An open-source framework for stress-testing non-invasive foetal ECG extraction algorithms

Fernando Andreotti[1,2], Joachim Behar[3], Sebastian Zaunseder[1], Julien Oster[2] and Gari D Clifford[4]

[1] Institute of Biomedical Engineering, Faculty of Electrical and Computer Engineering, Technische Universität Dresden, Dresden, Germany
[2] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
[3] Biomedical Engineering Faculty, Technion-IIT, Haifa, Israel
[4] Departments of Biomedical Informatics & Biomedical Engineering, Emory University & Georgia Institute of Technology, Atlanta, GA, USA

E-mail: fernando.andreotti@mailbox.tu-dresden.de

Abstract

Over the past decades, many studies have been published on the extraction of non-invasive foetal electrocardiogram (NI-FECG) from abdominal recordings. Most of these contributions claim to obtain excellent results in detecting foetal QRS (FQRS) complexes in terms of location. A small subset of authors have investigated the extraction of morphological features from the NI-FECG. However, due to the shortage of available public databases, the large variety of performance measures employed and the lack of open-source reference algorithms, most contributions cannot be meaningfully assessed.

This article attempts to address these issues by presenting a standardised methodology for stress testing NI-FECG algorithms, including absolute data, as well as extraction and evaluation routines. To that end, a large database of realistic artificial signals was created, totaling 145.8 h of multichannel data and over one million FQRS complexes. An important characteristic of this dataset is the inclusion of several non-stationary events (e.g. foetal movements, uterine contractions and heart rate fluctuations) that are critical for evaluating extraction routines. To demonstrate our testing methodology, three classes of NI-FECG extraction algorithms were evaluated: blind source separation (BSS), template subtraction (TS) and adaptive methods (AM). Experiments were conducted to benchmark the performance of eight NI-FECG extraction algorithms on the artificial database focusing on: FQRS detection and morphological analysis (foetal QT and T/QRS ratio).

The overall median FQRS detection accuracies (i.e. considering all non-stationary events) for the best performing methods in each group were 99.9% for BSS, 97.9% for AM and 96.0% for TS. Both FQRS detections and morphological parameters were shown to heavily depend on the extraction techniques and signal-to-noise ratio. Particularly, it is shown that their evaluation in the source domain, obtained after using a BSS technique, should be avoided. Data, extraction algorithms and evaluation routines were released as part of the *fecgsyn* toolbox on Physionet under an GNU GPL open-source license. This contribution provides a standard framework for benchmarking and regulatory testing of NI-FECG extraction algorithms.

Keywords: foetal ECG (FECG), foetal QRS (FQRS), morphological analysis, benchmark, blind source separation (BSS), template subtraction, adaptive filtering

(Some figures may appear in colour only in the online journal)

1. Introduction

Cardiotocography (CTG) has been the standard for assessing foetal cardiac activity since the 1960s. Despite its wide usage, randomised medical studies (Samueloff *et al* 1994, Bailey 2009) have casted doubt on CTG's efficacy in improving neonatal outcome. The foetal electrocardiogram (FECG), on the other hand, presents a viable alternative that can be recorded invasively or non-invasively. Invasive FECG technology uses a needle-like electrode attached to the foetal scalp, however, this technique presents three major drawbacks: (1) restricted usability (during labour only); (2) associated risk of infection; and (3) reduced number of available leads (usually one—which prevents a three-dimensional analysis of the myocard electrical activity (Behar *et al* 2014b)). The latter technique, i.e. non-invasive FECG (NI-FECG), makes use of surface electrodes placed onto the maternal abdomen and is usually applied from 20th week of gestation onwards. NI-FECG's undemanding recording setup comes at cost of a generally lower signal-to-noise ratio (SNR) for the FECG signal, since FECG overlaps both in time and frequency domain with the maternal ECG (MECG) and various noise sources (e.g. muscular artefacts). Extracting the foetal signal from the abdominal mixture remains a challenging task, which has hindered NI-FECG's further usage in the clinical practice.

A number of methods have been proposed to process abdominal mixtures (see Behar *et al* (2016) for a detailed review). However, due to the lack of annotated public databases and defined protocols for assessing these algorithms, available studies may be biased and of questionable reproducibility. The PhysioNet/Computing in Cardiology Challenge 2013 (Clifford *et al* 2014), here referred to as 'Challenge', addressed the topic of NI-FECG extraction. A few records of Challenge's database consisted of simulated abdominal signals using the synthetic foetal ECG synthetic simulator (*fecgsyn*—Behar *et al* (2014c)).

The present contribution aims at providing a standardised evaluation procedure for NI-FECG signal processing algorithms. For this purpose, a large simulated dataset was generated using the *fecgsyn* and three experiments were conducted: (1) assessing how the success of blind source separation algorithms is changing with the number of input recorded abdominal channels; (2) benchmarking a series of open-source state-of-the-art extraction methods, in terms of FQRS detection accuracy; (3) investigating the accuracy of morphological parameters evaluation (i.e. foetal QT-interval and T/QRS ratio) in the presence of non-stationarities and using different separation techniques. For the sake of standardisation and reproducibility,

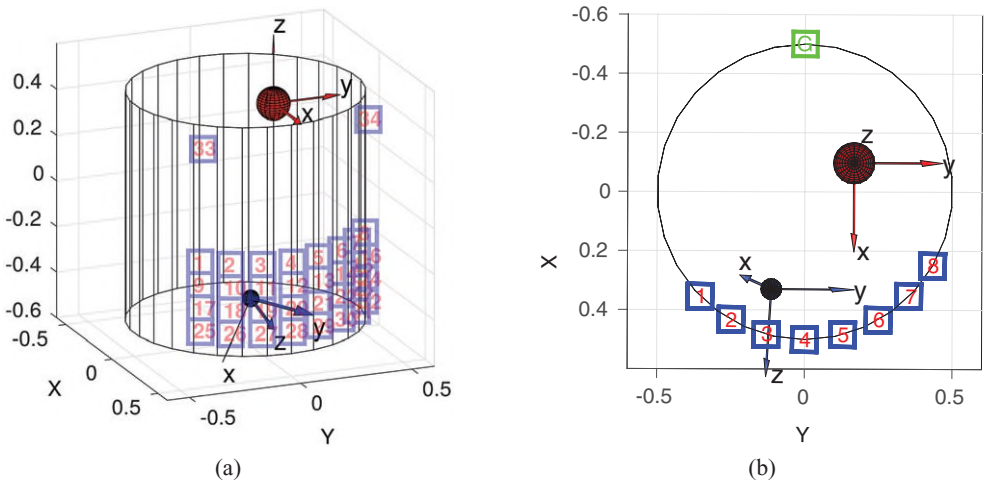

Figure 1. Side (a) and upper (b) view the of volume conductor. Positions for foetal (small sphere, blue) and maternal (larger sphere, red) hearts are shown.

the extraction routines, simulated data and the complete evaluation protocol used in this contribution were made freely available under a GNU GPL license[5]. The framework extends our previous works on the *fecgsyn*, offering fellow researchers uncomplicated data generation, evaluation and benchmarking of tools for evaluating their new algorithms.

2. Methods

2.1. Data simulation

Simulated data was generated using the *fecgsyn* (Behar *et al* 2014c). The *fecgsyn* is an extension of the original ECG model introduced by McSharry *et al* (2003) and later adapted for NI-FECG by Sameni *et al* (2007). The simulator represents maternal and foetal hearts as punctual dipoles with different magnitudes and spatial positions. Differently from the previous works on the simulator, this current version obtains foetal–maternal mixtures by treating each abdominal signal component (e.g. foetal/maternal ECG or noise signals) as an individual source, whose signal is propagated onto the observational points ('electrodes', see figure 1). This improved encapsulated design proposed by Behar *et al* (2014c), enables the modelling of a number of non-stationary physiological phenomena that affect the morphology and dynamics of the abdominal ECG by rotating, translating and modulating each available source. In this study, we are particularly interested in investigating the output of NI-FECG processing methods in the presence of such non-stationarities.

A total of seven physiological events (i.e. described in table 1) was considered. For each case, the heart dipole models (for mother and foetuses) were generated ten times by randomly selecting one of the nine vectorcardiograms available in the *fecgsyn* toolbox. Five different levels of additive noise were included (0, 3, 6, 9, and 12 dB). Simulations were repeated five times, re-generating noise signals on every iteration, to obtain a more representative database. Overall a total of $7 \times 10 \times 5 \times 5 = 1750$ synthetic signals were produced. Each simulation consisted of 5 minutes abdominal mixtures projected onto 34 channels

[5] Available at www.physionet.org/physiotools/ipmcode/fecgsyn/.

Table 1. Scenarios used for simulating pregnancy's pathophysiological events.

Case	Description
Baseline	Abdominal mixture (no noise or events)
Case 0	Baseline (no events) + noise
Case 1	Foetal movement + noise
Case 2	MHR /FHR acceleration / decelerations + noise
Case 3	Uterine contraction + noise
Case 4	Ectopic beats (for both foetus and mother) + noise
Case 5	Additional NI-FECG (twin pregnancy) + noise

Note: noise refers to muscular noise added as two independent sources situated on the lower half of the conductor volume. MHR/FHR represent the maternal/foetal heart rates.

Table 2. Model parameters used within this work, based on Behar *et al* (2014c).

Parameters	Definition	Range/type	Unit
fs	Sampling frequency	250	Hz
SNR_{fm}	Signal to noise ratio of the FECG relative to MECG	$\mathcal{N}(-9, 2)$	dB
SNR_{mn}	Signal to noise ratio of the MECG over noise	$\{0, 3, 6, 9, 12\}$	dB
fhr	Foetal heart rate	$\mathcal{N}(135, 25)$	bpm
mhr	Maternal heart rate	$\mathcal{N}(80, 20)$	bpm
facc	Foetal heart rate acceleration/ deceleration	$\mathcal{N}(30, 10)$	bpm
macc	Maternal heart rate acceleration/ deceleration	$\mathcal{N}(20, 10)$	bpm
fres	Foetal respiration frequency	$\mathcal{N}(0.90, 0.05)$	Hz
mres	Maternal respiration frequency	$\mathcal{N}(0.25, 0.05)$	Hz
mheart	Maternal heart position in polar coordinates	$\{2\pi/3, 0.2, 0.4\}$	—
fheart	Foetal heart position in polar coordinates	$\{\mathcal{U}(-\pi/10, \pi/10), \mathcal{U}(0, 0.1) + 0.25, \mathcal{U}(-0.4, -0.2)\}$	—

Note: $\mathcal{N}(\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2 and $\mathcal{U}(a, b)$ an uniform distribution between a and b. *mheart* was allowed to vary its position up to 1% of the conductor's volume in any direction.

(32 abdominal and two MECG reference channels), totaling 145.8 h of multichannel data and 1.1 million foetal peaks. Several parameters were required whilst generating these events. The most relevant ones are summarised in table 2, which summarises the ranges for the most relevant ones used in this study (for more details the reader may refer to Behar *et al* (2014c)).

2.2. NI-FECG extraction techniques

Despite the large number of NI-FECG extraction methods that have been proposed in the literature, very few studies provide open-source code for their algorithms. The Challenge promoted a considerable advance in the field by making a dataset and evaluation algorithms freely available, while some participants could voluntarily open-source their own code. Aside from the Challenge, another valuable source for NI-FECG algorithms is the Open-Source

Electrophysiological Toolbox (OSET) (Sameni 2010), which contains several algorithms for filtering, detecting and extracting foetal signals and was released under a GNU GPL license. In this contribution, some exemplary techniques have been included as a benchmark for other researchers, who can use them to compare their own algorithmic performance. In this work, we have selected extraction algorithms based on the following criteria:

(i) algorithm availability: in order to enable the integration of those algorithms in the open-source toolbox, extraction methods should be freely available under a license compatible with the GNU GPL v.3.0;
(ii) performance in the Challenge: the top-scoring entries in the Challenge were considered since they provided a fair comparison between several researchers and reflect the state-of-the-art of NI-FECG analysis;
(iii) input restrictions: aiming at a fair comparison between the algorithmic requirements, extraction methods were provided with a maternal QRS (MQRS) reference, up to one MECG reference lead and one or more preprocessed abdominal channels;
(iv) output restrictions: we focused on the best possible results for every extracted channel/component. Therefore, any channel/component selection steps were not taking into consideration when performing FQRS detection or morphological analysis;
(v) initialisation window: in this work, we allowed all extraction methods an initialisation time of up to 60 s. After this period, algorithms can be run online. For the sake of an objective comparison, no offline smoothing filter was applied.

Extraction methods may be divided into four main categories, namely blind source separation (BSS), template subtraction (TS) and adaptive methods (AM), or combination of those (so-called hybrid methods). The following sections briefly describe those classes of algorithms with exemplary applications from the literature.

2.2.1. Blind source separation. BSS techniques attempt to decompose the multichannel abdominal mixture into different components without *a priori* knowledge about the signal itself. In contrast, the sources present in the mixture are separated according to the statistical properties of the data, e.g. correlation or independence. Some of the widely used BSS methods in NI-FECG analysis are principal component analysis (PCA—Bacharakis *et al* (1996)), singular value decomposition (Callaerts *et al* 1990, Kanjilal *et al* 1997) and independent component analysis (ICA—Zarzoso *et al* (1997), De Lathauwer *et al* (2000)). In case additional information is required (e.g. MQRS locations) methods are referred to as semi-BSS, e.g. the periodic component analysis (πCA—Sameni *et al* (2008)). In this contribution, ICA and PCA were applied due to their recurrent usage in NI-FECG analysis, including top-scoring entries in the Challenge (Behar *et al* 2014a, Varanini *et al* 2014).

PCA is a simple and non-parametric method that aims to identify a meaningful basis to re-express a data set. PCA is often used for visualisation, dimensionality reduction and source separation. It assumes linearity (i.e. the mixture of signals are a linear combination of the sources), that (large) variance represent interesting structure (i.e. assessing second-order statistics) and that the different available sources can be well separated by projecting those onto the orthogonal principal components. These assumptions enable an analytical solution to the PCA problem (Shlens 2014).

ICA was introduced by Herault and Jutten (1986), later being more clearly stated by Comon (1994). The purpose of ICA is to find a linear transformation that minimises the statistical dependence between the components present in the given input data. In the context of NI-FECG, ICA attempts to obtain a demixing matrix that separates the multivariate abdominal signal into its additive sub-components, including foetal and maternal ECGs.

In this contribution, two different ICA algorithms were evaluated, namely JADE (Cardoso and Souloumiac 1993) and FAST-ICA (Hyvärinen 1999). JADE was applied using its default parameters. FAST-ICA's maximal number of iterations was set to $1000 \times N_{ch}$ (where N_{ch} is the number of channels), using a hyperbolic tangent non-linearity contrast function. Both symmetric and deflationary FAST-ICA approaches were evaluated.

BSS techniques generally assume that the signal mixture is stationary, i.e. the statistical properties of the signal do not vary over time. For clarity, BSS extraction of the NI-FECG using ICA and PCA are further referred to as $\mathbf{BSS}_{ica}$ and $\mathbf{BSS}_{pca}$, respectively. The mixing matrices for BSS techniques were estimated on every 60 s and used for processing the data in the following 60 s. This was done in order to allow the methods to partially cope with non-stationarities and be in conformity with the algorithm initialisation window requirement imposed.

2.2.2. Template subtraction. TS rely on building an average MECG cycle (the so-called 'template') by means of coherent averaging several maternal beats. This procedure heavily depends on accurate maternal QRS detection. After being constructed, templates are adapted and subtracted from each maternal cycle, leaving residual FECG and noise. A variety of TS techniques have been described in the literature (Cerutti *et al* 1986, Martens *et al* 2007, Ungureanu *et al* 2007, 2009, Vullings *et al* 2009, Di Marco *et al* 2013, Zaunseder *et al* 2013, Andreotti *et al* 2014, Behar *et al* 2014a, Lipponen and Tarvainen 2014). In this contribution, three methods were used. The first, $\mathbf{TS}_c$ (Cerutti *et al* 1986), simply adapts the template to each beat using a scalar gain. The $\mathbf{TS}_c$ was used during the Challenge by Podziemski and Gierałtowski (2013), Behar *et al* (2014a). The second, $\mathbf{TS}_{pca}$ (Kanjilal *et al* 1997), stacks MECG cycles, selects some of the principal components and, next, a back-propagation step takes place on a beat-to-beat basis, thus producing MECG estimates every cycle. $\mathbf{TS}_{pca}$ was applied during the Challenge by Behar *et al* (2014a) and Lipponen and Tarvainen (2014). The last method, i.e. $\mathbf{TS}_{ekf}$, is based on the extended Kalman filter as introduced for NI-FECG extraction by Sameni (2008). In contrast to the previous two methods, $\mathbf{TS}_{ekf}$ performs a continuous and adaptive sample-by-sample estimation of the MECG. Compared with the other approaches, $\mathbf{TS}_{ekf}$ is more adaptive, thus theoretically allowing a better estimation of the MECG in highly non-stationary scenarios. $\mathbf{TS}_c$ and $\mathbf{TS}_{pca}$ implementations were obtained from Behar *et al* (2014a), templates were built using 20 cycles and updated on every cycle. Meanwhile, the presented implementation of $\mathbf{TS}_{ekf}$ was adapted from the Challenge entry by Andreotti *et al* (2014), based on the works of Sameni (2008), (2010), and uses the first 30 MECG cycles for initialisation.

2.2.3. Adaptive methods. AM make use of one (or more) maternal reference channel(s), in order to estimate its projection onto each abdominal signal. Some of the methods that have been proposed in the literature include Strobach *et al* (1994), Widrow *et al* (1975), Rodrigues (2014), Behar *et al* (2014b), Ma *et al* (2016). In this work, three AMs using a single reference lead (channel 33—see figure 1) were applied: the least mean square ($\mathbf{AM}_{lms}$—Widrow *et al* (1975)), the recursive least square algorithm ($\mathbf{AM}_{rls}$) and the echo state neural network ($\mathbf{AM}_{esn}$—Behar *et al* (2014b)). $\mathbf{AM}_{lms}$ and $\mathbf{AM}_{rls}$ assume a linear relationship between reference and abdominal projected MECG, while $\mathbf{AM}_{esn}$ handles non-linearities. AM were not widely applied during the Challenge due to the lack of MECG reference leads. One exception was Rodrigues (2014), who made use of a Wiener filter, using three out of the four available abdominal leads as reference. Due to the lower scores obtained during the Challenge and multi-lead reference scheme, this method was excluded from this analysis. Moreover, Ma *et al* (2016) recently proposed multi-reference AMs which were outperformed by the AM_{esn}. An extensive review and performance evaluation on AM algorithms for NI-FECG extraction can be found in Behar *et al* (2014b), whose algorithmic implementation were adopted in this contribution for $\mathbf{AM}_{lms}$, $\mathbf{AM}_{rls}$ and $\mathbf{AM}_{esn}$.

2.2.4. Hybrid extraction. Hybrid methods are composed by combining methods from different classes (as presented in the previous sections). One example is the deflation procedure introduced by Sameni *et al* (2010). This general procedure transforms multichannel abdominal signals into the source-domain (by means of any BSS method), next the MECG interference is removed from the source-domain components by means of TS techniques, lastly, the denoised sources are back-propagated to the observational domain. This procedure is repeated a number of times until the output signals satisfy some predefined measure of signal separability. Sameni *et al* (2010) themselves applied the framework in NI-FECG extraction using a πCA and $\mathbf{TS}_{\text{ekf}}$ combination. These technique are out of the scope of the comparison performed in this paper due its low scores obtained during the Challenge (Haghpanahi and Borkholder 2014).

2.3. Statistical assessment

The *fecgsyn* toolbox provides the exact locations of the FQRS complexes as well as the propagated FECG signals, i.e. prior to the mixture with MECG and other noise sources. Therefore, the toolbox enables the assessment of both FQRS detection accuracy and the extraction method's ability to conserve FECG's morphological features. Both aspects were exploited throughout this work, whose respective performance evaluation are described in the following sections.

2.3.1. FQRS detection. In order to detect FQRS complexes, an adapted version (Behar *et al* 2014a) of the Pan and Tompkins algorithm (Pan and Tompkins 1985) was used. In accordance with the (ANSI/AAMI/ISO EC57 (1998/(R)2008)) guideline, sensitivity (SE) and positive predictive value (PPV) were reported as:

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP, FP and FN are the number of true positives (correctly identified FQRS), false positives (falsely detected non-existent peaks) and false negatives (missed FQRS detections), respectively. The classical adult acceptance interval[6] is 150 ms (ANSI/AAMI/ISO EC57 (1998/(R)2008)), however, to account for the higher FHR a matching window of 50 ms was used as in Andreotti *et al* (2014), Behar *et al* (2014b) and Zaunseder *et al* (2013). To summarise the results in the context of binary classification the F_1 accuracy measure, firstly suggested for FQRS detection by Behar *et al* (2014b), was used. F_1 is defined as:

$$F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{SE}}{\text{PPV} + \text{SE}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}}$$

notice that FN and FP equally affect the F_1 accuracy measure.

Beyond the presented metrics for evaluating FQRS's window-based accuracy (i.e. SE, PPV and F_1), a distance measure is necessary to discriminate between precise and imprecise detections, i.e. if any jitter occurs. This information is not captured by window-based metrics. In order to measure this distance, the mean absolute error (MAE) was used. MAE consists of the absolute time difference between the reference annotation (d_i) and detected annotation ($\hat{d}_i$). Only annotations considered as TP were considered for the MAE calculation, to make the criterion independent from the detection accuracy. Therefore, MAE is expressed as:

[6] i.e. the temporal difference allowed between an existent beat and an algorithmic detection, in which the latter is counted as correct detection (TP).

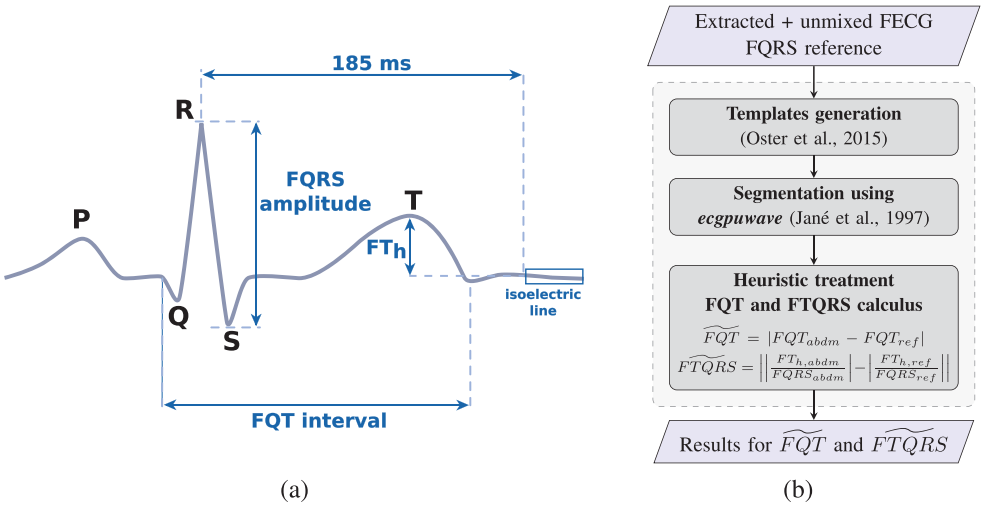

Figure 2. Experiment 3 schematics showing: (a) exemplary template beat with QT interval, foetal T-wave amplitude (FT_h), FQRS amplitude and the defined isoelectric line (starting 185 ms after R-peak); (b) signal processing steps for morphological analysis.

$$\mathrm{MAE} = \frac{1}{\mathrm{TP}} \cdot \sum_{i=1}^{\mathrm{TP}} |d_i - \hat{d}_i|.$$

2.3.2. Morphological analysis. Several studies and commercial NI-FECG equipment claim to obtain accurate FHR tracings (Behar *et al* 2016). However, beat-to-beat FQRS detection is only part of the potential benefits of NI-FECG over CTG. In adult electrocardiography, changes in the QT-interval are associated with myocardial ischemia (Murabayashi *et al* 2002) and sudden cardiac death (Piccirillo *et al* 2007). Early works on extracting morphological information from NI-FECG recordings have been published by Behar *et al* (2014d), Clifford *et al* (2011) and Reinhard *et al* (2014). These recent advances are very exciting, but the studies are still limited in number, population size and (patho) physiological conditions. To date, the STAN monitor (Neoventa Medical, Mölndal, Sweden) is the only commercial equipment performing morphological analysis of the foetal ECG in clinical environments (Clifford *et al* 2014). STAN provides FHR readings, a proxy measure for the ST segment deviation (the T/QRS amplitude ratio) and evaluates whether biphasic ST segments are present or not (Wolfberg 2012). However, for NI-FECG recordings, it is unclear how robust these measures are, particularly when accompanied by: (i) noise/artefacts; (ii) foetal movements; (iii) different electrode configurations; (iv) undesired distortions caused by extraction algorithms. Moreover, even when using the STAN as silver standard, it is not possible to assess how well the morphology of the foetal signal is preserved. This because the reference (invasive FECG) is based on a different lead, therefore, representing another projection of the cardiac electrical activity.

In this contribution, a standard analysis of morphology extraction techniques using simulated data is presented. The proposed standard enables research groups to directly compare their methodologies. For this purpose, two morphological parameters (see figure 2) were assessed: foetal QT (FQT) interval, i.e. the distance between Q-onset and T-offset, and the T/QRS ratio (FTQRS), which is the height of the T-peak compared to the isoelectric line over

the QRS amplitude (shown in figure 2). Details of the algorithmic implementation of the mentioned measures are described in the following section.

2.4. Experiments

Three experiments were conducted and are explained in detail within this section. Experiment 1 evaluated the performance of BSS techniques regarding the number of abdominal channels used and the presence of non-stationarities. Experiment 2 assessed FQRS detection accuracy for each of the selected NI-FECG extraction technique. Experiment 3 analyzed the impact of noise and the extraction procedures on the estimation of morphological parameters, such as FQT interval and FTQRS ratio.

Abdominal signals were preprocessed using low and high-pass Butterworth filters. A low-pass cutoff frequency of 100 Hz was used in all experiments. The first two experiments used a third-order low-pass and fifth-order high-pass filters with cutoff at 3 Hz. Meanwhile experiment 3 made use of a seventh-order low-pass filter and eighth-order high-pass filter with cutoff at 0.5 Hz, in order to preserve most of the foetal T-wave (both with 20 dB attenuation at the stop-band and 0.1 dB gain at the pass-band). The high-pass cutoff frequency applied in experiment 3 as well as the choice for zero-phase filtering were based on Kligfield *et al* (2007).

2.4.1. Experiment 1. The first experiment focused on preliminary considerations relevant to the usage of BSS techniques in the following experiments. In this experiment $\mathbf{BSS}_{\text{pca}}$ and $\mathbf{BSS}_{\text{ica}}$ were applied, foetal R peak detection results were reported using the F_1 measure. Several combinations of channels were used comprising two (11 and 22), four (1, 11, 22 and 32), six (1, 8, 11, 22, 25 and 32), eight (1, 8, 11, 14, 19, 22, 25 and 32) and 16 (1, 3, 6, 8, 9, 11, 14, 16, 17, 19, 22, 24, 25, 27, 30 and 32), as depicted in figure 1. The pre-selected electrode configurations were designed to span across the matrix of electrodes presented. When increasing the number of inputs (e.g. from two to four electrodes) all the channels from the previous iteration were maintained. This was done to avoid that our results depend on the quality of individual leads, i.e. if a new 'geometry' would have been chosen for each iteration, information would depend on other factors than only the number of electrodes. In order to demonstrate the highest achievable accuracy by each individual technique, the channel (or 'source') with highest F_1 was selected for every 1 min epoch and gross statistics were reported. The evaluation process presented does not take into account the permutation indeterminacy problem (Acharyya *et al* 2010), which is characteristic of BSS techniques. As a consequence, the challenge of component selection that affects the benchmark results was not addressed.

2.4.2. Experiment 2. This experiment consisted of comparing different NI-FECG extraction techniques (described in section 2.2) in stationary and non-stationary scenarios by means of F_1 and MAE. The experiment aims at obtaining an overview of how the presence of non-stationary mixtures affect extraction methods. The number of electrodes, as well as the implementation of the $\mathbf{BSS}_{\text{ica}}$ algorithm, used in this analysis were decided considering results from the previous experiment. Single-lead extraction methods (i.e. TS and AM) were applied to all available leads. In order to produce a fair comparison between BSS and the remaining techniques (TS and AM), as in experiment 1, results only consider the lead/component with the highest F_1. Gross statistics were calculated in a similar fashion to the first experiment for evaluating agreement (F_1) and distance (MAE) measures.

2.4.3. Experiment 3. This last experiment aims to assess how accurate morphological features (described in section 2.3.2) can be obtained. The experiment took into consideration the

presence of noise and effects of the different NI-FECG extraction methods on the estimated morphological measures. In this experiment, all extraction methods from experiment 2 were used, since an accurate FQRS detection does not necessarily imply that the FECG morphology is best preserved. For providing a meaningful pathophysiological analysis and clearer presentation of the results, we concentrate on a subset of the database containing Baseline and case 0 (noise only). These cases are similar in the sense that the Baseline can be considered as a case 0 with infinite SNR, since no noise is present.

Figure 2 depicts the evaluated features and signal processing key-steps for the morphological analysis. After performing the extraction (analogously to experiment 2 with a 0.5–100 Hz pass-band filter), foetal ECG templates were built for each channel using the FQRS reference annotations. The reference FQRS was used, in order to make this experiment independent from the FQRS detection accuracy obtained throughout the previous two experiments. Templates were built on a minute basis, hence five templates per channel were produced for each recording. These templates were created as in Oster *et al* (2015), which can be summarised as: (i) derivation of phase information $\in [-\pi, \pi]$ with QRS peak assigned to $-\pi/3$; (ii) individual beats are stretched/compressed into a pre-defined number of bins (250 bins); (iii) beats were clustered by calculating the normalised cross-correlation between individual beats; (iv) the cluster with highest number of members was selected; (v) beats within this cluster were averaged, resulting in a template. The most relevant parameters used for the template generation are the minimal number of cycles for a mode to be considered relevant (defined as 30) and the threshold for the correlation coefficient (starting at 0.90, decrementing with a 0.05 step and a minimal value of 0.60). Similar approaches have been proposed in the literature (Lagerholm *et al* 2000, Christov *et al* 2006, Starc and Schlegel 2006) to minimise averaging errors due to ectopic beats or false positive detections. If a template could not be built for either test or reference signal, the segment was discarded.

Each template was segmented using the *ecgpuwave* (Jané *et al* 1997), which is freely available in the WFDB toolbox (Goldberger *et al* 2000, Silva and Moody 2014). *ecgpuwave* attempts to return the fiducial points, i.e. locations for Q-onset, T-offset as well as T-peak. Reference FQRS, reference and extracted FECG signals were given as input to *ecgpuwave*. ECG signals were upsampled at 500 Hz and normalised to 2 mV to match *ecgpuwave*'s adult ECG parametrisation for sampling frequency and amplitude. A simple heuristic treatment was given to *ecgpuwave*'s output as follows: (i) if any of these fiducials are missing for a given template, no morphological analysis was carried out; (ii) if the T-wave was detected as biphasic, the T-peak with maximum absolute value was selected; (iii) templates with FQT < 100 ms or FQT > 500 ms were excluded due to physiological infeasibility (Behar *et al* 2016). The isoelectric line, required for determining foetal T-wave height ($\mathrm{FT_h}$), was defined as the median amplitude of the segment starting 185 ms after the R-peak and finishing at the end of the template (see figure 2) as in Clifford *et al* (2011). This alternative definition is due to the generally low $\mathrm{FT_h}$ amplitude. Lastly, in order to assess the fidelity of the extracted morphological parameters, the absolute error between reference and extracted FECG was evaluated using FQT ($\widetilde{\mathrm{FQT}}$) and the FTQRS ratio ($\widetilde{\mathrm{FTQRS}}$), see figure 2.

3. Results

3.1. Results for experiment 1: BSS techniques and number of sources

Preliminary tests showed a continuous decrease in F_1 when the number of input sources N_{ch} exceeds eight. This decrease in accuracy is justifiable, since ICA assumption of square mixing does not hold when the number of channels is greater than the number of underlying sources.

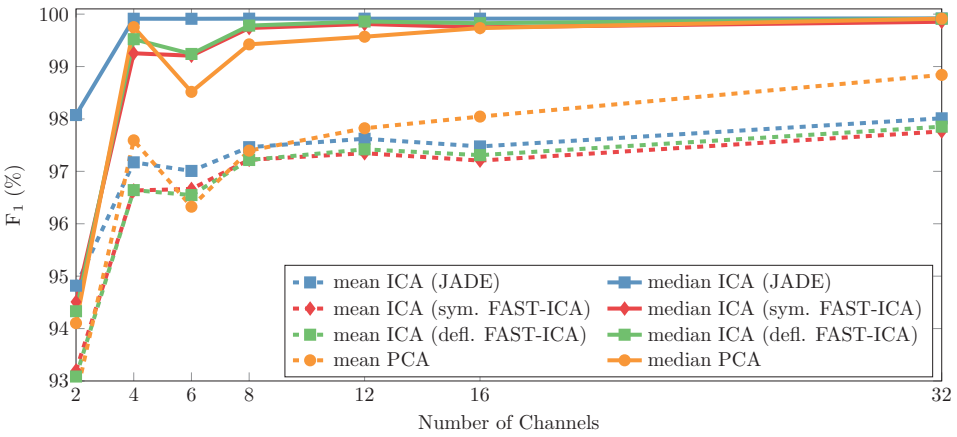

Figure 3. $\mathbf{BSS}_{\mathrm{pca}}$ and $\mathbf{BSS}_{\mathrm{ica}}$ performance with respect to the number of abdominal channels. ICA's performance was assessed using two different algoritmic implementations (JADE and FAST-ICA). Two variants of FAST-ICA, namely symmetric or deflationary, were evaluated, median results were very similar for these variants. A PCA dimension reduction step was used before the usage of $\mathbf{BSS}_{\mathrm{ica}}$, as specified in section 2.4.1.

The issue of determining the number of optimal underlying sources, referred to as model order selection problem, is well documented in the literature (Penny *et al* 2000, James and Hesse 2005). In order to partially overcome this difficulty, a PCA dimension reduction step was applied. The PCA step was designed to only keep principal components featuring 99.9% of the data variance, thus disregarding components with small eigenvalues. For instance, by including the dimension reduction step and using $N_{\mathrm{ch}} = 16$ average ICA, F_1 increased from 95.3% to 97.3%. In this example the number of final components ranged between three and eight depending on the dataset and simulated case.

Figure 3 shows the F_1 results of $\mathbf{BSS}_{\mathrm{pca}}$ and various $\mathbf{BSS}_{\mathrm{ica}}$ algorithms with respect to the number of available abdominal channels. Best average results using eight channels were $\mathbf{BSS}_{\mathrm{pca}}$ (97.40%), $\mathbf{BSS}_{\mathrm{ica}}$ with FAST-ICA (97.22%—both deflationary and symmetric) and $\mathbf{BSS}_{\mathrm{ica}}$ with JADE (97.46%). For this reason, JADE was selected to represent $\mathbf{BSS}_{\mathrm{ica}}$ techniques in experiments 2 and 3. Since for $N_{\mathrm{ch}} \geqslant 8$, little increase in F_1 was achieved, further experiments only used the 8 channels electrode configuration.

3.2. Results for experiment 2: Benchmarking of various extraction methods

A case-by-case overview on the performance of each method is shown in table 3. The highest median F_1 for each category of methods was achieved by $\mathbf{BSS}_{\mathrm{ica}}$ (99.9%), $\mathbf{AM}_{\mathrm{esn}}$ (97.9%) and $\mathbf{TS}_{\mathrm{pca}}$ (96.0%). MAE results for most methods were similar, except for $\mathbf{TS}_{\mathrm{ekf}}$, which obtained the best median results with 3.8 ms. Figure 4 provides detailed view considering both metrics (F_1 and MAE) for each technique, cases and SNR levels. Using a Kruskal–Wallis test, we found a significant effect of the SNR for most of the methods (see figure 4). Furthermore, by using a two-tailed Friedman test we evaluated the effect of the different cases and methods considering each SNR level separately. Regarding F_1, low SNRs (i.e. $\{0, 3\}$ dB) exhibited extremely significant ($p < 0.001$) differences between cases; highly significant ($p < 0.01$) for intermediate SNRs ($\{6, 9\}$ dB); whereas for a high SNR (12 dB) no significant difference was found ($p > 0.05$). Regarding MAE, extremely significant differences were found in most SNR levels ($\{0, 6, 9, 12\}$ dB) aside from for SNR = 3 dB, where it was highly significant.

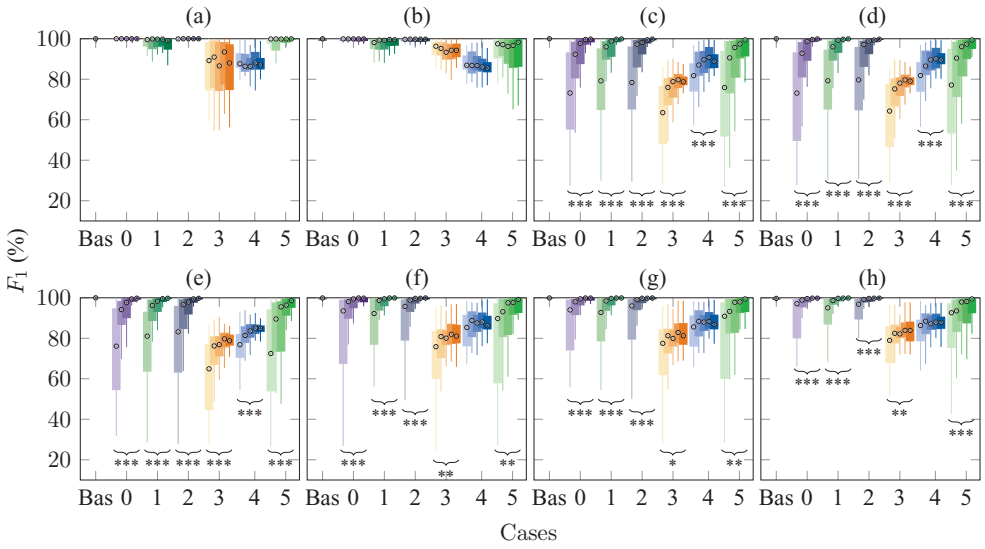

Figure 4. Detailed F_1 results for experiment 2 using different extraction methods. For each case (x-axis) there are five boxplots, one for each SNR level (improving from left to right), portrayed using different colour contrasts (the darker its contrast, the higher is the SNR). A Kruskal–Wallis test was performed to evaluate statistically significant differences across different SNR levels, where $*$ indicates $p < 0.05$, *** $p < 0.01$ and *** $p < 0.001$. Outliers were omitted for visualisation purposes. Bas = baseline case; (a) $\mathbf{BSS}_{\text{ica}}$; (b) $\mathbf{BSS}_{\text{pca}}$; (c) $\mathbf{TS}_{\text{c}}$; (d) $\mathbf{TS}_{\text{pca}}$; (e) $\mathbf{TS}_{\text{ekf}}$; (f) $\mathbf{AM}_{\text{lms}}$; (g) $\mathbf{AM}_{\text{rls}}$; (h) $\mathbf{AM}_{\text{esn}}$.

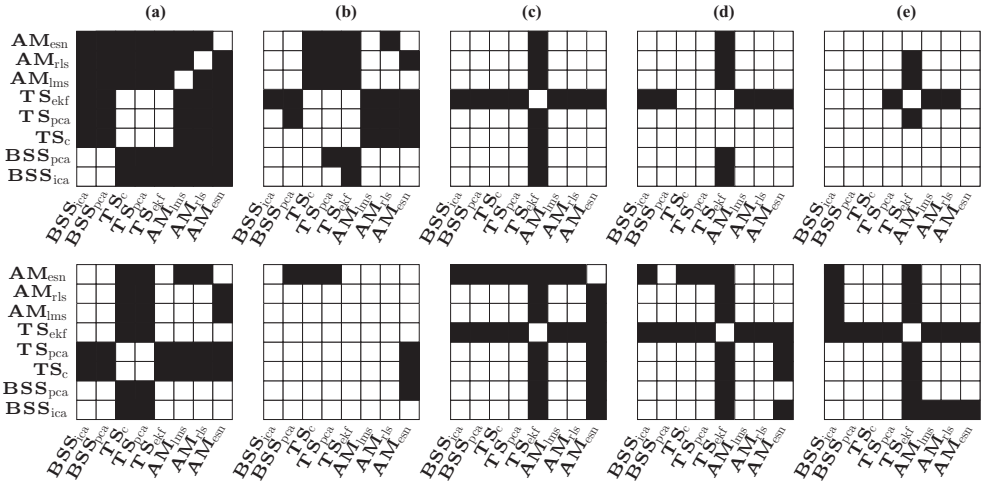

Figure 5. Post hoc analysis for experiment 2, performed using the Sign test across extraction methods. The first row shows tests regarding F_1, while the second MAE. For this analysis, the Baseline case was excluded due to its independence of the SNR level. Black squares accuse highly significant differences ($p < 0.01$) and white non-significant ($p > 0.5$). (a) SNR = 00 dB. (b) SNR = 03 dB. (c) SNR = 06 dB. (d) SNR = 09 dB. (e) SNR = 12 dB.

Similarly, the effects of different methods were tested and indicated extremely significant differences on every SNR level for both F_1 and MAE. At last, a *post hoc* test was performed using the Sign test for evaluating paired differences between methods (shown in figure 5).

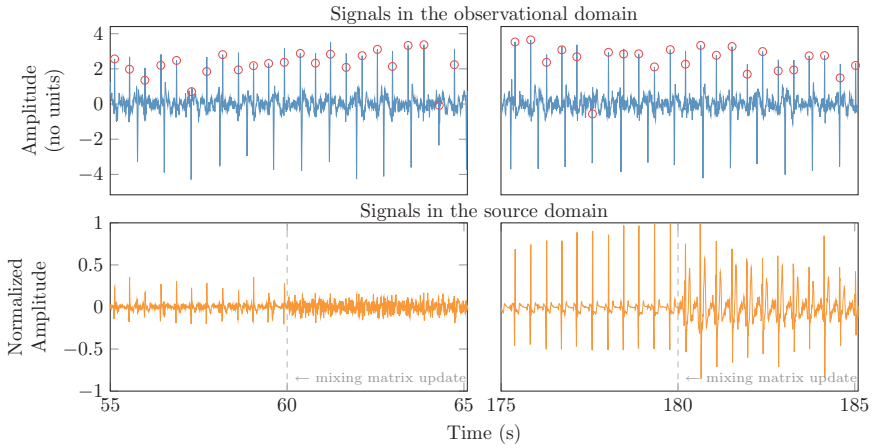

Figure 6. Two exemplary segments are portrayed during which $\mathbf{BSS}_{\text{ica}}$'s mixing matrix is updated (i.e. at 60 s and 180 s—dashed line) for a dataset containing foetal movement. Depicted above are the abdominal mixtures for channel 14, below are the selected components with highest F_1, eight channels were used as input. The plots on the first row depict abdominal mixture and FQRS locations (marked with ○), the ones in the inferior row show the selected BSS_{ica} components.

3.3. Results for experiment 3: morphological analysis

Figure 6 shows exemplary selected independent components (output from $\mathbf{BSS}_{\text{ica}}$) around segments where the mixing matrix was updated. The dataset presented in this figure includes a highly non-stationary event (case 1—foetal movement) and demonstrates the expected difficulties of analysing the morphology in case such non-stationarities would to be considered in our morphological analysis. Due to the current state of signal processing techniques to foetal morphological analysis, several beats were excluded either during template generation, segmentation using *ecgpuwave* or due to distortions by the extraction methods. The percentage of excluded beats increased with an decrease in the SNR level ranging from 8 to 78% for TS methods, 19 to 48% for AM, meanwhile it was relatively constant for BSS (14–20%). For performing a fair comparison on the morphological trustworthy, only segments on which template beats could be obtained across all methods were used in our further analysis. Therefore, the number of usable beats monotonically decreased from 69.7% on the baseline case to 11.6% on case 0 (with 0 dB noise).

Figure 7 exhibits the correlations between FQT intervals and FTQRS obtained in the FECG reference and extracted channels in the presence and absence of noise. In this figure the methods with best coefficient of determination (r^2) from each class of methods are presented. The median FQT/FTQRS was taken across all channels and segments which could be obtained for all methods. An overview on the results in terms of $\widetilde{\text{FQT}}$ and $\widetilde{\text{FTQRS}}$ for different methods, SNR levels and cases (baseline and case 0) is presented in figure 8. Similarly to the analysis in the previous experiment, the difference between median results were statistically tested (see figure 8). In this figure, a Kruskal–Wallis test was performed in evaluating if differences in the median $\widetilde{\text{FQT}}$ and $\widetilde{\text{FTQRS}}$ results for the various SNR levels were significant. It has to be kept in mind that the percentage of missing templates due to failed segmentation or failure in construction is not represented on these plots and increase with a decreasing SNR.

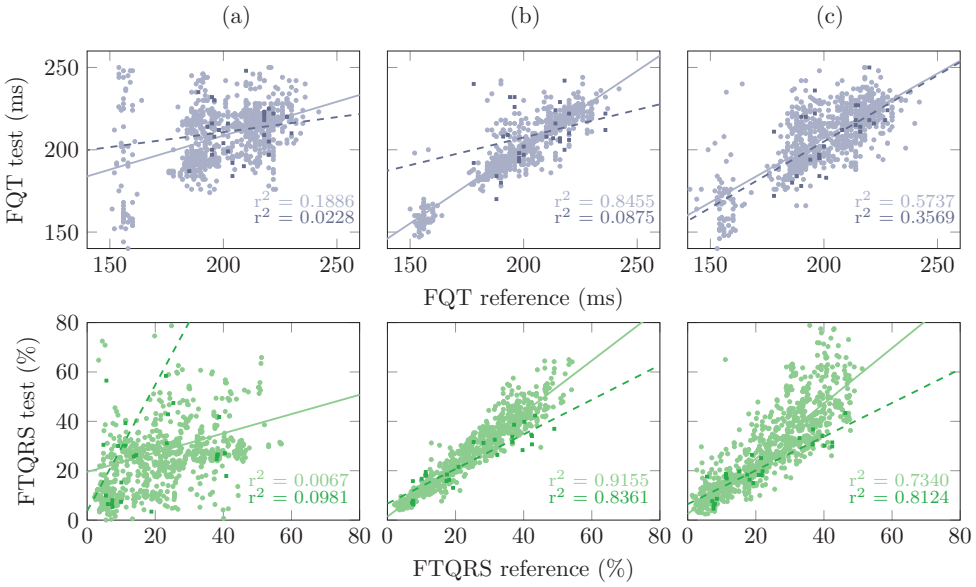

Figure 7. Differences in measured FQT interval (first row) and FTQRS (second row) between extracted channels/components (FQT/FTQRS test) and reference propagated FECG signal (FQT/FTQRS reference). Extraction methods with highest coefficient of determination (r^2) for each category are shown: (a) $\mathbf{BSS}_{\mathrm{ica}}$, (b) $\mathbf{TS}_{\mathrm{c}}$ and (c) $\mathbf{AM}_{\mathrm{esn}}$. Results are shown for the baseline (lighter colors, solid lines) and case 0 with SNR $= 0$ dB (darker colors, dashed lines). A few outliers occurred in (a—FTQRS), where $\mathrm{FTQRS}_{\mathrm{test}} \gg \mathrm{FTQRS}_{\mathrm{ref}}$, which are not shown for visualisation purposes.

4. Discussion

4.1. Experiment's results

In experiment 1, $\mathbf{BSS}_{\mathrm{ica}}$ showed superior F_1 results for a small number of channels than $\mathbf{BSS}_{\mathrm{pca}}$ (see figure 3). Similar findings using real data were obtained by Behar *et al* (2014a), where ICA outperformed PCA's FQRS detection accuracy by 10%. Moreover, JADE produced slightly better results than FAST-ICA. However, as the number of channels increases, particularly for $N_{\mathrm{ch}} \geqslant 16$, PCA's average accuracy steadily increased, indicating a decrease in the number of outliers. This consistent increase on average performance by PCA might be due to the energy threshold chosen for the dimension reduction step, i.e. a fixed threshold at 99.9% of the data's variance might provide adequate performances for eight channels but not for $N_{\mathrm{ch}} \geqslant 8$. The results for all BSS methods drastically improved by using $N_{\mathrm{ch}} > 2$. This outcome suggests that by increasing N_{ch}, one increases the chances to find a FECG component. An important consideration was the dimensionality reduction step. This step substantially improved the convergence of the FAST-ICA algorithm and promoted an increase ICA's performance for $N_{\mathrm{ch}} > 8$ (instead of a sharp decrease as preliminary tests have shown). Therefore, this highlights the fact that applications using BSS techniques in clinical data should carefully consider the model order selection problem. Another aspect to be considered is that the longer the evaluated segment (in our case 1 min), the weaker the assumption of stationarity becomes. In instances where an algorithm is applied on a long segment, BSS techniques are expected to underperform. Meanwhile, if this segment is too short, it may not contain sufficient statistical

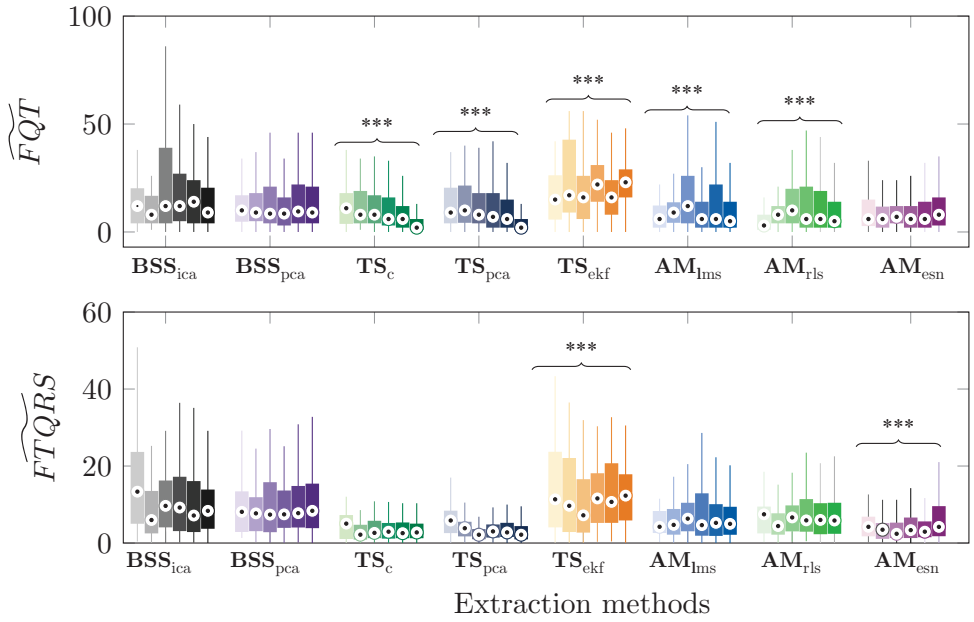

Figure 8. Results for experiment 3 showing accuracy of FQT and FTQRS extraction for the different methods and when considering different SNR levels (increasing from left to right, i.e. baseline case furthest to the right). A Kruskal–Wallis test was performed to evaluate significant differences across different SNR levels. As mentioned, as the SNR level increases, more template beats could be evaluated. Outliers were omitted for visualisation purposes.

information to represent MECG/FECG's features (e.g. non-Gaussianity), therefore preventing a satisfactory source separation.

The second experiment dealt with a general comparison between eight open-source algorithms' performance. In this test, BSS (especially **BSS**$_{ica}$) outperformed all other techniques (see figure 4 and table 3). Meanwhile, **TS**$_{pca}$ and **AM**$_{esn}$ obtained the best results in their respective categories (similar findings for the AM category using real data were presented in Behar *et al* (2014b) and Ma *et al* (2016)). MAE results were very similar for all methods, notwithstanding, temporal techniques consistently obtained the best results (see table 3). Such results were only possible, since the MAE was used as distance measure and disregarded FP and FN detections (regarded by F_1 measure). In further works, the results presented in tables 3(a) and (b) should always come in pairs to account for both accuracy and precision of FQRS detections, therefore avoiding incomplete benchmarks. It is important to mention that MAE was calculated using an FQRS reference, which was not aligned to each individual channel, therefore, a slight systematic error is expected for all methods as the R-peak location varies slightly from channel to channel. These results, however, have to be taken with caution since key algorithmic steps of BSS that can cause accuracy to decrease, were not accounted for. Particularly, the component selection, representing the foetal signal, is a problematic and decisive step for its performance. In this work, the component with highest F_1 was selected, disregarding the component selection step. Figure 4 shows that **BSS**$_{ica}$ can better extract the foetal component than AM/TS techniques in the presence of strong uterine contractions (case 2) and multiple foetuses (case 5). On the other hand, in cases with high non-stationarity of the sources such as foetal movement (case 1) and ectopic beats (case 4) **BSS**$_{ica}$ is outperformed by temporal techniques. These results for cases 1 and 4 are due to the violation of the ICA stationarity

Table 3. Case-by-case results for experiment 2, shown as median (interquartile range).

(a) F_1 (%)

Method	Baseline	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Overall
BSS$_{ica}$	100.0 (0.1)	**100.0 (0.1)**	99.7 (3.9)	**100.0 (0.1)**	89.4 (23.2)	87.1 (6.6)	**99.9 (1.9)**	**99.9 (6.4)**
BSS$_{pca}$	100.0 (0.3)	99.9 (1.1)	99.3 (4.8)	99.8 (1.4)	**94.5 (7.2)**	86.4 (7.7)	97.2 (11.3)	98.7 (7.9)
TS$_{c}$	100.0 (0.2)	97.7 (13.3)	99.0 (8.5)	98.5 (9.6)	77.4 (11.1)	87.8 (11.0)	94.1 (23.6)	95.0 (18.8)
TS$_{pca}$	**100.0 (0.0)**	98.3 (11.8)	99.2 (7.6)	98.7 (9.3)	77.5 (10.2)	**88.4 (10.8)**	94.9 (23.2)	96.0 (18.5)
TS$_{ekf}$	100.0 (0.2)	98.1 (9.9)	98.8 (8.2)	98.2 (9.6)	77.8 (10.0)	83.4 (7.3)	94.5 (24.8)	94.5 (19.8)
AM$_{lms}$	100.0 (0.4)	99.1 (6.9)	99.5 (4.9)	99.5 (3.7)	80.2 (10.6)	87.1 (9.8)	96.0 (19.2)	97.1 (15.0)
AM$_{rls}$	99.9 (0.4)	99.2 (6.0)	99.6 (4.2)	99.6 (3.4)	80.6 (11.1)	87.6 (9.4)	96.3 (17.4)	97.3 (14.6)
AM$_{esn}$	99.7 (1.1)	99.6 (2.9)	**99.7 (2.7)**	99.6 (2.0)	82.5 (8.8)	87.4 (8.3)	97.0 (14.6)	97.9 (12.5)

(b) MAE (ms)

Method	Baseline	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Overall
BSS$_{ica}$	4.0 (1.1)	4.0 (1.1)	4.2 (1.9)	4.0 (0.1)	5.2 (2.4)	5.2 (3.2)	4.0 (0.8)	4.1 (1.9)
BSS$_{pca}$	4.0 (1.8)	4.2 (1.6)	4.7 (2.0)	4.0 (1.0)	4.6 (2.2)	5.2 (1.4)	4.3 (1.9)	4.4 (1.7)
TS$_{c}$	4.0 (1.8)	4.2 (2.0)	4.1 (0.6)	4.1 (0.7)	4.7 (1.1)	5.6 (0.9)	4.3 (1.8)	4.4 (1.5)
TS$_{pca}$	4.0 (2.0)	4.2 (2.1)	4.2 (0.6)	4.1 (0.8)	4.7 (1.1)	5.6 (0.8)	4.3 (1.8)	4.4 (1.6)
TS$_{ekf}$	4.5 (6.3)	3.8 (4.4)	**3.1 (0.9)**	**3.1 (1.1)**	**3.9 (1.6)**	**4.8 (1.4)**	**3.5 (3.7)**	**3.8 (2.7)**
AM$_{lms}$	3.9 (1.8)	4.0 (1.8)	3.9 (0.7)	3.9 (0.6)	5.0 (1.5)	5.3 (1.1)	4.1 (1.5)	4.2 (1.5)
AM$_{rls}$	3.9 (2.2)	4.0 (1.8)	3.9 (0.8)	3.9 (0.8)	4.9 (1.4)	5.3 (1.2)	4.2 (1.6)	4.2 (1.6)
AM$_{esn}$	**3.8 (1.9)**	**3.8 (1.5)**	3.8 (0.6)	3.7 (0.6)	4.6 (1.1)	5.2 (1.1)	3.9 (1.1)	4.0 (1.5)

Note: cases' description available in table 1, the best performing method for each case is highlighted.

assumption. The Friedman test performed demonstrated highly significant effects of different methods' and SNR levels for both F_1 and MAE. These results were further clarified by the Kruskal–Wallis test presented in figure 4. The test demonstrates that BSS techniques are robust against varying SNR levels, while TS and AM techniques show highly significant differences in their results when varying SNR levels (results improve for higher SNRs—see figure 4). Figure 5 shows that regardless of the performance measure, the performance of the methods varies for low SNR, but the differences decrease with better SNR levels. In both cases TS_{ekf} behaves differently from other methods, regarding F_1 it performs worse than most methods, while for MAE it has the best scores. Figure 6 shows the remarkably distinct morphologies of $\mathbf{BSS}_{\text{ica}}$ output components, which depend on the calculated mixing matrix.

Experiment 3 demonstrates that considerable efforts need to be made in improving currently available techniques, so that clinical relevant information can be obtained from the FECG morphology. The expressive number of excluded beats were either due to problems in the template generation, segmentation or due to the applied methods themselves, therefore further studies should focus on improving these individual steps. Moreover, in order to perform fair morphological comparisons, studies should make sure to use the same segments and report their failure rate, as in this contribution. Figure 7 demonstrates that morphological analysis in the source domain (after applying BSS techniques) will inevitably lead to inaccurate FQT and FTQRS values, hence the results from this paper suggest that such analysis should be avoided. This claim is supported by figure 6, on which differences in the morphology of the output signals at different time-instants are evident. An alternative would be to back-propagate specifically selected FECG components from the source domain to the observation domain, in order to perform this analysis. However, as mentioned, component selection is a challenging task, which was not carried out in this study. Figure 7 demonstrates that BSS techniques are unable to provide satisfactory FQT or FTQRS measures. Meanwhile TS and AM produced highly correlating FTQRS measures (even in the presence of noise), but their FQT measures are less robust to noise. The $\mathbf{TS}_{\text{c}}$, $\mathbf{TS}_{\text{pca}}$ and $\mathbf{AM}_{\text{esn}}$ techniques delivered the best FQT and FTQRS estimates (see figure 8). These findings reiterate the importance of using excluding bad quality ECG segments from the analysis, e.g. using signal quality measures as in Behar *et al* (2013), Li *et al* (2008). The results suggest that the less adaptive the method, the fewer the distortions in the output FECG estimate. In contrast, more sophisticated methods may possess some implementation details, which skew morphological analysis. This result is consistent with best FQT estimation scores obtained in the Challenge by Podziemski and Gierałtowski (2013), who also made use of $\mathbf{TS}_{\text{c}}$. However, these results should be validated on a larger database with clinical (pathological) recordings. From figure 8 it is evident that TS techniques heavily depend on the amount of noise present in the measurement, while BSS and AM methods showed a lower dependency to noise. Moreover, our analysis showed a negative correlation between the number of templates that were unsuccessfully generated or segmented for TS and AM techniques. The $\mathbf{TS}_{\text{ekf}}$ algorithm did not perform as well as expected, this may be attributed to its simple model, which falsely models the FECG signal as a white Gaussian noise. A promising solution to this modelling problem are variations of the $\mathbf{TS}_{\text{ekf}}$ technique, which takes into consideration both maternal and foetal heart models, as suggested by Behar *et al* (2014d) and Niknazar *et al* (2013). However, accurate FQRS detections are pre-requirements for these techniques. For this reason these methods have not been included in this contribution. The separate analysis of the results for the baseline and case 0 (see figure 7, enables a clear understanding of each methods performance, because ectopic beats and other non-stationarities may cause the algorithm to fail in different routines (e.g. during template construction). This figure suggests rather high FQT errors for all extraction methods, and technical improvements are therefore required before being tested on more

difficult cases. Several explanations are possible for this low performance, e.g. the extraction techniques were optimised for FQRS detection and might therefore distort other segments of the FECG, construction of the template might average out small amplitude components (P or T wave), but also the segmentation was designed for adult ECG and its performance is likely to be sub-optimal for FECG signals. Nevertheless, figure 7 suggests that TS techniques are more suited for FTQRS analysis.

4.2. Limitations of this study

The highly accurate FQRS detections obtained in experiment 1 (even in the presence of non-stationary events) are partially due to the simplicity of the *fecgsyn*'s dipole model, which models the ECG using three orthogonal projections. Following the idea behind this dipole model, one would expect to find $3 + 3 = 6$ cardiac sources on top of which noise sources are added. Sameni *et al* (2007) have empirically identified that four to eight components can well represent the adult ECG, whereas the number of sources representing the FECG was found to be between one and three (the other components being dominated by the maternal signals and noise). This high number of components representing the adult ECG can be explained by the fact that the heart is not a punctual source, but a combination of several distributed micro-sources. Therefore, the modelling performed by the simulator limits the overall complexity of the heart signal. For this purpose, the model could be expanded to simulate such micro-sources in a similar fashion as in van Oosterom (2004). Another limitation of the model is the linear phase applied to the modelled beats, which leads to a simple stretching of those beats. This simplification has the drawback of disregarding physiological variations, such as T-wave prolongation and ST segments variation. The modelled FECG signal was acquired from adult ECG signals, which produces non-physiological FQT intervals. As highlighted in Behar *et al* (2016), the range of the FQT is roughly in the range 200–340 ms, depending on the studies, whereas the ranges obtained in this study are lower due to modelling constraints. Similarly, the amplitude values of the P-QRS-T waves and their respective ratios have been modelled by Behar *et al* (2014c) disregarding physio/pathological values. Most of the variation in beat-to-beat amplitude of the ECG components is due to the pseudo-periodic respiratory-related movement of the source dipoles with respect to the measurement locations (sensors), or non-stationarity axis shifts due to postural changes. In the absence of pathological changes (such as ectopy), such changes are the predominant shape modifiers. Small beat-to-beat changes in the P, QRS and T-wave morphologies due to autonomic changes do exist. However, with the exception of some basic relationships for QT hysteresis, and some evidence for partial respiratory sinus arrhythmia-like modulation of the PR interval, there are no well-documented studies that would provide sufficient information to describe these changes accurately. Moreover, this effect is second order compared to that due to translation and rotation. The *fecgsyn* model could be extended to provide such beat-to-beat amplitude and width variations, by modifying the amplitude and phase information of the Gaussians for each beat *if* a realistic dynamic model for modulating these parameters is later identified. However, there are no studies known to the authors describing such pathophysiological behaviours, particularly for FHR or foetal ECG morphological signals. These considerations enable BSS techniques to achieve better performance with a low number of channels.

The extraction techniques applied in our methodology have been often used in the literature, including top-scoring entries during the Challenge. Several other methods have been proposed in the literature and can easily be benchmarked using the proposed scheme, e.g. different BSS approaches as the tensor decomposition (Niknazar *et al* 2014) and combinations of different classes of algorithms as the deflation algorithm proposed in Sameni *et al*

(2010) or πCA (Sameni *et al* 2008). Besides the extraction procedure, other aspects of the NI-FECG's signal processing should be separately considered, e.g. preprocessing, channel/component selection, smoothing techniques, FQRS detection and FECG enhancement. The authors strongly recommend that researchers compare their own methodologies performance against the provided database and techniques.

In this study, simulated data were used and silver-standard reference FQT and FTQRS values were obtained using the open-source segmentation software *ecgpuwave*. Despite its wide use, in this study *ecgpuwave* often did not return relevant fiducials (e.g. T-peak or T-end). Still, *ecgpuwave* is one of the very few existent open-source algorithms for ECG segmentation freely available in the WFDB toolbox (Silva and Moody 2014). Several alternatives have been proposed in the literature (Martinez *et al* 2004, Schmidt *et al* 2014), which have reported better performance than *ecgpuwave*. However, current algorithms are mainly designed and trained using adult ECG databases and are likely to be sub-optimal for FECG analysis.

Preprocessing is a crucial aspect of morphological analysis. In this work the signal bandwidth was configured as recommended by the American Heart Society for adult electrocardiography (Kligfield *et al* 2007). A clinical trial is required to confirm if such standards can indeed be adopted for FECG analysis. Regarding the template generation, there are no current standards in ECG analysis. For example, signal-averaging is not a consesus among researchers, despite its wide usage in the literature (e.g. Clifford *et al* 2011, Farrell *et al* 1991, Gomes *et al* 2001, Zaman *et al* 2000). However, for signals with low SNR such as the NI-FECG such a step is imperative. Template generation played an important role in producing our results and a comprehensive study comparing a number of template construction strategies should be conducted to investigate at which depth its low-pass effect may hinder the morphological analysis. In our preliminary work, the template construction method proposed by Oster *et al* (2015) obtained better empirical results than simply using mean or medians, even when poorly correlated beats were excluded. It is important to remark that the findings presented here need to be validated using clinical data and expert annotations.

5. Conclusion

This contribution features a well-defined analysis framework, reproducible statistical measures and a substantial dataset for benchmarking algorithms in NI-FECG analysis in terms of FQRS detection and morphological analysis. Data, extraction algorithms and evaluation routines used in this study were released as part of the *fecgsyn* toolbox available at www.physionet.org/physiotools/ipmcode/fecgsyn/. Three experiments were conducted to benchmark the performance of eight state-of-the-art NI-FECG extraction methods in their capacity to evaluate the FQRS locations, FQT length and FTQRS ratio. For that purpose, a large dataset comprising simulated abdominal signals modelling different non-stationary scenarios was developed. Indeed, no method performed systematically better on every experiments or non-stationary cases (see table 3 and figure 4). Therefore, a combination of different extraction techniques may be beneficial. Moreover, it is important to consider non-stationary cases when evaluating NI-FECG extraction algorithms, since a given algorithm can perform well in some specific instances and fail in the case of some non-stationarity (see table 3—no method performs systematically best). In general the following items need to be further studied: pre-filtering, template generation and segmentation. For BSS techniques important considerations such as the number of channels, dimensionality reduction step and the limitations of source domain morphological analysis were addressed. The *fecgsyn* toolbox is the largest open-source collection

of NI-FECG extraction algorithms known to the authors, which also provides a large artificial database and the necessary code for evaluating algorithmic performances for the purpose of FHR extraction and morphological analysis.

Acknowledgments

FA is financially supported by the Conselho Nacional de Desenvolvimento Tecnologico (CNPq—Brazil) and TU Dresden's Graduate Academy. JB is supported in part at the Technion by the Aly Kaufman Fellowship. JO was supported by the Royal Society under a Newton Fellowship, grant number 793/914/N/K/EST/DD PF/tkg/4004642.

References

Acharyya A, Maharatna K, Al-Hashimi B M and Mondal S 2010 Robust channel identification scheme: solving permutation indeterminacy of ICA for artifacts removal from ECG **2010** 1142–5

Andreotti F, Riedl M, Himmelsbach T, Wedekind D, Wessel N, Stepan H, Schmieder C, Jank A, Malberg H and Zaunseder S 2014 Robust fetal ECG extraction and detection from abdominal leads *Physiol. Meas.* **35** 1551–67

ANSI/AAMI/ISO EC57 (1998/(R)2008) Testing and reporting performance results of cardiac rhythm and ST-segment measurement algorithms

Bacharakis E, Nandi A and Zarzoso V 1996 Foetal ECG extraction using blind source separation methods *Proc. of the European Signal Processing Conf. (EUSIPCO) (Trueste, Italy)* pp 395–8

Bailey R E 2009 Intrapartum fetal monitoring *Am. Family Physician* **80** 1388–96

Behar J, Andreotti F, Oster J and Clifford G D 2014d A Bayesian filtering framework for accurate extracting of the non invasive FECG morphology *Proc. of the IEEE Conf. on Computing in Cardiology (Boston, USA, 2014)* pp 53–6

Behar J, Andreotti F, Zaunseder S, Li Q, Oster J and Clifford G D 2014c An ECG model for simulating maternal-foetal activity mixtures on abdominal ECG recordings *Physiol. Meas.* **35** 1537–50 (Code freely available at: www.physionet.org/physiotools/ipmcode/fecgsyn/)

Behar J, Andreotti F, Zaunseder S, Oster J and Clifford G D 2016 A practical guide to non-invasive foetal electrocardiogram extraction and analysis *Physiol. Meas.* **37** R1–35

Behar J, Johnson A, Clifford G D and Oster J 2014b A comparison of single channel foetal ECG extraction methods *Ann. Biomed. Eng.* **42** 1340–53

Behar J, Oster J and Clifford G D 2014a Combining and comparing benchmarking methods of foetal ECG extraction without maternal or scalp electrode data *Physiol. Meas.* **35** 1569–89

Behar J, Oster J, Li Q and Clifford G 2013 ECG signal quality during arrhythmia and its application to false alarm reduction *IEEE Trans. Biomed. Eng.* **60** 1660–6

Callaerts D, De Moor B, Vandewalle J, Sansen W, Vantrappen G and Janssens J 1990 Comparison of SVD methods to extract the foetal electrocardiogram from cutaneous electrode signals *Med. Biol. Eng. Comput.* **28** 217–24

Cardoso J F and Souloumiac A 1993 Blind beamforming for non-gaussian signals *IEE Proc. F: Radar Signal Process.* **140** 362–70

Cerutti S, Baselli G, Civardi S, Ferrazzi E, Marconi A M, Pagani M and Pardi G 1986 Variability analysis of fetal heart rate signals as obtained from abdominal electrocardiographic recordings *J. Perinat. Med.* **14** 445–52

Christov I, Gómez-Herrero G, Krasteva V, Jekova I, Gotchev A and Egiazarian K 2006 Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification *Med. Eng. Phys.* **28** 876–87

Clifford G D, Silva I, Behar J and Moody G B 2014 Non-invasive fetal ECG analysis *Physiol. Meas.* **35** 1521–36

Clifford G, Sameni R, Ward J, Robinson J and Wolfberg A J 2011 Clinically accurate fetal ECG parameters acquired from maternal abdominal sensors *Am. J. Obstet. Gynecol.* **205** 47

Comon P 1994 Independent component analysis, a new concept? *Signal Process.* **36** 287–314

De Lathauwer L, De Moor B and Vandewalle J 2000 Fetal electrocardiogram extraction by blind source subspace separation *IEEE Trans. Biomed. Eng.* **47** 567–72
Di Marco L, Marzo A and Frangi A 2013 Multichannel foetal heartbeat detection by combining source cancellation with expectation-weighted estimation of fiducial points *Proc. of the Computing in Cardiology Conf.* pp 329–32
Farrell T G, Bashir Y, Cripps T, Malik M, Poloniecki J, Bennett E, Ward D E and Camm A 1991 Risk stratification for arrhythmic events in postinfarction patients based on heart rate variability, ambulatory electrocardiographic variables and the signal-averaged electrocardiogram *J. Am. Coll. Cardiol.* **18** 687–97
Goldberger A L, Amaral L A, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals *Circ.* **101** e215–20
Gomes J, Cain M and Buxton A 2001 Prediction of long-term outcomes by signal-averaged electrocardiography in patients with unsustained ventricular tachycardia coronary artery disease, and left ventricular dysfunction *Circulation* **104** 436–41 (PMID: 11468206)
Haghpanahi M and Borkholder D A 2014 Fetal QRS extraction from abdominal recordings via model-based signal processing and intelligent signal merging *Physiol. Meas.* **35** 1591–605
Herault J and Jutten C 1986 Space or time adaptive signal processing by neural network models *AIP Conf. Proc. Neural Network for Computer* (*AIP Publishing American Institute of Physics Inc. Snowbird, USA*) vol 151 pp 206–11
Hyvärinen A 1999 Fast and robust fixed-point algorithms for independent component analysis *IEEE Trans. Neural Netw.* **10** 626–34
James C J and Hesse C W 2005 Independent component analysis for biomedical signals *Physiol. Meas.* **26** R15
Jané R, Blasi A, García J and Laguna P 1997 Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database *Proc. of the IEEE Conf. on Computers in Cardiology* pp 295–8
Kanjilal P, Palit S and Saha G 1997 Fetal ECG extraction from single-channel maternal ECG using singular value decomposition *IEEE Trans. Biomed. Eng.* **44** 51–9
Kligfield P *et al* 2007 Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American heart association electrocardiography and arrhythmias committee, council on clinic *J. Am. Coll. Cardiol.* **49** 1109–27
Lagerholm M, Peterson C, Braccini G, Edenbrandt L and Sörnmo L 2000 Clustering ECG complexes using hermite functions and self-organizing maps *IEEE Trans. Biomed. Eng.* **47** 838–48
Lipponen J A and Tarvainen M P 2014 Principal component model for maternal ECG extraction in fetal QRS detection *Physiol. Meas.* **35** 1637–48
Li Q, Mark R G and Clifford G D 2008 Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter *Physiol. Meas.* **29** 15–32
Martens S M M, Rabotti C, Mischi M and Sluijter R J 2007 A robust fetal ECG detection method for abdominal recordings *Physiol. Meas.* **28** 373–88
Martinez J P *et al* 2004 A wavelet-based ECG delineator: evaluation on standard databases *IEEE Trans. Biomed. Eng.* **51** 570–81
Ma Y, Xiao Y, Wei G and Sun J 2016 A multichannel nonlinear adaptive noise canceller based on generalized FLANN for fetal ECG extraction *Meas. Sci. Technol.* **27** 15703
McSharry P E, Clifford G D, Tarassenko L and Smith L A 2003 A dynamical model for generating synthetic electrocardiogram signals *IEEE Trans. Biomed. Eng.* **50** 289–94
Murabayashi T, Fetics B, Kass D, Nevo E, Gramatikov B and Berger R D 2002 Beat-to-beat QT interval variability associated with acute myocardial ischemia *J. Electrocardiol.* **35** 19–25
Niknazar M, Becker H, Rivet B, Jutten C and Comon P 2014 Blind source separation of underdetermined mixtures of event-related sources *Signal Process.* **101** 52–64
Niknazar M, Rivet B and Jutten C 2013 Fetal ECG extraction by extended state Kalman filtering based on single-channel recordings *IEEE Trans. Biomed. Eng.* **60** 1345–52
Oster J, Behar J, Sayadi O, Nemati S, Johnson A and Clifford G 2015 Semisupervised ECG ventricular beat classification with novelty detection based on switching Kalman filters *IEEE Trans. Biomed. Eng.* **62** 2125–34
Pan J and Tompkins W J 1985 A real-time QRS detection algorithm *IEEE Trans. Biomed. Eng.* **32** 230–36

Penny W D, Roberts S and Everson R M 2000 *Independent Component Analysis: Principles and Practice* (Cambridge: Cambridge Univeristy Press) chapter 12, pp 299–314
Piccirillo G *et al* 2007 QT variability strongly predicts sudden cardiac death in asymptomatic subjects with mild or moderate left ventricular systolic dysfunction: a prospective study *Eur. Heart J.* **28** 1344–50
Podziemski P and Gierałtowski J 2013 Fetal heart rate discovery: algorithm for detection of fetal heart rate from noisy, noninvasive fetal ECG recordings *Comput. Cardiol.* **40** 333–6
Reinhard J, Hayes-Gill B, Yuan K, Schiermeier S and Louwen F 2014 Intrapartum ST segment analyses (STAN) using simultaneous invasive and non-invasive fetal electrocardiography: a report of 6 cases *Z. Geburtshilfe Neonatol.* **218** 122–7
Rodrigues R 2014 Fetal beat detection in abdominal ECG recordings: global and time adaptive approaches *Physiol. Meas.* **35** 1699–711
Sameni R, Clifford G D, Jutten C and Shamsollahi M B 2007 Multichannel ECG and noise modeling: application to maternal and fetal ECG signals *EURASIP J. Appl. Signal Process.* **2007** 94
Sameni R, Jutten C and Shamsollahi M B 2008 Multichannel electrocardiogram decomposition using periodic component analysis *IEEE Trans. Biomed. Eng.* **55** 1935–40
Sameni R, Jutten C and Shamsollahi M B 2010 A Deflation Procedure for Subspace Decomposition *IEEE Trans. Signal Process.* **58** 2363–74
Sameni R 2008 Extraction of fetal cardiac signals from an array of maternal abdominal recordings *PhD Thesis* Sharif University of Technology—Institut National Polytechnique de Grenoble (http://hal.univ-grenoble-alpes.fr/tel-00373361/document)
Sameni R 2010 'The open-source electrophysiological toolbox (OSET), version 21 (http://spc.shirazu.ac.ir/products/Featured-Products/oset/)
Samueloff A, Langer O, Berkus M, Field N, Xenakis E and Ridgway L 1994 Is fetal heart rate variability a good predictor of fetal outcome? *Acta Obstet. Gynecol. Scand.* **73** 39–44
Schmidt M, Baumert M, Porta A, Malberg H and Zaunseder S 2014 Two-dimensional warping for one-dimensional signals—conceptual framework and application to ECG processing *IEEE Trans. Signal Process.* **62** 5577–88
Shlens J 2014 A tutorial on principal component analysis (arXiv:14041100)
Silva I and Moody G 2014 An open-source toolbox for analysing and processing physionet databases in matlab and octave *J. Open Res. Softw.* **2** e27
Starc V and Schlegel T T 2006 Real-time multichannel system for beat-to-beat QT interval variability *J. Electrocardiol.* **39** 358–67
Strobach P, Abraham-Fuchs K and Härer W 1994 Event-synchronous cancellation of the heart interference in biomedical signals *IEEE Trans. Biomed. Eng.* **41** 343–50
Ungureanu G M, Bergmans J W M, Oei S G, Ungureanu A and Wolf W 2009 The event synchronous canceller algorithm removes maternal ECG from abdominal signals without affecting the fetal ECG *Comput. Biol. Med.* **39** 562–7
Ungureanu M, Bergmans J, Oei S and Strungaru R 2007 Fetal ECG extraction during labor using an adaptive maternal beat subtraction technique *Biomed. Tech.* **52** 56–60
van Oosterom A 2004 ECGSIM: an interactive tool for studying the genesis of QRST waveforms *Heart* **90** 165–8
Varanini M, Tartarisco G, Billeci L, Macerata A and Balocchi R 2014 An efficient unsupervised fetal QRS complex detection from abdominal maternal ECG *Physiol. Meas.* **35** 1607–19
Vullings R, Peters C H L, Sluijter R J, Mischi M, Oei S G and Bergmans J W M 2009 Dynamic segmentation and linear prediction for maternal ECG removal in antenatal abdominal recordings *Physiol. Meas.* **30** 291–307
Widrow B, Glover J R Jr, McCool J, Kaunitz J, Williams C, Hearn R, Zeidler J, Eugene Dong J and Goodlin R 1975 Adaptive noise cancelling: principles and applications *Proc. IEEE* **63** 1692–16
Wolfberg A J 2012 The future of fetal monitoring *Rev. Obstet. Gynecol.* **5** e132–6
Zaman A, Archbold R, Helft G and Paul E 2000 Atrial fibrillation after coronary artery bypass surgery a model for preoperative risk stratification *Circulation* **101** 1403–8
Zarzoso V, Nandi A and Bacharakis E 1997 Maternal and foetal ecg separation using blind source separation methods *Math. Med. Biol.* **14** 207–25
Zaunseder S, Andreotti F, Cruz M, Stepan H, Schmieder C, Malberg H, Jank A and Wessel N 2013 Fetal QRS detection by means of Kalman filtering and using the event synchronous canceller *IJBEM* **15** 83–9