct myocardial depression play an increasing role in the deceleration response [39], [40]. Consequently, the latency of the fetal response is expected to increase with hypoxia.

Our finding that models of pathological cases tend to have higher VAF is also consistent with clinical expectation. In a healthy fetus, contractions cause little disruption in oxygen delivery and little need for compensatory FHR changes; a fetus in distress, however, will tend to be less resistent to the insults of labour, especially the highly compressive contraction events. Their compensatory mechanisms may be quite compromised, causing them to follow rather than resist the onslaught of the stimulus. As a result, their response is a more predictable phenomenon that can be modelled more precisely.

Clinical UP and FHR are very noisy signals prone to frequent sensor disturbance; to our knowledge, CTG signals collected under these conditions have never been subject to the analysis of an approach such as ours. The pathological cases were especially prone to artifact, suggesting that these cases were more likely subject to sensor disturbance and clinical intervention. Despite these conditions, our modelling successfully generated valid models for around a third of the data (including artifactual data). Furthermore, the proportion of artifact-free epochs resulting in models was higher for pathological cases (one-half) compared to the normal cases (one-third). We think that this success rate is very acceptable: normal cases often exhibit significant FHR response to UP only very late in labour (i.e., there is often a weak or non-existent UP/FHR model). Consequently, it is to be expected that no model would be found for many epochs where there is no response; these should not raise an alarm because they are consistent with the behaviour of healthy fetuses. Pathological cases tend to have an FHR response to UP that is more significant and occurring earlier in labour, but they may also have no response due to severe loss of compensatory mechanisms (which may be indicative of chronic rather than acute injury). Therefore, even if the data were artifact-free, we would not expect to generate models for every epoch.

C. Classification

Our results suggest that the model parameters could be used successfully to automatically classify the fetal state. Indeed, in other studies, we have successfully used our model, in conjunction with measures of FHR variability and baseline, for classification purposes. Using the system identification model alone for epochs near delivery, we reported a false positive rate of 10% (i.e., a specificity of 90 %)

with a sensitivity of ~75% (i.e., a false negative rate of 25%) [17]. This classification considered epochs in isolation; for a more robust classifier, we also considered the history of epoch classifications [18]. With this approach, we observed a mean area under the receiver-operator characteristic (ROC) of 0.801. This ROC was generated by varying the number of single epoch pathological classifications required to generate an overall pathological classification (i.e. varying from one to six epochs). For comparison, we also generated similar classifiers that used measures of FHR baseline and variability as features. These features were generated by linear fitting for the baseline and an autoregressive (AR) model for the variability. A study of individual pathological cases indicated that some fetuses were classified by one or the other, but not both of the classifiers. Of the 31 pathological cases tested (selected to have at least three system identification models), 5 were correctly classified by the system identification model classifier alone, 4 were correctly classified by the baseline-variability classifier alone, 18 were correctly classified by both and 4 were missed by both. These results clearly demonstrate the potential clinical significance of our models as well as the complementarity of the system identification model and measures of FHR variability and baseline.

D. Future work

We expect that better UP measurement techniques, such as uterine electromyography [41], would improve discrimination of the two classes further. Indeed, further insights could be obtained by applying our approach to data collected with an improved acquisition protocol.

The generally low VAFs can be attributed to that fact that UP is not the only influence on FHR. Background intrinsic FHR variability unrelated to UP can be of significant amplitude [42]. Also, our linear model assumes stationarity within epochs (which can be violated) and does not account for non-linear interactions such as 1) the higher-frequency harmonics present in some very sharp decelerations and 2) variability that is coincident with decelerations [43]. We are also aware that certain severely injured fetuses may display FHR with abnormal variability or levels of baseline [29], [30]. We view our very-low-frequency system identification model as complementary to measures of FHR variability and baseline for assessment of the fetal state.

We also hope to classify intermediate cases, which may develop hypoxia quite late in delivery. A survey of clinical FHR studies [5] concluded that fetal acidemia usually develops over one hour. Given fetal-state evolution over this time frame, there is great potential for the combination of better data collection and model-based analysis to improve current clinical practice by identifying fetuses at risk before they become injured.

References

[1] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney, "Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate for normal and hypoxic fetuses," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 6, pp. 1587–1597, June 2009.

[2] ACOG, *Neonatal Encephalopathy and Cerebral Palsy: Defining the Pathogenesis and Pathophysiology*. ACOG Task force on Neonatal Encephalopathy and Cerebral Palsy, January 2003.

[3] N. Badawi, J. Kurinczuk, J. Keogh, L. Alessandri, F. O'Sullivan, P. Burton, P. Pemberton, and F. Stanley, "Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study," *BMJ*, vol. 317, pp. 1549–1553, 1998.

[4] E. Draper, J. Kurinczuk, C. Lamming, M. Clarke, D. James, and D. Field, "A confidential enquiry into cases of neonatal encephalopathy," *Arch Dis Child Fetal Neonatal Ed*, vol. 87, pp. F176–F180, 2002.

[5] J. T. Parer, T. King, S. Flanders, M. Fox, and S. J. Kilpatrick, "Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association?" *Journal of Maternal-Fetal & Neonatal Medicine*, vol. 19, no. 5, pp. 289–294, May 2006.

[6] S. Ransom, D. Studdert, M. Dombrowski, J. Mello, and T. Brennan, "Reduced medicolegal risk by compliance with obstetric clinical pathways: A case-control study," *Obstet Gynecol*, vol. 101, 2003.

[7] C. Saphier, E. Thomas, D. S. D, T. Brennan, and D. Acker, "Applying no-fault compensation to obstetric malpractice claims," *Prim Care Update Ob Gyns*, vol. 5, pp. 208–9, 1998.

[8] B. Stalnaker, J. Maher, G. Kleinman, J. Macksey, L. Fishman, and J. Bernard, "Characteristics of successful claims for payment by the Florida Neurologic Injury Compensation Association Fund," *Am J Obstet Gynecol*, vol. 177, pp. 268–71, 1997.

[9] G. Berry and P. Martin, "Perinatal risks. risk management foundation harvard medical institutions forum," Harvard University, Tech. Rep., March 2001.

[10] K. Maeda, M. Utsu, A. Makio, M. Serizawa, Y. Noguchi, T. Hamada, K. Mariko, and F. Matsumoto, "Neural network computer analysis of fetal heart rate," *Journal of Maternal-Fetal Investigation*, vol. 8, no. 4, pp. 163–171, Dec. 1998.

[11] J. Skinner, J. Garibaldi, J. Curnow, and E. Ifeachor, "Intelligent fetal heart rate analysis," in *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, 2000, pp. 14–21.

[12] F. Lunghi, G. Magenes, L. Pedrinazzi, and M. Signorini, "Detection of fetal distress though a support vector machine based on fetal heart rate parameters," in *Computers in Cardiology, 2005*, 2005, pp. 247–250.

[13] P. Warrick, E. Hamilton, and M. Macieszczak, "Neural network based detection of fetal heart rate patterns," in *Neural Networks, 2005. Proceedings. 2005 IEEE International Joint Conference on*, vol. 4, 2005, pp. 2400–2405.

[14] H. Cao, D. Lake, I. Ferguson, J.E., C. Chisholm, M. Griffin, and J. Moorman, "Toward quantitative fetal heart rate monitoring," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 1, pp. 111–118, 2006.

[15] G. Georgoulas, D. Stylios, and P. Groumpos, "Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 5, pp. 875–884, 2006.

[16] F. Cowan, M. Rutherford, F. Groenendaal, P. Eken, E. Mercuri, G. M. Bydder, L. C. Meiners, L. M. Dubowitz, and L. S. de Vries, "Origin and timing of brain lesions in term infants with neonatal encephalopathy," *The Lancet*, vol. 361, no. 9359, pp. 736–742, Mar. 2003.

[17] P. A. Warrick, E. F. Hamilton, R. E. Kearney, and D. Precup, "Classification of normal and hypoxic fetuses using

system identification from intra-partum cardiotocography," in *ICML2008 Workshop on Machine Learning for Health Care Applications*, 2008.

[18] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney, "Classification of normal and hypoxic fetuses from systems modelling of intra-partum cardiotocography," *Biomedical Engineering, IEEE Transactions on*, 2009, to be submitted.

[19] E. Hamilton, A. Dyachenko, C. Elliott, P. Warrick, and A. Ciampi, "Progression of intrapartum EFM patterns in births with symptomatic metabolic acidosis," *American Journal of Obstetrics and Gynecology*, vol. 197, no. 6, p. s182, 2007.

[20] J. Low, R. Victory, and E. Derrick, "Predictive value of electronic fetal monitoring for intrapartum fetal asphyxia with metabolic acidosis," *Obstet Gynecol*, vol. 93, pp. 285–291, 1999.

[21] A. MacLennan, "A template for defining a causal relation between acute intrapartum events and cerebral palsy: international consensus statement," *BMJ*, vol. 319, no. 7216, pp. 1054–1059, 1999.

[22] T. Vanner and J. Gardosi, "Intrapartum assessment of uterine activity," *Bailliere's Clinical Obstetrics And Gynaecology*, vol. 10, no. 2, pp. 243–257, June 1996.

[23] M. Miles, A.M., M. Monga, M., and M. Richeson, K.S., "Correlation of external and internal monitoring of uterine activity in a cohort of term patients," *American Journal Of Perinatology*, no. 03, pp. 137–140, 2001.

[24] J. Jezewski, K. Horoba, A. Matonia, and J. Wrobel, "Quantitative analysis of contraction patterns in electrical activity signal of pregnant uterus as an alternative to mechanical approach," *Physiological Measurement*, no. 5, p. 753, 2005.

[25] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[26] B. Aysin, L. Chaparro, I. Grave, and V. Shusterman, "Orthonormal-basis partitioning and time-frequency representation of cardiac rhythm dynamics," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 5, pp. 878–889, 2005.

[27] I. W. Hunter and R. E. Kearney, "Two-sided linear filter identification," *Medical & Biological Engineering & Computing*, vol. 21, pp. 203–209, 1983.

[28] S. Cerutti, S. Civardi, A. Bianchi, M. Signorini, E. Ferrazzi, and G. Pardi, "Spectral analysis of antepartum heart rate variability," *Clin. Phys. Physiol. Meas.*, vol. 10, pp. 27–31, 1989.

[29] International Federation of Gynecology and Obstetrics (FIGO), "Guidelines for the use of fetal monitoring," *Int J Gynaecol Obstet*, vol. 25, pp. 159–67, 1987.

[30] National Institute of Child Health and Human Development (NICHD), "Electronic fetal heart rate monitoring: research guidelines for interpretation. national institute of child health and human development research planning workshop." *Am J Obstet Gynecol.*, vol. 177, no. 6, pp. 1385–90, 1997.

[31] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton, "System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses," in *Computers in Cardiology 2006. Proceedings.*, vol. 33, 2006, pp. 937–940.

[32] D. T. Westwick and R. E. Kearney, *Identification of nonlinear physiological systems*. Hoboken NJ: Wiley-Interscience, 2003.

[33] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[34] S. Mallat, *A wavelet tour of signal processing*. San Diego, CA: Academic Press, 1999.

[35] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney, "Detecting the temporal extent of the impulse response function from intra-partum cardiotocography for normal and hypoxic fetuses," in *The 2008 IEEE Engineering in Medicine and Biology 30th Annual Conference*, 2008, pp. 2797–2800.

[36] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton, "Time progression of a parametric impulse response function

estimate from intra-partum cardiotocography for normal and hypoxic fetuses," in *Computers in Cardiology 2007*, 2007, pp. 693–696.

[37] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77–94, Sept. 1992.

[38] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, Aug. 2000.

[39] C. B. Martin, J. de Haan, B. van der Wildt, H. W. Jongsma, A. Dieleman, and T. H. M. Arts, "Mechanisms of late decelerations in the fetal heart rate : A study with autonomic blocking agents in fetal lambs," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 9, no. 6, pp. 361–373, Dec. 1979.

[40] L. Bennet, J. A. Westgate, Y.-C. J. Liu, G. Wassink, and A. J. Gunn, "Fetal acidosis and hypotension during repeated umbilical cord occlusions are associated with enhanced chemoreflex responses in near-term fetal sheep," *J Appl Physiol*, vol. 99, no. 4, pp. 1477–1482, 2005.

[41] M. Skowronski, J. Harris, D. Marossero, R. Edwards, and T. Euliano, "Prediction of intrauterine pressure from electrohysterography using optimal linear filtering," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 10, pp. 1983–1989, 2006.

[42] J. Westgate, B. Wibbens, L. Bennet, G. Wassink, J. Parer, and A. Gunn, "The intrapartum deceleration in center stage: a physiologic approach to the interpretation of fetal heart rate changes in labor." *American Journal of Obstetrics and Gynecology*, vol. 197, no. 3, pp. 236.e1 – 236.e11, 2007.

[43] M. Romano, P. Bifulco, M. Cesarelli, M. Sansone, and M. Bracale, "Foetal heart rate power spectrum response to uterine contraction," *Medical and Biological Engineering and Computing*, vol. 44, pp. 88–201, March 2006.

CHAPTER 6
Journal article III: Fetal-state classification

Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Classification of normal and hypoxic fetuses from systems modelling of intra-partum cardiotocography. *IEEE Transactions on Biomedical Engineering*, October 2009. accepted

6.1 Introduction

This paper addresses the main goal of the thesis, that is, automated assessment of the fetal state during labour and delivery. Using the system identification models of the previous paper, we classified 20-min epochs as normal or pathological. At the same time, we did similar classification using the complementary FHR information of baseline and variability parameters. By combining these epoch classifiers and observing their history over time, we constructed a detector of fetal pathology. The results indicate that half of the pathological cases could be detected, with low false positive rates (7.5%), early enough, (1.5 hours before delivery) to allow appropriate clinical intervention.

Classification of Normal And Hypoxic Fetuses from Systems Modelling of Intra-Partum Cardiotocography

Philip A. Warrick[1], Emily F. Hamilton[3,4], Doina Precup[2] and Robert E. Kearney[1]

Abstract

Recording of maternal uterine pressure (UP) and fetal heart rate (FHR) during labour and delivery is a procedure referred to as cardiotocography. We modelled this signal pair as an input-output system using a system identification approach to estimate their dynamic relation in terms of an impulse response function. We also modelled FHR baseline with a linear fit and FHR variability unrelated to UP using the power spectral density computed from an auto-regressive model. Using a perinatal database of normal and pathological cases, we trained support-vector-machine classifiers with feature sets from these models. We used the classification in a detection process. We obtained the best results with a detector that combined the decisions of classifiers using both feature sets. It detected half of the pathological cases, with very few false positives (7.5%), one hour and forty minutes before delivery. This would leave sufficient time for an appropriate clinical response. These results clearly demonstrate the utility of our method for the early detection of cases needing clinical intervention.

I. Introduction

BETWEEN 1 and 7 in 1000 fetuses experience oxygen deprivation during labour that is severe enough to cause fetal death or brain injury [1]–[3]. Currently, non-invasive methods to measure directly the fetal acid-base status and cerebral oxygenation during labour and delivery do not exist, leaving clinicians with only indirect measures of oxygen delivery and neurological function. A standard approach is cardiotocography (CTG), which measures maternal uterine pressure (UP) and fetal heart rate (FHR).

The clinical relevance of some FHR signal characteristics is well understood. The average FHR level, or baseline, reflects the cardiac output; a value within the range 120-160 beats per minute (bpm) indicates

Corresponding author: Philip Warrick (e-mail: philip.warrick@mcgill.ca). The authors acknowledge the financial support of this work by LMS Medical Systems, Inc. and the Natural Sciences and Engineering Research Council of Canada (NSERC).

[1]Biomedical Engineering Department, McGill University, Montreal, Quebec, Canada. [2]School of Computer Science, McGill University, Montreal, Quebec, Canada. [3]Department of Obstetrics and Gynecology, McGill University, Montreal Quebec, Canada. [4]LMS Medical Systems, Inc., Montreal, Quebec, Canada.

that blood delivery is adequate. Small, random fluctuations around baseline (normally 5-15 bpm), known as FHR variability (HRV), indicate that the central nervous system is intact and providing a healthy modulating influence [1], [4].

Temporary decreases in FHR (from 15 s to several minutes in duration and >15 bpm in amplitude) are known as decelerations and reflect events such as compression of the umbilical cord by uterine contractions, malfunction of the fetal heart muscle, or premature separation of the placenta. Generally, larger insults are indicated by recurring episodes of deep, long decelerations whose onsets occur late with respect to the uterine contractions. Temporary increases in FHR (>15 s, >15 bpm), called accelerations, accompany fetal movement and are generally thought to indicate a healthy state.

Visual pattern recognition and inference are the basis of clinical interpretation, but are inconsistently applied and have low specificity [5]. Significant hypoxia is rare while false alarms are common, so physicians often disregard truly abnormal signals [3], [6]. Thus, there is great motivation to find better methods that discriminate between healthy and hypoxic conditions.

In a previous paper [7], we described a new approach to FHR analysis that focuses on the dynamic relationship between UP (as input) and FHR (as output). We used a non-parametric system-identification approach to estimate system dynamics in terms of an impulse response function (IRF) that describes the linear, dynamic relation between very-low-frequency FHR energy (i.e., <30 mHz) and UP. This approach provided robust parameter estimates from noisy CTG collected under clinical conditions.

In this paper, we evaluate the relative merits of parameters estimated from system identification, baseline and HRV modelling for the early detection of pathological fetuses. HRV was characterized quantitatively by using signal processing to subdivide it into distinct energy bands: low frequency (LF, 30-150 mHz), movement frequency (MF, 150-500 mHz) and high frequency (HF, 0.5-1.0 Hz) [8]. These bands are associated physiologically with neural sympathetic and parasympathetic fetal activity (LF), fetal movement (MF) and fetal breathing (HF) and their power spectral densities have been used in previous research for the discrimination of fetal pathology [9], [10].

Using supervised learning, these parameters were used as inputs for classification of sliding window data segments (or *epochs*) from normal and pathological cases. These classifications were used in a detector of pathology. The results from this approach demonstrate that these are discriminating parameters that enable the detection of pathological fetuses early in labour. From a clinical perspective, this capability provides a useful new tool for assisting clinicians in their intra-partum decision-making.

II. Methods

A. Data

We used a database of intrapartum CTG recordings for pregnancies having a birth gestational age greater than 36 weeks and no known genetic malformations [11]. All data was provided in compliance with institutional regulations.

Data collection was performed by clinicians using standard clinical fetal monitors to acquire the CTG. The monitors reported at uniform sampling rates of 4 Hz for FHR (measured in beats per minute (bpm)) and 1 Hz for UP (measured in mmHg), which we up-sampled to 4 Hz by zero-insertion and low-pass filtering. In the majority of cases, the UP or FHR sensors were attached to the maternal abdomen; the FHR was acquired from an ultrasound probe and the UP was acquired by tocography. In a few cases, they were acquired internally via an intra-uterine (IU) probe and/or a fetal-scalp electrode.

B. Overall model

CTG data comprises the UP signal, u and the FHR signal, f. Conceptually, we view FHR generation as originating from three sources: 1) baseline heart rate (producing average cardiac output), 2) response to maternal uterine contractions, and 3) variability due to sympathetic-parasympathetic modulation. Consequently, we modelled f as the sum of three components, $f_{BL} + f_{SI} + f_{HRV}$, as shown in Fig. 1. These components represent the signal energy from three FHR frequency bands modelled as follows:

- the baseline signal f_{BL} (0-4.5 mHz) modelling the lowest frequency FHR by a linear trend.
- the system identification signal f_{SI} (4.5-30 mHz) modelling the fetal response to the maternal uterine contractions as the convolution of the measured input UP u with an impulse response function.
- the variability signal f_{HRV} (30mHz-1.0 Hz) modelling the highest frequency FHR as the output of an auto-regressive model driven by some (unknown) Gaussian white input noise v.

The computation of these three estimates is elaborated below.

C. Preprocessing

The CTG data was recorded in a clinical setting, so it was subject to specific types of noise. The loss of sensor contact can temporarily interrupt the UP or FHR signals, and interference from the (much lower) maternal heart rate can corrupt the FHR. These both appeared in the signal as a sharp drop to much lower amplitude followed by a sharp signal restoration. As described in [7], we preprocessed the data to bridge interruptions lasting less than 15 s and remove segments containing longer interruptions.

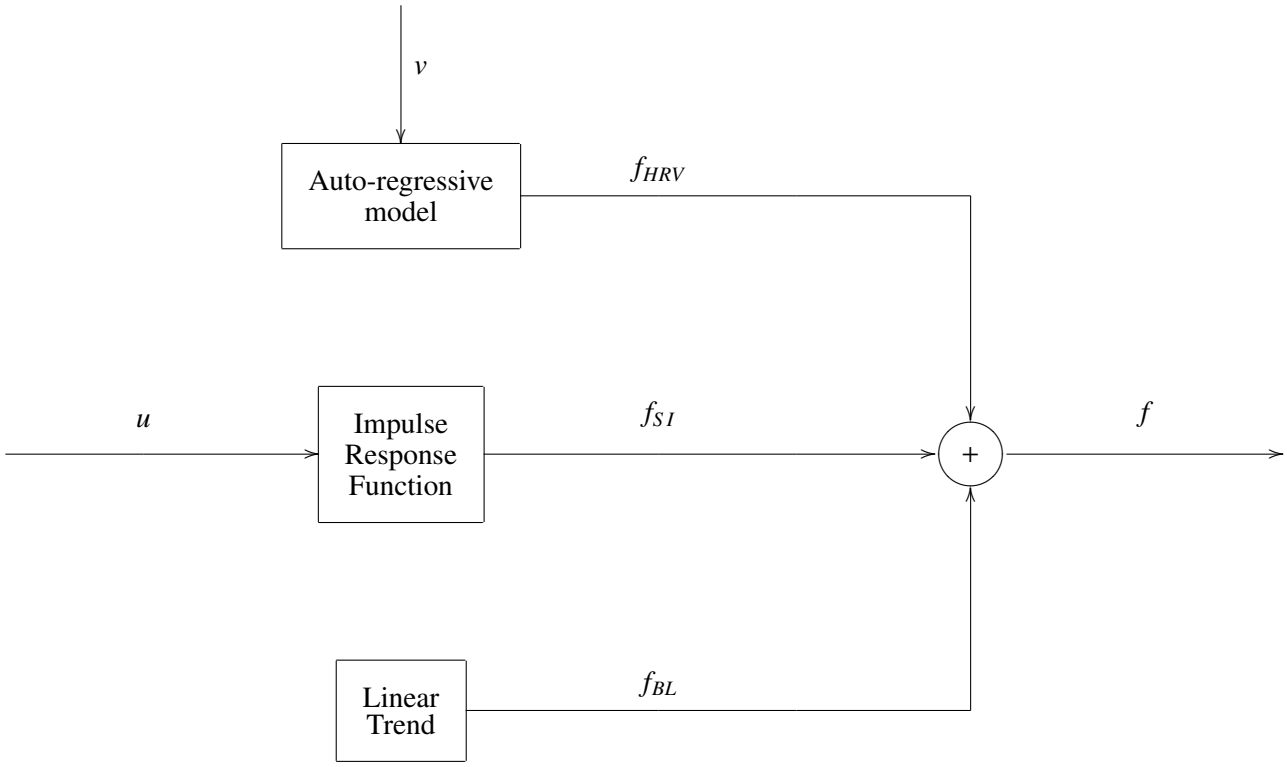


Fig. 1. Modelling FHR f. The FHR baseline f_{BL} is modelled as a linear trend of the lowest frequency FHR. The fetal response to uterine pressure f_{SI} is modelled as the convolution of the measured uterine pressure u with an impulse response function. The heart-rate variability f_{HRV} is modelled as the response of an auto-regressive model to an (unknown) Gaussian white noise source v.

In a stationary system, the quality of model estimates obtained by system identification can be expected to increase with data length. However, the state of the fetus changes with time [5]. Thus, there is a trade-off between selecting a record length that is long enough for successful system identification but short enough to avoid significant non-stationarities. To resolve this, we extracted 20-min epochs with 10-min overlap between successive epochs. This epoch length typically spanned several UP contraction-FHR deceleration pairs. As many epochs as possible were extracted starting from the beginning of each clean (artifact-free) segment; to include any remaining data at the end of the segment (i.e., < 10 min), the overlap was increased for the last epoch to a maximum of 18 min.

D. Baseline

The FHR was digitally low-pass filtered with a cut-off frequency of 4.5 mHz, to remove power due to decelerations, and decimated to a sampling rate of 0.5 Hz. This filter and the HRV filter, described later, were designed using the Parks and McClennan FIR algorithm [12] with a -70 dB stopband. We estimated baseline by performing a linear fit of the filtered FHR, as illustrated in Fig. 3(a), and characterized by the offset and slope.

E. System Identification

To detrend the FHR in preparation for system identification, the signal was filtered with the high-pass filter complementary to the low-pass filter used to estimate the baseline and decimated to 0.5 Hz. The same processing was applied to the UP. The IRF **h** was estimated from the resulting signals.

As described in [7], let the input, UP, and output, FHR, at time sample n ($n = 1 \ldots N$) be denoted by u_n and f_n, respectively. The linear dynamic response f_n to an arbitrary input signal u_n is given by the convolution sum:

$$f_n = \sum_{i=d}^{d+M-1} (h_i \Delta t) u_{n-i} = \mathbf{h} * \mathbf{u}_n \tag{1}$$

where Δt is the sampling period, and **h** is the IRF beginning at sample d, and of length M. d is also called the delay and M is called the memory length. $\mathbf{u}_n$ is the length-M vector of input samples $[u_{n-d-M+1} \ \ldots \ u_{n-d-1} \ u_{n-d}]$ used to compute f_n at sample n. For causal (physically realizable) systems, $d \geq 0$, but in the presence of an input measurement delay, d may be negative [13]. The UP tocography signal depends on the pressure-sensor contact and abdominal tissue thickness [14], which may result in a measurement delay, so negative d is possible. In contrast, we expect the physiological response to have a positive delay. The combination of these two delays can produce an FHR response that occurs before or after the measured UP contraction onset (i.e., d could be negative or positive).

Several processing steps were used to obtain a good **h** estimate, as described in [15]. We first obtained the least-squares estimate

$$\hat{\mathbf{h}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{f} \approx \boldsymbol{\Phi}_{uu}^{-1}\boldsymbol{\phi}_{uf} \tag{2}$$

where **U** is an $N \times M$ Toeplitz matrix formed from $\mathbf{u}_n$, **f** is the length-N measured output, and for $N \gg M$, $\mathbf{U}^T\mathbf{U}$ and $\mathbf{U}^T\mathbf{f}$ are estimated by the input autocorrelation matrix $\boldsymbol{\Phi}_{uu}$ and the input-output cross-correlation $\boldsymbol{\phi}_{uf}$, respectively [13]. We used a pseudo-inverse method to represent the IRF by its most significant principal components, in order to suppress noise [7], [16].

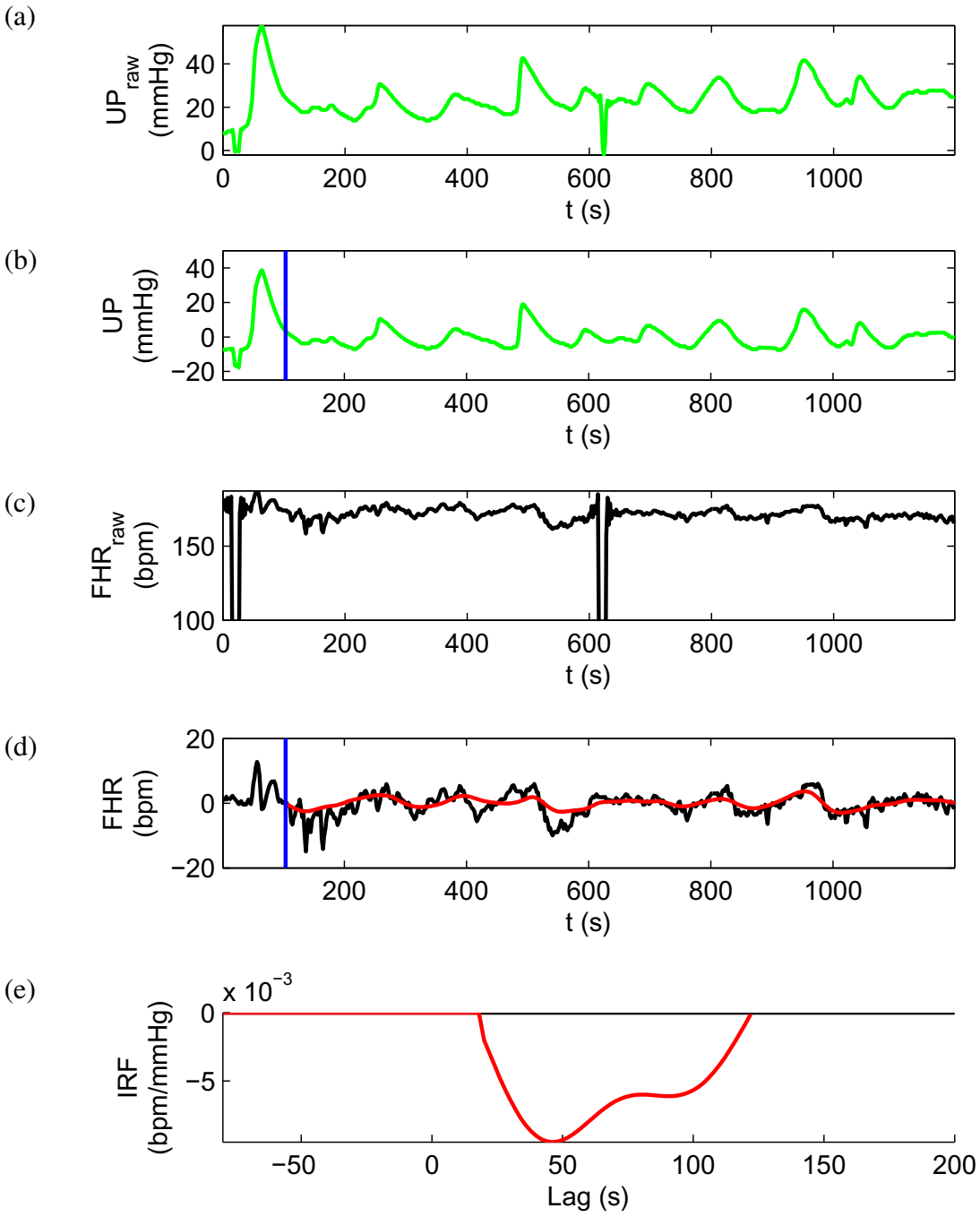


Fig. 2. System identification modelling results for an typical epoch of a pathological case. (a) raw input UP (b) pre-processed UP (c) raw output FHR (d) preprocessed (black) and predicted (red) output FHR (e) final IRF. The IRF delay d was 20 s, the gain G was -0.32 bpm/mmHg and the VAF of the model was 44.0.

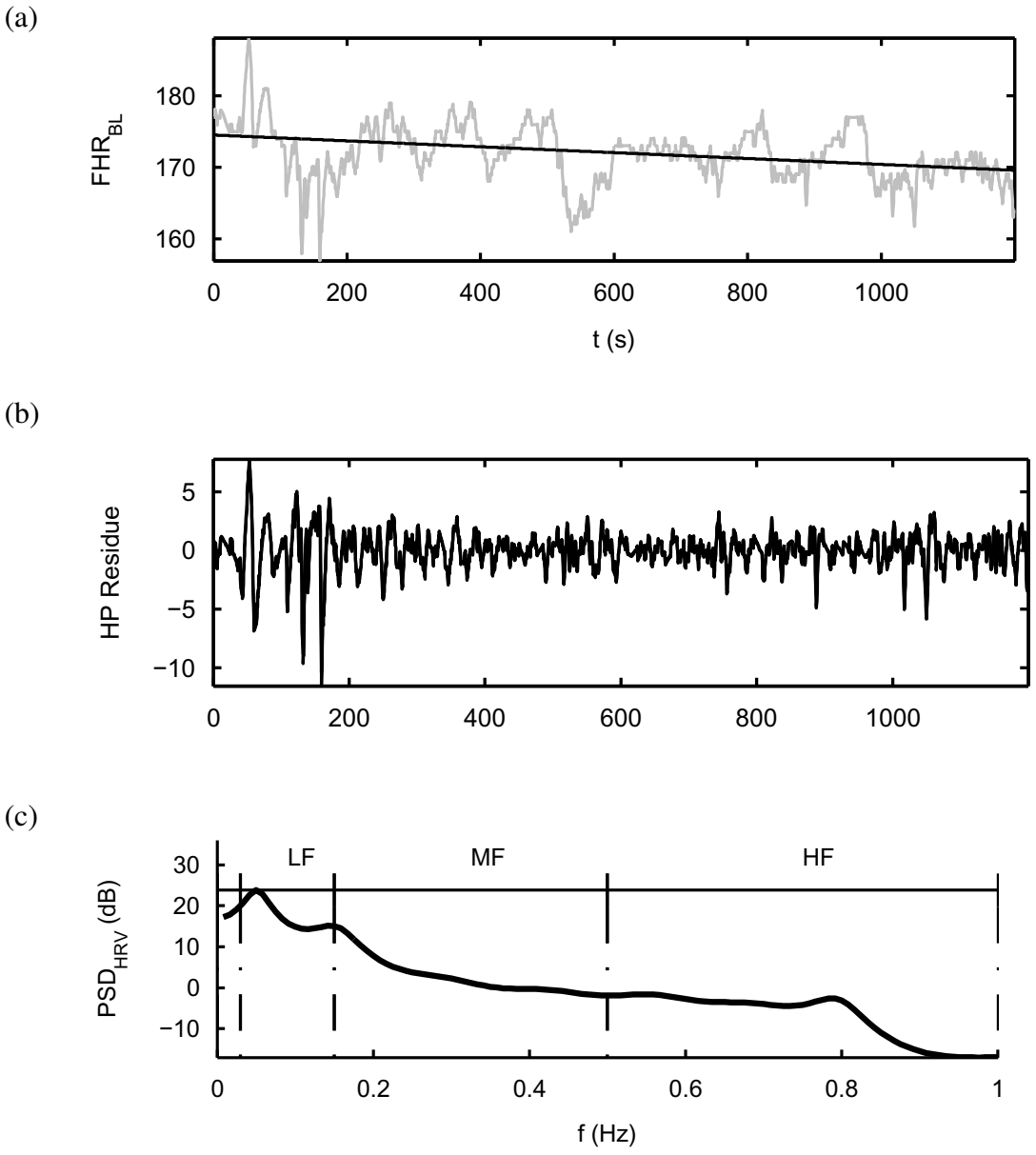


Fig. 3. Model estimates from an epoch of a typical pathological case for (a) the fetal heart-rate baseline FHR_{BL} (black line); this is a linear fit of the lowest frequency band of the FHR (0-4.5 mHz), shown in grey, (b) the HRV band signal; the estimated HRV was virtually identical with a VAF=93.8%, and (c) HRV measured as the power spectral density PSD_{HRV} over the LF (30-150 mHz), MF (150-500 mHz) and HF (0.5-1.0 Hz) frequency bands. Note that we considered the lowest band, VLF (0.0-30 mHz), in the system identification model; before estimating the HRV, we attenuated this band by a 30 mHz high-pass filter.

The input UP of the CTG data was narrow-band and the IRF memory was a significant proportion of the epoch, leading to noisy IRF estimates. To ensure that our model captured true system dynamics rather than a spurious input-output correlation due to noise, we also estimated a large number (99) of models from surrogate FHR. Since the FHR spectrum was far from white, the surrogates were generated using the amplitude-adjusted Fourier-transform method (AAFT) [17], [18], which approximates the original spectrum but scrambles the phase. Models were generated for both original and surrogate data. Only models of the original data that ranked in the top five in terms of *variance accounted for* (VAF) [19] were retained. Under these conditions, there was only a 5% probability that a top ranking occurred by coincidence, giving greater confidence that we retained valid models of system dynamics.

Fig. 2 shows the input and preprocessed UP and FHR, the IRF and the model-predicted FHR for an epoch from a typical pathological case. For this example, the IRF delay d was 20 s. We also estimated the *steady state gain* $G = \sum_{i=0}^{M-1} h_i$, reflecting the magnitude of the response. For this example, G was -0.32 bpm/mmHg. The VAF of this model was 44.0%.

F. Heart-rate variability (HRV)

The FHR was detrended by a high-pass filter with cutoff frequency 30 mHz, corresponding to the lower limit of the LF band of fetal HRV [9], and decimated to 2.0 Hz to include the 1.0 Hz upper limit of the HF band of fetal HRV.

As shown in Fig. 1, we model the fetal HRV f_{HRV} as the output of an auto-regressive (AR) model given some unknown Gaussian white input noise v. We estimated the AR coefficients using linear prediction and chose the order p using the minimum-description length (MDL) criterion. Over the entire set of cases, the minimum, median and maximum p selected in this way were 6, 23 and 40, respectively. Power spectral densities (PSD) of the AR models were computed following [8] and sampled at 120 frequencies to resolve the majority of the resonances.

We summarized the PSD by the energy of three bands: LF (30-150 mHz), MF (150-500 mHz) and HF (0.5-1.0 Hz) [8], [9]. Relative band energies (in units of bpm^2) were calculated by summing the spectrum over each frequency band; these sum to 1 because of the assumption that unity variance Gaussian white noise is the input to the AR model. This result was scaled by the variance of the filtered HRV to calculate absolute band energies. The square root of these energies gave the parameters HRV_{LF}, HRV_{MF} and HRV_{HF}, in more interpretable units of bpm.

Fig. 3(b) shows the HRV signal and Fig. 3(c) shows its estimated PSD. The relative band energies were proportional to the indicated areas for LF, MF and HF. The expected low-pass nature of the spectrum [8]

is evident. Note that the very-low frequency (VLF) band (0.0-30 mHz) was accounted for in the system identification model, so it was attenuated by the 30 mHz high-pass filter. Hence it did not have an impact on the HRV estimate.

G. Classification

1) Data: Each recording was labelled by outcome according to the fetal, arterial, umbilical-cord, base deficit and associated neonatal indications of neurological impairment. An elevated base deficit is an important indicator of metabolic acidosis large enough to cause neurological injury [5], [20], [21]. The majority of the recordings were from normal fetuses (base deficit $<$ 8 mmol/L); these were collected from consecutive births at an urban university teaching hospital, which used CTG extensively and routinely measured umbilical artery blood gases at birth. The rest were severely pathological (base deficit $\geq$ 12 mmol/L and death or evidence of hypoxic ischemic encephalopathy). The very low natural incidence of pathology necessitated collecting cases from a number of hospitals and medico-legal files.

All CTG records comprised at least three hours of recording; the selected 263 records were a subset of a larger database of several thousand records. The recording was included if there were at least three epochs in which an IRF could be identified. Only those epochs having a valid system identification model were considered, in order to allow a comparison of the relative information content of the models derived from system identification and baseline-HRV. Of the 220 normal and 43 pathological cases in the initial subset, 187 normal and 26 pathological cases fulfilled these criteria.

2) Attributes: We generated attributes for classification from the model parameters. We chose attributes that showed statistically significant differences between the normal and the pathological cases. Statistical significance was assessed by comparing class distributions with the Kolmogorov-Smirnov (K-S) hypothesis test, which uses the maximum difference of the empirical cumulative distribution functions (cdfs) as a test statistic. We rejected the null hypothesis at the $p < 0.05$ significance level.

Two attributes were retained from the system identification models: the delay d, which measures the timing of the FHR response to UP, and the steady-state gain G, which reflects the strength of the response to contractions. Three attributes were included from the baseline and HRV models: the offset of the linear fit FHR_{BL}, the low frequency HRV band level HRV_{LF} and the movement band level HRV_{MF}.

3) Per-epoch classification: The attributes for each epoch and the outcome labels from the database were used to generate instances for classification. Instances were selected only from epochs with valid system identification models; consequently, the number of the instances per fetus varied. The label given to all instances of a fetus was the known fetal state at birth (despite the fact that the state can change

during labour).

We chose a discriminative classification approach because it was simple, intuitive and permitted efficient and robust learning. Specifically, we generated two support-vector machine (SVM) classifiers: one using the system identification attributes and another using the baseline-HRV attributes. SVM classification was chosen because it gives state-of-the-art performance, is robust to overfitting, and computationally efficient implementations are available. SVM learning algorithms process labelled data to determine an optimal decision boundary. This boundary is represented using a subset of the data called the support vectors. The decision boundary can be linear or nonlinear, depending on the kernel chosen. We used a Gaussian kernel to allow a nonlinear boundary.

These per-epoch classifiers predict the fetal state at each epoch using attributes computed from models of that epoch. In addition, the decisions of these two classifiers were logically ORed at each epoch to generate a combined classification.

4) Detection of pathological fetuses: It is reasonable to assume that the more often the classifier predictions are pathological, the higher is the likelihood that the fetus is truly pathological. To measure this, we implemented a detector that computed a running sum of the per-epoch pathological classifications; once this count attained a certain threshold, a pathological detection was deemed to have occurred. We applied this detection process to the system identification, baseline-HRV and combined classifiers. The combined detector should lead to earlier detection if the decisions of the two classifiers are independent.

5) Empirical setup: Multiple ten-fold cross-validation simulations were performed. The folds were selected on a per-fetus basis; that is, all instances generated from a particular fetus were used either in training or testing, but not both. Each fold was partitioned into test and training data, and a subset of the training data was used as validation data to choose the SVM hyper-parameters, as we describe below. For each cross validation, the mean and standard error of the receiver-operating characteristic (ROC) was computed for classifiers over all cross-validation folds. We report the mean values of these statistics over five repetitions of cross validation.

Each SVM had two hyperparameters to select: the Gaussian kernel width σ and the misclassification cost parameter α. We chose σ using the per-epoch classifiers and α using the subsequent detection step.

To choose the kernel width σ, we first generated ROC curves for each value of σ by varying the α hyperparameter over a range that favoured either sensitivity or specificity. The hyper-parameter σ that generated the ROC with the highest area-under-curve (AUC) was selected, as measured on the validation data. Then, with σ fixed, we generated detector ROCs for each value of α by varying the threshold from 1 to 6 epochs. We chose the value of α that generated the detector ROC with the highest AUC, as

measured on the validation data. In both cases, the AUC was calculated over a limited range of clinically useful false positive rates (i.e., < 0.3). We used these selected σ and α to train SVMs on the entire training data and used the resulting classifiers in the final detectors.

III. Results

A. Time progression of features

Fig. 5 shows the average model parameters at each epoch for the two groups. Delivery is considered to have occurred at epoch 0. All features had significant differences in their class statistics at some epochs but not at others. Differences tended to occur earlier in labour and delivery and diminish near the end of delivery, although the delay, d, was consistently higher for pathological cases throughout. All parameters tended to be larger for pathological cases. At least one parameter was significantly different more often for system identification (13 epochs) than for baseline-HRV (8 epochs). Note that the K-S test that is used to determine statistical significance is very conservative, so the differences may be significant more often than indicated.

B. Illustration of the classification using system identification features

Fig. 6 shows all the instances (including the support vectors outlined in black) from one fold of training data, the value of the classifier decision function $\mathcal{H}$ (as shown by shades of gold and turquoise), and the decision boundary $\mathcal{H} = 0$ based on the system identification feature set.

We note that there are two regions in which instances are classified as pathological; the most heavily populated (lower right) is characterized by long delay and large negative gain; a smaller population in the upper left region are instances characterized by short delay and large positive gain. Between these pathological regions there is a region in which instances are classified as normal. At the boundaries of these regions are the support vectors, where classification is less certain. The large proportion of support vectors (684 of 1499 training samples) indicates that the classification problem is difficult.

The trajectory of one pathological case, not included in the training set, is shown by the red arrows. It began in one of the support vector regions, in which intuitively classification is somewhat uncertain, then moved into the normal region, passed through the other support vector region and finally ended in the pathological region. This suggests that this case deteriorated from a normal to a pathological state over time. We observed other pathological cases with similar behaviour. We also observed normal cases that started normal and ended close to or within the pathological region near delivery.

C. Per-epoch system identification and baseline-HRV classifiers

Fig. 4 shows ROCs for the per-epoch system identification and baseline-HRV classifiers. The system identification classifier (AUC = 0.140±0.020) performed marginally better than the baseline-HRV classifier (AUC = 0.131±0.018), but the differences are not compelling and neither classifier did particularly well. This is not surprising since the state of the fetus degrades over time; some normal cases may have transitioned towards a pathological state late in delivery while some pathological cases may have been in a normal state early in labour and delivery. Unfortunately, it is impossible to obtain a better labelling of the data based on the available clinical information. Hence, we used a detection approach to improve accuracy.

D. Detection of pathological fetuses

The detectors used the history of per-epoch classifications for each fetus to detect pathology. Six detectors were examined using different per-epoch classifiers and thresholds, as shown in Table I. Only thresholds 1 and 2 are shown because detectors with higher thresholds performed worse. Fig. 7 shows the performance of the 6 detectors in terms of pathological detection (sensitivity) and normal detection (specificity) over time. Higher detection is better for both measures. Error bars are omitted because there was little plotting overlap and a clear ranking can be observed. It is apparent that pathological detection is more conservative (i.e., delayed, as indicated by a shift to the right) for the higher threshold. The combined detectors (C1 and C2) identified pathological cases earlier and consistently better than equivalent individual detectors.

Detector	**Per-epoch classifier**	**Threshold (epochs)**
S1	System identification	1
S2	System identification	2
B1	Baseline-HRV	1
B2	Baseline-HRV	2
C1	System identification ∪ Baseline-HRV	1
C2	System identification ∪ Baseline-HRV	2

TABLE I

DETECTOR DESCRIPTIONS

Selecting the best performing detector must consider both performance measures. We consider C2 to

be the best detector because it had close to the best detection of pathological cases and close to the best false positive rates, especially in the first half of the three hour record when a clinical response is most important. C2 detected half of the pathological cases with a false positive rate of 7.5% at epoch -10 (i.e., roughly one hour and forty minutes before the original time of delivery). In comparison, while C1 had that best detection of pathological cases, it had the worst false positive rates.

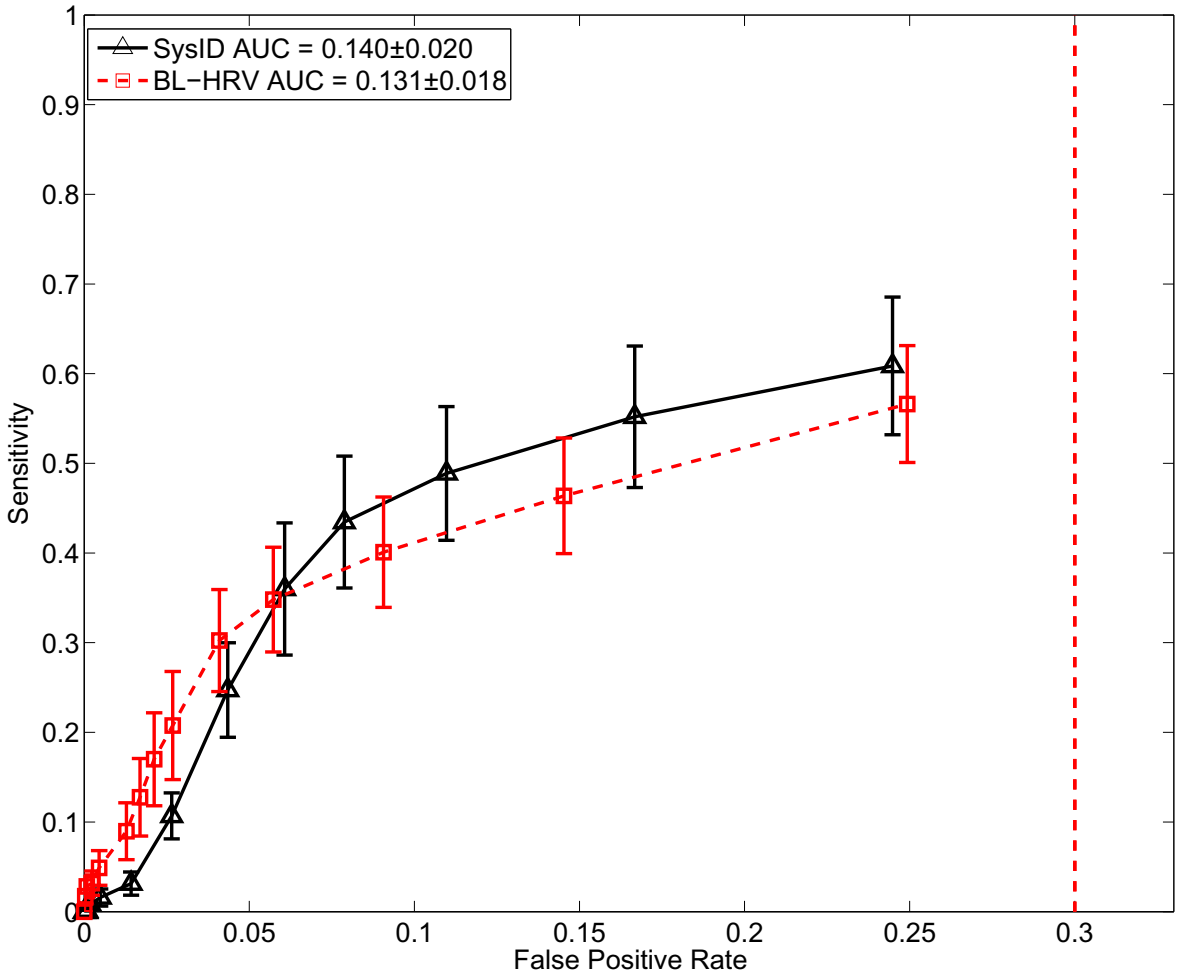


Fig. 4. Receiver-operator characteristic (ROC) comparison for system identification and baseline-HRV per-epoch classifiers. Each ROC was generated using a constant γ of the SVM Gaussian kernel and varying the cost associated with misclassifying normal or pathological examples. The bars indicate the standard error of the ROC estimate over the ten cross-validation folds. The area under the curve (AUC) was calculated for false positive rates less than 0.3.

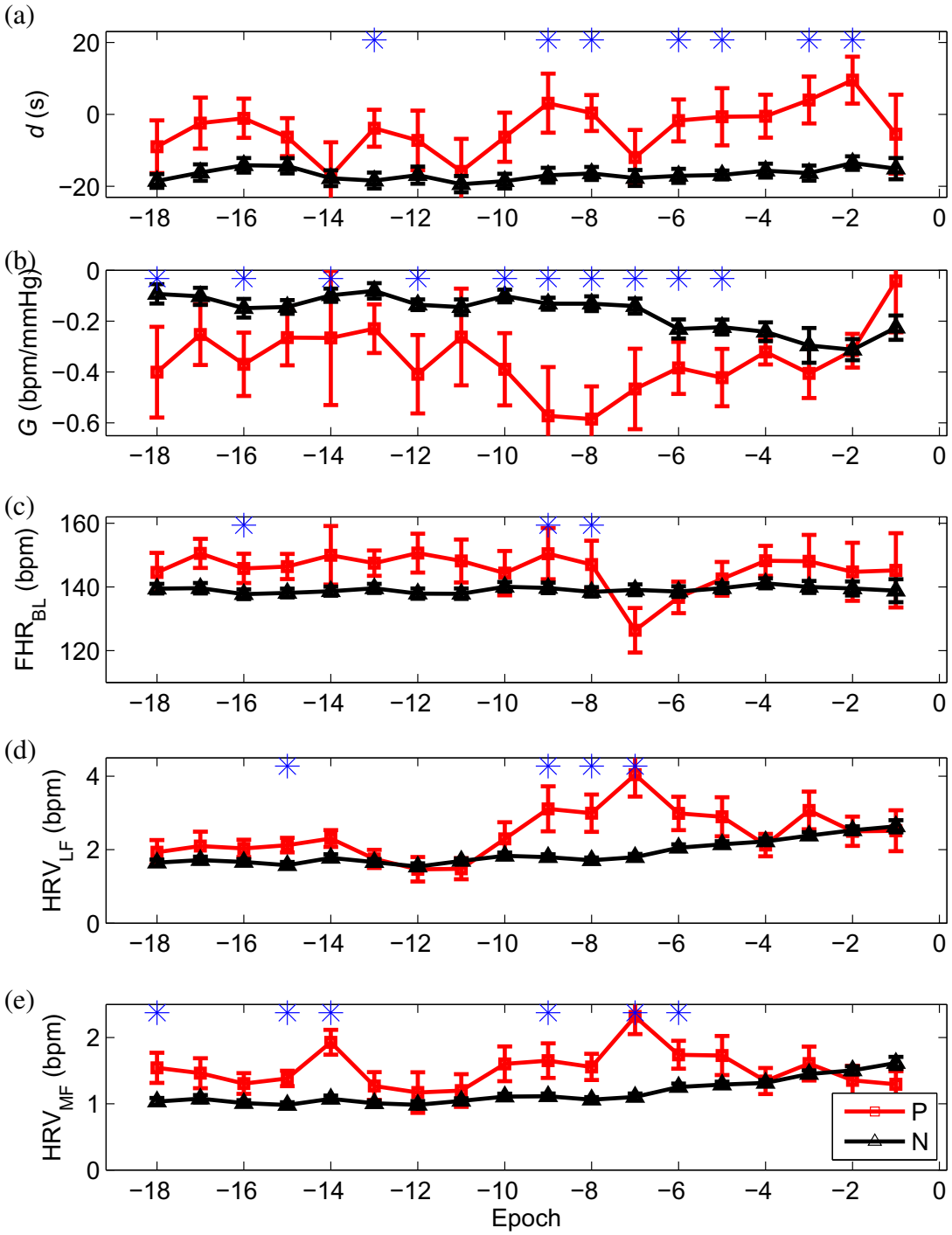


Fig. 5. Time progression of model parameters in the last three hours of labour and delivery for normal (N, black triangles) and pathological (P, red squares) cases. The mean and standard error are plotted by epoch for (a) delay d, (b) gain G (c) baseline FHR_{BL}, (d) low-frequency band heart-rate variability HRV_{LF} and (e) movement-frequency band heart-rate variability HRV_{MF}. For each parameter, epochs where hypothesis tests indicated significant differences between classes ($p < 0.05$) are indicated by stars.

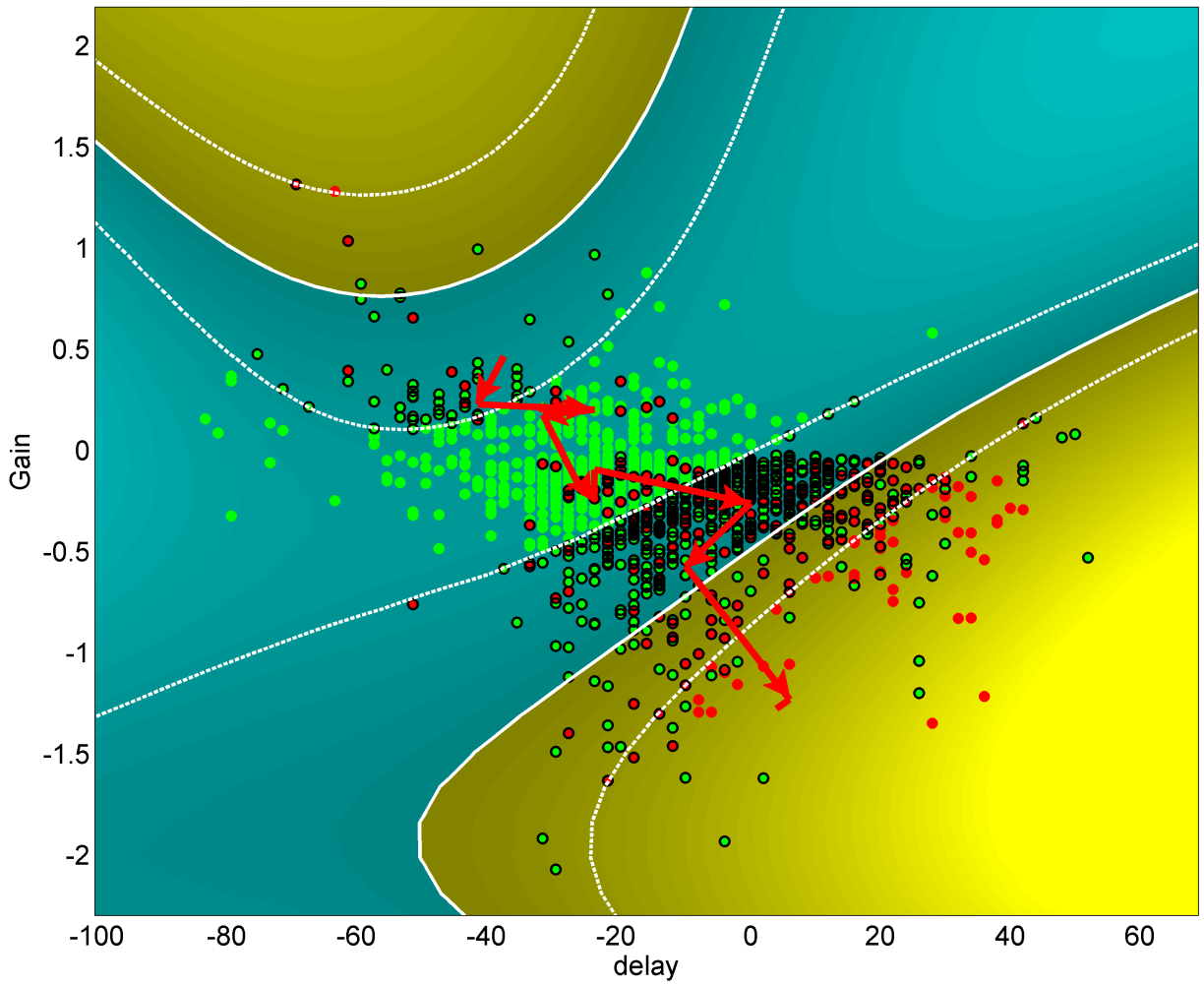


Fig. 6. Feature and decision space for one fold of a system identification classifier. Training data are labelled by green (normal) and red (pathological) circles. The classifier decision function $\mathcal{H}$ is indicated by turquoise ($\mathcal{H} < 0$, normal) and gold ($\mathcal{H} > 0$, pathological) regions separated by solid white lines ($\mathcal{H} = 0$). Most of the support vectors (684 of 1499 training samples, outlined in black) lie the two regions between the contours $\mathcal{H} = \pm 1$ (dotted white lines), where classification is less certain. The trajectory highlighted by a solid red line is a non-training pathological case whose classification transitions from normal to pathological. The arrows indicate increasing time.

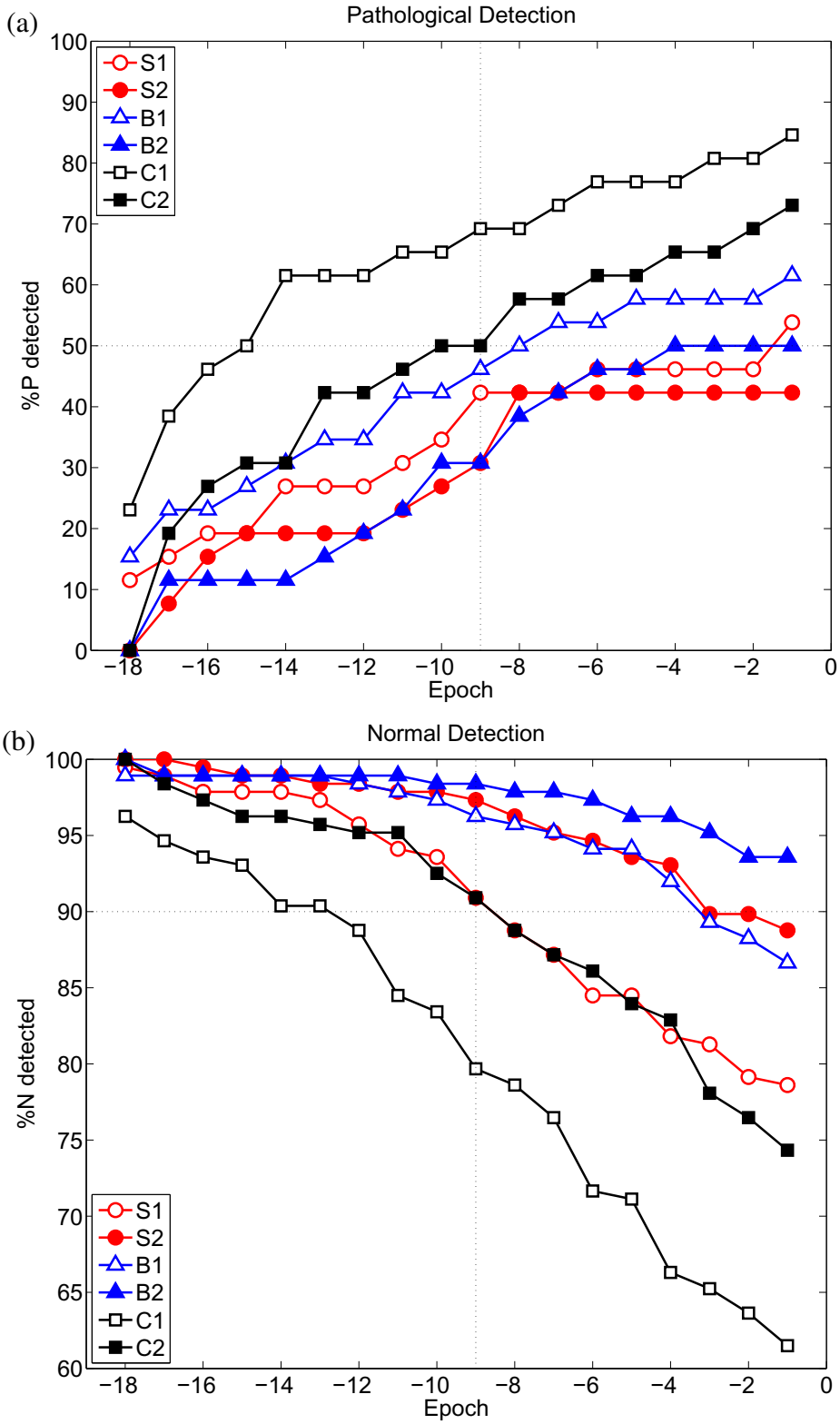


Fig. 7. (a) Pathological and (b) normal detection over time for selected system identification (S1 and S2, red circles), baseline-HRV (B1 and B2, blue triangles) and combined (C1 and C2, black squares) detectors. The cumulative count is indicated by open (threshold = 1) and filled (threshold = 2) markers. The vertical dotted lines indicate the time of 90 minutes before delivery. The horizontal dotted lines indicate the 50% pathological and 90% normal detection levels.

IV. Discussion

The major finding of this study is that a combination of system identification and baseline-HRV features can be incorporated successfully in a very good detector of fetal distress. It detected correctly half of the pathological cases, with acceptable false positive rates (7.5%), early enough to permit clinical intervention. This detector was superior to alternatives using either feature set by itself.

By definition, the pathological cases in our database had been missed by clinicians. Therefore, it is very significant that our automated approach detected half of these cases near the midpoint of the three hour records. It is interesting that this corresponds well to the clinical fact that approximately 50 % of birth-related brain injuries are deemed preventable, with incorrect CTG interpretation leading the list of causes [3], [6], [22], [23].

Timing of detection is very important given that fetal state evolves; detecting fetal distress near the time of delivery has less potential to improve clinical outcomes, while an advance warning of one hour and forty minutes is very significant clinically. This is a relatively long time for treatment to occur and improve outcome; typically, the interval between a decision to intervene and Cesarian birth is less than 30 minutes [24]. Furthermore, the cost of believing these decisions (i.e., a rate of unnecessary Cesarian sections of 7.5%) is acceptable clinically.

A. Non-stationarity

The non-stationarity of both the attributes and the fetal state poses several challenges to the detection of pathology. The later a detection occurred, the more likely it was that the fetus was a normal case that had deteriorated with time. This is because near delivery, attributes from normal and pathological cases tended to become closer in value and false positive rates rose, as shown in Figs. 5 and 7. Early detection gives better confidence that a case is indeed pathological. On the other hand, pathological cases frequently began with parameters that were normal and deteriorated over time, as indicated by their trajectory in decision space. This created problems for per-epoch classification, meaning that several instances (epochs) may be mislabelled. However, this is the best that can be done given the data we have, since the true state is not observed during labour and delivery.

We addressed this problem by introducing a detector of pathology defined by a threshold of accumulated pathological classifications. This resulted in a more consistent interpretation of the classifications over time despite the non-stationary and noisy nature of the data. The detector avoids the confusion of decision oscillations by restricting decision transitions to be from a normal decision to a pathological one.

It might also be useful to incorporate knowledge of the trajectory in feature space. One approach worth investigating would be to relabel cases according to the decisions of our detector, and re-train the classifiers. This approach is more akin to semi-supervised learning: one would essentially treat the examples with a more certain classification as labelled, and the ones with uncertain classifications as unlabelled data, whose labels are iteratively improved by using a sequence of classifiers. We have a database of intermediate cases, not used in this study, for which such an approach would be useful. An alternative approach that we had considered was to use expectation maximization in order to fit a probabilistic model of evolution from a normal to a pathological state. However, given the small number of pathological cases, and the difficulty of choosing initial parameters, we decided to switch to using discriminative classification. The approach we presented in the paper is more data-efficient and robust to noise than any of the probabilistic architectures we tried.

It is notable that there were two pathological decision regions: the lower right region of Fig. 6 associated with deep and long-delay decelerations and the upper left region reflecting acceleration response to contractions. These infrequent cases of acceleration response occurred early and gradually evolved to decelerations, as in the example shown. Further investigation is required to determine whether such acceleration responses could be precursors to pathology.

B. Data

We obtained these very encouraging results despite the use of noisy data collected under clinical conditions and therefore subject to frequent artifacts related to sensor disturbances. We also defined pathology to include only severely compromised fetuses; other databases used for fetal-state classification have used milder definitions of pathology (c.f., [25]–[30]). Therefore, we have relatively few of these cases, and a slight variation in the case selection can affect results. Our database is nevertheless unique, and required significant time for compilation; the relatively low number of pathological cases (as we have defined pathology) is actually high compared to other known databases.

Because of the prevalence of data artifact, decisions were often based on a small proportion of the data record. However, with a more rigorous data-collection protocol generating cleaner and more informative data (e.g., using better calibrated UP [7] and using R-R intervals for HRV estimates [31]), we would expect better results in terms of accuracy and timing since we could obtain better models and a decision could be made at more epochs.

C. Clinical relevance

The system identification and baseline-HRV detectors gave comparable results, and combining these feature sets further improved results, as shown in Fig. 7. This is consistent with clinical expectation on the importance of both the deceleration response to UP and baseline-HRV as indicators of fetal well-being [4].

It is natural to inquire why half of the cases were not detected. We lumped all pathological cases together using the criteria of base deficit value and indications of neurological impairment. However, these cases likely represent more than one population corresponding to different pathological conditions during labour and delivery. Some may have experienced severe, acute events near delivery (giving little time to respond) while others may have had different underlying chronic conditions during pregnancy (e.g. intrauterine infection) making the fetus more vulnerable during labour and delivery [4]. Unfortunately, this information was not consistently available in our database; hence we cannot label these sub-populations and perform more precise classification.

Our methodology uses the simplicity and robustness of discriminatory classification as its foundation. However, because of the non-stationarity of these attributes and the evolution of the state of the fetus during labour and delivery, using classification alone is problematic. To overcome this, our thresholding approach for detection incorporates timing in a way that is simple and intuitive. It is also sensible clinically since it is known that persistence of "non-reassuring" states is associated with pathology [5]. Implementing the detector in a clinical setting will require more research into the issue of what is the best way to present the detection results to the clinician.

References

[1] ACOG, *Neonatal Encephalopathy and Cerebral Palsy: Defining the Pathogenesis and Pathophysiology*. ACOG Task force on Neonatal Encephalopathy and Cerebral Palsy, January 2003.

[2] N. Badawi, J. Kurinczuk, J. Keogh, L. Alessandri, F. O'Sullivan, P. Burton, P. Pemberton, and F. Stanley, "Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study," *BMJ*, vol. 317, pp. 1549–1553, 1998.

[3] E. Draper, J. Kurinczuk, C. Lamming, M. Clarke, D. James, and D. Field, "A confidential enquiry into cases of neonatal encephalopathy," *Arch Dis Child Fetal Neonatal Ed*, vol. 87, pp. F176–F180, 2002.

[4] R. Freeman, T. Garite, and M. Nageotte, *Fetal Heart Monitoring*. Philadelphia, PA: Lippincott Williams and Wilkins, 2003.

[5] J. T. Parer, T. King, S. Flanders, M. Fox, and S. J. Kilpatrick, "Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association?" *Journal of Maternal-Fetal & Neonatal Medicine*, vol. 19, no. 5, pp. 289–294, May 2006.

[6] S. Ransom, D. Studdert, M. Dombrowski, J. Mello, and T. Brennan, "Reduced medicolegal risk by compliance with obstetric clinical pathways: A case-control study," *Obstet Gynecol*, vol. 101, 2003.

[7] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney, "Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate for normal and hypoxic fetuses," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1587–1597, June 2009.

[8] S. Cerutti, S. Civardi, A. Bianchi, M. Signorini, E. Ferrazzi, and G. Pardi, "Spectral analysis of antepartum heart rate variability," *Clin. Phys. Physiol. Meas.*, vol. 10, pp. 27–31, 1989.

[9] M. Signorini, G. Magenes, S. Cerutti, and D. Arduini, "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 365–374, 2003.

[10] M. Ferrario, M. Signorini, G. Magenes, and S. Cerutti, "Comparison of entropy-based regularity estimators: application to the fetal heart rate signal for the identification of fetal distress," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 1, pp. 119–125, 2006.

[11] E. Hamilton, A. Dyachenko, C. Elliott, P. Warrick, and A. Ciampi, "Progression of intrapartum EFM patterns in births with symptomatic metabolic acidosis," *American Journal of Obstetrics and Gynecology*, vol. 197, no. 6, p. s182, 2007.

[12] L. Rabiner, J. McClellan, and T. Parks, "FIR digital filter design techniques using weighted Chebyshev approximations," *Proc. IEEE*, vol. 63, 1975.

[13] I. W. Hunter and R. E. Kearney, "Two-sided linear filter identification," *Medical & Biological Engineering & Computing*, vol. 21, pp. 203–209, 1983.

[14] T. Vanner and J. Gardosi, "Intrapartum assessment of uterine activity," *Bailliere's Clinical Obstetrics And Gynaecology*, vol. 10, no. 2, pp. 243–257, June 1996.

[15] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton, "Linear models of intrapartum uterine pressure-fetal heart rate interaction for the normal and hypoxic fetus," in *Proceedings of the 2006 IEEE Engineering in Medicine and Biology 28th Annual Conference*, 2006, pp. 6434–6437.

[16] ——, "System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses," in *Computers in Cardiology 2006. Proceedings.*, vol. 33, 2006, pp. 937–940.

[17] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77–94, Sept. 1992.

[18] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, Aug. 2000.

[19] D. T. Westwick and R. E. Kearney, *Identification of nonlinear physiological systems*. Hoboken NJ: Wiley-Interscience, 2003.

[20] J. Low, R. Victory, and E. Derrick, "Predictive value of electronic fetal monitoring for intrapartum fetal asphyxia with metabolic acidosis," *Obstet Gynecol*, vol. 93, pp. 285–291, 1999.

[21] A. MacLennan, "A template for defining a causal relation between acute intrapartum events and cerebral palsy: international consensus statement," *BMJ*, vol. 319, no. 7216, pp. 1054–1059, 1999.

[22] C. Saphier, E. Thomas, D. S. D, T. Brennan, and D. Acker, "Applying no-fault compensation to obstetric malpractice claims," *Prim Care Update Ob Gyns*, vol. 5, pp. 208–9, 1998.

[23] B. Stalnaker, J. Maher, G. Kleinman, J. Macksey, L. Fishman, and J. Bernard, "Characteristics of successful claims for payment by the Florida Neurologic Injury Compensation Association Fund," *Am J Obstet Gynecol*, vol. 177, pp. 268–71, 1997.

[24] S. Bloom, K. Leveno, C. Spong, S. Gilbert, J. Hauth, M. Landon, M. Varner, A. Moawad, S. Caritis, M. Harper, R. Wapne, Y. Sorokin, M. Miodovnik, M. O'sullivan, B. Sibai, O. Langer, S. Gabbe, and the National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network, "Decision-to-incision times and maternal and infant outcomes." *Obstet Gynecol.*, vol. 108, no. 1, pp. 6–11, July 2006.

[25] K. Maeda, M. Utsu, A. Makio, M. Serizawa, Y. Noguchi, T. Hamada, K. Mariko, and F. Matsumoto, "Neural network computer analysis of fetal heart rate," *Journal of Maternal-Fetal Investigation*, vol. 8, no. 4, pp. 163–171, Dec. 1998.

[26] T. K. H. Chung, M. P. Mohajer, Z. J. Yang, A. M. Z. Chang, and D. S. Sahota, "The prediction of fetal acidosis at birth by computerised analysis of intrapartum cardiotocography," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 102, no. 6, pp. 454–460, 1995.

[27] J. Skinner, J. Garibaldi, J. Curnow, and E. Ifeachor, "Intelligent fetal heart rate analysis," in *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, 2000, pp. 14–21.

[28] G. Georgoulas, C. Stylios, G. Nokas, and P. Groumpos, "Classification of fetal heart rate during labour using hidden markov models," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 3, 2004, pp. 2471–2475 vol.3.

[29] M. Signorini, M. Ferrario, L. Pedrinazzi, and G. Magenes, "Analysis of echographic and heart rate time and frequency domain parameters for the antepartun fetal surveillance," in *Computers in Cardiology, 2005*, 2005, pp. 679–682.

[30] H. Cao, D. Lake, I. Ferguson, J.E., C. Chisholm, M. Griffin, and J. Moorman, "Toward quantitative fetal heart rate monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 1, pp. 111–118, 2006.

[31] G. Clifford and L. Tarassenko, "Quantifying errors in spectral estimates of HRV due to beat replacement and resampling," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 630–638, 2005.

CHAPTER 7
Appendix to Journal article III

The following table shows the sensitivity and specificity values (given as false positive rates) used to generate the plots for all the classifiers of Fig. 7 in the previous chapter.

Detector	Epoch	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
S1	Sens	11.5	15.4	19.2	19.2	26.9	26.9	26.9	30.8	34.6	42.3	42.3	42.3	46.2	46.2	46.2	46.2	46.2	53.8
	FP	0.5	1.1	2.1	2.1	2.1	2.7	4.3	5.9	6.4	9.1	11.2	12.8	15.5	15.5	18.2	18.7	20.9	21.4
S2	Sens	0.0	7.7	15.4	19.2	19.2	19.2	19.2	23.1	26.9	30.8	42.3	42.3	42.3	42.3	42.3	42.3	42.3	42.3
	FP	0.0	0.0	0.5	1.1	1.1	1.6	1.6	2.1	2.1	2.7	3.7	4.8	5.3	6.4	7.0	10.2	10.2	11.2
B1	Sens	15.4	23.1	23.1	26.9	30.8	34.6	34.6	42.3	42.3	46.2	50.0	53.8	53.8	57.7	57.7	57.7	57.7	61.5
	FP	1.1	1.1	1.1	1.1	1.1	1.1	1.6	2.1	2.7	3.7	4.3	4.8	5.9	5.9	8.0	10.7	11.8	13.4
B2	Sens	0.0	11.5	11.5	11.5	11.5	15.4	19.2	23.1	30.8	30.8	38.5	42.3	46.2	46.2	50.0	50.0	50.0	50.0
	FP	0.0	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.6	1.6	2.1	2.1	2.7	3.7	3.7	4.8	6.4	6.4
C1	Sens	23.1	38.5	46.2	50.0	61.5	61.5	61.5	65.4	65.4	69.2	69.2	73.1	76.9	76.9	76.9	80.8	80.8	84.6
	FP	3.7	5.3	6.4	7.0	9.6	9.6	11.2	15.5	16.6	20.3	21.4	23.5	28.3	28.9	33.7	34.8	36.4	38.5
C2	Sens	0.0	19.2	26.9	30.8	30.8	42.3	42.3	46.2	50.0	50.0	57.7	57.7	61.5	61.5	65.4	65.4	69.2	73.1
	FP	0.0	1.6	2.7	3.7	3.7	4.3	4.8	4.8	7.5	9.1	11.2	12.8	13.9	16.0	17.1	21.9	23.5	25.7

Table 7–1: Detector performances. FP = false positive rate = 100 − specificity, Sens = sensitivity

CHAPTER 8
Conclusions and discussion

8.1 Thesis approach and rationale

The three journal papers of this thesis clearly met the objectives of the thesis approach and rationale. We reiterate these objectives below. The research shall:

- be *data-driven*: model the data as much as possible rather than choose features *a-priori* and in so doing extract as much of the information content as possible. As such, FHR estimates from the model should account for most of the original measured FHR signal energy, apart from noise and distortion.
- obtain the most informative *input-output* models from UP and FHR. While coarse measures of UP-FHR interaction have been described, this thesis is the first to propose a system-identification approach the problem.
- report performance using an independent test set, using multiple k-fold cross-validation simulations whenever possible so that all performance estimates also include confidence levels.
- use outcome-class labelling that is as objective as possible. Base deficit and signs of neonatal neuropathology are some of the best early neonatal indicators of fetal hypoxic distress.

All of these objectives have been met. First, a signal modelling approach was used throughout. The feature extraction approach of journal paper I (chapter 4) modelled the FHR using singular spectrum analysis for the purposes of deceleration detection. Journal paper II (chapter 5) focused on modelling in creating a system identification model of the input UP and output FHR data. Journal paper III (chapter 6) also used a modelling approach to estimating FHR parameters (for baseline and heart-rate variability) complementary to the system identification parameters.

In Journal paper III, we used 10-fold cross-validation to report all classification performance results, and included a validation set to ensure that hyperparameter selection was unbiased by test results. Finally, in Journal papers II and III, we used base deficit value and indications of neonatal encephalopathy for objective outcome class labelling.

All the papers shared the common goal of FHR signal estimation. Early thesis work used this estimation for feature detection (e.g. for deceleration detection in Journal paper I). Later in the thesis, we shifted the focus of the estimation from FHR alone to the UP-FHR signal pair to establish the novel system identification model (Journal papers II and III). The purpose of this modelling went beyond feature detection and directly to the main goal of the thesis, that is, automated decision support, by using machine learning techniques to associate these stimulus-response models with the fetal state during labour and delivery.

8.2 Contributions to the field

8.2.1 Novel application of system identification modelling to CTG signals

The thesis made several important new contributions to the field of intra-partum fetal surveillance. The core of the thesis is the novel system identification model of the dynamic relationship between UP and FHR. This is a marked departure from most approaches based on feature detection. While system identification has been applied in numerous biomedical problems, this is the first reported use with CTG signals. Our application of system identification to CTG provided a vehicle for introducing several noise reduction steps into the modelling process. Use of the pseudo-inverse method, order selection and surrogate model testing gave very good confidence that models had captured significant UP-FHR interaction despite the prevalence of noise in both these signals. Most importantly, the models were discriminating; we showed that the models for normal and pathological cases is our database had parameters with statistically significant differences, with pathological cases having stronger, more delayed and more predictable responses, consistent with clinical expectation.

8.2.2 Creation of a fetal distress detector suitable for real-time application

Journal paper III addressed the main goal of the thesis, that is, automated assessment of the fetal state during labour and delivery. Using the system identification models of the previous paper, we classified 20-min epochs as normal or pathological. At the same time, we did similar classification using the complementary FHR information of baseline and variability parameters. By combining these epoch classifiers and observing their history over time, we constructed a detector of fetal pathology. The results indicate that half of the pathological cases in the database were detected, with low false positive rates (7.5%), early enough (1.5 hours before delivery) to allow appropriate clinical intervention. These results are also significant because these detected pathological cases in our retrospective study, by the definition of our database, had originally been undetected by obstetricians.

These results are one of the few in the field that consider the real-time problem of detecting pathology early enough to allow appropriate clinical intervention. Other studies most often examine retrospective data as a whole before delivery before making a decision. Under clinical conditions, it is not feasible to expect that all data before delivery is available to form opinions about the state of the fetus. Indeed, implementing the detector in a clinical setting will require more research into the issue of what is the best way to present the detection results to the clinician.

As an addendum to this article, we note that for the cases considered for detection, the incidence of cesarian section was higher for the pathological cases (10/26=38%) compared to the normal cases (21/187=11%). In Fig. 8–1, we compare the sensitivities and specificities of vaginal and cesarian-section births and the overall population reported in journal article III. We note that the specificity is considerably lower for the cesarian sectioned cases, likely indicating that the associated FHRs tended to be less reassuring, and that intervention improved outcomes. Consequently, this confounded our overall specificity estimate, but in a way that underestimated performance; indeed, the specificity for the vaginal birth population is slightly better than the overall rate. As well, the sensitivities were greater

for the vaginal cases compared to both the overall and cesarian cases. This supports our contention that the system can detect cases early that were not acted upon by clinicians.

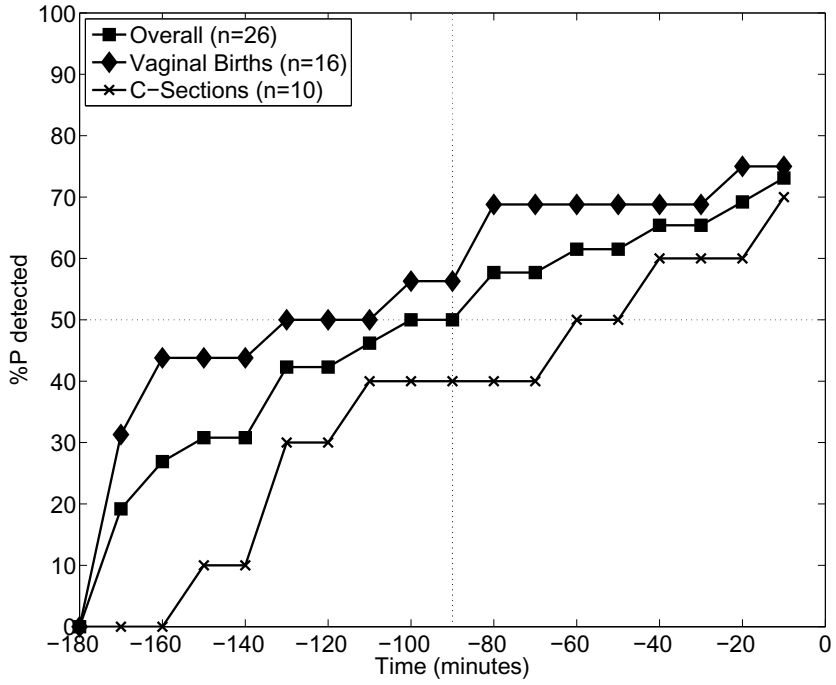
It should be clarified that while the article states that the pathological cases were *missed* by clinicians, a number of factors could have contributed to the adverse outcome, including refusal of intervention by the mother, non-availability of the operating room and non-presence of the obstetrician. Our database does not provide this specific information consistently. However, numerous reviews cited in the thesis [13, 15] show that the large majority (80%) of hypoxic injuries are acute in nature (i.e., they occur during labour and delivery). Therefore, it is very likely that more timely intervention could have improved outcome for some of the cesarian-sectioned pathological cases.

8.2.3 Selecting objective, orthogonal features for feature detection

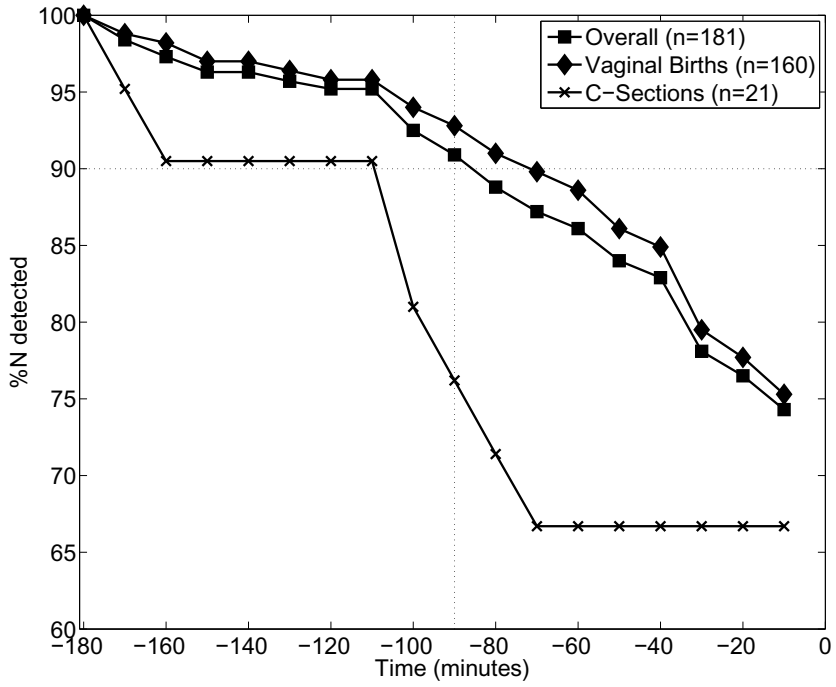
Rather than mimic the clinician by selecting features that correspond to methods of visual inspection, Journal paper I used a principal components approach to obtain more objective and orthogonal features. This approach to input feature selection has been used in other biomedical classification problems, but this thesis is one of the few studies that apply it to CTG analysis. Manually selected features are often non-orthogonal (i.e. correlated) which complicates classification procedures by elevating the number of input parameters used. Increasing the number of inputs evokes the so-called 'curse of dimensionality', because increasing input dimensions require increasingly complicated decision boundaries to separate classes.

8.2.4 Future work

We obtained these very encouraging results despite the use of noisy data collected under clinical conditions and therefore subject to frequent artifacts related to sensor disturbances. Because of the prevalence of data artifact, decisions were often based on a small proportion of the data record. However, with a more rigorous data-collection protocol generating cleaner and more informative data (e.g., using better calibrated UP [85] and using R-R intervals for HRV estimates [11]), we would expect better results in terms of accuracy



(a)



(b)

Figure 8–1: (a) Pathological and (b) normal detection over time for the C2 detector, comparing the rates for vaginal (diamonds) and cesarian births (crosshatches) to the overall rates (squares) reported in journal article III. The sample sizes are indicated for each population.

and timing since we could obtain better models and a decision could be made at more epochs.

The generally low VAFs of the system identification models can be attributed to that fact that UP is not the only influence on FHR. Background intrinsic FHR variability unrelated to UP can be of significant amplitude [93]. Also, our linear model assumes stationarity within epochs (which can be violated) and there are approaches that better handle these non-stationarities [24]. As well, our model does not account for non-linear interactions such as 1) the higher-frequency harmonics present in some very sharp decelerations and 2) variability that is coincident with decelerations [60]. Incorporating these nonlinearities into the model could improve modelling, but more importantly, could capture discriminating signal characteristics that are not present in the linear model.

We also hope to classify intermediate cases, which may develop hypoxia quite late in delivery. A survey of clinical FHR studies [56] concluded that fetal acidemia usually develops over one hour. Given fetal-state evolution over this time frame, there is great potential for the combination of better data collection and model-based analysis to improve current clinical practice by identifying fetuses at risk before they become injured.

Appendix A: Summary of publications

Journal articles

- **Feature extraction**
  - Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of singular spectrum analysis. *Methods of Information in Medicine*, 46(2):196–201, 2007
- **System identification**
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Identification of the dynamic relationship between intra-partum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *IEEE Transactions on Biomedical Engineering*, 56(6):1587–1597, June 2009
- **Fetal-state classification**
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Classification of normal and hypoxic fetuses from systems modelling of intra-partum cardiotocography. *IEEE Transactions on Biomedical Engineering*, October 2009. accepted
  - C.H Elliott, P.A. Warrick, E. Graham, and E.F. Hamilton. Graded classification of fetal heart rate tracings:association with neonatal metabolic acidosis and neurologic morbidity. *American Journal of Obstetrics and Gynecology*, 2009. accepted

Refereed conferences & workshops

- **Feature extraction**
  - Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of singular spectrum analysis. In *The 5th International Workshop*

On Biosignal Interpretation. Proceedings., pages 151–154, 2005. (3rd-prize in student competition)

(3rd-prize in student competition)

- Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection from the discrete cosine transform spectrum. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 5555–5558, 2005
- E. Hamilton, A. Dyachenko, C. Elliott, P. Warrick, and A. Ciampi. Progression of intrapartum EFM patterns in births with symptomatic metabolic acidosis. *American Journal of Obstetrics and Gynecology*, 197(6):s182, 2007

- **System identification**
  - Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Linear models of intrapartum uterine pressure-fetal heart rate interaction for the normal and hypoxic fetus. In *Proceedings of the 2006 IEEE Engineering in Medicine and Biology 28th Annual Conference*, pages 6434–6437, 2006
  - Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Low-order parametric system identification for intrapartum uterine pressure-fetal heart rate interaction. In *The 2007 IEEE Engineering in Medicine and Biology 29th Annual Conference*, pages 5043–5046, 2007
  - Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Time progression of a parametric impulse response function estimate from intrapartum cardiotocography for normal and hypoxic fetuses. In *Computers in Cardiology 2007*, pages 693–696, 2007
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Detecting the temporal extent of the impulse response function from intrapartum cardiotocography for normal and hypoxic fetuses. In *The 2008 IEEE Engineering in Medicine and Biology 30th Annual Conference*, pages 2797–2800, 2008

- **Fetal-state classification from system identification**
  - Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses. In *Computers in Cardiology 2006. Proceedings.*, volume 33, pages 937–940, 2006
  - Philip A. Warrick, Emily F. Hamilton, Robert E. Kearney, and Doina Precup. Classification of normal and hypoxic fetuses using system identification from intra-partum cardiotocography. In *ICML2008 Workshop on Machine Learning for Health Care Applications*, 2008

Book Chapters

- **System identification and clinical relevance**
  - Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. *Brain Hypoxia Ischemia Research Progress*, chapter Perinatal Hypoxic Brain Injury, pages 257–269. Nova Science Publishers, Inc., Hauppauge, NY, 2008

References

[1] ACOG. *Harming Patient Access To Care: The Impact Of Excessive Litigation Statement Of The ACOG To The Subcommittee On Health, Committee On Energy And Commerce, United States House Of Representatives.* ACOG, February 2003.

[2] ACOG. *Neonatal Encephalopathy and Cerebral Palsy: Defining the Pathogenesis and Pathophysiology.* ACOG Task force on Neonatal Encephalopathy and Cerebral Palsy, January 2003.

[3] N. Badawi, J. Kurinczuk, J. Keogh, L.M. Alessandri, F. O'Sullivan, P.R. Burton, P.J. Pemberton, and F.J. Stanley. Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study. *BMJ*, 317:1549–1553, 1998.

[4] G. Baselli and S. Cerutti. Identification techniques applied to processing of signals from cardiovascular systems. *Med. Inf.*, 10:223–35, 1985.

[5] G. Berry and P. Martin. Perinatal risks. Risk management foundation Harvard Medical Institutions forum. Technical report, Harvard University, March 2001.

[6] Steven L. Bloom, Catherine Y. Spong, Elizabeth Thom, Michael W. Varner, Dwight J. Rouse, Sandy Weininger, Susan M. Ramin, Steve N. Caritis, Alan Peaceman, Yoram Sorokin, Anthony Sciscione, Marshall Carpenter, Brian Mercer, John Thorp, Fergal Malone, Margaret Harper, Jay Iams, Garland Anderson, the National Institute of Child Health, and Human Development Maternal-Fetal Medicine Units Network. Fetal pulse oximetry and Cesarean delivery. *N Engl J Med*, 355(21):2195–2202, November 2006.

[7] H. Cao, D.E. Lake, II Ferguson, J.E., C.A. Chisholm, M.P. Griffin, and J.R. Moorman. Toward quantitative fetal heart rate monitoring. *IEEE Transactions on Biomedical Engineering*, 53(1):111–118, 2006.

[8] Shelley Cazares, Mary Moulden, Christopher W. G. Redman, and Lionel Tarassenko. Tracking poles with an autoregressive model: a confidence index for the analysis of the intrapartum cardiotocogram. *Medical Engineering & Physics*, 23(9):603–614, November 2001.

[9] S. Cerutti, S. Civardi, A. Bianchi, M.G. Signorini, E. Ferrazzi, and G Pardi. Spectral analysis of antepartum heart rate variability. *Clin. Phys. Physiol. Meas.*, 10:27–31, 1989.

[10] T. K. H. Chung, M. P. Mohajer, Z. J. Yang, A. M. Z. Chang, and D. S. Sahota. The prediction of fetal acidosis at birth by computerised analysis of intrapartum

cardiotocography. *BJOG: An International Journal of Obstetrics and Gynaecology*, 102(6):454–460, 1995.

[11] G.D. Clifford and L. Tarassenko. Quantifying errors in spectral estimates of HRV due to beat replacement and resampling. *IEEE Transactions on Biomedical Engineering*, 52(4):630–638, 2005.

[12] Antonia Costa, Diogo Ayres-de Campos, Fernanda Costa, Cristina Santos, and Joao Bernardes. Prediction of neonatal acidemia by computer analysis of fetal heart rate and st event signals, June 2009.

[13] Frances Cowan, Mary Rutherford, Floris Groenendaal, Paula Eken, Eugenio Mercuri, Graeme M Bydder, Linda C Meiners, Lilly MS Dubowitz, and Linda S de Vries. Origin and timing of brain lesions in term infants with neonatal encephalopathy. *The Lancet*, 361(9359):736–742, March 2003.

[14] G.S. Dawes, C.R.S. Houghton, and C.W.G. Redman. Baseline in human fetal heart-rate records. *Br J Obstet Gynaecol*, 89:270–275, 1982.

[15] E.S. Draper, J.J. Kurinczuk, C.R. Lamming, M. Clarke, D. James, and D. Field. A confidential enquiry into cases of neonatal encephalopathy. *Arch Dis Child Fetal Neonatal Ed*, 87:F176–F180, 2002.

[16] C.H. Elliott, P.A. Warrick, E. Graham, and E.F. Hamilton. Computerized analysis of fetal heart rate tracings and identification of metabolic acidosis and neonatal neurologic morbidity. In *Royal College of Obstetricians and Gynaecologists 7th International Meeting*, 2008. (in press).

[17] C.H Elliott, P.A. Warrick, E. Graham, and E.F. Hamilton. Graded classification of fetal heart rate tracings:association with neonatal metabolic acidosis and neurologic morbidity. *American Journal of Obstetrics and Gynecology*, 2009. accepted.

[18] M. Ferrario, M.G. Signorini, and G. Magenes. Complexity analysis of the fetal heart rate for the identification of pathology in fetuses. In *Computers in Cardiology, 2005*, pages 989–992, 2005.

[19] M. Ferrario, M.G. Signorini, G. Magenes, and S. Cerutti. Comparison of entropy-based regularity estimators: application to the fetal heart rate signal for the identification of fetal distress. *IEEE Transactions on Biomedical Engineering*, 53(1):119–125, 2006.

[20] J.M. Freeman and K. Nelson. Intrapartum asphyxia and cerebral palsy. *Pediatrics*, 82:240–9, 1988.

[21] R.K. Freeman, T.J. Garite, and M.P. Nageotte. *Fetal Heart Monitoring*. Lippincott Williams and Wilkins, Philadelphia, PA, 2003.

[22] Thomas J. Garite, Gary A. Dildy, Helen McNamara, Michael P. Nageotte, Frank H Boehm, Eric H. Dellinger, Robert A. Knuppel, Richard P. Porreco, Hugh S. Miller, Shiraz Sunderji, Michael W. Varner, and David B. Swedlow. A multicenter controlled

trial of fetal pulse oximetry in the intrapartum management of nonreassuring fetal heart rate patterns. *American Journal of Obstetrics & Gynecology*, 183(5):1049–1058, November 2000.

[23] G.G. Georgoulas, C.D. Stylios, G. Nokas, and P.P. Groumpos. Classification of fetal heart rate during labour using hidden markov models. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 3, pages 2471–2475 vol.3, 2004.

[24] A. Goli, D. M. McNamara, and A. K. Ziarani. A novel method for decomposition of multicomponent nonstationary signals. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 255–258, 2007.

[25] Miller SP. Gonzalez FF. Does perinatal asphyxia impair cognitive function without cerebral palsy? *Arch Dis Child Fetal Neonatal Ed.*, 91(6):F454–9, 2006.

[26] N. Guler, O.T. Yildiz, F. Gurgen, F. Varol, and E. Alpaydin. Discriminant functions and decision tree induction techniques forantenatal fetal risk assessment. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, volume 4, pages 2712–2717 vol.4, 2001.

[27] E. Hamilton, A. Dyachenko, C. Elliott, P. Warrick, and A. Ciampi. Progression of intrapartum EFM patterns in births with symptomatic metabolic acidosis. *American Journal of Obstetrics and Gynecology*, 197(6):s182, 2007.

[28] B.H. Hon and S.T. Lee. Electronic evaluation of fetal heart rate. patterns preceding fetal death: further observations. *Am. J. Obstet. Gynecol*, 87:814–826, 1963.

[29] W. Jarisch and J.S. Detwiler. Statistical modeling of fetal heart rate variability. *IEEE Transactions on Biomedical Engineering*, BME-27(10):582–9, October 1980.

[30] Arne Jensen, Yves Garnier, and Richard Berger. Dynamics of fetal circulatory responses to hypoxia and asphyxia. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 84(2):155–172, June 1999.

[31] M. Jezewski, J. Wrobel, P. Labaj, J. Leski, N. Henzel, K. Horoba, and J. Jezewski. Some practical remarks on neural networks approach to fetal cardiotocograms classification. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5170–5173, 2007.

[32] Yoshitaka Kimura, Kunihiro Okamura, Takanori Watanabe, Nobuo Yaegashi, Shigeki Uehara, and Akira Yajima. Time-frequency analysis of fetal heartbeat fluctuation using wavelet transform. *Am J Physiol Heart Circ Physiol*, 275(6):H1993–1999, 1998.

[33] P. Laguna, G.B. Moody, and R.G. Mark. Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals. *IEEE Transactions on Biomedical Engineering*, 45(6):698–715, 1998.

[34] J.A. Low. The role of blood gas and acid-base assessment in the diagnosis of intrapartum fetal asphyxia. *Am J Obstet Gynecol.*, 159(5):1235–40, 1988.

[35] J.A. Low. Metabolic acidosis and fetal reserve. *Baillires Clinical Obstetrics and Gynecology*, 10(2):211–24, 1996.

[36] J.A. Low. The current crisis in obstetrics. *J Obstet Gynaecol Can.*, 27(11):1031–7, 2005.

[37] J.A. Low, R. Victory, and E.J. Derrick. Predictive value of electronic fetal monitoring for intrapartum fetal asphyxia with metabolic acidosis. *Obstet Gynecol*, 93:285–291, 1999.

[38] F. Lunghi, G. Magenes, L. Pedrinazzi, and M.G. Signorini. Detection of fetal distress though a support vector machine based on fetal heart rate parameters. In *Computers in Cardiology, 2005*, pages 247–250, 2005.

[39] Alastair MacLennan. A template for defining a causal relation between acute intrapartum events and cerebral palsy: international consensus statement. *BMJ*, 319(7216):1054–1059, 1999.

[40] K. Maeda, M. Utsu, A. Makio, M. Serizawa, Y. Noguchi, T. Hamada, K. Mariko, and F. Matsumoto. Neural network computer analysis of fetal heart rate. *Journal of Maternal-Fetal Investigation*, 8(4):163–171, December 1998.

[41] G. Magenes, L. Pedrinazzi, and M.G. Signorini. Identification of fetal sufferance antepartum through a multiparametric analysis and a support vector machine. In *Engineering in Medicine and Biology Society, 2004. EMBC 2004. Conference Proceedings. 26th Annual International Conference of the*, volume 1, pages 462–465 Vol.1, 2004.

[42] G. Magenes, M.G. Signorini, and D. Arduini. Detection of normal and pathological fetal states by means ofneural and fuzzy classifiers applied to ctg parameters. In *[Engineering in Medicine and Biology, 1999. 21st Annual Conf. and the 1999 Annual Fall Meeting of the Biomedical Engineering Soc.] BMES/EMBS Conference, 1999. Proceedings of the First Joint*, volume 2, page 936 vol.2, 1999.

[43] G. Magenes, M.G. Signorini, and D. Arduini. Classification of cardiotocographic records by neural networks. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 3, pages 637–641 vol.3, 2000.

[44] G. Magenes, M.G. Signorini, D. Arduini, and S. Cerutti. Fetal heart rate variability due to vibroacoustic stimulation: linear and nonlinear contribution. In *Methods Inf. Med. (Germany)*, volume 43, pages 47–51, Como, Italy, 2004. Schattauer GmbH.

[45] G. Magenes, M.G. Signorini, M. Ferrario, L. Pedrinazzi, and D. Arduini. Improving the fetal cardiotocographic monitoring by advanced signal processing. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 3, pages 2295–2298 Vol.3, 2003.

[46] G. Magenes, M.G. Signorini, and R. Sassi. Automatic diagnosis of fetal heart rate: comparison of different methodological approaches. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 2, pages 1604–1607 vol.2, 2001.

[47] R Mantel, HP Van Geijn, FJM Caron, EE van Woerden, and H Jongsma. Computer analysis of antepartum fetal heart rate. I. Baseline determination. *Int. J. Biomed. Comput.*, 25:261–272, 1990.

[48] R. Mantel, H. P. Van Geijn, and van Woerden E. E. & Jongsma H. W. Caron F. J. M. Computer analysis of antepartum fetal heart rate. ii. detection of accelerations and decelerations. *Int. J. Biomed. Comput.*, 25:273–286., 1990.

[49] Vasilis Z. Marmarelis. *Nonlinear Dynamic Modeling of Physiological Systems*. Wiley Interscience, Hoboken, NJ, 2004.

[50] J.P. Marques de Sa. Characterization of fetal heart rate using approximate entropy. In *Computers in Cardiology, 2005*, pages 671–673, 2005.

[51] C. B. Martin, J. de Haan, B. van der Wildt, H. W. Jongsma, A. Dieleman, and T. H. M. Arts. Mechanisms of late decelerations in the fetal heart rate : A study with autonomic blocking agents in fetal lambs. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 9(6):361–373, December 1979.

[52] M.D. Miles, A.M., M.D. Monga, M., and M.D. Richeson, K.S. Correlation of external and internal monitoring of uterine activity in a cohort of term patients. *American Journal Of Perinatology*, (03):137–140, 2001.

[53] Steven P. Miller, Vijay Ramaswamy, David Michelson, A. James Barkovich, Barbara Holshouser, Nathaniel Wycliffe, David V. Glidden, Douglas Deming, J. Colin Partridge, Yvonne W. Wu, Stephen Ashwal, and Donna M. Ferriero. Patterns of brain injury in term neonatal encephalopathy. *The Journal of Pediatrics*, 146(4):453–460, April 2005.

[54] H. Norn, S. Blad, A. Carlsson, A. Flisberg, A. Gustavsson, H Lilja, M. Wennergren, and H. Hagberg. Stan in clinical practice–the outcome of 2 years of regular use in the city of gothenburg. *American Journal of Obstetrics and Gynecology*, 195(1):7–15, July 2006.

[55] N.S. Padhye, Z. Duan, and M.T. Verklan. Response of fetal heart rate to uterine contractions. In *Engineering in Medicine and Biology Society, 2004. EMBC 2004. Conference Proceedings. 26th Annual International Conference of the*, volume 2, pages 3953–3955 Vol.6, 2004.

[56] J. T. Parer, T. King, S. Flanders, M. Fox, and S. J. Kilpatrick. Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association? *Journal of Maternal-Fetal & Neonatal Medicine*, 19(5):289–294, May 2006.

[57] A. Peliowski and N.N. Finer. *Effective Care of the Newborn Infant*, chapter Birth asphyxia in the term infant, pages 249–79. Oxford University Press, Oxford, 1992.

[58] S M Pincus and R R Viscarello. Approximate entropy: a regularity measure for fetal heart rate analysis. *Obstetrics And Gynecology*, 79(2):249–255, February 1992.

[59] S.B. Ransom, D.M. Studdert, M.P. Dombrowski, J.D. Mello, and T.A. Brennan. Reduced medicolegal risk by compliance with obstetric clinical pathways: A case-control study. *Obstet Gynecol*, 101, 2003.

[60] M. Romano, P. Bifulco, M. Cesarelli, M. Sansone, and M. Bracale. Foetal heart rate power spectrum response to uterine contraction. *Medical and Biological Engineering and Computing*, 44:88–201, March 2006.

[61] B. Rosen, D. Soriano, T. Bylander, H. Ortiz-Zuazaga, and B. Schiffrin. Training a neural network to recognize artefacts and decelerations in cardiotocograms. In *AAAI Spring Symposium on Artificial Intelligence in Medicine*, 1996.

[62] M. G. Ross, L. D. Devoe, and K. G. Rosen. St-segment analysis of the fetal electrocardiogram improves fetal heart rate tracing interpretation and clinical decision making. *The Journal of Maternal-Fetal & Neonatal Medicine*, 15(3):181–185, 2004.

[63] Michael G. Ross and Rageev Gala. Use of umbilical artery base excess: Algorithm for the timing of hypoxic injury. *American Journal of Obstetrics and Gynecology*, 187(1):1–9, July 2002.

[64] A. M. Rudolph. The fetal circulation and its response to stress. *Journal of Developmental Physiology*, 6:11–19, 1984.

[65] E. Salamalekis, E. Hintipas, I. Salloum, G. Vasios, C. Loghis, N. Vitoratos, Ch. Chrelias, and G. Creatsas. Computerized analysis of fetal heart rate variability using the matching pursuit technique as an indicator of fetal hypoxia during labor. *Journal of Maternal-Fetal & Neonatal Medicine*, 19(3):165–169, March 2006.

[66] E. Salamalekis, P. Thomopoulos, D. Giannaris, I. Salloum, G. Vasios, A. Prentza, and D. Koutsouris. Computerised intrapartum diagnosis of fetal hypoxia based on fetal heart rate monitoring and fetal pulse oximetry recordings utilising wavelet analysis and neural networks. *BJOG: An International Journal of Obstetrics and Gynaecology*, 109(10):1137–1142, 2002.

[67] C.J. Saphier, E.J. Thomas, D. Studdert D, T.A. Brennan, and D. Acker. Applying no-fault compensation to obstetric malpractice claims. *Prim Care Update Ob Gyns*, 5:208–9, 1998.

[68] M. Shevell. Cerebral palsy: Defining the problem. *Seminars in Pediatric Neurology*, 11(1), 2004.

[69] M.G. Signorini. Nonlinear analysis of heart rate variability signal: physiological knowledge and diagnostic indications. In *Engineering in Medicine and Biology Society, 2004.*

EMBC 2004. Conference Proceedings. 26th Annual International Conference of the, volume 2, pages 5407–5410 Vol.7, 2004.

[70] M.G. Signorini, A. de Angelis, G. Magenes, R. Sassi, D. Arduini, and S. Cerutti. Classification of fetal pathologies through fuzzy inference systemsbased on a multiparametric analysis of fetal heart rate. In *Computers in Cardiology 2000*, pages 435–438, 2000.

[71] M.G. Signorini, M. Ferrario, L. Pedrinazzi, and G. Magenes. Analysis of echographic and heart rate time and frequency domain parameters for the antepartun fetal surveillance. In *Computers in Cardiology, 2005*, pages 679–682, 2005.

[72] M.G. Signorini, G. Magenes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Transactions on Biomedical Engineering*, 50(3):365–374, 2003.

[73] M.G. Signorini, F. Marchetti, and S. Cerutti. Applying nonlinear noise reduction in the analysis of heart-rate variability. *Engineering in Medicine and Biology Magazine, IEEE*, 20(2):59–68, 2001.

[74] J.F. Skinner, J.M. Garibaldi, J. Curnow, and E.C. Ifeachor. Intelligent fetal heart rate analysis. In *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, pages 14–21, 2000.

[75] B.L. Stalnaker, J.E. Maher, G.E. Kleinman, J.M. Macksey, L.A. Fishman, and J.M. Bernard. Characteristics of successful claims for payment by the Florida Neurologic Injury Compensation Association Fund. *Am J Obstet Gynecol*, 177:268–71, 1997.

[76] S.B. Thacker, D. Stroup, and M. Chang. Continuous electronic heart rate monitoring for fetal assessment during labor. In *(Cochrane Review). In: The Cochrane Library*, volume 4. John Wiley & Sons, Ltd., 2004.

[77] I. Tsoulos, G. Georgoulas, D. Gavrilis, C. Stylios, J. Bernardes, and P. Groumpos. Introducing grammatical evolution in fetal heart rate analysis and classification. In *Intelligent Systems, 2006 3rd International IEEE Conference on*, pages 560–565, 2006.

[78] C. Ulbricht, G. Dorffner, and A. Lee. Neural networks for recognizing patterns in cardiotocograms. *Artif Intell Med*, 12:271284, 1998.

[79] T Vanner and J Gardosi. Intrapartum assessment of uterine activity. *Bailliere's Clinical Obstetrics And Gynaecology*, 10(2):243–257, June 1996.

[80] P. Warrick, E. Hamilton, and M. Macieszczak. Neural network based detection of fetal heart rate patterns. In *Neural Networks, 2005. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2400–2405, 2005.

[81] Philip A. Warrick, Emily F. Hamilton, Robert E. Kearney, and Doina Precup. Classification of normal and hypoxic fetuses using system identification from intra-partum

cardiotocography. In *ICML2008 Workshop on Machine Learning for Health Care Applications*, 2008.

[82] Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. *Brain Hypoxia Ischemia Research Progress*, chapter Perinatal Hypoxic Brain Injury, pages 257–269. Nova Science Publishers, Inc., Hauppauge, NY, 2008.

[83] Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Detecting the temporal extent of the impulse response function from intra-partum cardiotocography for normal and hypoxic fetuses. In *The 2008 IEEE Engineering in Medicine and Biology 30th Annual Conference*, pages 2797–2800, 2008.

[84] Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Classification of normal and hypoxic fetuses from systems modelling of intra-partum cardiotocography. *IEEE Transactions on Biomedical Engineering*, October 2009. accepted.

[85] Philip A. Warrick, Emily F. Hamilton, Doina Precup, and Robert E. Kearney. Identification of the dynamic relationship between intra-partum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *IEEE Transactions on Biomedical Engineering*, 56(6):1587–1597, June 2009.

[86] Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Linear models of intrapartum uterine pressure-fetal heart rate interaction for the normal and hypoxic fetus. In *Proceedings of the 2006 IEEE Engineering in Medicine and Biology 28th Annual Conference*, pages 6434–6437, 2006.

[87] Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses. In *Computers in Cardiology 2006. Proceedings.*, volume 33, pages 937–940, 2006.

[88] Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Low-order parametric system identification for intrapartum uterine pressure-fetal heart rate interaction. In *The 2007 IEEE Engineering in Medicine and Biology 29th Annual Conference*, pages 5043–5046, 2007.

[89] Philip A. Warrick, Robert E. Kearney, Doina Precup, and Emily F. Hamilton. Time progression of a parametric impulse response function estimate from intra-partum cardiotocography for normal and hypoxic fetuses. In *Computers in Cardiology 2007*, pages 693–696, 2007.

[90] Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection from the discrete cosine transform spectrum. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 5555–5558, 2005.

[91] Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of

singular spectrum analysis. In *The 5th International Workshop On Biosignal Interpretation. Proceedings.*, pages 151–154, 2005. (3rd-prize in student competition).

[92] Philip A. Warrick, Doina Precup, Emily F. Hamilton, and Robert E. Kearney. Fetal heart rate deceleration detection using a discrete cosine transform implementation of singular spectrum analysis. *Methods of Information in Medicine*, 46(2):196–201, 2007.

[93] J. Westgate, B. Wibbens, L. Bennet, G. Wassink, J. Parer, and A. Gunn. The intrapartum deceleration in center stage: a physiologic approach to the interpretation of fetal heart rate changes in labor. *American Journal of Obstetrics and Gynecology*, 197(3):236.e1 – 236.e11, 2007.

[94] JA Westgate, L Bennet, H De Haan, and AJ Gunn. Fetal heart rate overshoot during repeated umbilical cord occlusion in sheep. *Obstetrics & Gynecology*, 97(3):454–459, 2001.

[95] David T. Westwick and Robert E. Kearney. *Identification of nonlinear physiological systems*. Wiley-Interscience, Hoboken NJ, 2003.

[96] Chung-Hsien Wu, Chen-Hsiang Yu, and Ang-Tai Lin. Computer-aided fetus analysis and diagnosis system using fetal heart rate. *Chinese Journal of Medical and Biological Engineering*, 18(4):233–43, December 1998.

[97] Cheul-Sung Yoo, Sang-Hoon Yi, and Gye-Rok Jeon. Approximate entropy of the short-term fetal heart rate variability as an efficient measure of fetal distress. *Sae Mulli*, 50(3):193–202, 2005.