

Comparison of Cardiotocography and Fetal Heart Rate Estimators Based on Non-Invasive Fetal ECG

Rasmus G Sæderup¹, Henrik Zimmermann¹, Dagbjört H Eiríksdóttir², John Hansen², Johannes J Struijk², Samuel Schmidt²

¹ Viewcare A/S, Søborg, Denmark

² Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

Abstract

Non-invasive fetal ECG (NI-FECG) extraction algorithms enable long-term continuous beat-to-beat monitoring of the fetal heart rate (FHR), as opposed to the gold standard in FHR monitoring, cardiotocography (CTG). We investigate how NI-FECG extraction algorithms selected from the CinC 2013 Challenge (CinC13) perform on data with low quality signals and how performance can be evaluated using CTG, when FQRS annotation is not possible.

Four-channel NI-FECG was recorded simultaneously with a CTG trace on 22 pregnant women, gestational age 29-41 weeks. Seven algorithms were tested: The winning CinC13 entry from Varanini et al. and six algorithms from the unofficial top-scoring CinC13 entry by Behar et al. Two accuracy measures were used: 1) The RMSE between the FECG-based FHR and CTG traces; 2) The Pearson correlation coefficient r between the FECG-based FHR and CTG trace and its average over all recordings, \bar{r} .

The algorithms with the lowest RMSE's are Behar's FUSE-SMOOTH, a constant FHR, and Varanini, while the Varanini algorithm delivers the best correlation with the CTG trace ($\bar{r} = 0.73$) with 41% of the recordings having $r > 0.8$, whereas the other algorithms have $\bar{r} \leq 0.59$ and $\leq 29\%$ of the recordings with $r > 0.8$. FHR was estimated accurately in some recordings and poorly in others, believed to be due to large differences in signal quality.

1. Introduction

The development of non-invasive fetal ECG (NI-FECG) extraction algorithms is of great interest in the field of fetal monitoring, as these algorithms enable new and improved diagnostic possibilities in comparison to the current gold standard in fetal heart rate (FHR) monitoring, cardiotocography (CTG). NI-FECG, as opposed to CTG, provides a stable interface for long-term continuous beat-to-beat monitoring, and can even provide morphological analysis of the FECG. [1]

Despite these immediate advantages, developing robust NI-FECG algorithms is of great difficulty - primarily due to the low SNR, the overlap of the maternal and fetal ECGs in both time and frequency [2], and a lack of ground truth annotations.

1.1. Previous Work

In order to promote research in NI-FECG, the Computing in Cardiology 2013 Challenge (CinC13) was given, in which participants had to estimate the FHR from a dataset of 4-channel abdominal ECG mixtures. The winning entry by Varanini *et al.* [3] obtained an accuracy of 99%, although on a subset of the 75-recordings training set. The overall best-performing entry to CinC13, was by Behar *et al.* [4] who implemented a wide range of FECG separation methods, including template subtraction (TS), blind source separation such as principal component analysis (PCA) and independent component analysis (ICA), as well as combinations of these. The FUSE-SMOOTH algorithm which combined all of the above was the best-performing method from [4], with an F_1 score of 96%¹.

In addition, a range of algorithms were tested on artificially generated signals [5] in different physiological scenarios such as fetal movement, uterine contractions, ectopic beats and twin pregnancy, with the SNR ranging from 0-12 dB. In most of these cases, the median F_1 score was very high for all algorithms ($> 99\%$).

At first sight, the high F_1 scores indicate that the algorithms perform very well on real-life signals, and even better on simulated signals. However, these results might not be generalizable to a clinical setting:

In the CinC13 challenge, the FECGs had a signal quality which ensured that manual annotation of the fetal QRS (FQRS) was possible [2]. In clinical practice, manual annotation of the signal is often not possible, as e.g. An-

¹The Behar algorithm performed better than Varanini in CinC13 despite having lower F_1 scores because a modified version of the FUSE algorithm was used, which was designed to maximize the CinC13 scoring system.

dreotti *et al.* [6] found that only 4.1 % of 259 multichannel FECGs had a clearly distinguishable FQRS (see Figure 1), indicating a clinical reality that calls for new methods to test NI-FECG algorithms. Also, the use of simulated data is problematic, if only because of the over-performance of the algorithms on simulated data in comparison with measured data.

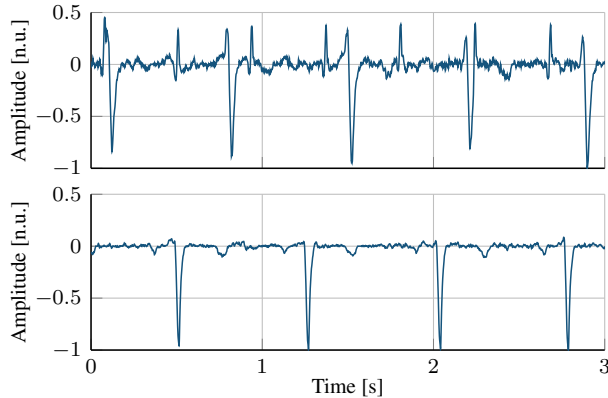


Figure 1. Two three-second excerpts from abdominal ECG recordings with good FECG quality (top) and poor quality (bottom).

These observations give rise to two questions which we answer in this paper:

- How do NI-FECG algorithms perform on clinical data containing many low quality signals?
- How can the performance of NI-FECG algorithms be evaluated when manual FQRS annotation is not possible due to poor signal quality?

2. Methods

We compared the FHR estimated from NI-FECG extraction algorithms to the FHR estimated using the CTG, using two performance measures: The root-mean-squared-error and Pearson’s correlation coefficient. Such a comparison has been done with different performance measures by [7], but to the best of our knowledge, no direct comparison of the CTG and published state-of-the-art NI-FECG algorithms has previously been made.

2.1. Data Acquisition

The data used for this study originates from 34 recordings from 22 pregnant women with singleton pregnancies having body mass indices (BMI) ranging from 20-42 kg/m², and gestational ages (GA) ranging from 29 to 41 weeks with a median GA of 36 weeks.

In each recording, four ECG traces were obtained by placing six Ag/AgCl electrodes on the abdomen of the subject. The signals were recorded with 24 bit resolution at a

sample rate of 8 kHz and have a median duration of 290 s (range 56 to 1830 s).

Simultaneous recordings with Doppler CTG were carried out and digitally saved, which resulted in a FHR sampled at a rate of 4 Hz, and smoothed by a built-in moving average window with a size of 2 s [7].

2.2. FHR estimation

Prior to applying the FECG extraction algorithms, the signals were downsampled to 1 kHz as this is the sample rate used in the CinC13 challenge [2]. The FECG extraction algorithms used in this study are listed below and consists of the algorithm by [3] and the best-performing algorithms from [4], and are all available at archive.physionet.org/challenge/2013/sources

- Varanini [3]
- FUSE [4]
- FUSE_{SMOOTH} [4]
- TS_{CERUTTI} [4]
- TS-PCA [4]
- TS-PCA-ICA [4]
- Constant FHR (143 BPM) [4]

The output of these algorithms are the locations of the R-peaks in the FECG, from which we calculate the (instantaneous) FHR in beats per minute (BPM) at the i 'th R-peak location and denote it $FHR_{NI-FECG}[i]$.

2.3. Post-processing

After performing FHR estimation, four post-processing steps were used to enable the comparison of the FHRs from CTG and NI-FECG:

1. The FHR segments which contained valid CTG data were selected, as a visual inspection of the CTG data showed that a valid FHR is not present at all times. CTG outliers were therefore removed if they did not satisfy the following criteria:
 - $100 < FHR_{CTG} < 190$
(to avoid confusion with the maternal HR)
 - $\Delta FHR_{CTG} < 10$
(an instantaneous FHR change larger than 10 BPM is not realistic on a smoothed FHR)
2. Since the NI-FECG algorithms are making beat-by-beat FHR estimates, the outputs were rather noisy, as opposed to the smoothed CTG trace. As visual inspection showed that much of the noise came from outliers, a median filter with order $N=11$ was applied to smoothen the FHR estimates, where $N=9,11,13,15$ had been tested but $N=11$ gave the overall highest performance across algorithms in terms of the performance measures in Section 2.4.
3. The FHR estimates from the NI-FECG were interpolated to the CTG sample points instead of the R-peak locations, using linear interpolation, as the per-

formance measures require the FHR traces to be sampled simultaneously.

- As the CTG and ECG are recorded on separate devices without synchronized clocks, their sample rates might be out of sync, which shows as a drift between the FHR traces on long recordings (> 15 min. in our case based on visual inspection of the FHR traces). To correct this drift, Dynamic Time Warping was used to align the time-indices between the FHR traces.

An example of a post-processed FHR trace from both CTG and NI-FECG algorithms can be seen in Figure 2.

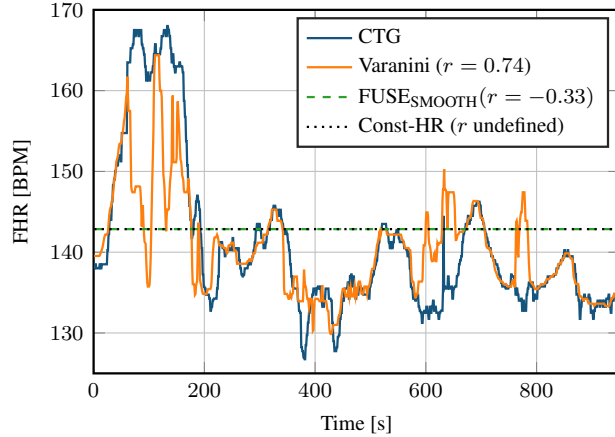


Figure 2. Fetal heart rate (FHR) estimated from CTG and NI-FECG extraction algorithms from a single recording.

2.4. Performance Measures

Two performance measures were used to compare the CTG to each NI-FECG extraction algorithm: The root-mean-squared-error and Pearson's correlation coefficient, which were chosen based on their two different ways of describing the performance. The RMSE measures the overall ability to estimate the FHR as it captures both the bias and variance, while the correlation coefficient describes the algorithm's ability to correctly capture the variability of the FHR trace i.e. differences in their overall shape.

The mean squared error (MSE) between the FHR traces from the CTG and each NI-FECG algorithm was estimated for each recording m , from which the square root of the MSE averaged across all M recordings, RaMSE, was found as:

$$\text{MSE}_m = \frac{1}{N_m} \sum_{n=0}^{N_m-1} (\text{FHR}_{\text{CTG},m}[n] - \text{FHR}_{\text{NI-FECG},m}[n])^2 \quad (1)$$

$$\text{RaMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^M \text{MSE}_m} \quad , \quad (2)$$

with N_m being the number of valid CTG samples in the m 'th recording.

The Pearson correlation coefficient between the CTG and each NI-FECG algorithm was calculated for the m 'th recording as:

$$r_m = \frac{1}{N_m - 1} \sum_{n=0}^{N_m-1} \left(\frac{\text{FHR}_{\text{CTG},m}[n] - \overline{\text{FHR}}_{\text{CTG},m}}{s_{\text{CTG},m}} \right) \cdot \left(\frac{\text{FHR}_{\text{NI-FECG},m}[n] - \overline{\text{FHR}}_{\text{NI-FECG},m}}{s_{\text{NI-FECG},m}} \right), \quad (3)$$

where $\overline{\text{FHR}}_{\{\cdot\}}$ and $s_{\{\cdot\}}$ denotes FHR sample mean and sample SD over all N_m samples, respectively.

We also set a threshold of $r > 0.8$ for the correlation between CTG and NI-FECG FHRs to be acceptably strong for the FHR estimate to be useful.

Using the Fisher z -transformation to correct for the non-normality of the sampling distribution of r_m , the average correlation across recordings is found as:

$$\bar{r} = \tanh \left(\frac{1}{M} \sum_{m=1}^M \text{arctanh}(r_m) \right). \quad (4)$$

3. Results

In Table 1, RaMSE and \bar{r} are given for each of the 7 implemented NI-FECG algorithms, along with the smallest, biggest and median correlation coefficients r . Three algorithms have significantly lower RaMSE than the other, namely Varanini, FUSE_{SMOOTH} and CONST-HR. See also Figure 3, where a box-plot of the RaMSE for each of the tested algorithms is shown.

For the averaged correlation coefficients \bar{r} , Varanini performs moderately better ($\bar{r} = 0.73$) than the others, which all have similar performances of $0.5 \leq \bar{r} \leq 0.6$, except from the const-HR which, being a constant, has an undefined correlation coefficient. Varanini makes useful ($r > 0.8$) estimates in 41 % of the recordings, while it is less than 30 % for the other algorithms, see Table 1. In 58 % of the signals, all 7 algorithms had $r < 0.8$.

4. Discussion

The Varanini, CONST-HR and FUSE_{SMOOTH} algorithms performed very well in terms of RaMSE. The good performance of CONST-HR and FUSE_{SMOOTH} is due to these algorithms being biased towards the mean fetal heart rate of around 143 BPM [4], although the high performance of a constant estimate is rather surprising.

CONST-HR performed poorly in terms of correlation, leading to the Varanini algorithm clearly being the best performing algorithm when considering both performance measures.

Table 1. The RaMSE and correlation coefficient performance measures for each of the 7 NI-FECG algorithms.

Algorithm	RaMSE [BPM]	\bar{r} (95% CI)	med(r) (r_{\min}, r_{\max})	% recordings w. $r > 0.8$
Varanini	14.2	0.73 (0.52, 0.85)	0.48 (-0.37, 0.99)	41
FUSE	51.4	0.56 (0.33, 0.73)	0.27 (-0.31, 0.99)	26
FUSE _{SMOOTH}	12.8	0.53 (0.28, 0.71)	0.28 (-0.35, 0.99)	26
TS _{CERUTTI}	52.7	0.53 (0.31, 0.70)	0.28 (-0.37, 0.99)	29
TS-PCA	49.5	0.57 (0.33, 0.74)	0.37 (-0.36, 0.99)	29
TS-PCA-ICA	49.2	0.59 (0.37, 0.74)	0.37 (-0.34, 0.99)	29
Const-HR	13.0	Undefined	Undefined	0

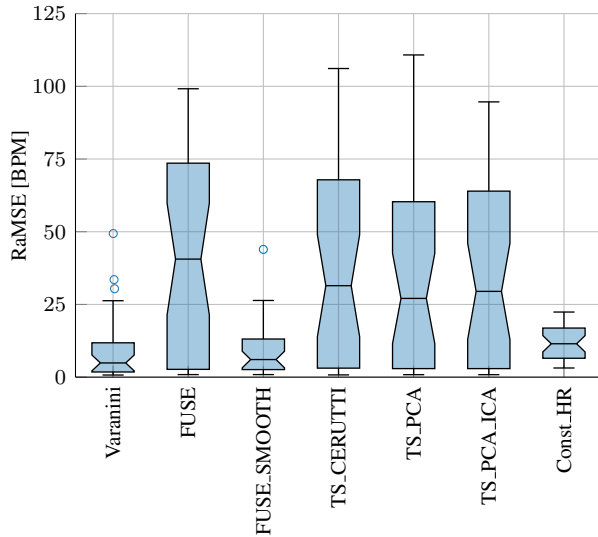


Figure 3. A boxplot of the RaMSE between the CTG and each of the 7 NI-FECG algorithms ($M=34$).

A very large range of correlation coefficients was seen among the recordings for all algorithms, from near-perfect ($r = 0.99$) to practically useless negative values. On some signals, all algorithms but CONST-HR delivered $r > 0.98$, while on other signals, all algorithms failed. Further work should investigate the relationship between signal quality and algorithm performance, as the large performance variations among recordings can most likely be ascribed to varying signal quality among recordings similar to that described by [6]. Also, development of automated signal quality measures as described by [6] might assist in quantifying the reliability of FHR estimates.

5. Conclusion

While fetal heart rate estimation using NI-FECG may be a promising alternative to the CTG, a satisfactory algorithm for extraction of the FHR from the NI-FECG is not available among the tested algorithms. The Varanini algorithm performs the best, but still the FHR is only estimated accurately in less than half of the recordings and poorly in

others, believed to be due to varying signal quality. Future algorithms should therefore be developed and tested on clinically representative datasets and quantify the reliability of the FHR estimate by e.g. estimating signal quality.

Conflict of Interest

R. G. Sæderup and H. Zimmermann are full time employees at Viewcare A/S. The current research was sponsored by Viewcare A/S.

References

- [1] Behar J, Andreotti F, Zaunseder S, Oster J, Clifford GD. A practical guide to non-invasive foetal electrocardiogram extraction and analysis. *Physiological measurement* 2016; 37(5):R1.
- [2] Clifford GD, Silva I, Behar J, Moody GB. Non-invasive fetal ecg analysis. *Physiological measurement* 2014;35(8):1521.
- [3] Varanini M, Tartarisco G, Billeci L, Macerata A, Pioggia G, Balocchi R. An efficient unsupervised fetal qrs complex detection from abdominal maternal ecg. *Physiological measurement* 2014;35(8):1607.
- [4] Behar J, Oster J, Clifford GD. Combining and benchmarking methods of foetal ecg extraction without maternal or scalp electrode data. *Physiological measurement* 2014;35(8):1569.
- [5] Andreotti F, Behar J, Zaunseder S, Oster J, Clifford GD. An open-source framework for stress-testing non-invasive foetal ecg extraction algorithms. *Physiological measurement* 2016; 37(5):627.
- [6] Andreotti F, Gräßer F, Malberg H, Zaunseder S. Non-invasive fetal ecg signal quality assessment for multichannel heart rate estimation. *IEEE Transactions on Biomedical Engineering* 2017;64(12):2793–2802.
- [7] Reinhard J, Hayes-Gill BR, Yi Q, Hatzmann H, Schiermeier S. Comparison of non-invasive fetal electrocardiogram to doppler cardiocogram during the 1st stage of labor. *Journal of perinatal medicine* 2010;38(2):179–185.

Address for correspondence:

Rasmus Gundorff Sæderup
Viewcare A/S
Generatorvej 41, 2860 Søborg, Denmark
E-mail: rgs@viewcare.com