Automatic Classification of Intrapartal Fetal Heart-Rate Recordings – Can It Compete with Experts?

Václav Chudáček[1], Jiří Spilka[1], Michal Huptych[1], George Georgoulas[5], Petr Janků[2], Michal Koucký[3], Chrysostomos Stylios[4], and Lenka Lhotská[1]

[1] Dept. of Cybernetics CTU in Prague,
[2] Dept. of Obstetrics and Gynaecology FN Brno,
[3] Dept. of Obstetrics and Gynaecology 1.LF CUNI in Prague
[4] Dept. of Communications, Informatics and Management, TEI of Epirus, Artas, Greece
[5] Dept. of Computer Applications in Finance and Management, TEI of Ionian Islands, Lefkas, Greece
chudacv@fel.cvut.cz

Abstract. Fetal heart rate (fHR) is used to evaluate the fetal well-being during the delivery. It provides information of fetal status and allows doctors to detect ongoing hypoxia. Routine clinical evaluation of intrapartal fHR is based on description of macroscopic morphological features of its baseline. In this paper we show, that by using additional features for description of the fHR recordings, we can improve the classification accuracy. Additionally since results of automatic signal evaluation are easily reproducible we can objectify the whole process, thus enabling us to focus on the underlying reasons for high expert inter-observer and intra-observer variability.

Keywords: fetal heart rate, intrapartum, classification, inter-observer variability.

1 Introduction

Correct evaluation of fetal status from the available information is crucial when difficulties occur during delivery after an otherwise normal pregnancy. Even though the baby is equipped with a defense mechanism to tackle the stress conditions during the delivery, in some cases only timely intervention can prevent the long-term consequences resulting from prolonged oxygen insufficiency - consequences such as cerebral palsy, neonatal encephalopathy or even death [1].

Instrumental evaluation of the fetal well-being during delivery is more than hundred years old. Auscultation, sensing of the fetal heart rate (fHR) using a fetal stethoscope, introduced by Pinard in 1876, was replaced in 1960's by electronic fetal monitoring (EFM) by cardiotocography (CTG - recording of fetal heart rate and force/pressure of contractions) as the most important representative.

Although introduction of the EFM was accompanied by large expectation, since it offered continuous fetal surveillance, meta-analysis of large multicentre studies [1] did not prove any significant improvements in the delivery outcomes. Some studies additionally disapproved any evidence of advantages of continuous monitoring when compared to intermittent one. Moreover, EFM became the main suspect for increased rate of cesarean sections [1].

S. Khuri, L. Lhotská, and N. Pisanti (Eds.): ITBAM 2010, LNCS 6266, pp. 57–66, 2010.

In order to improve interpretation and thus lower the number of asphyxiated neonates CTG guidelines were introduced by International Federation of Gynecology and Obstetrics (FIGO) [2]. Even though the guidelines are available for more than twenty years poor interpretation of CTG still persists with inter-observer as well as intra-observer variations in CTG evaluation [3, 4].

First attempts of automatic CTG analysis followed FIGO guidelines that basically describe morphological changes in CTG.

Attempts to evaluate CTG using computer are as old as the guidelines. FIGO morphological features became fundamental features in most of clinically oriented systems. Automatic extraction of morphological features was proposed by prof. Bernades and resulted in development of SisPorto [5] and later CAFE [6] by Berdinas – both automatic "expert-like" systems for CTG analysis.

In many papers, including this one, only fHR signal is used since it is the signal containing direct information about the fetal state. For fHR description different features 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 works of Magenes [7] and Goncalves [8]. Another approach to fHR analysis examined frequency content by spectral analysis and paper [9] gives a short overview of works where fHR spectrum was analyzed. The fHR was also analyzed by various wavelets with different properties [10]. Other works analyzed nonlinear properties of fHR such as fractal dimension of reconstructed attractor and waveform fractal dimension. Different estimations of fractal dimension were reviewed by [11]. The most successful nonlinear methods for HRV analysis, so far, are approximate and sample entropy. They are often used for examination of nonlinear systems and proved their applicability also in fHR analysis [12].

The paper is further structured as follows: First the data used throughout the paper are described, than follows signal pre-processing and presentation of methods for feature extraction. Afterwards selection of the useful features for the data description is presented and finally results achieved by expert as well as automatic classification are compared and discussed.

2 Data Used

The fHR signals used in this work were measured either externally using Doppler ultrasound or internally using scalp electrode. Our data set consisted of 476 delivery recordings. The data were obtained at the Obstetricians ward of General Teaching Hospital in Prague on Neoventas' STAN S21 device. Two modes of acquisition are depicted on Figure 1. In 60% (280) of all cases the measurement of fHR was done using ultrasound, in the rest electrode was attached to the scalp of the fetus. Signals were then annotated by five experts with at least five years of praxis as obstetricians; the process is described in Section 4.

Recordings obtained with the scalp electrode have usually fewer missed values and are in general less noisy.

All the recordings had to be checked for patient anamnesis and only one fold pregnancies delivered during 38^{th} – 42^{nd} week of pregnancy were chosen for the final database.

Additionally umbilical artery pH was obtained as and objective evaluation of hypoxia. The definition of neonatal acidemia is defined in most textbooks when pH

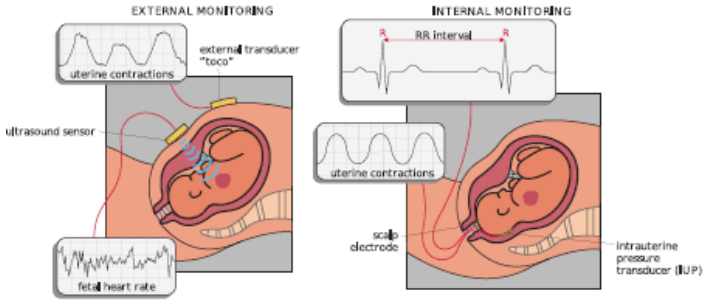

Fig. 1. Recording of the fetal heart rate and uterine activity [13]

value is below 7.05, but many recent works suggest otherwise [14]. Since pathological recordings are very hard to get and based on recommendation of cooperating obstetricians we decided to classify the fetuses as normal (i.e. without sustained hypoxia) if having pH above or equal to 7.15 and abnormal otherwise.

3 Automatic Classification Process

3.1 Signal Preprocessing

Relevance of extracted features usually highly depends on the quality of the preprocessing steps - in our work following steps were used: artifacts removal; interpolation; choice of appropriate segment; and de-trending of the signal.

The fHR signal contains a lot of artifacts caused by mother and fetal movements or displacements of the transducer. Usually between 20% and 40% of all data are affected by artifacts.

We employed the artifacts removal previously used in [5], where all abrupt changes in fHR are removed and replaced by line - see Figure 2.

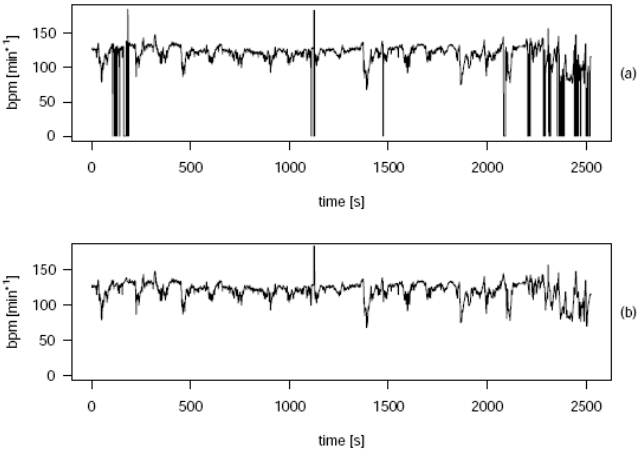

Fig. 2. Example of raw fHR signal (a) and signal after preprocessing (b)

It is important to emphasize the way the data are replaced since the results of analysis are affected by the way we treat gaps. We have used Hermit interpolation of missing data which is in general correct for fHR, but holds only for gaps of duration less then *20s*.

The fetal heart rate is non-uniformly sampled. This could affect the results of nonlinear methods, such as fractal dimension and entropy, though, since sampling is deterministically non-uniform, no problems were expected.

During the segment selection phase we tried to choose the segment as close as possible to delivery. But usually fHR directly preceding the delivery is largely contaminated with artifacts and noise. Therefore, only segments with less then 20% of erroneous signal were selected. The final segments were 20 minutes long meaning 4800 samples for signal sampled at 4 Hz.

Physiological time series are generally considered as non-stationary, i.e. statistical properties of physiological signal (mean, variance, and correlation structure) vary during time. We work with segments of short duration. Therefore, we could carefully detrend signal using third order polynomial and consider it stationary.

3.2 Feature Extraction

Since establishment of FIGO guidelines, various types of features were tested in the technical papers. Nevertheless only the classical - morphological and macroscopic features - are used in the clinical setting. In this section we thoroughly describe nonlinear features that could be used for automatic fHR classification and could bring added value to the decision process in comparison to use of classical feature set alone.

3.2.1 Classical Features

Morphological features introduced by the FIGO guidelines describe the shape and changes of the baseline such as: baseline mean (**baselineMean**); standard deviation (**baselineSD**); median of long-term (**medianLTV**) and short term variability (**STV**); and mean of interval index (**meanII**). More detailed description of the classical features is not purpose of this paper and can be found e.g. in [10].

The second category of features, that can be called classical are those used routinely in adult HRV evaluation and are standardized in [15] such as **NN50**, **RMSSD**, as well as the frequency features describing the amounts of energy in different energy bins – for example – energy in the low frequencies bin 0.05-0.15Hz (**LF**) and ratios of energies in different bins e.g. (**ratLF_HF**) [12].

3.2.2 Nonlinear Features

The nonlinear approach may reveal relevant clinical information of fHR not apparent in the time and frequency domain. Since heart-beat fluctuates on different time scales and is self-similar, fractal dimension is useful estimator of fHR dynamics. Detailed description of methods for estimation of fractal dimension entropy and complexity measures can be found in [16-19] and we will use the limited space to bring to attention the specifics of applying the methods on the fHR data.

Fractal dimension of attractor

In order to reconstruct an attractor we have estimated embedding parameters (time delay, τ, and embedding dimension, m) using mutual information approach and Cao's

method. We used only correlation dimension, *D2*, for attractor analysis since it is robust to noise. To be able to reliably estimate the fractal dimension a certain data length is necessary. There exist many theories about required data length with general agreement that the data length needed increases exponentially with data dimension. We followed data size requirements as suggested in [17]; for estimating a dimension d, a minimum data length required is $N_{min} = 10^{d/2}$.

Fractal dimension of waveform

The former approach considers signal in R^2 as a geometric object and directly uses it without any further transformation. Thus estimated dimension is always in range (1;2) because the geometrical representation of signal is more complicated than line but never covers the whole 2D space. We used following methods for waveform dimension estimation: box counting (**fdBoxCountDx**), Higuchi's (**fdHiguchiDx**), and variance (**fdVarianceDx**). Note that both methods estimate Hurst exponent, H, that is related to fractal dimension $D = E + 1 - H$, where E is the Euclidian dimension which equals to one for time series. Additionally detrended fluctuation analysis (**DFA**) is used for spectral slope estimation α that is related Hurst exponent $H = 1 - \alpha$.

We used Weierstrass cosine function $w(t)$ and examined how the data length affects estimate of fractal dimension for each method. The Higuchi method provides good estimates for all data lengths while variance and box counting methods offer biased estimates of fractal dimension. The DFA method converges to the theoretical value for increasing N. The different algorithms for estimation of the fractal dimension were probed by [18]. They also concluded that Higuchi's method is the most reliable but, also, the most sensitive to noise. Since fHR certainly contains noise we will also take advantage of the variance method offering robustness to noise.

The estimated dimension is not only dependent on data length but also on the dimensionality of data. We examined bias of the estimated fractal dimension by varying the dimension D of $w(t)$ in $(1;2)$ interval. The results showed that dimension estimated by the box counting method is biased. The Higuchi's and the variance method provide unbiased estimate of the fractal dimension only for dimensions $D>1.5$. The DFA method offers biased estimation for $D<1.3$ and $D>1.7$.

Next, we estimated the two scaling regions as were described by [18]. The authors suggested two scaling regions on the log-log plot of some measurement function, e.g. number of boxes, versus size of region (e.g. size of the box). Higuchi named the time where the curve bends as critical time τ_c. To standardize estimated dimension we determined the τ_c for all methods and finally we set it to $\tau_c = 3s$.

The used data meets demands on required length. The limitation of waveform fractal analysis lies in biased estimate for low dimension - in our work we used 4800 samples long signals. Analysis performed on synthetic Weierstrass function suggests that we can reliably estimate the fractal dimension by Higuchi's method. It is necessary to point out, that Weierstrass function does not completely reflects nonlinear and stochastic properties of fetal heart rate. Hence, we have to carefully interpret results of estimated waveform fractal dimensions.

Entropy

Kolmogorov–Sinai entropy is not usable for a noisy time series of finite length; therefore, approximate entropy **(ApEn)** [19] was designed and proved its usefulness for a short and noisy time series.

However, it provides biased estimation and is dependent on data length. Sample entropy **(SampEn)** eliminated drawbacks of ApEn by reducing bias and data length dependence. Since fHR records have finite length, the entropy estimation in terms of length is of a major interest. In [19] the Pincus showed that ApEn is broadly applicable for data series of length $N > 100$. Nevertheless, this value was suggested for wide spectrum of applications and in our case, meaningful data length for ApEn is about 1000 samples. The choice of parameter m was proposed by [20]. They concluded, the best results are achieved when $m = 2$. However, this holds only in cases when a dynamical system is not purely deterministic. The tolerance r was set to $r = 0.2 *$ standard deviation.

Lempel Ziv Complexity

Lempel Ziv Complexity **(LZC)** [21] examines reoccurring patterns in time series. The more reoccurring patterns - the less complex signal is. The LZC method estimates complexity of encoded signal so that dynamical changes of signal are replaced with particular character. We used binary encoding in order to avoid dependence of results on quantification criteria and normalization procedures. The required data length for LZC was examined by [22] and concluded that the minimum length is 1000 samples for binary encoded data.

3.3 Feature Selection

To determine which features to use for classification purposes feature selection algorithms were used to identify the ones containing the most of useful information. Features were ranked and since cross-validation was used only features ranked within the first ten best are presented in Table 1.

Techniques used to asses the features were computed in Weka [23]:

- **Information Gain Evaluation** (InfoGain), which evaluates attributes by measuring their information gain with respect to the class. First we discretize numeric attributes. Then based on entropy computation we define Information Gain as a criterion of impurity in a training set. Measure that reflects additional information about Y provided by X that represents the amount by which the entropy of Y decreases is called Information Gain.
- **One Rule Evaluation** uses the simple accuracy measure adopted by the One Rule classifier. It uses the minimum-error attribute for prediction, discretizing numeric attributes.
- **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, relearning.

Based on the results of the single selectors **meta-selection** was performed to acquire one representative set of features that was used further in classification. The meta-selection was performed by simple majority voting, where only features selected by at least two selectors were picked into the final set. The results of each selection method as well as the final meta-selected set are presented in Table 1.

Table 1. Features selected by different selection methods, and the final set of features used for further classification. Features abbreviations are described in Section 3.2.1 and 3.2.2. Description of the feature selection method is in Section 3.3.

Selection method	Features selected
InfoGain	baselineSD, deltaTotal, meanLTV, medianLTV, fdBoxCountDs, ApEn, meanII, baselineMean, medianII
One Rule	baselineSD, deltaTotal, LF_HFMF, medianLTV, meanSTV, meanII, fdBoxCountDs, FdHiguchiDs, baselineMean, stdLTV
SVM	baselineSD, meanII, FdHiguchiDl, baseline_Mean, ApEn, DFA_1, FdBoxCountDs, deltaTotal, LF_HF
MetaSelection	baselineSD, deltaTotal, meanII, medianLTV, fdBoxCountDs, ApEn baselineMean,

4 Results of Expert Classification

Even though objective annotation was obtained with the data, our preliminary tests showed that sole post delivery pH measurement does not give full picture of the clinical evaluation of ongoing delivery. Therefore some tool to allow the expert obstetricians to annotate the signal was needed.

The annotator application, developed for this purpose, runs in the Java runtime environment. The application adopts the most commonly used display layout of CTG machines, therefore poses no difficulty to adjust. The centimeter grid is always preserved irrespective of display resolution - see Figure 3.

After clicking on the Java web start reference in the browser the application checks its directory in the documents folder of MyDocuments directory, unzips the data for annotation and presents them to the expert.

Expert annotation is based on three FIGO classes – normal; pathological and suspect. One of the biggest advantages of the system is its collection system, where an automatic upload is performed after every hundred annotated records. The information about annotations is then submitted to the ftp site, from which further analysis can be made.

Results of annotation depicting the sensitivity and specificity of each individual expert against the collectively built up Gold standard (computed using majority voting of three experts) is presented in Table 2.

The Table 2 also presents resulting intra-observer (IaOV) and inter-observer variability (IeOV). Finally we use kappa statistics to compare expert agreement against that which might be expected by chance – value of 0.36 corresponds to 36% above chance agreement of experts.

Fig. 3. Main window of the application – in the middle fHR signal on the grid where one square represents exactly one centimeter (on screen)

Table 2. Final results of expert evaluation relative to Gold standard computed as rounded average of three expert evaluations

All in [%]	**Expert1**	**Expert2**	**Expert3**
Accuracy	76	75	84
Sensitivity	75	72	79
Specificity	78	74	76
IaOV	71	56	77
IeOV		80	
Kappa statistics		36	

5 Results of Automatic Classification

When computing the results of automatic classification we have utilized 10-fold cross-validation using following classifiers: Naive Bayes, Support Vector Machine (SVM), where the polynomial kernel and penalty parameters $C = 1$ were used, and C4.5 decision tree. All methods were implemented in WEKA software [23]. Short description and further references for these methods can be found in [23].

The classification results are presented in Table 3. From all performance measures, the specificity, computed as a ratio of true positives to sum of true positive and false negative, is of major importance since a classifier with higher specificity causes lower number of false alarms that leads to lower rate of unnecessary intervention. Regarding the specificity, the SVM performed best. However, statistical tests revealed that difference between individual classifiers is statistically insignificant on $p>0.01$ confidence level.

Table 3. Final results of automatic classification

All in [%]	NaiveBayes	SVM	C4.5 Tree
Accuracy	73	72	65
Sensitivity	84	78	74
Specificity	64	70	57
AUC	79	74	69

The same stands for the results of the automatic classification, that are well comparable to inter-observer variability and differences to expert classification were found insignificant in all three experts on confidence level $p>0.05$.

6 Discussion and Conclusion

We have evaluated 41 computed features from the classical (morphological and HRV) and non-linear domains. The best results were achieved on the meta-selected feature set based on automatic feature selection. Results achieved by automatic classification of more than 70% sensitivity and specificity are comparable with the inter-observer variability in expert evaluation.

We can conclude that automatic evaluation of the intrapartum fHR works sufficiently well. It supports our decision to use it as the first building block in the future work. Experience with the way clinicians decide on the class of the fHR record clearly suggests that additional clinical information about the patient is need to put the fHR into perspective. In the future work we will try to develop decision support system that will be built on signal processing/evaluation and clinical context.

Acknowledgments. The authors would like to thank the clinical experts who beside helping with evaluation of the signals also contributed with useful comments. Namely prof. Binder from the 2nd LF CUNI in Prague, and dr. Vít from the Bulovka Teaching Hospital.

This work was supported by the research program No. MSM 6840770012 "Transdisciplinary 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.

References

1. Steer, P.J.: Has electronic fetal heart rate monitoring made a difference. Seminars in Fetal and Neonatal Medicine 13, 2–7 (2008)
2. FIGO. Guidelines for the use of fetal monitoring. International Journal of Gynecology & Obstetrics 25, 159–167 (1986)
3. Blix, E., Sviggum, O., Koss, K.S., Oian, P.: Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. In: BJOG, Nordic School of Public Health, Gothenburg, Sweden, vol. 110(1), pp. 1–5 (2003)
4. Bernardes, J., Costa-Pereira, A., de Campos, D.A., van Geijn, H.P., Pereira-Leite, L.: Evaluation of interobserver agreement of cardiotocograms. Int. J. Gynaecol Obstet, epartamento de Ginecologia e Obstetrícia, Hospital de S. Jo?o, Faculdade de Medicina do Porto, Oporto, Portugal 57(1), 33–37 (1997)

5. Bernardes, J., Moura, C., de Sa, J.P., Leite, L.P.: The Porto system for automated cardiotocographic signal analysis. J. Perinat. Med. 19(1-2), 61–65 (1991)
6. Guijarro-Berdinas, B., Alonso-Betanzos, A., Fontenla-Romero, O.: Intelligent analysis and pattern recognition in ctg signals using a tightly coupled hybrid system. Artif. Intell. 136, 1–27 (2002)
7. Magenes, G., Pedrinazzi, L., Signorini, M.G.: Identification of fetal sufferance Intepartum through a multiparametric analysis and a support vector machine. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., Dipartimento di Informatica e Sistemistica, Pavia Univ., Italy, vol. 1, pp. 462–465 (2004)
8. Goncalves, H., Rocha, A.P., Ayres-de Campos, D., Bernardes, J.: Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. Med. Bio. Eng. Comput. 44, 847–855 (2006)
9. Laar, J., Porath, M.M., Peters, C.H.L., Oei, S.G.: Spectral analysis of fetal heart rate variability for fetal surveillance: review of the literature. Acta Obstet. Gynecol. Scand. 87(3), 300–306 (2008)
10. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: Feature extraction and classification of fetal heart rate using wavelet analysis and support vector machines. International Journal on Artificial Intelligence Tools 15, 411–432 (2005)
11. Hopkins, P., Outram, N., Löfgren, N., Ifeachor, E.C., Rosén, K.G.: A comparative study of fetal heart rate variability analysis techniques. Conf. Proc. IEEE Eng. Med. Biol. Soc. 1, 1784–1787 (2006)
12. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. IEEE Trans. Biomed. Eng., Laboratory for Automation and Robotics 53(5), 875–884 (2006)
13. Sundstrom, A., Rosen, D., Rosen, K.: Fetal surveillance - textbook, Gothenburg
14. Cao, L.: Practical method for determining the minimum embedding dimension of a scalar time series. Physica D 110, 43–50 (1997)
15. Task-Force. 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 (1996)
16. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information. Physical Review A 33(2), 1134–1140 (1986)
17. Esteller, R., Vachtsevanos, G., Echauz, J., Lilt, B.: A comparison of fractal dimension algorithms using synthetic and experimental data. In: Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, ISCAS 1999, vol. 3, pp. 199–202 (1999)
18. Peng, C.K., Havlin, S., Stanley, H.E., Goldberger, A.L.: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5, 82–87 (1995)
19. Pincus, S.: Approximate entropy (ApEn) as a complexity measure. Chaos 5(1), 110–117 (1995)
20. Pincus, S.M., Viscarello, R.R.: Approximate entropy: a regularity measure for fetal heart rate analysis. Obstet. Gynecol. 79(2), 249–255 (1992)
21. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on Information Theory, IT 22(1), 75–81 (1976)
22. Ferrario, M., Signorini, M.G., Cerutti, S.: Complexity analysis of 24 hours heart rate variability
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)