fetal monitoring,
baseline estimation,
acceleration, deceleration

Janusz JEZEWSKI[1], Krzysztof HOROBA[1], Dawid ROJ[1], Janusz WROBEL[1],
Tomasz KUPKA[1], Adam MATONIA[1]

A NOVEL APPROACH TO COMPARISON OF THE FETAL HEART RATE BASELINE ESTIMATION ALGORITHMS

A number of algorithms for estimating the so called fetal heart rate baseline was proposed so far. However, there is no reference pattern enabling their objective evaluation, and thus no methodology of comparing the competing algorithms still exists. In this paper we propose a method for evaluation of automatically determined baseline in reference to a group of experts, basing on ten separate groups of signals comprising typical patterns observed in the fetal heart rate. For the purpose of quantitative assessment of the estimated baseline a new synthetic inconsistency coefficient is presented. The proposed methodology was applied to evaluate ten well-known algorithms. We believe that the method will be a valuable tool for assessment of the existing algorithms, as well as for developing new ones.

1. INTRODUCTION

Automated analysis of fetal heart rate signal (FHR) is today one of the most common diagnostic tools used in perinatal medicine. It relies on assessment of some characteristic patterns, such as baseline, acceleration and deceleration episodes, oscillations and others. Among these patterns, the one most crucial for reliable interpretation of recording is the baseline – a line indicating the long-term changes of the fetal heart rate in time. The baseline definition provided by the FIGO guidelines says: ”Baseline is the mean level of the FHR when this is stable, accelerations and decelerations being absent” [12]. The impact of the baseline on detection of acceleration and deceleration patterns (A/D) is evident and results directly from the FIGO guidelines, which define these episodes as temporary deviations of FHR curve from the baseline. Acceleration is defined as an increase of the FHR with an amplitude of at least 15 beats per minute (bpm) and lasting more than 15 seconds. For deceleration the thresholds are equal to 15 bpm and 10 seconds.

Presence or absence of an appropriate number and type of A/D episodes in the FHR signal is a basis for evaluation of fetal wellbeing. Therefore, reliable detection of these episodes is a crucial issue in automated analysis of FHR recordings. It is, however, infeasible without correct baseline estimation, and even small inaccuracy may lead to misdetection of the fetal distress [4].

Automated methods of baseline estimation can be classified into three main groups:

[1]Institute of Medical Technology and Equipment ITAM, 118 Roosevelt Str., 41-800 Zabrze

- Filtering – based on filtering approach, most often relying on digital low-pass filtering. Also very common is nonlinear filtering, with filter parameters being modified according to the history of input and/or output signal.
- Statistical – they are based on statistical measures of central tendencies, being determined in successive segments of signal or for signal as a whole.
- Statistical and filtering mix – utilizes nonlinear filtering technique with parameters being modified according to statistical properties of the signal (as a whole or in overlapping windows).

As there is no strict definition of an algorithm for the FHR baseline estimation, the aim of all existing methods is to simulate the clinical expert's behavior. Therefore, the correctness of estimation can only be assessed involving clinical experts of considerable experience in this matter. The problem in this approach is that a perfect expert, and thus a reference baseline, does not exist. It was proven in the literature [9] that significant differences in baseline interpretation are observed between experts. What is more, those differences occur even between baselines estimated by the same expert after some time. For this reason it is necessary to use a pseudo-reference of baseline interpretation, coming from a group of experts. In this approach the final rating of an algorithm is determined by averaging the results obtained with respect to each individual expert's interpretation.

The baseline automatically estimated by a given algorithm should be as close as possible to the baseline assumed as a reference. However, direct assessment of the difference between two baselines does not provide any information on whether differences occur in a segment of FHR signal significant for further A/D episodes detection or not. From the perspective of FHR signal analysis it is important to evaluate inconsistency of automatically determined baseline especially in these segments corresponding to clinically significant deviations of FHR signal [7]. Therefore, the assessment of inconsistency should be based on differences between acceleration and deceleration episodes detected with respect to both baselines being compared.

2. METHODS

Method of inconsistency evaluation between two baselines has already been proposed in [9] and relies on a set of coefficients. The authors assumed that the coefficients had to reflect three main components of inconsistency between episodes determined for two baselines: difference in total number of detected episodes (AIN), difference in location (AIL) and difference in area (AIA) of matching episodes. The proposed coefficients were normalized to the range $0 \div 1$. Value of one indicates a full inconsistency, value of zero – perfect agreement of episodes. The resulting inconsistency coefficient for accelerations (detected automatically on the basis of two baselines) is calculated using three components of inconsistency, according to the equation:

$$AIR = 1 - \sqrt[3]{(1 - AIN) \cdot (1 - AIL) \cdot (1 - AIA)} \tag{1}$$

Similarly the DIR coefficients are determined for decelerations. The final inconsistency between baselines (ADIR) is calculated using AIR and DIR coefficients as:

$$ADIR = (AIR + 2 \cdot DIR)/3 \tag{2}$$

It should be noticed that acceleration inconsistency AIR comes to the final coefficient with lower weighting factor. It is caused by the fact that decelerations are clinically more significant, with bigger impact on the fetal state assessment.

The weak side of this coefficient may be that the size of the episode (its area) is not taken into account if the episode does not have corresponding episode detected on the basis of the other

baseline (so called matching episode). This means that a significant deviation of one baseline in relation to the reference baseline, which results in misdetection of large acceleration, will have the same impact on the final coefficient as misdetection of a small acceleration, slightly exceeding detection thresholds, caused by a small deviation of the baseline.

Keeping this in mind we proposed an alternative synthetic baseline inconsistency coefficient (SI), directly based on the area of A/D episodes. To overcome the issue of lack of matching episode detected on the basis of the other baseline, in the proposed method a virtual matching episodes with zero-area are created during the computation process. Both in our approach and in ADIR coefficient calculation, a set of features describing A/D episodes is used: namely their location in time, duration and area [10].

First, an agreement matrix Q* is created for episodes detected on the basis of L and L' baselines, separately for acceleration and deceleration episodes. The rows of the matrix correspond to the N episodes detected on the basis of the L baseline, and the columns correspond to the N' episodes from the L' baseline. The elements of the matrix take binary values. If the q_{km} element equals one, it means that the k-th episode detected using L baseline has location consistent with the location of m-th episode detected using L' baseline (matching episodes). If the q_{km} equals zero the locations of episodes are not matching. Two episodes are considered as matching each other if they overlap in time with at least 5 seconds.

$$[Q]_{N,N'} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} & \cdots & q_{1N'} \\ q_{21} & q_{22} & \cdots & \cdots & \cdots & q_{2N'} \\ \vdots & \vdots & \ddots & & & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{km} & \cdots & q_{kN'} \\ \vdots & \vdots & & & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{Nm} & \cdots & q_{NN'} \end{bmatrix}$$

where: $q_{km} =$
1 - if location of S_k matches S'_m
0 - if location of S_k does not match S'_m

(3)

The idea of an alternative (synthetic) baseline inconsistency coefficient is based on assumption that a given baseline can be described by a vector, whose elements correspond to the successive A(D) detected on the basis of this baseline. The elements of this vector are equal to the area of particular episodes (S_i). The other features describing the episodes are only used to create the agreement matrix (location and duration of the episode). When comparing two baselines it is quite common that an acceleration or deceleration detected using L baseline has no matching episode detected using L' baseline, or has more than one matching episode. Therefore, the V* and V*' vectors, representing events detected using L and L' baselines, have usually different lengths and comprise only a certain number of matching episodes.

The vectors determined independently using the L and L' baselines must be then extended, so that each k-th element of vector V is matching k-th element of vector V'. Thus, for each acceleration which has no matching acceleration in the other vector (row or column of Q* matrix with only zero elements) a matching acceleration of zero-area is created. On the other hand, if a given acceleration has N matching accelerations in the other vector (row or column of Q* matrix with N ones), this acceleration must be duplicated N-1 times, so that each episode has only one matching episode. This duplicating approach may seem quite pessimistic regarding the value of inconsistency coefficient, since a large acceleration detected using L baseline will be separately compared with all matching smaller accelerations detected using L'. However, we have to keep in mind that in our approach there is no separate location inconsistency component, which exists in the previous approach (AIL). The successive steps in determination of matching episodes are presented in Fig.1. It should be noted that the agreement matrix for the resulting vectors V and V' is a unit matrix.

The resulting V and V' vectors are used as a basis for calculating SI inconsistency coefficient.

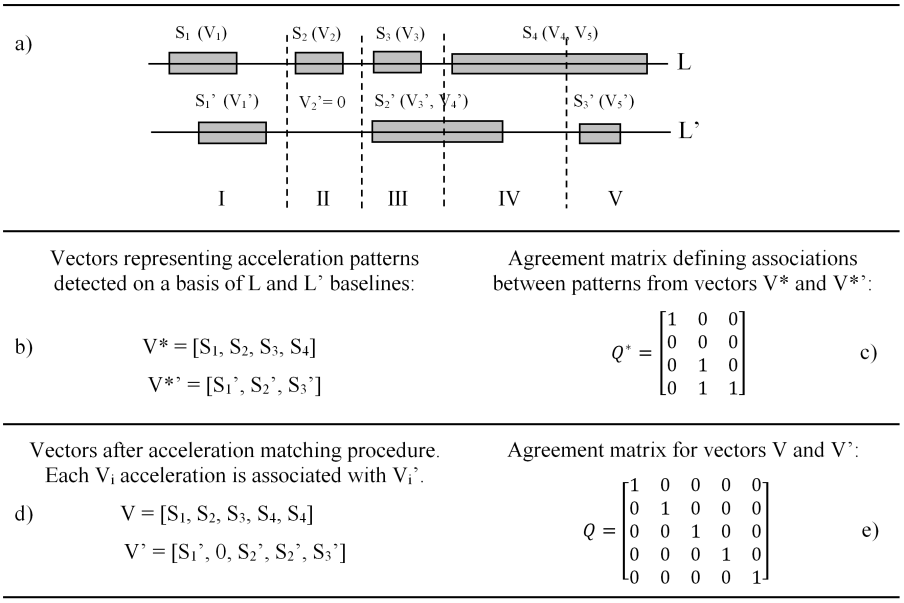

Fig. 1. The algorithm for creating the matched acceleration vectors. In the example below, there is one acceleration based on the L baseline with no matching acceleration based on L'. What is more, in each baseline there is one acceleration matching two accelerations in the other baseline (a). For both baselines initial vectors are created, consisting of acceleration pattern areas (b), and the agreement matrix is formed, where accelerations based on L baseline correspond to rows and accelerations based on L' correspond to columns (c). Matching accelerations are marked with '1' value. Finally vectors are extended so that each k-th element of V exclusively corresponds to the k-th element of V' (d). The agreement matrix for extended vectors is always a unit matrix (e). V_2' is the zero-area acceleration.

The proposed formula was chosen in such way that its value equals zero if the compared vectors are identical and equals one if there are no matching episodes. As a measure of similarity we applied a function of distance between two feature vectors (in our case consisting of A/D areas) in form of the Euclidean distance:

$$d(V, V') = \sqrt{\sum_{i=1}^{D} (V_i - V_i')^2} \tag{4}$$

where V and V' are two vectors of dimension D.

When using the distance function, it is assumed that the smaller the distance between two vectors, the more similar the vectors are. Another important feature of the distance measure is that the attributes of the highest value dominate over the other, smaller ones. It means that the bigger the difference between areas of matching episodes (or the bigger the episode with no matching episodes) the higher is the distance value.

To meet the requirement of coefficient's value between 0 and 1, the resulting distance must be normalized. There are many solutions to normalize the distance, but we had to choose a method providing inconsistency of 1 if there are no matching episodes in both vectors. This requirement is fulfilled by a following equation (for acceleration episodes):

$$ASI = \frac{\sqrt{\sum_{i=1}^{D} (V_i - V_i')^2}}{\sqrt{\sum_{i=1}^{D} max(V_i, V_i')^2}} \tag{5}$$

The synthetic inconsistency coefficient for decelerations (DSI) is calculated in the same way. To calculate the cumulative inconsistency coefficient SI we used the same convention as for the ADIR coefficient:

$$SI = (ASI + 2 \cdot DSI)/3 \tag{6}$$

In this work we have selected 10 different baseline estimation algorithms to illustrate the evaluation and comparison procedure: by Dalton (A) [5], by Dawes (D) [6], by van Alphen (G) [14], by Jezewski (K) [8], by Mantel (M) [11], by Bernardes (P) [3], by Arduini (R) [2], by Searle (S) [13], by Aeyels (Y) [1] and by Wrobel (N) [15]. According to the detailed descriptions provided in the literature, we have created appropriate functions for baseline estimation in the LabView environment.

The research material consists of 41 signals, collected between 31st and 41st week of gestation as a part of routine clinical examinations, using a computer-aided fetal monitoring system. The length of the records varied from 31 to 60 min. The five clinical experts have drawn their baselines on the printouts of FHR recordings, obtained from the system archive. The printout was scanned and dedicated software was created to convert the baseline into digital form. The recordings were then assigned to 10 different groups, each of them related to a typical pattern occurring in the FHR signal. It turned out that the biggest group (comprising the most common patterns) consisted of seven recordings, whereas the smallest – only two recordings.

Appropriate program for quantifying the baseline inconsistency, using both described coefficients, was created in the LabView environment. The program automatically determines the acceleration and deceleration episodes (strictly according to FIGO guidelines) both for the experts' baselines, as well as for the baseline being assessed. On the basis of these functions a framework was created for assessment of a baseline provided by an algorithm in relation to a set of experts' baselines. The individual results obtained with regard to the experts' interpretations were averaged and only the final value was provided.

3. RESULTS AND DISCUSSION

The results were obtained for 41 recordings and 10 different algorithms for the baseline estimation. The inconsistency values obtained for particular recordings were averaged within the established groups to facilitate presentation and interpretation of the data. It also reduced the influence of unequal group sizes.

In the first stage of investigation we paid attention to the signal groups, to find out which ones are the most difficult for automated methods of baseline estimation. In the Figure 2 the range of values obtained by all algorithms is presented (as min-max values) together with the median value of inconsistency for all ten algorithms. The median value is interpreted as "an average algorithm" and can be the reference in assessing whether the method performs "better than average". The groups in Figure 2 are presented in the ascending order of the median value of SI coefficient.

It can be noted that the least problematic groups (based on both coefficients) are the groups I and VII, as the median value for all algorithms is the lowest. In turn, the most difficult signals definitely belong to the group III, for which the median inconsistency is by far the highest. The median value of ADIR in this group was $\geq 53.5\%$. At the same time, one of the algorithms (K) provided baselines with average inconsistency in this group equal to only 26%, which proves that even the most difficult group can be effectively interpreted using the automated methods. The smallest dispersion of inconsistency values for ADIR was noted in group X, whereas for SI coefficient – in group VI. The highest dispersion was observed in group III, both for ADIR and SI coefficients.

In the next stage of analysis the relationship of two different inconsistency measures presented in this paper was evaluated (Fig.3). A strong linear correlation between both measures was observed, which is confirmed by high value of Pearson's correlation coefficient r = 0.904 ($p < 0.01$). It is worth noting that the new SI coefficient takes values in the whole range

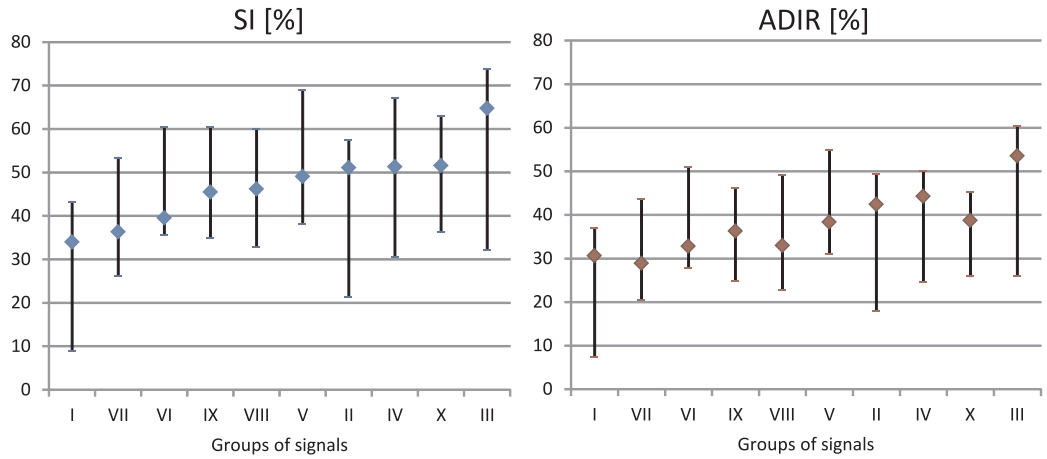

Fig. 2. Comparison of signal groups in terms of difficulty of baseline determination. The plots depict the range of inconsistency coefficients obtained for all analyzed algorithms in a given group. The results are expressed as minimum-median-maximum values of SI and ADIR coefficients.

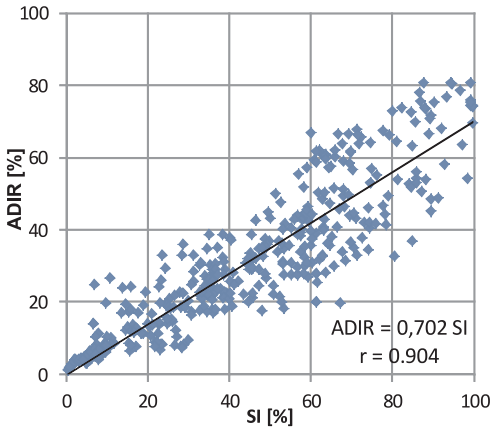

Fig. 3. The relationship between ADIR and SI inconsistency coefficients. Pairs of resulting coefficients for all 41 recordings and 10 algorithms (410 pairs) are represented as dots in the figure.

between 0 and 1 ($0 \div 100\%$). The values close to 100% inconsistency can be easily interpreted as nearly total lack of matching episodes. At the same time, the inconsistency evaluated using ADIR only slightly exceeded 80% in the worst cases. This observation illustrates the advantage of the synthetic inconsistency coefficient, which is easier to interpret and the provided range of values is more intuitive.

The detailed results describing the effectiveness of the analyzed algorithms within the established groups of recordings are presented in Table 1. It can be noted that the best results (mean inconsistency according to both coefficients) were obtained for the algorithm K, which also proved to be the most accurate in groups II, III and V. It is also important that the differences between the top three algorithms (K, R, M) are very small. On the other hand, the algorithm P provided baselines with the highest inconsistency in 6 out of 10 groups.

Interesting results were also provided by the S algorithm, which turned out to be the most accurate in groups VI, VII and VIII, whereas in the overall ranking it took a distant fifth place. It means that its accuracy in the other groups was much lower than the accuracy of the leading algorithms (K, R, M). This observation confirms the validity of the proposed approach, based on separate groups of characteristic patterns in the FHR signals. It also brings the conclusion, that in the comparison the most versatile algorithms should be preferred.

Table 1. Evaluation of effectiveness of different algorithms within the established signal groups. The values express mean inconsistency coefficients calculated for all recordings belonging to a given group. The best and the worst results within the group are marked in bold (the best are underlined). Algorithms are presented in the order of increasing mean inconsistency.

	SI [%]									
Group	K	R	M	D	S	N	G	Y	A	P
I	12.91	12.78	**8.83**	25.14	32.44	35.61	35.91	37.97	**43.22**	36.77
II	**21.31**	27.98	24.55	32.33	48.34	53.88	54.84	**57.36**	55.44	54.43
III	**32.22**	38.48	35.53	42.38	62.80	66.77	70.41	67.15	69.42	**73.86**
IV	32.04	38.26	35.63	**30.65**	49.96	52.65	57.34	56.80	**67.08**	64.78
V	**38.04**	44.68	45.73	39.78	43.79	52.48	55.75	57.16	59.74	**69.00**
VI	38.49	36.63	37.71	40.58	**35.51**	36.92	44.85	44.62	56.58	**60.46**
VII	36.31	36.27	35.66	39.51	**26.10**	28.86	36.36	39.90	47.81	**53.46**
VIII	45.41	43.41	48.10	54.08	**32.76**	36.82	46.97	45.21	50.60	**59.89**
IX	37.37	**34.89**	39.70	46.78	44.23	43.06	50.55	49.61	**60.47**	56.13
X	36.46	**36.21**	43.96	48.18	49.87	53.34	56.24	53.92	61.02	**62.92**
Mean	33.06	34.96	35.54	39.94	42.58	46.04	50.92	50.97	57.14	59.17

Finally, to conclude the evaluation procedure, a brief criterion should be chosen to decide on a final assessment of a given algorithm: a parameter enabling to objectively state that a given algorithm is more effective than another. Analysis of the obtained results suggests that the mean value from group inconsistency coefficients best reflects the quality of the algorithm. Additionally, the highest inconsistency from among the groups can also be valuable information, as it describes the algorithm more comprehensively. It allows you to determine if the algorithm performs well (is effective) in case of all kinds of cardiotocographic records.

4. CONCLUSIONS

The proposed synthetic inconsistency coefficient SI follows the ADIR coefficient in reflecting three components of inconsistency (number, location, area). However, in case of SI these three aspects are not assessed separately, but rather naturally included in the established formula. A very important advantage of the synthetic coefficient over ADIR index is that the area of each A/D episode (with or without matching episode) is taken into account when calculating its value. Paying attention to the area of all detected episodes is crucial as omitting an existing (or false detection of non-existent) large deceleration may have serious consequences for fetal state assessment. Therefore, the SI coefficient seems to be a better description of differences between baselines being compared. At the same time it is much simpler and much easier to interpret. The proposed methodology of algorithm evaluation, based on the inconsistency coefficients, does not provide a direct measure of a given algorithm's quality. The quantitative results can only be interpreted by comparison with the results obtained for the competing algorithms. Nevertheless, the proposed methodology seems to be a valuable tool supporting future research works in the baseline estimation area.

ACKNOWLEDGEMENT

This work was in part financed by the Polish National Centre for Research and Development under the Strategic Programme STRATEGMED.

BIBLIOGRAPHY

[1] AEYELS B., VAN DER PERRE G., PELLO L. On line processing of perinatal fetal heart rate and intra uterine pressure. Proc. of the 14th IEEE/EMBS Conf., 1992. Paris, pp. 2728–2729.

[2] ARDUINI D., RIZZO G., ROMANINI C. Computerized analysis of fetal heart rate. J. Perin. Med., 1994, Vol. 22. pp. 22–27. Article ID 485684.
[3] BERNARDES J., AYRES DE CAMPOS D., MOURA C. Computer recognition of fetal heart rate patterns by the porto system. Proc. 2nd World Congress of Perinatal Medicine, 1993. Rome, pp. 559–564.
[4] CHOURASIA V. S., TIWARI A. K. A review and comparative analysis of recent advancements in fetal monitoring techniques. Critical Reviews in Biomedical Engineering, 2008, Vol. 36. pp. 335–373.
[5] DALTON K. J., DAWSON A. J. A computer method of calculating baseline in fetal heart rate recordings. Int. J. Biomed. Comput., 1984, Vol. 15. pp. 311–317.
[6] DAWES G. S., MOULDEN M., REDMAN C. W. G. System 8000: Computerized antenatal fhr analysis. J. Perinat. Med., 1991, Vol. 19. pp. 47–51.
[7] JEZEWSKI J., HOROBA K., MATONIA A. Baseline and acceleration episodes – clinically significant nonstationarities in FHR signal: Part II. indirect comparison. Computer Recognition Systems, 2005, Vol. 30. Springer Verlag, pp. 535–542.
[8] JEZEWSKI J., WROBEL J. Global baseline determination in fetal heart rate records with the identification of local prominent rates. Proc. 1st Int. Conf. on Medical Physics and Biomedical Engineering, 1994, Vol. 2. Nicosia, pp. 365–369.
[9] JEZEWSKI J., WROBEL J., KUPKA T. Baseline and acceleration episodes – clinically significant nonstationarities in FHR signal: Part I. coefficients of inconsistency. Computer Recognition Systems, 2005, Vol. 30. Springer Verlag, pp. 527–534.
[10] LABAJ P., JEZEWSKI M., MATONIA A. New approach to quantitative description of deceleration of fetal heart rate for the patterns classification. Proc. 29th Annual International Conference of the IEEE EMBS, 2007. Lyon, pp. 3156–3159.
[11] MANTEL R., VERVERS I., COLENBRANDER G. J., VAN GEIJN H. P. Automated antepartum baseline FHR determination and detection of accelerations and decelerations. A critical appraisal of fetal surveillance, 1994. Elsevier Science B.V., pp. 333–348.
[12] ROOTH G. Guidelines for the use of fetal monitoring. Int. J. Gynaecol. Obstet., 1987, Vol. 25. pp. 159–167.
[13] SEARLE J. R., DEVOE L. D., PHILLIPS M. C. Computerized analysis of resting fetal heart rate tracings. Obstet. Gynecol., 1988, Vol. 71. pp. 407–411.
[14] VAN ALPHEN M., WAGENVOORT A. M., VAN GEIJN H. P. Quantitative intrapartum CTG analysis. Current progress in perinatal medicine, 1994. Parthenon Publishing Group, pp. 805–811.
[15] WROBEL J., HOROBA K., PANDER T. Improving the fetal heart rate signal interpretation by application of myriad filtering. Biocybern. Biomed. Eng., 2013, Vol. 33. pp. 211–221.