Non-invasive Fetal ECG Signal Quality Assessment for Multichannel Heart Rate Estimation

Fernando Andreotti, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder

***Abstract*—*Objective*: The noninvasive fetal ECG (NI-FECG) from abdominal recordings offers novel prospects for prenatal monitoring. However, NI-FECG signals are corrupted by various nonstationary noise sources, making the processing of abdominal recordings a challenging task. In this paper, we present an online approach that dynamically assess the quality of NI-FECG to improve fetal heart rate (FHR) estimation. *Methods*: Using a naive Bayes classifier, state-of-the-art and novel signal quality indices (SQIs), and an existing adaptive Kalman filter, FHR estimation was improved. For the purpose of training and validating the proposed methods, a large annotated private clinical dataset was used. *Results*: The suggested classification scheme demonstrated an accuracy of Krippendorff's alpha = 0.65 in determining the overall quality of NI-FECG signals. The proposed Kalman filter outperformed alternative methods for FHR estimation achieving 75.6% accuracy. *Conclusion*: The proposed algorithm was able to reliably reflect changes of signal quality and can be used in improving FHR estimation. *Significance*: NI-ECG signal quality estimation and multichannel information fusion are largely unexplored topics. Based on previous works, multichannel FHR estimation is a field that could strongly benefit from such methods. The developed SQI algorithms as well as resulting classifier were made available under a GNU GPL open-source license and contributed to the FECGSYN toolbox.**

***Index Terms*—Fetal ECG, signal quality, data fusion, Kalman filter, naive Bayes.**

I. Introduction

THE non-invasive fetal ECG (NI-FECG) derived from abdominal surface electrodes offers novel diagnostic possibilities for prenatal medicine by enabling long-term monitoring of the fetal cardiac activity. Despite its straightforward applicability, the NI-FECG signal is usually corrupted by many interfering sources, most significantly by the maternal ECG (MECG). The presence of additional highly non-stationary noise sources (e.g. muscular/uterine noise, electrode motion, etc.) further affects the signal-to-noise ratio (SNR) of the fetal signal, making the detection of the fetal QRS (FQRS) complexes a challenging signal-processing task, in which erroneous detections are unavoidable. Nonetheless, with the developments in signal acquisition and processing techniques over the last decades, the NI-FECG has regained the attention of several research groups [1]–[3]. This increasing interest culminated on Physionet/Computing in Cardiology Challenge (PCINCC) 2013 [2] on the topic accurate FQRS detection and FHR estimation. During the PCINCC 2013, various approaches for improving fetal heart rate detection were presented e.g. [4]–[6]. While most available FQRS detection methods are simple re-parametrized single-channel adult QRS detectors, FHR estimates are often based on heuristic rules for how a smooth FHR tracing should appear. Aside from that, multichannel FQRS/FHR information is often fused using simple majority voting or weighted average algorithms.

Despite the large progress over the recent years, particularly in case of low/varying SNR, present methods are still unable of producing reliable FQRS complexes. Therefore, the current state of NI-FECG research raises two relevant issues: 1) how can the quality of NI-FECG be quantified in scenarios of low/varying SNR? 2) how can one fuse the information of multiple NI-FECG channels when segments/channels containing erroneous detections are present? The first point relates to the usage of signal quality indices (SQIs), as previously addressed by many works on adult ECG monitoring [7], [8]. However, due to the presence of an additional pseudo-periodic signal (i.e. the MECG), conventional SQIs are expected to underperform. To the best of our knowledge, there is currently no contribution directed at NI-FECG quality estimation available in the literature. Regarding the latter question, given that suitable NI-FECG SQIs are at one's disposal, adult ECG approaches could be transferred to NI-FECG applications. A sophisticated framework that incorporates signal quality information in multichannel data-fusion is the Kalman Filter (KF) [9]. The filter makes use of both measurement and its intrinsic model in attributing adaptive weights to each channel's estimates, being applied in adult heart rate (HR) assessment by [10]. For a complete review on SQIs and their KF associated use, the reader is referred to Oster *et al.* [11].

Manuscript received October 18, 2016; revised February 17, 2017; accepted February 21, 2017. Date of publication March 1, 2017; date of current version November 20, 2017. The work of F. Andreotti was supported by the Conselho Nacional de Desenvolvimento Tecnologico (CNPq-Brazil) and Technische Universität Dresden Graduate Academy as part of the Excellence Initiative of the German Federal and State Governments. The work of F. Gräßer was supported by the Roland Ernst Stiftung für Gesundheitswesen (Project Therapieempfehlungssystem-Technische Universität Dresden). *(Corresponding author: Fernando Andreotti.)*

F. Andreotti is with the Institute of Biomedical Engineering, Technische Universität Dresden, Dresden 01069, Germany (e-mail: fernando.andreotti@tu-dresden.de).

F. Gräßer, H. Malberg, and S. Zaunseder are with the Institute of Biomedical Engineering, Technische Universität Dresden.

Digital Object Identifier 10.1109/TBME.2017.2675543

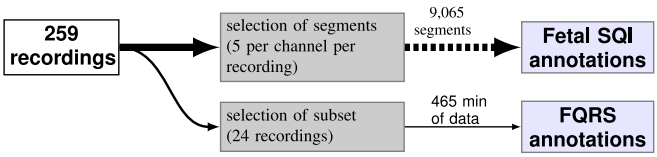

Fig. 1. Breakdown view of available data and annotation procedures carried out in this study. The line thickness represents the number of recordings used.

In this contribution, we aim at characterizing extracted NI-FECG signals by using signal quality measures. For this purpose, numerous novel quality indices specifically designed for dealing with NI-FECG recordings are proposed. These SQI metrics are thoroughly evaluated concerning their ability to mirror the overall quality of NI-FECG signals by using a Naive Bayes classifier. The resulting quality estimates are evaluated using manually annotated clinical data and are further incorporated into a multichannel KF framework for improving fetal heart rate (FHR) estimates. The developed SQI algorithms as well as resulting classifier were made available under a GNU GPL open-source license and contributed to the FECGSYN toolbox [12], [13] (available at www.fecgsyn.com).

II. Materials and Methods

A. Data and Annotation Procedure

Data was acquired at the Department of Obstetrics and Gynecology at the University Hospital of Leipzig (local ethics committee approval 348-12-24092012). A total of $n = 259$ multichannel recordings from 107 singleton pregnancies (diverse pathophysiological states and gestational weeks ranging from 17 to 39 weeks) were collected. Each record consisted of approximately 20 min on 1 maternal chest lead and 7 abdominal channels, as previously described in [4]. The electrode configuration consists of 4 external derivations (forming a larger circle around the maternal abdomen) and 3 internal (around the navel with reference electrode placed about the fundus of the uterus - see [4]). Two different annotations procedures were performed, detailed in Fig. 1 and further described in the following subsections.

1) Signal Quality Annotation: Five equidistant 5 s segments were extracted from each recorded abdominal channel for further annotation. The total of 9,065 segments (7 channels $\times$ 5 segments $\times$ 259 recordings) was carefully annotated by four experts for their FECG amplitude (4 classes - see Table I) and SNR levels (5 classes). Two observers annotated every segment, while the other two annotated in a complementary manner 72.6% and 37.0%, respectively, so that at least 3 annotations were available for each segment. A subset of 500 segments was annotated twice by every observer to evaluate intra-observer reliability. Preliminary results from the authors [14] showed a good intra-rater (> 0.65) and inter-rater (> 0.63) agreement on average (using Krippendorf's coefficient, further explained in Section II-D). A consensus was obtained for each segment using majority voting for each FECG amplitude and SNR criteria. For avoiding biases, ties were decided by randomly choosing one of the most voted alternatives. Further, an overall fetal peak visibility consensus was defined (see Table I), which is used in this work as gold-standard measure for the NI-FECG quality. For further reference, in our clinical database there was a class imbalance with the following relative frequencies for overall consensus: A (0.72 %), B (3.40 %), C (6.92 %), D (10.04 %) and E (78.93 %) – see Fig. 2. The amount of observations in class E, e.g. refers to segments which showed low quality. This number does not necessarily reflect the overall quality of our multichannel recordings, since individual channels are likely to contain different signal qualities. Nonetheless, as previously described, the exploratory clinical study herein presented comprised patients at diverse gestational weeks including those during which the well-known FECG quality degrading effect of the *vernix caseosa* is present.

TABLE I
Definition for Fetal Peak Visibility Consensus, Based on FECG Strength and SNR Consensus

SNR†	FECG‡	Meaning	Overall consensus
4,5	4	FECG clearly distinguishable,	A
4,5	3	FECG clearly distinguishable,	B
3	4	good SNR	
5	2	FECG distinguishable, adequate SNR	C
3	3	adequate SNR	
3,4	2	FECG distinguishable, low SNR	D
d.c.	1	No FECG distinguishable	E
1,2	d.c.		

Scoring was done visually, on a segment basis, one-channel at a time with the respective MECG chest lead serving as reference.
† 1 = unacceptable; 2 = bad; 3 = adequate;4 = good; 5 = excellent
‡ 1 = not present; 2 = low; 3 = moderate; 4 = high amplitude
d.c. Don't-care term

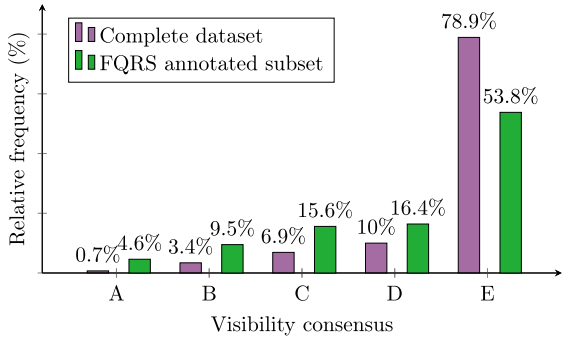

Fig. 2. Histogram showing overall annotated consensus for signal quality (see Table I). Both complete dataset and FQRS annotated subset are presented. "Complete dataset" refers to 9,065 segments annotated by experts, while "FQRS annotated subset" depicts the $7 \times 5 \times 24 = 840$ segments from the 24 recordings that had their FQRS complexes annotated (as explained in Fig. 1).

2) Fetal QRS Annotation: To validate the results for FHR estimation, a subset of the clinical dataset containing 24 recordings, where FQRS complexes were partially visible in at least one channel, were selected. These recordings were annotated by one and further corrected by another two experts for their individual FQRS and maternal QRS (MQRS) locations. The subset contained recordings with mixed quality as

reported in Fig. 2. This annotation procedure was carried out for each complete recording (i.e. approximately 20 min each) [4]. During the annotation procedure, experts had access to all 7 abdominal channels and the maternal chest lead. Annotators were asked to simultaneously make use of at least 2 abdominal channels (interchangeable during annotation) and the MECG lead in annotating visible FQRS complexes. As shown in Fig. 2 and discussed in [14], abdominal channels presented different qualities depending on several pathophysiological and recording variables such as fetal position, presence of the vernix, and electrode configuration. Therefore, fetal complexes may only be visible on a portion of the available channels. In cases where the FQRS were temporarily not clearly distinguishable in any channel, up to 2 consecutive peaks were annotated, if the expert could extrapolate their positions (e.g. during fetal/maternal overlaps). Longer periods without distinguishable fetal peaks were marked as a region of signal loss, where no FQRS annotation was placed. Annotation was performed on preprocessed abdominal signals (using simple bandpass filters as described in [4]) to avoid the bias by extraction methods on the annotation procedure. The annotated subset totalized 448 min (3.7 % excluded due to signal loss) and over 67,000 fetal complexes, with average FHR_{ref} across these recordings ranging from 139.1 to 154.4 bpm and standard deviation between 3.9 and 11.1 bpm. The FHR trace derived from these annotations was regarded as reference (FHR_{ref}) for evaluating the FHR estimates.

B. Automated Signal Quality Assessment

Improving signal quality estimation leads to a desirable increase on specificity of built-in signal processing techniques [11]. This topic's importance was confirmed by the recent PCINCCs that dealt with the determination of ECG's clinical acceptability (in 2011 [15]) and false alarm reduction (in 2015 [16]). Both competitions have in common that top-scoring entries ([7], [17]) made use of SQIs to improve the specificity of their results. One of the most significant works using SQI metrics was proposed by Li *et al.* [10], who suggested various SQIs for both ECG and blood pressure (BP) signals. As later described by the same authors [18], based on earlier works, which made use e.g. of the Karhunen-Loéve transform [19] or standard noise measurement methods [20], Li *et al.* [10] developed four SQI metrics: 1) the ratio of power in various bands of the spectrum, 2) the degree of agreement between different QRS detectors, 3) the degree of agreement between beat detection on different leads, and 4) the kurtosis of the ECG [18]. Since then, SQI metrics have been successfully applied in adult ECG [7], [8], [17], [21], BP [17], [21], [22] and PPG signals [23]–[25]. In contrast to the PCINCC 2011, [18] emphasized the need for multilevel signal quality classification approaches, rather than classifying ECGs as 'acceptable' or 'unacceptable'. Despite the abundance of works in adult ECG analysis, SQIs have not been conveyed to the context of NI-FECG. To resolve this issue, in this work we assume that the following signals are available:

i) maternal chest lead;
ii) reliable maternal MQRS annotations obtained as described in Section II-A2;
iii) N channels of preprocessed abdominal signals (Butterworth bandpass filters with a pass band of 3-100 Hz band were applied as in [13]);
iv) N extracted NI-FECG signals from (iii). In this contribution, the extraction was performed using the TS_{pca} technique explained in [13].

These input signals do not impose any restriction to our analysis since they are readily accessible after the FECG extraction is performed. In the following sections, the SQIs derived from these signals are presented, as well as a classification algorithm for combining those different metrics into one overall SQI value for each 5 s segment.

1) Feature Extraction - Signal Quality Indices:

The SQI metrics used are summarized in Table II. The metrics were divided into four classes of algorithms: time, frequency, detection-based methods and FECG-specific approaches. Amongst those metrics are adaptations of adult ECG SQI algorithms (derived from the literature), two indices proposed by the authors within a previous work [4] (i.e. $cSQI$ and $xSQI$) and four novel indices that are specific for NI-FECG analysis. These proposed novel FECG-specific indices, namely $mxSQI$, $mpSQI$, $mcSQI$ and $miSQI$, are intended to estimate how well the MECG suppression performs.

The conformity measure (*cSQI*), proposed in [4], is based on the construction of an averaged FECG template beat by simple coherent average. With that FECG template at hand, $cSQI$ is the average value for the correlation coefficient between each available beat and the FECG template within a segment. The extravagance (i.e. *xSQI* [4]) is a measure of contrast between the detected FQRS complex and the embedded noise. By using a window of ± 25 ms length around location of the individual FQRS, the FECG peak amplitude is compared against the power of the signal within three times that window length. The *mxSQI* is the analogous of $xSQI$ with focus on the amplitude of the MQRS peaks in comparison to the surrounding residuals of the extraction procedure. This metric uses the extracted abdominal signal and MQRS reference locations to estimate the strength of MECG residuals on the estimated FECG signal. The metric is obtained as $mxSQI = 1 - xSQI\big|_{MQRS,FECG}$. The *mpSQI* makes use of the magnitude-squared spectrum for the residuals and the median reference maternal HR (MHR) for each segment. Similarly to the approach in [27], the aim is to evaluate the relative power of MHR and its harmonics on the frequency-domain as an index for left-over maternal residuals. In order to do so, the MQRS fundamental frequency and its N_f first harmonics (within an empirically defined ± 0.3 Hz band) were compared with the power of up until the fifth harmonic ($mpSQI_a$) or all peaks within the $[0.5, 10]$ Hz band ($mpSQI_b$). The spectral coherence (or magnitude-squared coherence) is yet another metric that measures the cross-correlation between two frequency spectra that outputs values between $[0, 1]$. In this work, we aim at applying this measure to assess the extraction performance by using the involved signals, henceforth named *mcSQI*. We evaluated two variants of the $mcSQI$. The first measure ($mcSQI_a$) assesses the similarity between the MECG chest-lead and extracted FECG on the $[0, 100]$ Hz band, whereas the second metric ($mcSQI_b$) focuses on evaluating whether

TABLE II
SUMMARY OF FETAL SQI METRICS USED IN THIS CONTRIBUTION

Cat.	SQI	Mult.	Description	Reference
Time	$stdSQI$	✗	standard deviation of signal, i.e. $stdSQI = std(x(t)) = \sqrt{E\left[(x(t) - \bar{x}(t))^2\right]}$	[10], [7] (other moment)
	$sSQI$	✗	third moment (skewness) of signal, i.e. $sSQI = \frac{E\left[(x(t)-\bar{x}(t))^3\right]}{std(x(t))^3}$	[10], [7]
	$kSQI$	✗	fourth moment (kurtosis) of signal: $kSQI = \frac{E\left[(x(t)-\bar{x}(t))^4\right]}{std(x(t))^4}$	[10], [7]
Frequency	$pSQI$	✗	relative power in the FQRS complex: $pSQI = 1 - \int_5^{15}{}_{Hz}^{Hz} \lvert X(f)\rvert^2 df / \int_5^{45}{}_{Hz}^{Hz} \lvert X(f)\rvert^2 df$, where $X(f) = \mathcal{F}(x(t))$ is the Fourier transform of $x(t)$. Differently from [10], the inverse of the power was applied to represent MQRS suppression.	[10], [7] (inversed)
	$basSQI$	✗	relative power of baseline (bandwidth modified to $[0, 3]$ Hz to include most of the uterine contraction artefacts), i.e. $basSQI = 1 \int_0^{3}{}_{Hz}^{Hz} \lvert X(f)\rvert^2 df / \int_0^{100}{}_{Hz} \lvert X(f)\rvert^2 df$	[10], [7] (modified band)
Detection-based	$bSQI$	✗	percentage of beats commonly detected by two different QRS detectors. This metric is nothing more than an accuracy measure between those detectors. In this work F_1 metric [6] is used.	[10], [7], [26]
	$iSQI$	✓	percentage of beats detected on current lead that were detected on all other leads.	[10], [7]
	$rSQI$	✗	regularity of obtained FQRS intervals $rSQI = 1 - N_{out}/N_d$, where N_{out} is the number of outliers ($FHRV > 30$ bpm) and N_d the total number of detections in the segment	[6], [17]
	$cSQI$	✗	morphology conformity measure for FQRS similarity. Negative correlations were set to zero.	this work, (based on [4])
	$xSQI$	✗	extravagance of FQRS peaks compared to its surroundings.	this work, (based on [4])
FECG-specific	$mxSQI$	✓	analogous to $1 - xSQI$ considering the amplitude of MECG complexes residuals (100 ms window around MQRS reference annotations of ± 50 ms) in comparison with surrounding extracted abdominal signals.	this work
	$mpSQI$	✓	relative spectral power of the first five harmonics of the MHR ($mpSQI_a$) or all harmonics in the interval [0.5,10] Hz ($mpSQI_b$).	this work, (based on [27])
	$mcSQI$	✓	spectral coherence calculated between available signals. Two variants applied: $mcSQI_a$ uses MECG and FECG (0-100 Hz) and $mcSQI_b$ abdECG and FECG (60-100 Hz) as previously explained.	this work
	$miSQI$	✓	similar to $iSQI$ between current FQRS detection and MQRS reference: $miSQI = 1 - iSQI_{MQRS,FQRS}$, aims at exposing falsely detected MQRS residuals.	this work

"Cat." refers to category and "Mult." to the method requiring multiple channels or not (including MECG chest lead). Except for the time domain metrics, all other output belong to $\mathbb{R} \in [0, 1]$. For implementational details please refer to the open-source code [12].

the extraction method has included artifacts on higher frequent spectrum of the extracted signal. The last proposed metric, i.e. *miSQI*, is based on the $iSQI$ [10] and on the premise that in cases where the FECG extraction performs poorly, the MQRS residuals have larger amplitudes then the FQRSs themselves. As a consequence, the sub-optimal extraction leads to MQRS peaks being detected (instead of FQRS), which could be assessed by using the $iSQI$-like algorithm to compare the detected extracted abdominal signal and MQRS reference. Therefore, the metric is calculated using a $\pm$ 50 ms acceptance interval as: $miSQI = 1 - iSQI\big|_{MQRS,FQRS}$.

Since both detection-based SQIs and $miSQI$ metric (presented in Table II) make use of FQRS detectors, their outcomes are dependent on the FQRS detectors' performance. In this work, five publicly available QRS detectors were applied in generating fetal SQIs:

1) **maxsearch** (available from [28], using an expected FHR of 138 bpm as input parameter) searches for an absolute maximum within a pre-defined window. The overall accuracy of this detector on the FQRS annotated subset (see Fig. 1) using F_1 measure [26] was of $F_1 = 64.8\%$.
2) **jqrs** (available from [6]) implementation of the [29] peak energy detector specific for FQRS detection, based on filtering, adaptive thresholding as well as forward and backward search. Average performance of $F_1 = 52.9\%$.
3) **P&T algorithm** (available from [30]) alternative implementation of the Pan-Tompkins algorithm [29]. Average performance of $F_1 = 49.2\%$.
4) **gqrs** (available on PhysioNet [31], [32]) QRS matched filter with a custom-built set of heuristics (such as search back). Input signal was resampled before usage to match faster FHR. Average performance of $F_1 = 55.0\%$.
5) **wqrs** (available on PhysioNet [31]–[33]) detector based on low-pass filtering, a nonlinear curve length transformation and adaptive thresholding. Input signal was resampled before usage to match faster FHR. Average performance of $F_1 = 50.5\%$.

For further reference, the numbers listed above were used to describe variants of SQI metrics to which they are pertained, e.g. $bSQI_{12}$ represents the $bSQI$ evaluated using the maxsearch algorithm [28] and $jqrs$ [6] detectors. Therefore, some of the 14 main features listed in Table II have in fact 2, 5 or $5\times(5-1)/2$ subtypes, leading to a total of 45 features. Every feature was used in the following classification step.

2) Classification - Bayesian Probabilistic Classifier: The classification aims at generating a combined signal quality estimate based on the various SQI features available (described in the preceding section). The resulting SQI is further used within a KF framework to improve FHR estimates. For the purpose of classification, a Naive Bayes classifier was employed. By using the prior and attribute probability densities, the posterior probability for a given class can be modelled. During classification, the class with maximum posterior probability is usually selected [34]. However, in this contribution, we made use of the posterior probabilities to transform the classified values into continuous valued outputs for the following processing, later described in Section II-C.

In order to assess the expected classification performance of the trained Naive Bayes model, i.e. the efficiency in assigning the signal quality based on the used SQIs to the appropriate class, a 10-fold cross validation was performed. Here, all 9,065 observations in the dataset were used to generate training and test sets, respectively. The final Naive Bayes classification model was trained using all available observations as training set. To avoid the trained classifier to be bounded by the training data's class imbalance, the prior class probabilities were assumed to be equal for all classes during performance assessment and when training the final classifier.

C. Improved FHR Estimation

Regarding multichannel HR consensus, there are generally two types of possible combinations, namely channel selection or fusion. In channel selection, the lead with the best quality is selected. For example, Johnson *et al.* [17] applied SQI metrics on 10 s segments of the adult multimodal data during the PCINCC 2014, selecting the channel with highest SQI as output HR. As described by Oster and Clifford [11], several methods have been proposed for fusing multichannel HR estimates using SQI-like metrics. Amongst those are simplistic weighted averages and computationally demanding machine learning approaches. The use of KF is motivated by its well-defined paradigm, which has the advantage of incorporating knowledge of the FHR dynamics as well as the amount of uncertainty in its measurement and intrinsic model. Through their innovation, KF methods can identify trends and abrupt changes in the underlying features not requiring an intensive training period [11]. The use of KFs in improving heart/respiratory rate measures was firstly suggested by [10], [35]. The algorithm can be divided into two stages, namely single-channel HR estimation (using the unfiltered FHR estimates and the classified fetal SQI as input) and multichannel data fusion (using Kalman filtered single-channel estimates, fetal SQI and the innovation signal as byproduct of the KF algorithm). Those steps are briefly explained in the following sub-sections.

1) Single-Channel Kalman Filtering: With both information at hand (i.e. FHR rough estimates and combined SQI), the KF algorithm is applied on a channel basis to improve the single-channel FHR estimates. For the purpose of modelling FHR dynamics, first-order AR models were often applied in the literature [10], [35]. For completeness, we generalize this model with a pth order AR process, which allows the system to have memory. However, in this contribution we restrict ourselves on the first-order model (i.e. $p = 1$). An univariate pth order AR process is described as [25], [36]:

$$\begin{aligned} a_k &= \textstyle\sum_{i=1}^{p} \varphi_{i,k} \cdot a_{k-i} + w_k \,, \\ a_k &= \underline{\varphi}_k^T \cdot [a_{k-1}, \cdots, a_{k-p}]^T + w_k \,, \end{aligned} \tag{1}$$

where p denotes the order of the AR model, $\varphi_{i,k}$ are the time-dependent AR coefficients, and w_k is an additive zero-mean white Gaussian noise process. By applying this modelling into a KF framework, one aims to obtain a linear regression of the p previous FHR measures ($FHR_{k-1} \cdots FHR_{k-p}$) to estimate the current FHR_k, represented in (2):

$$\begin{aligned} \widehat{FHR}_k &= [FHR_{k-1} \cdot \varphi_1 + FHR_{k-2} \cdot \varphi_2 \\ &\quad + \ldots + FHR_{k-p} \cdot \varphi_p] \,. \end{aligned} \tag{2}$$

Equations (1)-(2) can be modelled into KF's space-state representation as in (3):

$$\begin{aligned} \underline{x}_k &= \mathbf{A} \cdot \underline{x}_{k-1} + \underline{w}_k \,, \\ \underline{y}_k &= \mathbf{H} \cdot \underline{x}_k + \underline{v}_k \,. \end{aligned} \tag{3}$$

The state variable $\underline{x}_k \in \mathbb{R}^{p \times 1}$ is defined as the signal of interest (i.e. the FHR p previous estimates), the measurement vector $\underline{y}_k \in \mathbb{R}^{q \times 1}$ contains the observations (rough FHR values), $\underline{w}_k$ is the process noise and $\underline{v}_k$ the observational noise. The filter noise covariance matrices are defined by $w \sim \mathcal{N}(0, \mathbf{Q}_k)$ and $v \sim \mathcal{N}(0, \mathbf{R}_k)$. The state transition matrix $\mathbf{A}$ is the $p \times p$ matrix, describing the expected dynamics of our state (see (4)), while the observational matrix $\mathbf{H}$ is a $q \times p$ null matrix, except for its first element which is unitary.

$$\underline{x}_k = \left[FHR_k, FHR_{k-1}, \cdots, FHR_{k-p+1} \right]^T$$

$$\text{and} \qquad \mathbf{A} = \begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_{p-1} & \varphi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \tag{4}$$

The KF's gain is mostly influenced by its defined process and observational noise covariance matrices ($\mathbf{Q}_k$ and $\mathbf{R}_k$). In order to integrate the information on the signal quality, Li *et al.* [10] proposed modulating the KF's measurement covariance matrix as described in (5).

$$\mathbf{R}_k \equiv \mathbf{R}_0 \cdot exp(1/SQI_k^2 - 1) \tag{5}$$

where SQI_k (with $SQI \in [0, 1]$) denotes the time-dependent estimated signal quality and $\mathbf{R}_0$ is the initial value of measurement covariance matrix, which is signal dependent. The employed nonlinear weighting function leads to $\mathbf{R}_k \to \mathbf{R}_0$ (if $SQI_k \to 1$) and $\mathbf{R}_k \to \infty$ (if $SQI_k \to 0$). The adaptive covariance matrix enhances the influence of cleaner estimates on the filter's outcome, providing the filter with a more rapid response to sudden changes in the signal quality [11]. In this work, to obtain continuous SQI_k values, we propose the use of posterior probabilities $P^+(i, k)$ obtained from the previously depicted classifier. Thus, the continuous-valued SQI output is generated by $SQI_k = \sum_{\{i=A,B,\ldots,E\}} P^+(i, k) * C_{i,k}$, where $C_{i,k}$ represents the numerical coded values for each annotation class $i = \{A, B, ..., E\}$ presented in Table I. The coding $C_i \in [0, 1]$ was defined based on the cumulative distributive function of each class, as further shown in Table III. The initial values for the measurement and process covariance matrices, were empirically determined using a grid-search algorithm (as in [25]) on the first minute of each annotated dataset, for avoiding overtraining. The calibration procedure resulted on $R_0 = 10^{-3}$ and $Q_0 = 1$.

TABLE III
STATISTICS OF ESTIMATED SQI_k VALUES ON SQI ANNOTATED SET

Class	C_i	$\mathbf{mean(SQI_k) \pm std(SQI_k)}$
A	1.00	0.97 ± 0.09
B	0.97	0.84 ± 0.26
C	0.80	0.60 ± 0.34
D	0.47	0.50 ± 0.33
E	0.00	0.08 ± 0.21

C_i represents the continuous coded values attributed to each class, $mean()$ and $std()$ represent the average and standard deviation operations, respectively.

2) Multichannel Data Fusion: After obtaining a Kalman filtered FHR estimate for each available channel, a sensor data fusion step takes place. This multichannel consensus is obtained by using both KF innovation ($\nu_k \equiv \underline{y}_k - \mathbf{H}\underline{\hat{x}}_{k|k-1}$) and consensus SQI signal such that:

$$\theta_k = \sum_{s=1}^{N} \left(\frac{\prod_{i=1, i\neq s}^{N} \sigma_{k,i}^2}{\sum_{i=1}^{N} \left(\prod_{j=1, j\neq i}^{N} \sigma_{k,j}^2 \right)} \cdot x_{k,s} \right) \tag{6}$$

where θ_k represents the resulting FHR estimate from the proposed fusion approach, $x_{k,s}$ is the current FHR estimate for each available channel s at time-step k and $\sigma_{k,s}^2 \equiv (\nu_{k,s}/SQI_{k,s})^2$. This approach was likewise proposed by [10]. For demonstrating the usefulness of innovation signals, the proposed fusion scheme was compared with a simpler approach, namely the weighted average of individual Kalman filtered channels by using solely their SQI values as follows:

$$\theta_k^w = \frac{\sum_{s=1}^{N} SQI_{k,s} \cdot x_{k,s}}{\sum_{s=1}^{N} SQI_{k,s}} \quad . \tag{7}$$

D. Performance Metrics

1) Classification Accuracy: Cohen's κ [37] coefficient is a widely used statistic for reporting classification accuracy. κ has the advantage of correcting for the expected agreement that occurs by chance alone. The κ coefficients are usually reported in the following categories of agreement [37]: "very good" (0.8 to 1.0), "good" (0.6 to 0.8), "moderate" (0.4 to 0.6), "fair" (0.2 to 0.4) and "poor" (< 0.2). However, κ is a measure of nominal agreement that for ordinal-scaled data, such as in this work, causes κ to penalize minor miss classifications (between two adjacent classes, e.g. B and C) with the same weight it penalizes e.g. classifying a value as A instead of E. To cope with those "less/more serious" classification errors, the Krippendorff's alpha coefficient (α) [38] was suggested. Krippendorff's coefficient can be used for any number of raters (not only two), is applicable for different kinds of variables (e.g. nominal, ordinal, interval), and can be used for incomplete or missing data [38]. For this reason, both κ and α are reported in this contribution.

2) FHR Estimation Performance: In this contribution, we aim at window-based FHR estimates (i.e. $\widehat{FHR}_k$) with a 5 seconds moving median window with 1 second overlap. The choice for 5 s window coincides with the length of the segments on the annotated training set, so that the values delivered by the fetal SQIs are similar.

TABLE IV
CONFUSION MATRIX FOR THE 10-FOLD CROSS VALIDATION PROCEDURE EXPLAINED IN SECTION II-B2

		Predicted Class				
		A	**B**	**C**	**D**	**E**
Actual Class	**A**	47	13	2	3	0
	B	116	101	10	63	18
	C	47	116	139	167	158
	D	33	149	76	378	274
	E	16	78	256	371	6434

The heartbeat detection rate (HDR) has been applied in adult HR detection as accuracy metric. HDR assesses the percentage of the HR values within ± 5 bpm tolerance (for adults) [39] of the reference HR annotations, regarded as true positives (TP). On the fetal case, we allowed a tolerance of $\pm$ 10 bpm to reflect the higher FHR (accelerations and decelerations of the FHR are usually defined by changes greater than 15 bpm [40]). SQI results are given in percent by dividing the number of TP by the total number of segments (N_s), as $HDR = 100 \cdot TP/N_s[\%]$ [25].

As for precision metric, the distance between FHR test and the reference values are commonly applied in the literature. Its use is particularly relevant when some averaging window is applied in producing the RR estimates. In this work, the straightforward root mean square error (RMSE) between $\widehat{FHR}_k$ and FHR_{ref} was evaluated [25].

III. RESULTS

A. SQI Evaluation

Table III reports the average and standard deviation statistics for the estimated SQI_k values compared to the annotated reference. The overall classification accuracy in predicting the 5 classes using a Naive Bayes classifier was on average $\kappa = 0.44 \pm 0.03$ (i.e. moderate) and $\alpha = 0.65 \pm 0.04$ (good). The resulting confusion matrix for the cross validation procedure is presented in Table IV). Fig. 3 illustrates the behaviour of some features as well as the resulting continuous-valued SQI, when muscular noise is present.

B. FHR Evaluation

The single channel FHR results before and after applying Kalman filter, as well as multichannel outcomes are presented in Table V. For evaluating the information obtained from the different leads, this procedure was divided into a 3-channel (i.e. using the internal leads configuration), 4-channel (i.e. circular lead system around abdomen) and 7-channel (i.e. all leads). An ideal best possible result is shown by always selecting the channel with maximal HDR and minimal RMSE before and after KF processing. An example of the FHR estimation using KF algorithm and classified SQI is shown in Fig. 4.

IV. DISCUSSION

In this study, a large dataset of annotated NI-FECG signals was used as gold standard. A subset of recordings contain-

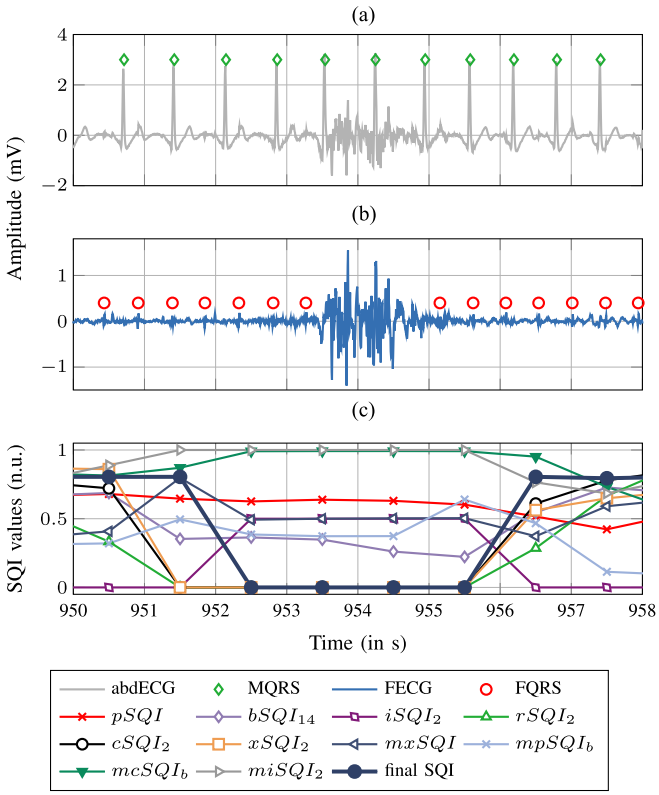

Fig. 3. Segment of clinical data showing a sudden muscular artefact at time 952.5 s. On the top the preprocessed channel is presented, in the middle the extracted fetal signal and below the respective SQI features. "abdECG" denotes the preprocessed abdominal ECG channel.

ing FQRS location annotations with 448 min duration and distinct FECG signal qualities was selected for assessing the performance of FHR estimation. Despite being larger than most datasets available in the literature for FHR analysis, one limitation of our data is the large disparity on the number of observations on each signal quality class, where most data has bad or low signal quality. Additionally, the very reduced number of fetal/maternal arrhythmic events and ectopic beats contained in the training data limits the applicability of the developed classifier in clinical practice. Future studies should address this issue by collecting a large dataset, in order to cover a representative spectrum of relevant arrhythmic situations. A further limitation is the fact that the quality of the NI-FECG recordings was annotated using abdominal recordings (prior to extraction of the FECG), while most SQIs (used to train the classifier) are based on the extracted FECG signals. Therefore, the annotation labels presented in Table I merely reflect if the FECG complexes can be visualized, but not if the FECG is separable/attainable from the abdominal mixture. The separability issue is a relevant topic, particularly when applying Blind Source Separation techniques, that remains for further works. However, such analysis would require the simultaneous recording of a gold-standard such as fetal a scalp electrode to guarantee that segments classified as being of good quality are in fact classified based on underlying FECGs. Consequently, the training data depends on the extraction method used, which is not ideal. Nevertheless, the method of choice for this work, i.e. TS_{pca}, is simple enough and should not produce any major distortion between the annotated and extracted signals. Furthermore, the authors recognize that the defined overall consensus (see Table I) is a subjective concept and may be sub-optimal. However, it was a necessary step for the analysis. Another possibility would be to have two separate sets of SQIs and classifiers, one dealing with the signal SNR and another with the amplitude of the FQRS peaks. Regarding the length of the segments used in this work, although some SQI algorithms may benefit from longer segments (e.g. for building FECG templates or estimating spectral content), the 5 s interval was considered as appropriate for the trade-off between window length and online capability of the proposed algorithm. Further works should focus on expanding these SQI indexes and applying them on a beat-to-beat basis. An exemplary framework for this purpose would be the FQRS detector based on evolutionary computing, previously proposed by the authors in [4].

The Naive Bayes classifier obtained good classification results during cross validation using Krippendorff's α coefficient, i.e. the most suitable metric considering the ordinal dataset used. Meanwhile, Cohen's κ produces moderate results, since it is a nominal agreement measure. Visual inspection of Table III and the confusion matrix (Table IV) confirms that most false classifications fall within neighbouring classes. The misclassifications appear to increase in better quality signals, which can be confirmed by grouping labels into 'acceptable' (labels 'A', 'B', and 'C') and 'unacceptable' ('D' and 'E') from Table IV. By doing so, an accuracy, sensitivity and positive predictive value of 92.1%, 49.3%, and 87.8% would be attained, respectively. This low sensitivity is likely to occur due to the significantly lower number of signals in classes A to C used in training. Furthermore, considering this class imbalance, the dimension of the dataset at hand (i.e. 45 features) in comparison with the number of available samples (i.e. 9065 segments) may affect the classification task unfavourably. Reducing the dimension by applying feature subset selection methods [41] has the potential to improve the classifier's performance. Moreover, it is important to mention that the Naive Bayes classifier assumes the features to be normally distributed and conditionally independent given a class, which is a strong assumption that does not hold for our data. Nevertheless, studies have shown [34] that Bayesian classifiers perform quite well in practice even when attribute dependencies are present. Furthermore, its use is justified by the transparent conversion from discrete to continuous-valued classification results, which was necessary for the further processing. The authors are aware that the underlying class distribution highly impacts the trained classifier in terms of prior class probability. However, the assumption having uniform class distribution is important for obtaining greatest generalization potential with the available data. Unfortunately, there is no annotated clinical database currently available to serve as a standard for comparing our classification results and provide further insights into the signal quality distribution.

Fig. 3 provides an initial intuition on the behaviour of different types of SQI features. From this figure it is visible that most features produce lower values in the presence of muscular noise, while $iSQI_2$ and FECG-specific metrics (i.e. $mxSQI$, $mpSQI_b$, $mcSQI_b$ and $miSQI_2$) output higher values during this period. On the other hand, $pSQI$ does not detect significant changes on the signal. All together, the resulting

TABLE V
COMPARISON FHR ACCURACY AND PRECISION, IN TERMS OF HDR AND RMSE

Method	HDR results (in %)			RMSE results (in bpm)		
	3 channels	4 channels	7 channels	3 channels	4 channels	7 channels
average of individual rough FHR estimates – i.e. $mean\,(\underline{FHR}_k)$	54.4 ± 16.2	58.5 ± 14.5	56.7 ± 13.5	15.5 ± 4.7	14.3 ± 4.2	14.8 ± 4.0
average of individual filtered channels using KF – i.e. $mean\,(\underline{x}_k)$, from (3)	59.1 ± 14.4	63.9 ± 12.1	61.8 ± 11.5	14.4 ± 4.2	13.1 ± 3.7	13.6 ± 3.6
multichannel fusion using weighted average – i.e. θ_k^w, from (7)	61.8 ± 18.0	71.6 ± 16.6	75.0 ± 14.1	13.9 ± 5.0	11.4 ± 4.2	10.8 ± 4.0
multichannel fusion using proposed KF approach – i.e. θ_k, from (6)	$\underline{64.2 \pm 16.0}$	$\underline{73.8 \pm 15.0}$	$\underline{75.6 \pm 13.4}$	$\underline{13.3 \pm 4.6}$	$\underline{10.8 \pm 4.1}$	$\underline{10.5 \pm 4.0}$
theoretical best individual unfilitered channel – i.e. $\max\,/\,\min\,(\underline{FHR}_k)$	59.3 ± 17.4	69.3 ± 17.7	71.2 ± 16.7	14.2 ± 5.0	11.5 ± 4.9	11.1 ± 4.7
theoretical best individual KF channel – i.e. $\max\,/\,\min\,(\underline{x}_k)$	65.0 ± 14.6	75.7 ± 13.0	77.0 ± 12.3	13.0 ± 4.5	10.1 ± 4.1	9.8 ± 4.0

Results presented as average $\pm$ standard deviation. The best mean results in each category are underlined. Theoretically best achievable results are shown at the bottom of the table, by taking the maximum/minimum results amongst different channels for each segment using the reference.

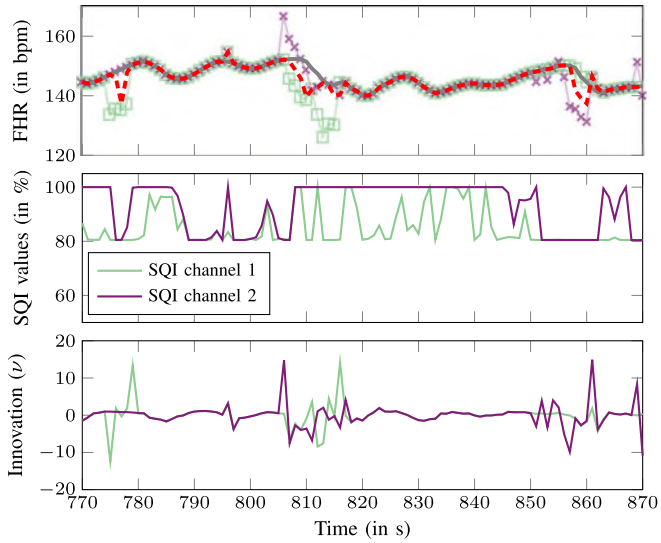

Fig. 4. Multichannel FHR estimation by means of Kalman filtering using one external and one internal channel. Above FHR estimation is shown with the reference FHR (——), estimated multichannel FHR (- - - -), single channel rough FHR estimates (□ and ×) and the respective single channel Kalman filtered estimates (—— and ——). In the middle and below are presented the SQI and innovation values for the 2 channels used, respectively.

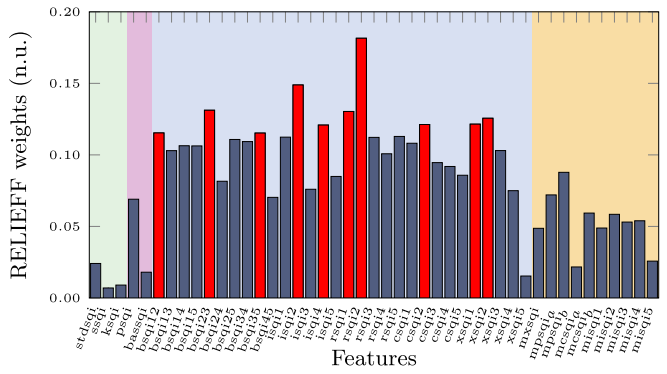

Fig. 5. Bar graph showing the feature importance using RELIEFF. The different background colors on the graphic denote the different groups of SQI metrics presented in Table II, while the 10 features with highest results are emphasized.

continuous-valued SQI showed to be very sensitive to such artefacts and clearly detected the abrupt change in quality. In order to obtain a further insight into the importance of individual variables independent from the applied classifier, the RELIEFF filter method [42] was applied (see Fig. 5). This algorithm, often employed for feature selection, assigns weights to the individual features according to their class separation capabilities. The number of nearest class hits and misses for the algorithm considered during weight computation was set to $k = 50$ nearest neighbours of each class (approximately the number of observations on the lowest frequent class) and prior was defined as uniform, so that our analysis does not depend on the presented class distribution (as fetal signal quality may depend on several factors, e.g. gestational week). From Fig. 5 it is evident that time and frequency metrics (see Table II) showed little importance, which can be explained by the well-known similarities and spectral overlap between abdominal ECG/MECG/FECG. Detection-based algorithms such as $bSQI$, $rSQI$, $cSQI$ and $xSQI$ (particularly when using $maxsearch$ or $jqrs$ detectors) were relevant. While applying $bSQI$, it is particularly important to use a combination of a more and other less predictive detector, e.g. $maxsearch/wqrs$ (similar to the results obtained in [17]). This result goes along with the authors' previous works and top scoring entry on the PCINCC 2013 [4], where features like $cSQI$, $xSQI$, and $rSQI$ were responsible for accurate FQRS detections. The proposed FECG-specific features (latter group in Fig. 5) showed modest importance. Amongst these, $mpSQI_b$ was deemed as most important by RELIEFF in the latter SQI category (see Fig. 5) that can be confirmed by its moderate reaction to the presence of noise (on Fig. 3). Despite being a vital part of machine learning, the aforementioned feature selection step was not included into this study's methodology because the focus of this contribution was the development of FECG-specific SQIs, rather than fine-tuning the classification method applied. Moreover, such additional step would imply further assumptions about the data or the classification algorithm used, which would reduce this work's generalizability.

The overall modest FHR accuracy results (see Table V), demonstrate how challenging FHR estimation on clinical data really is. Behar *et al.* [26] compared several extraction methods, including TS_{pca}, using 42 min of manually annotated abdominal signals from the Physionet's Non-Invasive Fetal Electrocardiogram Database [31] and 40 min of a private commercial database using a scalp electrode recording as reference. Results

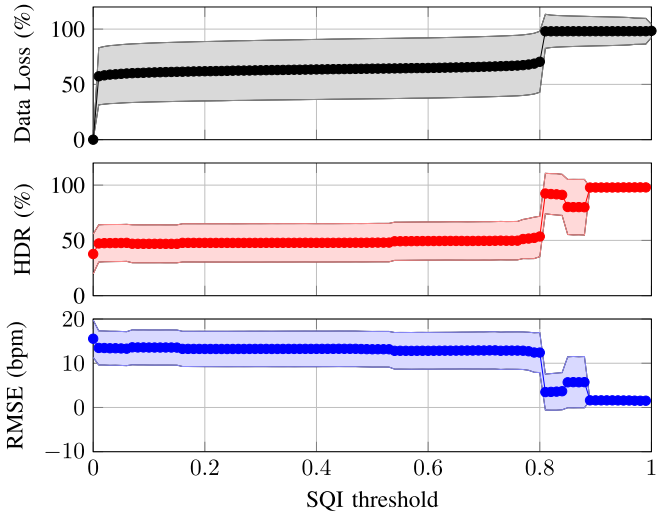

Fig. 6. HDR and RMSE average results for FHR estimation after excluding segments with SQI lower than a given threshold (x-axis). Above the amount of data loss, in the middle HDR and below RMSE results for the remaining segments.

for FHR accuracy applying TS_{pca} were 68.7 % and 73.9 % for each dataset (using a ± 5 bpm acceptance interval). Differently from many FHR studies that regard periods of "signal loss", in this contribution, segments/recordings with general bad quality were not discarded. In long-term recording scenarios, removing portions of data with bad signal quality from further analysis is desirable because FHR estimates during periods of low SQI would disregard the current measurements and follow the filter's dynamic equations (first-order AR process), therefore, delivering unreliable results. The manner with which selection of inadequate segments may be performed is another complex topic, which deserves its own study. Nonetheless, Fig. 6 provides initial insights on the potential that such application may have.

The multichannel estimation of FHR using Kalman filter showed the best performance both in terms of HDR and RMSE (see Table V). The filter's performance monotonically grew with the increase in the number of available leads, which shows how powerful the method is in incorporating additional information. Meanwhile, the average result from single channel estimation (with or without KF) does not show this trend. Large differences were found between the 3 and 4-lead (i.e. internal or external) schemes. The latter performed better, which can be attributed to a lower presence of noise for greater inter-electrode distance between exploring and indifferent electrodes. After the calibration procedure, the smaller value obtained for the initial observational noise covariance matrix $R_0 = 10^{-3}$ compared to $Q_0 = 1$ shows that the filter tends to "trust" its observations associated with the SQI metrics. Additionally, Fig. 4 provides a qualitative example using both external and internal leads with varying channel qualities. As it can be seen from this figure, the individual rough FHR estimates are inherently inaccurate. However, if the quality is sufficient in some channels, the proposed KF approach is able to reliably reflect the true FHR. Therefore, it is clear that the KF innovation in association with the proposed SQI metrics is able to improve FHR estimation, as it did in estimating adult heart rates [10].

In this work, we implemented multiple KFs running parallel for estimating each of individual channel FHR. The multichannel fusion was considered as an additional step following this single-channel estimation. Oster and Clifford [11] proposed combining both single-channel FHR estimation and data fusion steps into a single step, rather than implementing multiple KF and combining those later on. This is performed by considering the consensus amongst those different sensors as an additional Kalman state and allowing the transition and observational matrices ($\mathbf{A}_k$ and $\mathbf{H}_k$) to be time-variant and dynamically include the previously defined weighting factors $\sigma^2_{k,s}$ (see (6)).

V. Conclusion

In this contribution multiple fetal SQI metrics were investigated and applied in a Naive Bayes classifier for estimating the quality of fetal signals in 5 s segments. This classifier was then used in association with a Kalman filter algorithm to improve online FHR estimation from multichannel non-invasive fetal ECG recordings. Results indicate a set of SQI features that have more importance in our classification and suggests that the proposed SQI-Kalman filter fusion produces accurate FHR estimates. Furthermore, for instigating reproducible research in NI-FECG field, the attained classifier as well as SQI algorithms used in this study were released as part of the FECGSYN toolbox under http://www.fecgsyn.com/.

Acknowledgment

The authors would like to thank Prof. Holger Stepan, Dr. Alexander Jank, Sophia Schröder, Susanne Fritze, and Julia Kage from the University Hospital of Leipzig, for providing data and careful annotations.

References

[1] R. Sameni and Gari D. Clifford, "A review of fetal ECG signal processing; issues and promising directions," *Open Pacing Electrophysiol. Ther. J.*, vol. 3, p. 4, 2010.
[2] G. D. Clifford *et al.*, "Non-invasive fetal ECG analysis," *Physiol. Meas.*, vol. 35, no. 8, 2014, Art. no. 1521.
[3] J. Behar *et al.*, "A practical guide to non-invasive foetal electrocardiogram extraction and analysis," *Physiol. Meas.*, vol. 37, no. 5, pp. R1–R35, 2016.
[4] F. Andreotti *et al.*, "Robust fetal ECG extraction and detection from abdominal leads," *Physiol. Meas.*, vol. 35, no. 8, pp. 1551–1567, 2014.
[5] M. Varanini *et al.*, "An efficient unsupervised fetal QRS complex detection from abdominal maternal ECG," *Physiol. Meas.*, vol. 35, no. 8, pp. 1607–19, 2014.
[6] J. Behar *et al.*, "Combining and benchmarking methods of foetal ecg extraction without maternal or scalp electrode data," *Physiol. Meas.*, vol. 35, no. 8, pp. 1569–1589, 2014.
[7] G. D. Clifford *et al.*, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiol. Meas.*, vol. 33, no. 9, pp. 1419–1433, 2012.
[8] J. Behar *et al.*, "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1660–1666, Jun. 2013.
[9] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.
[10] Q. Li *et al.*, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter." *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, 2008.

[11] J. Oster and G. D. Clifford, "Signal quality indices for state space electrophysiological signal processing and vice versa," in *Advance State Space Methods Neural Clinical Data*, 2015, p. 345.
[12] J. Behar *et al.*, "An ECG model for simulating maternal-foetal activity mixtures on abdominal ECG recordings," *Physiol. Meas.*, vol. 35, no. 8, pp. 1537–1550, 2014. [Online]. Available: http://www.fecgsyn.com
[13] F. Andreotti *et al.*, "An open-source framework for stress-testing non-invasive foetal ECG extraction algorithms," *Physiol. Meas.*, vol. 37, no. 5, pp. 627–648, 2016. [Online]. Available: https://physionet.org/physiobank/database/fecgsyndb/
[14] D. Hoyer *et al.*, "Monitoring fetal maturation - Objectives, techniques and indices of autonomic function," *Physiol. Meas.*, vol. 38, 2017. [Online]. Available: http://iopscience.iop.org/article/10.1088/1361-6579/aa5fca.
[15] I. Silva *et al.*, "Improving the quality of ECGs collected using mobile phones: The physionet/computing in cardiology challenge 2011," in *Proc. 2011 Comput. Cardiol.*, 2011, pp. 273–276.
[16] G. D. Clifford *et al.*, "False alarm reduction in critical care," *Physiol. Meas.*, vol. 37, no. 8, pp. E5–E23, 2016.
[17] A. E. W. A. Johnson *et al.*, "Multimodal heart beat detection using signal quality indices," *Physiol. Meas.*, vol. 36, no. 8, pp. 1665–77, 2015.
[18] Q. L. *et al.*, "A machine learning approach to multi-level ECG signal quality classification," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 435–447, 2014.
[19] G. B. Moody and R. G. Mark, "QRS morphology representation and noise estimation using the Karhunen-Loeve transform," in *Proc. Comput. Cardiol.*, 1989, pp. 269–272.
[20] G. D. Clifford *et al.*, *Advanced Methods and Tools for ECG Data Analysis*. Norwood, MA, USA: Artech House, 2006.
[21] M. A. F. Pimentel *et al.*, "Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices," *Physiol. Meas.*, vol. 36, no. 8, pp. 1717–1727, 2015.
[22] Q. Li *et al.*, "Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator," *Biomed. Eng. Online*, vol. 8, 2009, Art. no. 13.
[23] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiol. Meas.*, vol. 33, no. 9, pp. 1491–1501, 2012.
[24] C. Orphanidou *et al.*, "Signal Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 832–838, May 2015.
[25] F. Andreotti *et al.*, "Improved heart rate detection for camera-based photoplethysmography by means of Kalman filtering," in *Proc. IEEE 35th Int. Conf. Electron. Nanotechnol.*, Kiev, Ukraine, 2015, pp. 428–433.
[26] J. Behar *et al.*, "A Comparison of Single Channel Fetal ECG Extraction Methods," *Ann. Biomed. Eng.*, vol. 42, no. 6, pp. 1340–1353, 2014.
[27] G. de Haan and V. Jeanne, "Robust pulse-rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.
[28] R. Sameni, "The open-source electrophysiological toolbox (OSET)," 2010. [Online]. Available: http://www.oset.ir
[29] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.
[30] D. Wedekind, "MATLAB script of the pan-tompkins QRS detector," 2014. [Online]. Available: https://github.com/danielwedekind/qrsdetector
[31] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, 2000.
[32] I. Silva and G. B. Moody, "An open-source toolbox for analysing and processing physionet databases in MATLAB and Octave," *J. Open Res. Softw.*, vol. 2, no. 1, p. e27, 2014.
[33] W. Zong *et al.*, "A robust open-source algorithm to detect onset and duration of QRS complexes," in *Proc. 2003 Comput. Cardiol.*, pp. 737–740, 2003.
[34] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Mach. Learn.*, vol. 29, pp. 103–130, 1997.
[35] L. Tarassenko *et al.*, "Multi-sensor fusion for robust computation of breathing rate," *Electron. Lett.*, vol. 38, no. 22, 2002, Art. no. 1314.
[36] M. Arnold *et al.*, "Adaptive AR modeling of nonstationary time series by means of Kalman filtering," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 5, pp. 553–562, May 1998.
[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[38] A. F. Hayes and K. Krippendorff, "Answering the Call for a Standard Reliability Measure for Coding Data," *Commun. Methods Meas.*, vol. 1, no. 1, pp. 77–89, 2007.
[39] "Cardiac monitors, heart rate meters, and alarms," American National Standards Institute, Association for the Advancement of Medical Instrumentation, Washington, DC, USA, Tech. Rep. EC13, 2002.
[40] American College of Obstetricians and Gynecologists, "ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles," *Obs. Gynecol*, vol. 114, no. 1, pp. 192–202, 2009.
[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[42] I. Kononenko *et al.*, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, 1997.

Fernando Andreotti received the Automation and Control Engineering degree from the Universidade Federal de Minas Gerais (Brazil) in 2011 and Ph.D. in Electrical Engineering degreee from the Technische Universitt Dresden (Germany) in 2017.

His research focuses on Kalman filtering, signal processing, and modelling of cardiac signals with particular focus to non-invasive fetal ECG analysis.

Felix Gräßer received the Diplomingenieur degree in mechatronics from Technische Universität Dresden (TUD), Dresden, Germany, in 2012. In 2015, he joined the Institute of Biomedical Engineering of TUD, where he is currently working toward the Ph.D. degree in biomedical engineering. His research interests include recommender systems and machine learning in health related applications, particularly clinical decision support systems.

Hagen Malberg received the Ph.D. degree in 1999. From 1999 to 2010, he was the Head of Biosignal Processing Group (Karlsruhe Research Center and KIT Karlsruhe Institute of Technology). Since 2010, he has been the Chair of Biomedical Engineering, Technische Universität Dresden, Dresden, Germany. His research interests include biosignal analysis, particularly related to the regulation of the cardiovascular system.

Sebastian Zaunseder received the Ph.D. degree in electrical engineering from Technische Universität Dresden (TUD), Dresden, Germany, in 2011. He joined the Institute of Biomedical Engineering of TUD, where he is currently the Head of the Group Medical Sensing and Signal Processing. His research interests include contact-free measurement systems, processing of biomedical signals and images to acquire robust information on vital signs, investigations on the cardio-respiratory autonomic modulation and research related to sleep.