Automatic evaluation of intrapartum fetal heart rate recordings: A comprehensive analysis of useful features

V Chudáček[1], J Spilka[1], P Janků[2], M Koucký[3], L Lhotská[1], M Huptych[1]

1 Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, the Czech Republic
2 Obstetrics and Gynaecology clinic, University Hospital in Brno, Czech Republic
3 Gynaecology and Obstetrics unit of the Charles University Hospital, Prague, the Czech Republic

E-mail: chudacv@fel.cvut.cz

Abstract. Cardiotocography (CTG) is the monitoring of fetal heart rate (FHR) and uterine contractions (TOCO) since 1960's used routinely by obstetricians to detect fetal hypoxia. The evaluation of the FHR in clinical settings is based on an evaluation of macroscopic morphological features and so far has managed to avoid adopting any achievements from the HRV research field.

In this work, most of the ever-used features utilized for FHR characterization, including FIGO, HRV, nonlinear, wavelet, and time and frequency domain features, are investigated and the features are assessed based on their statistical significance in the task of distinguishing the FHR into three FIGO classes.

We assess the features on a large data set (552 records) and unlike in other published papers we use three-class expert evaluation of the records instead of the pH values.

We conclude the paper by presenting the best uncorrelated features and their individual rank of importance according to the meta-analysis of three different ranking methods. Number of acceleration and deceleration, interval index, as well as Lempel-Ziv complexity and Higuchi's fractal dimension are among the top five features.

Keywords: cardiotocography, fetal heart rate, feature selection, expert annotation.

Accepted in *Physiological Measurement*, 2011

1. Introduction

Fetal heart activity is the prominent source of information about fetal well being during delivery. Cardiotocography (CTG) – recording of fetal heart rate (FHR) and uterine contractions enables obstetricians to detect possible ongoing fetal hypoxia which may occur even in a previously uncomplicated pregnancy.

Even though a fetus has its own natural defence mechanism to tackle the oxygen insufficiency during the delivery, in some cases only timely intervention can prevent adverse consequences (Steer 2008). Hypoxia, with prevalence lying in the region of

0.6% (Heintz et al. 2008) to 3.5% (Strachan et al. 2000), is considered to be the third most common cause of newborn death (d'Aloja et al. 2009).

Cardiotocography was introduced in late 1960s and is still the most prevalent method of intrapartum hypoxia detection. It did not however bring the expected improvements in the delivery outcomes in comparison to previously used intermittent auscultation (Alfirevic et al. 2006) and, moreover, continuous CTG is the main suspect for increased rate of cesarean sections for objective reasons (Steer 2008).

To improve the results of cardiotocography, the International Federation of Gynecology and Obstetrics (FIGO) introduced general guidelines (FIGO 1986). They are based on an evaluation of macroscopic morphological FHR features and their relation to the tocographic measurement. Even though the guidelines have been available for more than twenty years poor interpretation of CTG still persists (Steer 2008) with large inter-observer as well as intra-observer assessment variations (Blix et al. 2003, Bernardes et al. 1997). The goal of this paper is to contribute to the discussion about the feasibility of the automatic evaluation of the FHR.

Recently ST-analysis is getting much attention as an extension of the classical CTG measurements using additional information from the invasive measurement of the fetal ECG (Rosén 2005). Although most studies show that ST-analysis is performing better (Amer-Wåhlin & Maršál 2011), it is important to keep in mind that the first step to correctly interpret the ST ratio in ST-analysis is to correctly evaluate the CTG itself.

Attempts to use computer evaluation of the CTG are as old as the guidelines themselves. FIGO features became fundamental in most of the clinically oriented systems and automatically extracted morphological features have been integrated also into automatic systems for CTG analysis (de Campos et al. 2008, Guijarro-Berdinas & Alonso-Betanzos 2002).

In many papers only the FHR signal is used since FHR is the signal containing direct information about the fetal state. Our paper follows this assumption, also because of the inferior quality of the available electronically stored TOCO recordings.

Different features to describe FHR were investigated in the past, many of them heavily influenced by the research in adult heart rate variability (HRV) analysis.

Statistical description of CTG tracings was employed in the work of (Magenes et al. 2000) and in the following study of (Gonçalves et al. 2006*b*). Another approach to FHR analysis examined frequency content by spectral analysis and (Laar et al. 2008) gives a short overview of most of the works where FHR spectrum was analysed. The FHR was also analysed by wavelets with different properties (Salamalekis et al. 2002, Salamalekis et al. 2006). Other works analysed nonlinear properties of FHR such as fractal dimension of reconstructed attractor (Chaffin et al. 1991) and waveform fractal dimension (Felgueiras et al. 1998). Different estimations of fractal dimension were reviewed in (Hopkins et al. 2006). The most successful nonlinear methods for HRV analysis so far were approximate entropy (ApEn) and sample entropy (SampEn). They are often used for the examination of nonlinear systems and have proved their applicability also in FHR analysis (Gonçalves et al. 2006*b*, Georgoulas et al. 2006). Another method that performs well on the FHR recordings is Lempel-Ziv complexity employed e.g. in (Ferrario et al. 2009)

The main motivation for this paper was the persistent feeling of disconnection between the technical papers and clinical practice. In this paper we compare the features previously used by others in many different experimental settings (e.g. size of the database, preprocessing steps and pH threshold setting) and use them in one

clearly defined setting where their mutual relationship and overall usability can be assessed. The overall methodology is presented in Figure 1.

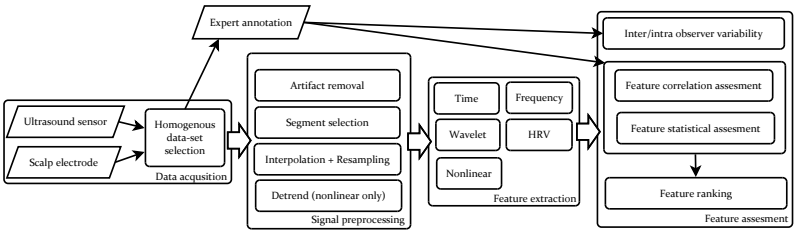

Figure 1. The overall scheme of the presented methodology of the paper.

The rest of the paper is structured as follows. Section 2 describes the data set involved in this study followed by Section 3 which details the annotation process. Individual steps of the FHR preprocessing are presented in section 4. Sections 5 and 6 present the extracted features and the evaluation process respectively, while Section 7 summarizes the results and Section 8 concludes the paper highlighting the merits of this research effort.

2. Data description

Data for this work was obtained at the Dept. of Obstetrics and Gynaecology, General Teaching Hospital in Prague from 2007 to 2009; all women signed informed consent. The FHR signals were measured on a Neoventa's STAN S21 system using an external ultrasound probe as well as an internal scalp electrode, see Figure 2. The differences between the methods of signal acquisition are mainly in the overall signal quality. For the common clinical purposes they are considered insignificant but for the use of automated FHR processing and analysis they were reported to be significant (Gonçalves et al. 2006*a*). Therefore, we included only external records.

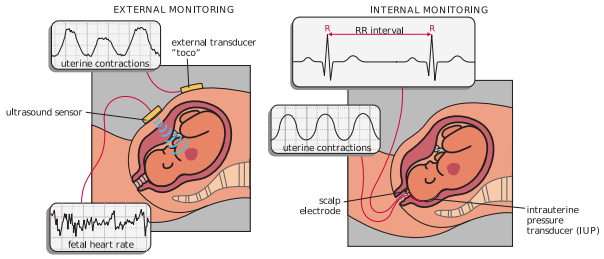

Figure 2. Recording of fetal heart rate and uterine activity (Sundström et al. 2000).

All recordings were checked for patient anamnesis and only one fold pregnancies delivered during the $38^{th} - 42^{nd}$ week of pregnancy were chosen. We have included the mature fetuses only, since the fetal heart rate and reaction of the fetal heart rate to the uterine pressure differs in the immature fetuses. Finally our database consisted of 552 delivery recordings altogether.The main characteristics of the set were: gestational age of 39.77 $\pm$ 1.03 (minimum 38, maximum 42); number of caesarean sections was 144

the rest was delivered vaginally; and Apgar score in the fifth minute was 8.98 ± 1.38 (minimum 6, maximum 10). Values of pH were not available for all records, however 90 records had pH lower than 7.15.

3. Data annotation

To be able to evaluate the features an annotation of the FHR signal is needed. In general there are two possible types of annotation. Objective annotation is the value of pH or base deficit of newborn umbilical artery blood measured on a clamped umbilical artery immediately after the delivery. Subjective annotation can be either expert annotation of the FHR signal or evaluation of the newborn (Apgar score) in the delivery room.

The exact relationship of umbilical pH after delivery to FHR is so far not fully understood, time between the recording and actual delivery plays a crucial role, and it seems that pH is only weakly correlated to clinical annotation (Valentin et al. 1993, Schiermeier et al. 2008). The best example is the timely Caesarean section (CS) due to suspect CTG – the CTG is suspect/pathological but the intervention prevented the baby going into real asphyxia that would be reflected in the pH value. Therefore, in this paper, expert annotation of the FHR recordings was used as a basis for feature evaluation. Expert annotation also has its drawbacks – it is much more subjective, and suffers from inter- and intra-observer variations, but it gives better insight into the real clinical decision making than the post-delivery numerical assessment.

For the purpose of data annotation a stand-alone application running on Java runtime environment was developed. The application adopts the most commonly used display layout of CTG machines, therefore poses no difficulty for clinicians to adjust. Annotation is based on three FIGO classes (normal, pathological, and suspect).

Annotations coming from three experts were used for the preparation of the "Gold standard" (GS) annotation. The GS was constructed based on simple majority voting. Records where experts totally disagreed were removed from the final data set – 9 recordings were excluded and therefore the final dataset consisted of 543 recordings.

Intra-observer agreement was computed from records that were presented two times or more for annotation. Each tenth record presented for annotation was selected randomly from the database and the experts were unaware of their reoccurrence. Intra-observer agreement was computed as a ratio of consistently annotated records to all annotations. Inter-observer agreement was computed as a ratio of equally annotated records among the three experts to all annotations. To describe evaluation agreement of the experts kappa statistics was used (Gwet 2010). It represents an index which compares the expert agreement on the data evaluation against that expected by chance. Possible values of kappa statistics ranges from +1 (perfect agreement) through 0 (no agreement above that expected by chance) to –1 (complete disagreement).

4. Signal preprocessing

Values of extracted features and their further usability are highly dependent on the quality of signal preprocessing. Our preprocessing process consisted of four main steps: segment selection, artefacts removal, interpolation, and signal detrend.

Segments were selected from the complete recordings, some of them up to 12 hours long, as close as possible to the actual delivery. Signal quality was evaluated

in relation to the segment position and the segment with the best score was selected. When available information allowed, we tried to set the end of the segment onto the beginning of the second stage of labour, where the quality of signal sharply decreases.

The FHR signal nearly always contains artefact caused by mother and fetal movements as well as artefacts caused by transducer displacements. In general, the amount of unusable data due to artefacts or missing values ranges between 20% and 40%. In our case we allow for computation only segments that have less than 20% of their signal missing. Therefore we selected segments which were a maximum of 24 minutes long and due to further preprocessing (gap interpolation and noisy segments removal) we truncated them to equal, 20 minute, long segments – 4800 samples when using 4 Hz sampling frequency. An example of one of the selected segments is shown in Figure 3.

The algorithm proposed by (Bernardes et al. 1991) was utilized for artefacts removal. First, the successive five beats with a difference lower than 10 bpm among them are considered as a stable segment. Then, whenever the difference between adjacent beats is higher than 25 bpm, the sample is substituted by linear interpolation between the previous beat and the new stable segment. Thus, all abrupt changes in FHR are removed and replaced. The result of artefacts removal is presented in Figure 3b.

We used cubic Hermite spline interpolation (Schumaker 2007) to replace missing data. Based on our experiments, if the length of missing data was 20 seconds and more we skipped the data and did not compute across the gap (Sprott 2003, Kim et al. 2009). The spline interpolation also introduces nonlinearity, however, the amount of nonlinearity should be approximately the same for normal and pathological FHR.

5. Features

Only features based on FHR signal were evaluated in the context of this work as already reasoned in Section 1.

Features used for purposes of this paper are an almost complete collection of features used for the evaluation of intrapartal/antepartal FHR in recently published papers. The interested reader can easily find the respective equations for features in the papers reviewed in Section 1 and in those referenced by the respective methods therefore, in this section, we present only the context information necessary to

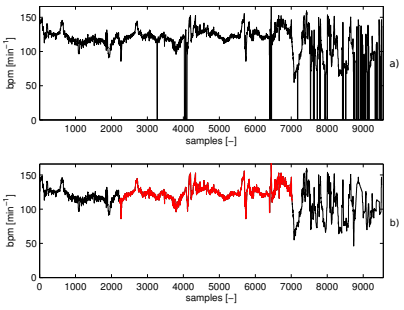

Figure 3. Artefacts removal. (a) Raw signal with artefacts, (b) signal after artefacts removal, where the red highlighted part of the signal marks the selected 20-minutes segment.

reproduce the analysis. Input to all feature extraction methods is 20-minute FHR segment.

5.1. Clinically used morphological features

Morphological features proposed in the FIGO guidelines are the features used in the obstetricians wards. These features describe the macroscopic – "visible" – properties of the FHR. A well known algorithm for feature extraction described in (Bernardes et al. 1991) was used in this study. The features extracted were: **Mean of the FHR baseline** without the influence of accelerations and decelerations; **Number of accelerations** – transient increase in heart rate above the baseline by 15 bpm or more, lasting 15 seconds or more; **Number of decelerations** – transient episode of slowing fetal heart rate below the baseline level by more than 15 bpm and lasting 10 seconds or more.

Among other features – approached by obstetricians in clinical practice as morphological – are standard deviation of FHR and short term variability (STV). Since these features are only very crudely estimated by clinicians, we have followed the separation of these features, as proposed in (Magenes et al. 2000), into the time domain subsection.

5.2. Time domain features

Two types of time domain features were computed. The first type deals with macroscopic features that are rarely assessed in a clinical setting. The second type assesses more subtle changes in FHR behaviour, that are impossible to spot with the naked eye. The equations can be found in (Georgoulas et al. 2006).

There are two time domain features describing the FHR baseline: **Median of the FHR baseline** and **standard deviation of the FHR baseline**. The rest of the time domain features are computed from the complete FHR signal segment: **Long term irregularity** (LTI); **Short term variability** (STV); **Interval index** (II); **Delta value** and **Total delta value**.

Many of the above mentioned features have been used in cases of antepartum signal evaluation and the effectiveness of many of them depends on their performance in the presence of accelerations and decelerations.

5.3. Frequency domain features

Various spectral methods have been used for the analysis of adult heart rate (Task-Force 1996). In the case of FHR analysis no standardized use of frequency bands exists. Therefore we used two slightly different partitionings of the frequency bands as was previously used in (Georgoulas et al. 2006).

First we divided the frequency range into 3 bands (Task-Force 1996) and calculated energy of the signal in each one of them: **Very Low Frequency** (VLF); **Low Frequency** (LF) referred to as Mayer waves and **High Frequency** (HF) corresponding to fetal movement. Additionally the **ratio of energies** in the bands: $Ratio3Band = \frac{LF}{HF}$ was computed. It is a standard measure in adults and expresses the balance of behaviour of the two autonomic nervous system branches.

The alternative frequency partitioning followed suggestions of (Signorini et al. 2003). They proposed following 4 bands: **Very Low Frequency** (VLF); **Low Frequency** (LF) correlated with neural sympathetic activity; **Movement**

Frequency (MF), related to fetal movements and maternal breathing; **High Frequency** (HF), marking the presence of fetal breathing. Similarly to the previous 3-band division **the ratio of energies** was computed: $Ratio4Band = \frac{LF}{MF+HF}$. It is supposed to quantify the autonomic balance control mechanism (in accordance with the LF/HF ratio normally calculated in adults).

5.4. HRV based statistical features

Fetuses suffering from any possible heart condition were excluded from the database, therefore all beats were considered as normal (N) – thus the distance between two beats was depicted as NN. Based on commonly used features in adult HRV we computed several statistical measures (Task-Force 1996): **Standard deviation of the NN intervals** (SDNN) computed on the complete FHR segment. SDNN reflects all the cyclic components responsible for variability in the period of recording; the **root of the mean squared differences** (RMSSD) of successive NN intervals; $NN50$ counts the number of consecutive NN pairs that differ more than 50ms. $pNN50$ gives the ratio of $NN50$ beats, to total number of beats; lengths of axes in Poincaré plot (**Poincaré SD1, SD2**)– derived from the method for geometric HRV representation in the form of a graph where each RR interval is plotted as a function of the previous one.

5.5. Wavelet features

Wavelet transform is a very popular technique in many fields of signal processing, and also has been used recently in FHR processing (Papadimitriou et al. 1997).

We decomposed the signal into five levels of decomposition using the Malat algorithm with Daubechies order 4 (db4) the mother wavelet. Based on the decomposition of the signal we computed the mean and standard deviation (e.g. **A5mean**, **A5std**) in all details and the last – 5^{th} approximation.

5.6. Nonlinear features

Almost all nonlinear methods used for FHR analysis have their roots in adult HRV research. Fractal dimension is one of the useful estimators of FHR dynamics. There are two approaches to estimate the dimension of time series either by direct measurement of waveform or by operating in a reconstructed state space.

Correlation dimension D_2 is based on estimation of the correlation sum $C(r)$ which gives the probability that two randomly chosen points are close to each other with a distance smaller than r (Grassberger & Procaccia 1983).

There are several methods for estimation of waveform fractal dimension: **box-counting dimension**, which expresses the relationship between the number of boxes that contain part of a signal and the size of the boxes; **the Higuchi method (FD_Hig)** (Higuchi 1988) where a curve length $\langle L(k) \rangle$ is computed for different steps k and is related to the fractal dimension by exponential formula; the **variance technique of fractal dimension estimation (FD_Var)** that is based on properties of fractional Brownian motion. The variance σ^2 is related to the time increments Δt of a signal $X(t)$ according to the power law (Kinsner 1994); estimate of fractal dimension proposed by **Sevcik** (Sevcik 1998).

Entropy describes the behaviour of a system in terms of randomness, and quantifies information about underlying dynamics. The **Approximate Entropy (ApEn)** is able to distinguish between low-dimensional deterministic systems, chaotic

systems, stochastic and mixed systems (Pincus 1995). ApEn(m,r) approximately equals the average of a natural logarithm of conditional probabilities that sequences of length m are close to each other, within a tolerance r, even if a new point is added.

A slightly modified estimation of approximate entropy was proposed by (Richman & Moorman 2000) and resulted in what is known as **Sample Entropy (SampEn)**. This estimation overcame the shortcomings of the ApEn mainly because the self-matches were excluded. Secondly, conditional probabilities are not estimated by a template approach. SampEn requires that only one template finds a match of length $m + 1$. Used parameters for ApEn and SampEn estimation: tolerance $r = (0.15; 0.2) \cdot SD$ (SD stands for standard deviation) and the embedding dimension $m = 2$ (Pincus & Viscarello 1992, Liu et al. 2011)

The last of the nonlinear features was the **Lempel Ziv Complexity (LZC)** (Lempel & Ziv 1976). This method examines reoccurring patterns contained in the time series irrespective of time. A periodic signal has the same reoccurring patterns and low complexity while in random signals individual patterns are rarely repeated and signal complexity is high.

6. Feature evaluation

To be able to select appropriate statistical tests, all features were tested for normal distribution using χ^2 test. Only features with normal distribution in all three subsets (normal, suspect, pathological) were considered to have normal distribution in general. Most of the features therefore did not have normal distribution mainly due to the distribution of the pathological class.

We have tested statistical significance of the features for distinguishing between the three classes. ANOVA test was used for normally distributed features. Kruskal-Wallis test, which makes no distributional assumptions and therefore is not as powerful as the ANOVA, was used for the rest of features with non-normal distribution.

We evaluated the statistical significance of the features against individual expert annotations as well as GS annotation, which was based on all three expert annotations. Additionally we have used three different feature selection techniques that enabled us to rank the features based on their performance in the potential classification process using 10-fold cross-validation. Based on our previous experience we have used the following techniques – each one based on a slightly different principle – these are described in larger detail in (Witten & Frank 2005, Blum & Langley 1997, Mitra et al. 2002):

- **Information Gain Evaluation** (InfoGain) evaluates attributes by measuring their information gain with respect to the class.
- **One Rule Evaluation** uses the simple minimum-error measure adopted by the One Rule classifier.
- **SVM Feature Evaluation** evaluates attributes using recursive feature elimination with a linear support vector machine. Attributes are selected one by one based on the size of their coefficients.

7. Results

In this paper we work with the database of 552 intrapartum FHR recordings 20 minutes long that were annotated by experts as described in Section 3. In total 9 cases were

Table 1. Final results of expert evaluation computed relatively to "Gold standard"

	Expert #1	Expert #2	Expert #3
Sensitivity	71.80	72.45	85.90
Specificity	92.72	92.72	67.55
Intra-observer agreement	70.83	56.20	76.67
Inter-observer agreement		80.61	
Kappa statistics		0.36	

excluded due to total expert disagreement in annotation, hence the database used for feature evaluation consisted of 543 cases.

Results of expert annotation depicting the sensitivity and specificity of each individual and collectively built-up Gold standard, computed using majority voting of three experts, are presented in Table 1. The measures were computed for the normal and pathological classes with the suspect class always classified as correct. The table also presents the resulting intra- and inter-observer agreement as described in Section 3. Finally we have used kappa statistics to compare expert agreement against an agreement which might be expected by chance – value of 0.36 corresponds to fair expert agreement. We should mention here that kappa value depends largely on the data used and can not be used for comparison with performance on different datasets (Gwet 2010).

Considering Gold standard annotation as the main one for our work 139 cases were annotated as Normal, 107 as Pathological, and 306 as Suspect.

Since we have obtained a large amount of features the correlation between them had to be considered. Based on the origin of the features and using experimentally set thresholds we have found the following inter-correlated groups, from which only the first one was selected as a representative:

- the meanHR correlated with the VLF and A5mean (correlations 0.924 and 0.911, respectively)
- LTV; Delta (0.96)
- ApEn; SampEn; Sevcik (0.948, 0.915)
- FD_HigD; FD_HigDs; FD_HigDl (0.971, 0.878)
- FD_BoxDl; FD_BoxD; FD_BoxDs (0.813, 0.854)
- PoincareSD2; A5std (0.912)

Chi-square test was performed prior to statistical testing of individual features. Most of the features were found to have not-normal distribution.

Appropriate statistical tests against the expert annotation were used as described in Section 6. The results of the tests are presented in Table 2. From 55 features only those having significance level $p < 0.01$ are presented. The statistical significance ($p < 0.01$) is depicted by checkmark. For instance, the number of accelerations (# Accel.) is significant to all experts including (GS). However, the number of decelerations (# Decel.) is only significant to Exp #3 and GS. In the two last columns we present the results of three different ranking algorithms to rank the significant features from the point of view of individual features and their combinations.

Table 2. Statistical significance of the features when tested against different types of annotation. Only features that were found significant ($p < 0.01$) are presented in the table. Annotations used were: individual experts; Gold standard (GS). The last two columns represents rank of the features when used for classification (class.) and when assessed individually (indiv.).

Domain	Features	Statistical significance of features					
		Exp #1	Exp #2	Exp #3	GS	Rank (indiv.)	Rank (class.)
Time	baselineSD	–	✓	–	–	10	9
	# Accel.	✓	✓	✓	✓	1	1
	# Decel.	–	–	✓	✓	4	2-3
	II	–	✓	–	✓	8	5
Frequency	VLF	✓	–	–	–	6	7-8
Wavelet	D2mean	–	✓	✓	✓	11	6
Nonlinear	ApEn	–	✓	–	✓	9	11
	LZc	–	–	✓	✓	3	2-3
	FD_BoxDl	✓	✓	✓	✓	7	10
	FD_HigD	✓	✓	–	✓	5	4
	FD_Var	✓	✓	✓	✓	12	12
	Poincaré SD2	✓	✓	✓	✓	2	7-8

8. Discussion

Although many of the building blocks, specifically the features, presented in this paper were used by others before e.g. (Georgoulas et al. 2006, Signorini et al. 2003) the overall aim of this paper was novel.

We decided to examine an almost complete set of ever-used features for FHR characterization on a fairly large database against three-class expert evaluation. This approach enabled us to examine the features from the point of view of clinical experts who are unaware of the final outcome of the delivery when assessing the ongoing FHR. More importantly, clinicians should act against adverse outcome of the delivery when pathological FHR occurs. Thus FHR might clearly be pathological (by expert judgement) but the final outcome after e.g. caesarean section can be normal (by pH assessment). We think that the simple use of pH values as a basis for classifier training proposed in many works in the past e.g. (Jezewski et al. 2008, Salamalekis et al. 2006) can be seen as one of the reasons behind the almost non-existent transfer of the recent ideas to the clinical practice in the obstetricians wards.

Evaluation of the expert decision making process presented in Table 1 shows that the expert decisions are quite inconsistent, even though all experts should follow the same guidelines. Especially the intra-observer agreement suggests that the resulting decision is made based to some extent on the "feelings" of the clinician. Such observation is consistent with the findings of (Bernardes et al. 1997) and encourage the conclusions of (Steer 2008) that an automatic decision support system for evaluation of the FHR/CTG recordings might be of great value in the future.

It is also important to mention that the know-how acquired during the process of obtaining expert annotation clearly demonstrates that the FHR (and CTG in general)

is almost never evaluated without specific clinical context. This fact is in clear contrast with the assumptions of the most of the papers involving evaluation of the features.

Statistical significance of the features used by experts only seldom crossed $p < 0.01$ threshold. When it did, the intuitive features that would be expected to perform the best (e.g. meanHR) were found in the main to be insignificant. Simultaneously the quantity of non-linear features found significant suggests that the "intuition" based part of the decision process is rather large. The general approach to the FHR/CTG assessment is indeed based on the official FIGO guidelines. But the guidelines contain crisp and clear thresholds and rules which are difficult to adhere to precisely in a clinical setting as shown in the values of expert agreement rate in Table 1 as well as in experiments of others (Bernardes et al. 1997). Two distinctive reasons can be identified – first many of the FIGO parameters are only roughly estimated e.g. variability of the FHR is estimated by clinician but not measured. Secondly the evaluation of the FHR/CTG does not occur without the actual clinical context – something the FIGO guidelines do not regulate for – which we believe is one of the main reasons for the large inter-observer variability.

The number of features that are significant when using Gold Standard is, as expected, highly consistent with the conjunction of the individual expert evaluations. The last but one column of Table 2 shows the individual performance of the features and the last column depicts average feature ranking. From the point of view of automatic serial assessment of features, the classical ones (number of acceleration and deceleration) were very distinctive and ended in the top half. The fact that many of the non-linear features are ranked to the bottom half can be justified by their correlation, where the additional features after using LZC and FD_HigD do not contribute significantly to improvement of the final score. The inter-correlation of the nonlinear features that are presented in the Table 2 was in the range of 0.53 - 0.78 – therefore it did not fulfill our condition for "high" correlation but the effect was pronounced in the ranking method results.

We have tried to make the results as general as possible – thus we used several iterations for feature significance and those features that did not fulfil our criteria in the majority of repetitions were excluded (since they did not perform consistently and doubts have arisen about their general application). For feature ranking 10-fold cross-validation was used.

It is very hard to compare our results to the results in other works. First of all our approach is unique in the way we obtain annotation of the data, whose parameters (inter- and intra-observer variability) are nevertheless comparable to works of (Bernardes et al. 1997). Also the fact that each paper uses a different data set limits the means of direct comparison of different sets of results. In general we can say that our paper is bringing a new view on the problematic nature of automatic evaluation of the features. It confirms the need for additional features than are the FIGO suggested macroscopic ones as suggested in (Schiermeier et al. 2008). We also confirmed in accordance with (Gonçalves et al. 2006*b*) that the most useful additional features are the nonlinear ones.

9. Conclusion

The goal of this work was to find out which features could be possibly useful for mimicking the obstetricians behaviour when dealing with intrapartum FHR recordings and thus helping them with the diagnosis. Our paper used the previous research on the

extraction of different types of features. We extended it by comparing directly all the different features on one database using the same preprocessing steps. Additionally, based on the clinical experience as documented in (Schiermeier et al. 2008), we do not fully agree with a simple and unconditional relationship between pH value and FHR/CTG recording. Instead we have used the expert evaluation of the features.

We can confidently say that the findings reported in this paper are in general consistent with findings of others – namely:

- There are other features with information value besides the FIGO guidelines suggested macroscopic features.
- The combination of the macroscopic(FIGO) features and non-linear features is especially worth using.
- The clinical evaluation of the signals suffers from fairly high inconsistency.
- The task of evaluation of the FHR without other clinical data can bring only partial improvements.

To conclude – for the first time, to the best of our knowledge, statistical assessment of the features was performed on a large dataset against expert annotation. We warn against ungrounded assumption of automatic large correlation between FHR and umbilical pH. We believe that certain relationship exists but the type of relationship was never shown in any study, partly due to low numbers of newborns with clearly pathological pH. Expensive follow-up studies would be necessary to link the assumed intrapartum asphyxia and its manifestation into the later stages of newborn's life.

Findings on inter- and intra-observer variability are consistent with previous works e.g. (Blix et al. 2003). Additionally in our case we can report that our experts based their decision on the most-easy-to-assess macroscopic features (number of acceleration, deceleration, variability) and the rest of their decision making seems to be based on their "intuition" – possibly correlated with the nonlinear features in Table 2.

The goal for any future work must clearly be to try to verify our findings using different data sets. We will also try to integrate additional knowledge into the system that would take into account the clinical context of the test in an attempt to provide a working practical decision support system.

Acknowledgments

This work was supported by the research programs No. NT11124-6/2010 Cardiotocography evaluation by means of artificial intelligence of the Ministry of Health Care, No. MSM 6840770012 Trans-disciplinary Research in the Field of Biomedical Engineering II of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.

The authors would like to thank the clinical experts who beside helping with evaluation of the signals also contributed with very useful comments. Namely assoc. prof. Binder from the 2^{nd} Medical Faculty of Charles University in Prague, and dr. Vít from the Bulovka Teaching Hospital.

References

Alfirevic Z, Devane D & Gyte G M L 2006 Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour *Cochrane Database Syst Rev* **3**, CD006066.

Amer-Wåhlin I & Maršál K 2011 ST analysis of fetal electrocardiography in labor *Seminars in Fetal and Neonatal Medicine* **16**(1), 29–35.

Bernardes J, Costa-Pereira A, de Campos D A, van Geijn H P & Pereira-Leite L 1997 Evaluation of interobserver agreement of cardiotocograms *Int J Gynaecol Obstet* **57**(1), 33–37.

Bernardes J, Moura C, de Sa J P & Leite L P 1991 The Porto system for automated cardiotocographic signal analysis *J Perinat Med* **19**(1-2), 61–65.

Blix E, Sviggum O, Koss K S & Oian P 2003 Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. *BJOG* **110**(1), 1–5.

Blum A I & Langley P 1997 Selection of relevant features and examples in machine learning *Artificial Intelligence* **97**, 245 – 271.

Chaffin D G, Goldberg C C & Reed K L 1991 The dimension of chaos in the fetal heart rate *Am J Obstet Gynecol* **165**(5 Pt 1), 1425–1429.

d'Aloja E, Müller M, Paribello F, Demontis R & Faa A 2009 Neonatal asphyxia and forensic medicine. *J Matern Fetal Neonatal Med* **22 Suppl 3**, 54–56.

de Campos D A, Sousa P, Costa A & Bernardes J 2008 Omniview-SisPorto 3.5 - A central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis *Journal of Perinatal Medicine* **36**(3), 260–264.

Felgueiras C S, de Sá J P, Bernardes J & Gama S 1998 Classification of foetal heart rate sequences based on fractal features *Med Biol Eng Comput* **36**(2), 197–201.

Ferrario M, Signorini M & Magenes G 2009 Complexity analysis of the fetal heart rate variability: Early identification of severe intrauterine growth-restricted fetuses *Medical and Biological Engineering and Computing* **47**(9), 911–919.

FIGO 1986 Guidelines for the Use of Fetal Monitoring *International Journal of Gynecology & Obstetrics* **25**, 159–167.

Georgoulas G, Stylios C D & Groumpos P P 2006 Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines *IEEE Trans Biomed Eng* **53**(5), 875–884.

Gonçalves H, Rocha A P, de Campos D A & Bernardes J 2006*a* Internal versus external intrapartum foetal heart rate monitoring: the effect on linear and nonlinear parameters. *Physiol Meas* **27**(3), 307–319.

Gonçalves H, Rocha A P, de Campos D A & Bernardes J 2006*b* Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery *Med Biol Eng Comput* **44**(10), 847–855.

Grassberger P & Procaccia I 1983 Measuring the strangeness of strange attractors *Physica D: Nonlinear Phenomena* **9**, 189–208.

Guijarro-Berdinas B & Alonso-Betanzos A 2002 Empirical evaluation of a hybrid intelligent monitoring system using different measures of effectiveness *Artif Intell Med* **24**(1), 71–96.

Gwet K L 2010 *Handbook of Inter-Rater Reliability* Advanced Analytics, LLC.

Heintz E, Brodtkorb T H, Nelson N & Levin L A 2008 The long-term cost-effectiveness of fetal monitoring during labour: a comparison of cardiotocography complemented with ST analysis versus cardiotocography alone. *BJOG* **115**(13), 1676–1687.

Higuchi T 1988 Approach to an irregular time series on the basis of the fractal theory *Phys. D* **31**(2), 277–283.

Hopkins P, Outram N, Lofgren N, Ifeachor E C & Rosen K G 2006 *in* 'Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS '06' pp. 1784–1787.

Jezewski J, Kupka T & Horoba K 2008 Extraction of fetal heart-rate signal as the time event series from evenly sampled data acquired using Doppler ultrasound technique. *IEEE Trans Biomed Eng* **55**(2 Pt 1), 805–810.

Kim K K, Kim J S, Lim Y G & Park K S 2009 The effect of missing RR-interval data on heart rate variability analysis in the frequency domain. *Physiol Meas* **30**(10), 1039–1050.

Kinsner W 1994 Batch and real-time computation of a fractal dimension based on variance of a time series. Technical report Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Canada.

Laar J V, Porath M M, Peters C H L & Oei S G 2008 Spectral analysis of fetal heart rate variability

for fetal surveillance: review of the literature. *Acta Obstet Gynecol Scand* **87**(3), 300–306.

Lempel A & Ziv J 1976 On the complexity of finite sequences *IEEE Transactions on Information Theory* **IT-22 (1)**, 75–81.

Liu C, Liu C, Shao P, Li L, Sun X, Wang X & Liu F 2011 Comparison of different threshold values r for approximate entropy: application to investigate the heart rate variability between heart failure and healthy control groups *Physiological Measurement* **32**(2), 167.

Magenes G, Signorini M G & Arduini D 2000 *in* 'Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000' Vol. 3 pp. 637–641 vol.3.

Mitra P, Murthy C A & Pal S K 2002 Unsupervised feature selection using feature similarity *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 301312.

Papadimitriou S, Gatzounas D, Papadopoulos V, Tzigounis V & Bezerianos A 1997 Denoising of the fetal heart rate signal with non-linear filtering of the wavelet transform maxima. *Int J Med Inform* **44**(3), 177–192.

Pincus S 1995 Approximate entropy (ApEn) as a complexity measure *Chaos* **5 (1)**, 110–117.

Pincus S M & Viscarello R R 1992 Approximate entropy: a regularity measure for fetal heart rate analysis *Obstet Gynecol* **79**(2), 249–255.

Richman J S & Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *Am J Physiol Heart Circ Physiol* **278**(6), H2039–H2049.

Rosén K G 2005 Fetal electrocardiogram waveform analysis in labour. *Curr Opin Obstet Gynecol* **17**(2), 147–150.

Salamalekis E, Hintipas E, Salloum I, Vasios G, Loghis C, Vitoratos N, Chrelias C & Creatsas G 2006 Computerized analysis of fetal heart rate variability using the matching pursuit technique as an indicator of fetal hypoxia during labor. *J Matern Fetal Neonatal Med* **19**(3), 165–169.

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

Schiermeier S, von Steinburg S P, Thieme A, Reinhard J, Daumer M, Scholz M, Hatzmann W & Schneider K T M 2008 Sensitivity and specificity of intrapartum computerised FIGO criteria for cardiotocography and fetal scalp pH during labour: multicentre, observational study. *BJOG* **115**(12), 1557–1563.

Schumaker L L 2007 *Spline Functions: Basic Theory* Cambridge University Press.

Sevcik C 1998 A Procedure to Estimate the Fractal Dimension of Waveforms *Complexity International* **5**, –.

Signorini M G, Magenes G, Cerutti S & Arduini D 2003 Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings *IEEE Trans Biomed Eng* **50**(3), 365–374.

Sprott J C 2003 *Chaos and Time-Series Analysis* Oxford University Press.

Steer P J 2008 Has electronic fetal heart rate monitoring made a difference *Semin Fetal Neonatal Med* **13**(1), 2–7.

Strachan B K, van Wijngaarden W J, Sahota D, Chang A & James D K 2000 Cardiotocography only versus cardiotocography plus PR-interval analysis in intrapartum surveillance: a randomised, multicentre trial. FECG Study Group. *Lancet* **355**(9202), 456–459.

Sundström A, Rosén D & Rosén K 2000 'Fetal Surveillance - textbook' [online] Gothenburg, Sweden: Neoventa Medical AB.

Task-Force 1996 Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology *Eur Heart J* **17**(3), 354–381.

Valentin L, Ekman G, Isberg P E, Polberger S & Maršál K 1993 Clinical evaluation of the fetus and neonate. Relation between intra-partum cardiotocography, Apgar score, cord blood acid-base status and neonatal morbidity. *Arch Gynecol Obstet* **253**(2), 103–115.

Witten I H & Frank E 2005 *Data Mining: Practical machine learning tools and techniques* Morgan Kaufmann, San Francisco.