PREDICTION OF MISSING DATA IN CARDIOTOCOGRAMS USING THE EXPECTATION MAXIMIZATION ALGORITHM

G. Nokas
Department of Electrical and Computer Engineering, University of Patras, 26500 Rion, Patras

A. Koutras, I. Christoyanni, G. Georgoulas, Ch. Stylios[1] and P. Groumpos
Department of Electrical and Computer Engineering, University of Patras
[1]Computer Science Department, University of Ioannina

1. SUMMARY

The Cardiotocogram (CTG), the continuous recording of fetal heart rate and maternal contractions during labor, has been introduced in the clinical routine eliminating fetus mortality and warning obstetricians about the health status of fetus and the occurrence of problems during antenatal as well as during labour. The instantaneous FHR (beats/min) can either be obtained by Doppler ultrasound or directly from the fetal electrocardiogram via scalp electrodes. The uterine activity is measured using an external tocodynamometer or with the use of an intra-uterine pressure catheter (mmHg) [1]. Today, almost all cardiotocographs in use are operating using Doppler ultrasound and external tocodynamometer (CTG is acquired using an Doppler ultrasound that is applied externally on the pregnant's womb and it records the movements of the fetus). This method of CTG acquisition often leads to a discontinuous trace, arising the problem of missing data. In this paper we propose a technique, based on the maximum likelihood framework, for the data forecasting between these discontinuities of the CTG signal. More specific, the probability density function (pdf) of the CTG signal is assumed to be mixture of Gaussian, and the parameters are estimating on-line using the Expectation-Maximization algorithm. Following the statistics of the signal's source, we predict the missing samples under the constraint that the distance between adjacent values is relative short in the time domain. The evaluation of our method shows improvement of the prediction accuracy when compared with a linear interpolation method.

2. INTRODUCTION

In the last 25 years, since CTG has been introduced, it has been a reliable indicator of fetal condition. CTG signal is consisted of two independent signals: the Fetal Heart Rate signal (FHR), whose values can highly oscillate in time showing a strongly nonlinear behavior and the uterine contractions. The variation of FHR values contains valuable information that must be extracted and analyzed. For this reason the effort of many researchers is focused on the development of computerized systems in order to evaluate CTG data [2][3][4]. Usually a recording of CTG trace for 20 minutes is enough to extract the basic features: baseline of FHR, variability, accelerations, decelerations and contractions that are classified according to several clinical terms. Nowadays researchers are developing expert systems to extract these features and to assist clinical decision-making [5][6][7].

This paper deals with the problem concerning the acquisition of CTG data. During the acquisition of the CTG using Doppler ultrasound the movements of the fetus often lead to a discontinuous trace, arising the problem of missing data. In figure 4, the case of a FHR with missing data is depicted. Although the missing data don't provoke problems to simple eye inspection, it leads to wrong results especially when further digital processing is applied to the FHR signal. Missing data mislead the algorithm for feature extraction that makes necessary a preprocessing of the signal for prediction of the missing signal values.

In this paper a technique, based on the maximum likelihood framework, is proposed for the prediction and replacement of the missing FHR signal data. More specifically, we model the FHR statistics using a mixture of Gaussian Power Density Function (pdf), and we estimate the parameters (c,m,σ equation 1) on-line. For the estimation of the parameters the Expectation Maximization algorithm is used. Mixture models have been utilized to design radial-basis-function neural networks and can efficiently model nonlinear systems, so they are adequate for the modeling of FHR, which is strongly nonlinear as mentioned before. Having the statistical characteristics of the signal's source we can estimate the missing data, generating new samples and filling the gaps of the missing signal values. The only constraint, regarding the memory of our method is that the distance between adjacent values must be relative short in the time domain.

The proposed prediction method is compared to two other methods for the filling of the missing beats, using the same experimental data: a) the uniform pdf where all the possible values were equiprobable and b) linear interpolation. The evaluation of the proposedmethod shows improvement of the prediction accuracy when compared with both the other two methods.

The structure of this paper is: In section 3 the data model and the estimation of the pdf parameters with EM algorithm are described. In section 4, the EM algorithm is implemented to predict missing data and the results are compared to the other methods. Section 5 concludes the paper.

3. ESTIMATION OF THE DATA MODEL

Suppose we have a set of N observations from a cardiotocogram $X=\{x_1,x_2,\ldots x_N\}$. We model the underlying distribution of these observations with a mixture of Gaussian pdf. The probability of the observation x thus becomes:

$$\mathrm{p}(x \mid \Theta) = \sum_{i=1}^{M} \mathrm{c}_i . \mathrm{N}(x, m_i, \sigma_i) \qquad (1)$$

Where N() is the Normal distribution:

$$\mathrm{N}(x \mid m, \sigma) = \frac{1}{\sigma \ \sqrt{2.\pi}} \exp\left\{ -\frac{1}{2} \frac{(x-\mathrm{m})^2}{\sigma^2} \right\} \qquad (2)$$

An efficient way to estimate the parameter set $\Theta = \{c_i,\sigma_i,m_i,i=1..M\}$, where M is the number of mixtures, is the Expectation Maximization (EM) algorithm [8], which is shortly described in the next paragraph.

3.1 EM algorithm

We assume that we have a set of independent and identically distributed (i.i.d) data samples with distribution p. The resulting density for the samples is:

$$p(X\,|\,\Theta) = \prod_{i=1}^{N} p(x_i\,|\,\Theta) = L(\Theta\,|\,X). \qquad (3)$$

The function $L(\Theta|X)$ is called the likelihood of the parameters given the data and is thought of as a function of the parameters Θ where the data X is fixed. In the maximum likelihood framework the goal is to find the Θ that maximizes L. That is we want to find Θ^* where

$$\Theta^* = \arg\max_{\Theta} L(\Theta\,|\,X) \qquad (4)$$

In the case of mixture of Gaussian pdf the solution of the above function, by setting the derivative equal to zero, leads to nonlinear, coupled equations. So we must resort to more elaborate techniques. Such a technique is the EM algorithm, which can solve nonlinear equations reaching a local maximum.

The EM algorithm leads to an iterative solution with the following update equations [9] for the parameter set $\Theta = \{c_i,\sigma_i,m_i,i=1..M\}$:

$$m_i^{(t+1)} = \frac{\sum_{k=1}^{N} \pi^{(t)}(\theta_i\,|\,x_k)x_k}{\sum_{k=1}^{N} \pi^{(t)}(\theta_i\,|\,x_k)} \qquad (5)$$

$$\sigma_i^{(t+1)} = \left\{ \frac{\sum_{k=1}^{N} \pi^{(t)}(\theta_i\,|\,x_k)(x_k - m_i)^{1/2}}{\sum_{k=1}^{N} \pi^{(t)}(\theta_i\,|\,x_k)} \right\}^2 \qquad (6)$$

$$c_i^{(t+1)} = \frac{1}{N}\sum_{k=1}^{N} \pi^{(t)}(\theta_i\,|\,x_k) \qquad (7)$$

Where,

$$\pi^{(t)}(\theta_i \mid x) = \frac{c_i N(x \mid \theta_i)}{\sum_{i=1}^{M} c_i N(x \mid \theta_i)}$$, Are the posterior probabilities.

Starting with initial values from the data range, the algorithm is guaranteed to converge after little iteration to a local maximum.

4. EXPERIMENTAL SETUP

The proposed method has been evaluated in several sets of data. First of all we will evaluate the method using a set of data where we know exactly the missing data. We take 2200 samples from a FHR signal without discontinuities, and we randomly select an interval of 200 samples that is been set to zero representing a discontinuity (figure 2). We implement the EM algorithm in the signal depicted on figure 2 and we estimate the parameters of the three Gaussian mixtures that are depicted on figure 1. The algorithm converged in 8 iterations using as stopping criterion of the algorithm the negligible change of parameters.

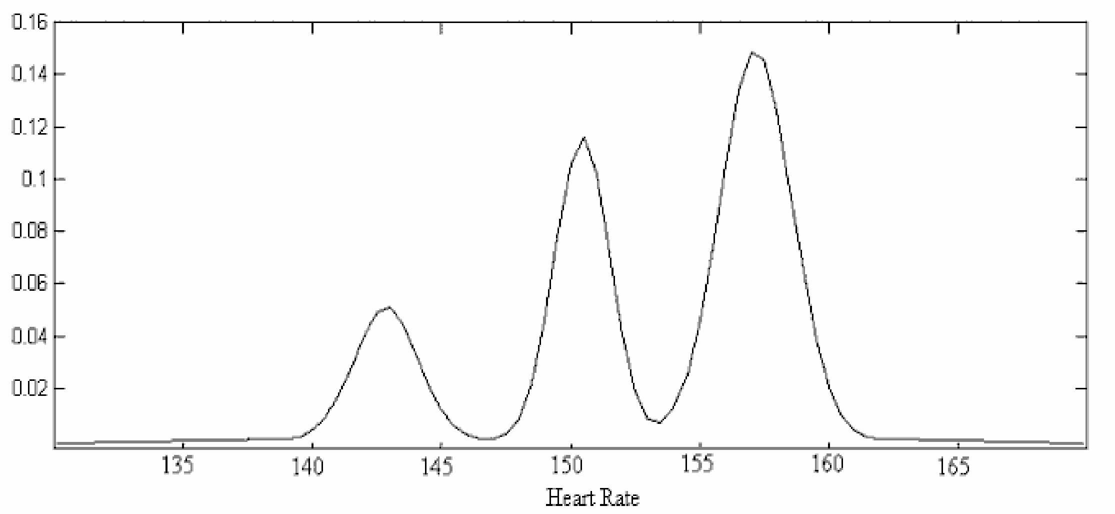

Figure1. Pdf of the available data.

Next, we generate new samples, following the statistics of the estimated pdf filling the gaps of the missing data. During the prediction we set the constraint that the absolute difference between adjacent samples must not be greater than a given threshold (3 in our case).

Figure 3 illustrates the original signal with a solid line, the estimated signal with a dot line and the estimated signal using linear interpolation with a dash-dot line. From the figure is concluded that the prediction signal with the proposed method has many similarities with the real one. The mean distance between the estimated and the real samples was found to be 7,023 for the interpolation method and only 4,850 for the EM algorithm. Due to the stochastic manner of our approach the same experiment was repeated 20 times. The average of the mean distance, between the predicted and the real samples has been computed. We also repeated the experiment using as pdf the uniform distribution. The results are illustrated in table1. We notice that the Gaussian mixtures based predictor, gives the better results. Although linear interpolation results are not significantly worse, the method has the weakness to extract a signal with dissimilar characteristics to the original. Finally, the uniform distribution gives the worst results as expected.

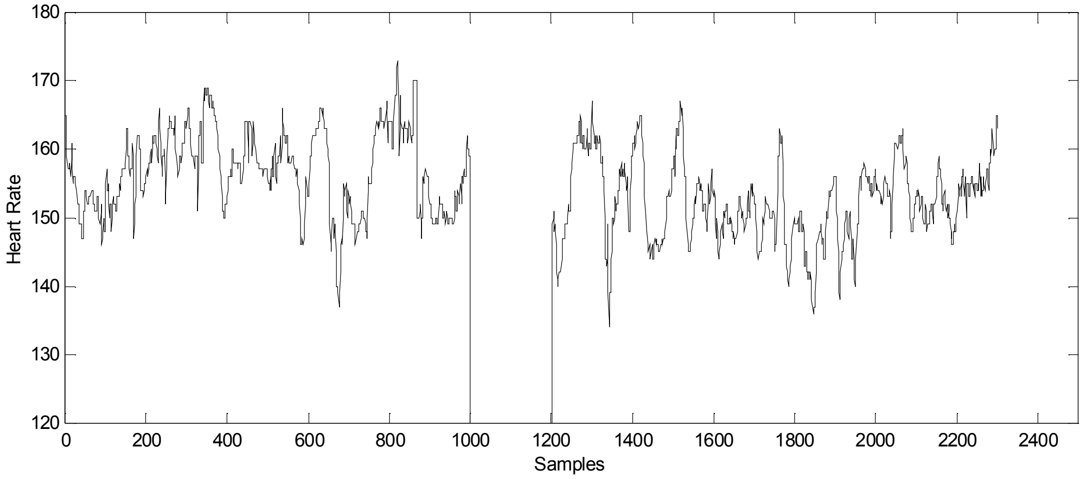

Figure2. Test signal with an artificial discontinuity.

Prediction method	Average Mean Distance
Linear Interpolation	7,023
Uniform Distribution	9,371
Mixture of Gaussian	6,772

Table 1. Average mean distance between predicted and real data in 20 repetitions.

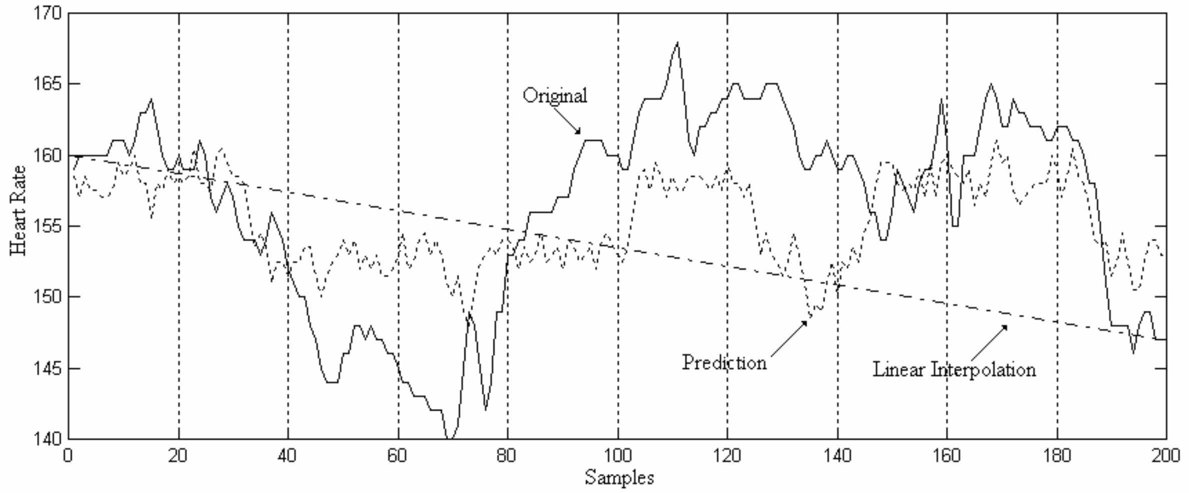

Figure 3. Illustration of original and predicted signal.

After having implemented and tested the proposed algorithm in a known missing set of data we will test it using an FHR signal that has missing data. The FHR signal that has been received by the carditocograph is depicted on figure 4 it has been sampled at 4 Hz and has many discontinuities. During discontinuities the signal is not zero because of the existing noise and measurement from other signals.

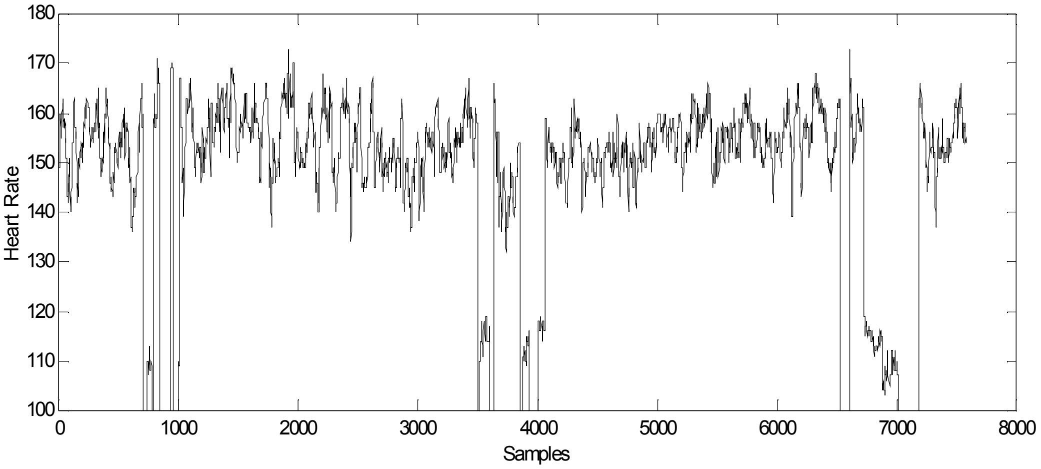

Figure 4. FHR with missing data

The preprocessing of the signal consists of two steps. First, we localize the discontinuities and we reject all the not acceptable values, using a heuristic threshold of 25 bpm. Then we set the rejected samples to zero as it is illustrated on figure 5.

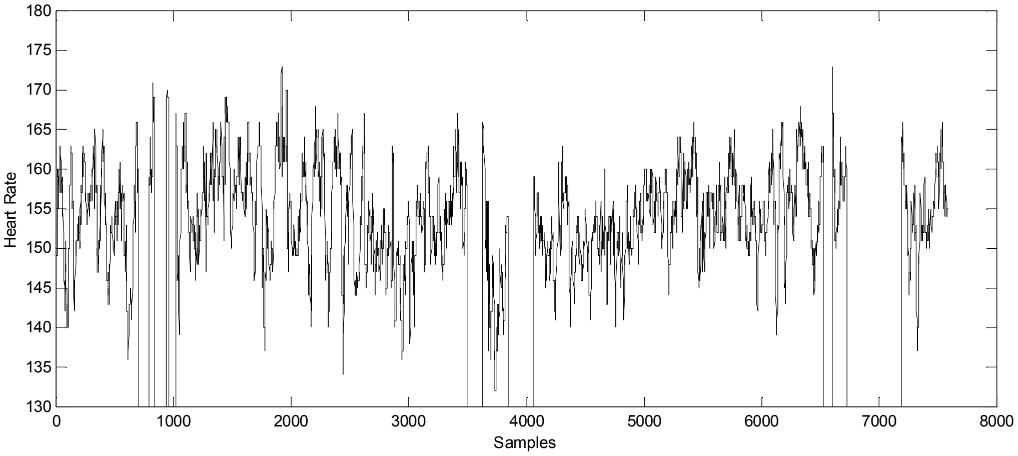

Figure 5. The FHR signal with discontinuities set to zero

Then we implement the EM algorithm to estimate the missing values of data and the restored signal, after Gaussian based prediction is illustrated on figure 6.

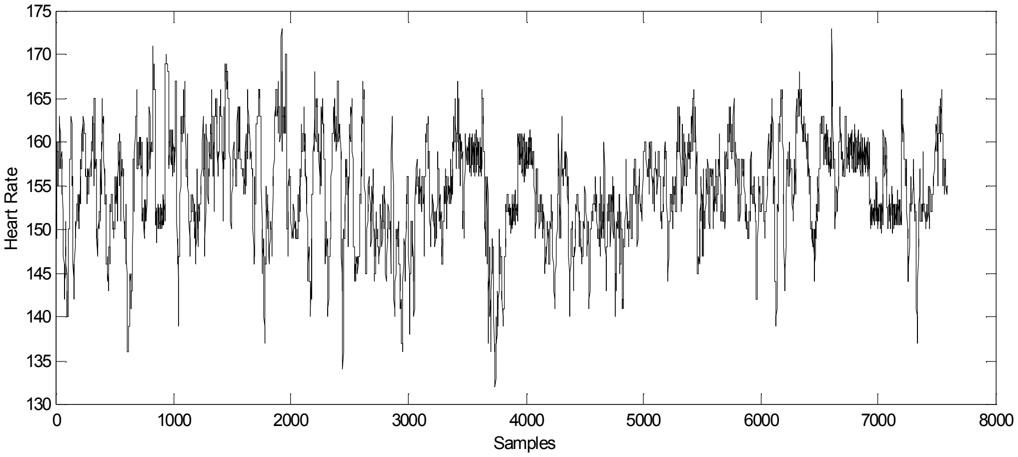

Figure 6. Reconstruction of the distorted signal after Gaussian based prediction.

5. CONCLUSION

In this paper a method for predicting the missing data in cardiotocograms has been proposed. The prediction is based on the characteristics of the existing data, from which the pdf (Mixture of Gaussian) is estimated. This estimation method can be implemented on-line using the EM algorithm.

The prediction signal is generated, following the estimated statistics under the constraint, that the distance between adjacent values is relative short in the time domain. Although our approach is memory-less, the results are promising and the method outperforms the linear interpolation. The algorithm can be implemented in real time and its convergence is guaranteed.

6. ACKNOWLEDGMENT

This research was supported by the Greek GSRT and European Social Fund under PENED'99 project 99ED514.

7. REFERENCES

[1] Carter M. C., Present-day performance qualities of cardiotocographs, Br J Obstet Gynaecol 100,9,10-14 (1993)

[2] J Bernardes, C Moura, JP Marques-de-Sá, HP van Geijn, L Pereira-Leite. “The Porto System.” In: HP van Geijn and FJA Copray (eds): A Critical Appraisal of Fetal Surveillance. Elsevier Science, Amsterdam 1994;315-24

[3] D Arduini, G Rizzo,G Pianna, A Bonalumi,P Brambilla, C Romanini, “Computerized Analysis of Fetal Heart Rate: I. Description of the Sustem (2CTG).” J Matern Fetal Invest 1993;3:159-163

[4] R Mantel,I Ververs, GJ Colenbrander and HP van Geijn, “Automated antepartum baseline FHR determination and detection of accelerations and decelerations,” In: HP Van Geijn and FJA Copray (eds): A Critical Appraisal of Fetal Surveillance. Elsevier Science, Amsterdam 1994;333-348

[5] Dawes GS, Moulden M, Redman CW, “System 8000: Computerized antenatal fetal heart rate analysis” J Perinat Med 1991;19:47-51

[6] Skinner, J.F.; Garibaldi, J.M.; Curnow, J.; Ifeachor, E.C, "Intelligent fetal heart rate analysis". Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476) , 2000 Page(s): 14-21

[7] Signorini, M.G.; de Angelis, A.; Magenes, G.; Sassi, R.; Arduini, D.; Cerutti, S Classification of fetal pathologies through fuzzy inference systems based on a multiparametric analysis of fetal heart rate Computers in Cardiology 2000 , 2000 Page(s): 435 –438 [8] Dempster A. P., Laird N. M. and Rubin D.B. "Maximum likelihood from incomplete data via the EM algorithm" Discussion on the paper by Dr S.J Haberman, J. Royal Statist. Soc. Ser. B.,39, 1977.

[9] Nasser Kehtarnavaz, Eliji Nakamura:"Generalization of the EM algorithm for mixture density estimation"Pattern Recognition Letters 19 (1998) pp 133-140.