33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Assessment of features for automatic CTG analysis based on expert annotation

Václav Chudáček, Jiří Spilka, Lenka Lhotská, Petr Janků, Michal Koucký, Michal Huptych and Miroslav Burša

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.

Annotation derived from the panel of experts instead of the commonly utilized pH values was used for evaluation of the features on a large data set (552 records).

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.

I. 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.

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 [1] and, moreover, continuous CTG is the main suspect for increased rate of cesarean sections for objective reasons [25].

To improve the results of cardiotocography, the International Federation of Gynecology and Obstetrics (FIGO) introduced general guidelines [8]. 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 [25] with large inter-observer as well as intra-observer assessment variations [4], [2].

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 [6] mostly for antepartum CTG evaluation. 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. Extension of this work proposed towards the full CTG (with TOCO included) should be straightforward in the future. Different FHR features were investigated in the past, many of them heavily influenced by the research in adult heart rate variability (HRV) analysis. Morphological features were used by the founders of the automatic CTG evaluation Bernardes et al. [2], [6]. Statistical description of CTG tracings was employed in the study of Gonçalves [10]. Another approach to FHR analysis examined frequency content by spectral analysis and the paper of Laar [15] gives a short overview of most of the works where FHR spectrum was analyzed. The FHR was also analyzed by wavelets with different properties [20]. A comprehensive evaluation of nonlinear methods was performed in [24]. Approximate and sample entropy – the most successful nonlinear methods used for examination of nonlinear systems – were used in works of [9], [10]. Another method that performs well on the FHR recordings is Lempel-Ziv complexity employed in [7].

In recent papers the features are used usually evaluated against umbilical artery pH measurement as an annotation. Although pH of the umbilical artery blood is certainly the objective value, there are many studies showing the relation between the CTG/FHR signal recorded during delivery and outcome of delivery to be rather weak [12], [26]. Therefore, we have used the experts as the source of signal annotations. Our recent paper presents the expert annotation and feature analysis in more detail [5].

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.

Václav Chudáček, Jiří Spilka, Lenka Lhotská, Michal Huptych and Miroslav Burša are with the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic chudacv@fel.cvut.cz

Petr Janků is with the Obstetrics and Gynaecology clinic, University Hospital in Brno, Czech Republic pjanku@fnbrno.cz

Michal Koucký is with the Gynaecology and Obstetrics unit of the Charles University Hospital, Prague, Czech Republic michalkoucky@seznam.cz

II. Data

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 Neoventa's STAN S21 system using external ultrasound probes.

All recordings were checked for patient anamnesis and only one fold pregnancies delivered during 38^{th} – 42^{nd} week of pregnancy were chosen for the database, which finally consisted of 552 deliveries. 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.

In this paper, expert annotation of the FHR recordings was used as a basis for feature evaluation. Expert annotation has also 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.

Annotations, acquired by our annotation software, 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. Three measures were used for evaluation: intra-observer agreement as a percentage of consistently annotated records to all annotations, inter-observer agreement as a percentage of equally annotated records among the three experts to all annotations, and the kappa statistics describing agreement among the experts.

III. Signal preprocessing

The preprocessing consisted of four main steps as described further: 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. Segments were a maximum of 24 minutes long maximally 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 the selected segment is shown in Figure 1.

The FHR signal almost always contains artefacts 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%.

The algorithm proposed by Bernardes et al. [3] was utilized for artefacts removal – all abrupt changes in FHR were removed and replaced.

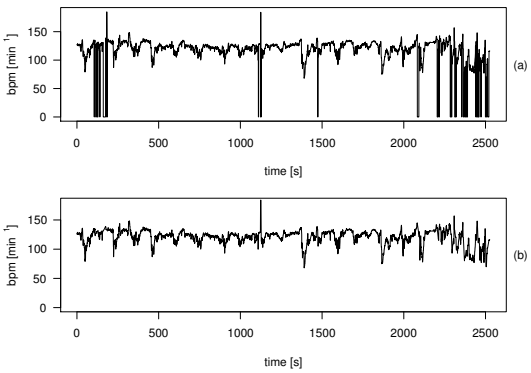

Fig. 1. Artefacts removal. (a) Raw signal, (b) signal after artefacts removal.

We used cubic Hermite spline interpolation [13] to interpolate over short gaps. We did not compute across a gap when the length of the missing data was 20 seconds or more – the value obtained based on our experiments.

Physiological time series are generally considered as non-stationary, i.e. statistical properties of physiological signal (mean, variance, and correlation structure) vary during time. For purposes of the frequency and non-linear features computation we have detrended the signal using third order polynomial, estimated, so that uninteresting trend was removed but interesting dynamics are preserved.

IV. Extracted Features

Features used for purposes of this paper were almost complete collection of all features used for evaluation of intrapartal/antepartal FHR in recently published papers.

Morphological features proposed in the FIGO guidelines are the features used in the obstetricians wards. A well known algorithm for feature extraction described in [3] was used for the macroscopic feature extraction. The features extracted were: **Mean of the FHR baseline**; **standard deviation of the FHR baseline**; **Number of accelerations**; **Number of decelerations**.

The time-domain features were computed according to [9]: **Long term irregularity** (LTI); **Short term variability** (STV); **Interval index** (II); **Delta value** of the FHR signal and **Total delta value**.

To analyze the the spectrum of the FHR partitioning into 4 bands was used following suggestion of [23].

For wavelet features we have decomposed the signal into five levels of decomposition using the Malat algorithm with Daubechies order 4 (db4) mother wavelet. Based on the decomposition of the signal we computed the mean (e.g. **D2mean**) and standard deviation in all details and the last – 5^{th} approximation.

Correlation dimension D_2 is based on estimation of correlation sum $C(r)$ which gives the probability that two randomly chosen points are close to each other with distance smaller than r. There are several waveform fractal dimensions estimated by the different methods: **box-counting dimension**, **the Higuchi dimension (FD_Hig)** [11], **the**

TABLE I

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

dimension of variance (FD_Var) [14], and estimate of fractal dimension proposed by **Sevcik** [22].

Entropy describes the behaviour of a system in terms of randomness, and quantifies information about underlying dynamics. The **Approximate Entropy (ApEn)** [17] is 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 [19] and resulted in **Sample Entropy (SampEn)**. Used parameters for ApEn and SampEn estimation were: tolerance $r = (0.15; 0.2) \cdot SD$ and the embedding dimension $m = 2$ [18].

The last of the two nonlinear features used were the **Lempel Ziv Complexity (LZC)** [16] and lengths of axes in Poincaré plot (**Poincaré SD1, SD2**).

V. EVALUATION METHODS

The statistical significance of the features for distinguishing between the three classes was tested using ANOVA test, for normally distributed features, and Kruskal-Wallis test 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 [27]:

- **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.

VI. RESULTS

Considering Gold standard annotation as the main one for our work 149 cases were annotated as Normal, 115 as Pathological, and 275 as Suspect. The high prevalence of the pathological class was because of our effort to collect as much cases with low neonatal pH values (88 records did have pH lower then 7.15). The suspect class resulted from the above mentioned fact that the clinicians were left without any additional clinical information, which is otherwise routinely used in their decision making process – thus any uncertainty usually resulted in the suspect class.

The features were selected first based on their mutual correlation. When the features' correlation coefficient was higher then 0.9 only one feature was used for further computation. The omitted feature was the one with average higher correlation among the rest of the features. Overall results are presented in the Table I. 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.

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 I 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.

TABLE II

FINAL RESULTS OF EXPERT EVALUATION COMPUTED RELATIVELY TO "GOLD STANDARD"

All in [%]	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	

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 II. 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. The kappa statistics was used to compare expert agreement against an agreement which might be expected by chance – value of 0.36 corresponds to fair expert agreement.

VII. CONCLUSION

For the first time we have evaluated the intrapartum FHR against the expert annotation encouraged by works of e.g. [21]. We have found features that seem to be useful for mimicking the obstetricians behavior when dealing with intrapartum FHR recordings. We have compared directly all the different features on one database using the same preprocessing steps. 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.

In the future we plan to work on lowering the variability of the clinician's decision by means of automatic evaluation of the selected features and their presentation to the doctor while using additional information about the actual status of the delivery – necessary step to really get the clinicians on board.

REFERENCES

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

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

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

[4] Ellen Blix, Oddvar Sviggum, Karen Sofie Koss, and Pal Oian. 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, Jan 2003.

[5] Václav Chudáček, Jiří Spilka, Petr Janků, Michal Koucký, Lenka Lhotská, and Michal Huptych. Automatic evaluation of intrapartum fetal heart rate recordings: A comprehensive analysis of useful features. *Physiological Measurement*, In Press:–, 2011.

[6] D. Ayres de Campos, P. Sousa, A. Costa, and J. Bernardes. 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, 2008.

[7] M. Ferrario, M.G. Signorini, and G. Magenes. 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, 2009.

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

[9] George Georgoulas, Chrysostomos D Stylios, and Peter P Groumpos. 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, May 2006.

[10] Hernâni Gonçalves, Ana Paula Rocha, Diogo Ayres de Campos, and João Bernardes. 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, Oct 2006.

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

[12] Janusz Jezewski, Tomasz Kupka, and Krzysztof Horoba. 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, Feb 2008.

[13] D. K. Kahaner, C. Moler, and S.G. Nash. *Numerical Methods and Software*. Prentice-Hall, 1989.

[14] Witold Kinsner. 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, 1994.

[15] J. Van Laar, M. M. Porath, C. H L Peters, and S. G. Oei. Spectral analysis of fetal heart rate variability for fetal surveillance: review of the literature. *Acta Obstet Gynecol Scand*, 87(3):300–306, 2008.

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

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

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

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

[20] 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. *J Matern Fetal Neonatal Med*, 19(3):165–169, Mar 2006.

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

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

[23] Maria G Signorini, Giovanni Magenes, Sergio Cerutti, and Domenico Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Trans Biomed Eng*, 50(3):365–374, Mar 2003.

[24] Jiří Spilka, Václav Chudáček, Michal Koucký, Lenka Lhotská, Michal Huptych, Petr Janků, George Georgoulas, and Chrysostomos Stylios. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, In Press:–, 2011.

[25] Philip J Steer. Has electronic fetal heart rate monitoring made a difference. *Semin Fetal Neonatal Med*, 13(1):2–7, Feb 2008.

[26] P.A. Warrick, E.F. Hamilton, D. Precup, and R.E. Kearney. Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *IEEE Transactions on Biomedical Engineering*, 57(4):771–779, 2010.

[27] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.