SPRINGER BRIEFS IN BIOENGINEERING

Christina Orphanidou

Signal Quality Assessment in Physiological Monitoring

State of the Art and Practical Considerations

Springer

SpringerBriefs in Bioengineering

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include: A timely report of state-of-the art analytical techniques, a bridge between new research results, as published in journal articles, and a contextual literature review, a snapshot of a hot or emerging topic, an in-depth case study, a presentation of core concepts that students must understand in order to make independent contributions.

More information about this series at http://www.springer.com/series/10280

Christina Orphanidou

Signal Quality Assessment in Physiological Monitoring

State of the Art and Practical Considerations

Springer

Christina Orphanidou
Department of Electrical and Computer Engineering
University of Cyprus
Nicosia
Cyprus

ISSN 2193-097X ISSN 2193-0988 (electronic)
SpringerBriefs in Bioengineering
ISBN 978-3-319-68414-7 ISBN 978-3-319-68415-4 (eBook)
https://doi.org/10.1007/978-3-319-68415-4

Library of Congress Control Number: 2017954278

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Signal Quality Assessment in Physiological Monitoring: Requirements, Practices and Future Directions

Abstract The emergence of telehealth and telemedicine systems that support the continuous monitoring of patients via wearable sensors has altered the landscape of healthcare, providing solutions to many of the open challenges of disease management. Signal Quality Assessment (SQA) systems aim to improve the reliability of physiological measurements obtained from signals recorded via wearable sensors which are more prone to artefacts. This chapter begins by making the case for SQA systems, as healthcare monitoring rapidly advances towards a wireless digital future. A review of system requirements and considerations is then provided, before a discussion on the future challenges in the design of SQA systems which need to be overcome such that the performance of such systems may be improved.

Keywords Signal quality assessment · Continuous monitoring · Wearable sensors · Telehealth · Telemonitoring

Before attempting to build a signal quality assessment system for physiological signals it is important to first understand the underlying principles and practical considerations for the implementation of such systems in a real-world healthcare environment. In this chapter, we will first make the case for Signal Quality Assessment (SQA), as healthcare monitoring rapidly advances towards a wireless digital future. The main requirements of SQA systems in the context of physiological monitoring will then be discussed, focusing on the major application of such systems, which is the continuous monitoring of vital signs via telemetry. An overview of the framework for developing current approaches will then be given which can be applied to many physiological signals collected in real-life hospital environments. Although each type of signal requires its own quality appraisal tools and techniques, the key ideas that underpin them are common. Later in the chapter, we will discuss the technical challenges of SQA systems and the future directions of SQA systems such that these challenges can be overcome, including recent approaches using data fusion of multiple simultaneously recorded signals. These approaches build on the heterogeneity of the failure mechanisms of the different signals, to improve system performance and

C. Orphanidou, *Signal Quality Assessment in Physiological Monitoring*,
SpringerBriefs in Bioengineering, http://doi.org/10.1007/978-3-319-68415-4_1

increase the reliability of vital sign measurements. Overcoming the existing challenges has the potential to yield results with a genuine impact on the healthcare management of the future.

1.1 The Case for Signal Quality Assessment

It is widely accepted that the current model of healthcare delivery is unsustainable: 37% of the European population is expected to be aged 60 or over by 2050 [1] and the prevalence of chronic diseases is ever-increasing. The proportion of wealth consumed by healthcare in developed nations is projected to be unsustainable, mainly due to chronic illness, which will account for 80% of the growth [2]. The constraints imposed on public finances by these changes put enormous pressure on healthcare systems to deliver more and better care, with reduced resources. For this to be achieved, care systems need to be redesigned and reorganised; technology has a vital role to play in catalysing the change that is urgently required. Of primary importance are (a) the shift from treatment towards prevention and early intervention, i.e., intervening before a condition becomes unavoidably serious, or providing specialist intervention early in the development of a long-term condition and (b) the shift of care away from the hospital to the home which will result in a reduction in healthcare costs, improved health outcomes and an improvement in the quality of life of the individual.

1.1.1 Continuous Physiological Monitoring

Technologies allowing the continuous recording of physiological signals, either via telemonitoring or via wired technologies at the bedside have emerged in the recent years as a response to the need for improved healthcare delivery and as a result of advancements in information and communication technologies. Wearable sensors, which enable the delivery of continuous vital sign measurements, while granting users the freedom of movement (and thus enabling a quicker recovery [2]) will undoubtedly be a key point of delivering healthcare in the future. For example, the development of efficient and effective home-based telemonitoring systems will provide a solution to address the challenges of sustainable healthcare and are particularly suitable for the management of chronic diseases. Via the frequent monitoring of patients using unobtrusive sensors, early prediction of deteriorating health is made possible outside the hospital environment, triggering interventions for the avoidance of severe episodes.

Wearable sensors have been used as part of routine healthcare for decades. In as early as 1906, Einthoven transmitted the first electrocardiographic representation (what is now known as the Electrodardiogram (ECG)) by directly connecting his string galvanometer to telephone lines [3]. In 1947, Norman Holter, transmitted a

recording of his ECG via radio, whilst cycling on a stationary bicycle. C*ardiac telemetry* devices, like the Holter monitoring device, are now routinely used to monitor the heart's electrical activity for patients who are at elevated risk of developing cardiac problems by identifying, for example, instances of cardiac arrhythmia, such as atrial fibrillation (AF). Nowadays, advances in information and communication technologies and electronics, as well as the widespread adoption of wireless networking technologies have led to the development of a new generation of wearable devices that can transmit data either via Wi-Fi or Bluetooth, vastly increasing the number of settings and environments in which they can be used.

Outside the Intensive Care Unit (ICU), the traditional monitoring of a person's health status has, until recently, consisted of periodic measurements of vital signs, commonly heart rate (HR), peripheral oxygen saturation (SpO_2), blood pressure (BP), breathing rate (BR) and temperature (T). While the frequency of taking vital sign observations by nursing staff varies, in a typical medium-to-high acuity hospital ward the frequency is typically every 4 hours [4]. Abnormal vital signs developing in the periods between these checks, which may be precursors of adverse events could possibly be missed.

In the hospital environment, continuous monitoring systems via wearable sensors address this clinical need for patients who are high-risk. Via the use of wearable monitoring systems which process signals and transmit measurements wirelessly in real-time, patients can benefit from the continuous monitoring of their heart and respiratory activity [5], while maintaining the freedom of movement [2].

Outside the hospital environment, wireless technology facilitates the development of remote health monitoring systems [6], monitoring individuals at home who either live in remote locations, or require regular checks, such as the elderly, people with chronic conditions or people with disability. M-health monitoring systems have also experienced a boom; these systems integrate smartphone applications with health sensors to track levels of physical activity, cardiac activity, mental health and Parkinson's disease, amongst others [7, 8]. Continuous physiological monitoring has also found applications in health surveillance for military, rescue and sports personnel [5, 9].

Although the ECG has long been considered the "gold standard" for the measurement of HR, the photoplethysmogram (PPG) is also a well-established source of HR (or pulse rate) information. Since the ability to obtain multiple vital signs from a single, unobtrusive, peripheral sensor is desirable [10] and SpO_2 is also measured from the PPG signal, many wearable monitoring systems now rely on the measurement of the PPG signal; that is the technology mainly used in currently marketed wrist-worn physiological monitors. Breathing rate (BR) is currently being measured using techniques which either interfere with normal breathing (nasal thermistors, carbon dioxide sensors) or which may cause discomfort (transthoracic inductance, pneumography and impedance plethysmography) [11]. These techniques are unsuitable for ambulatory monitoring; as a result, a lot of research activity has been recently directed to developing techniques for indirectly deriving respiration rate from the ECG and PPG signals. Especially relevant to wrist-worn monitors, achieving reliable derivation of BR from the PPG would be highly

beneficial, since all main vital signs would be able to be derived from a single unobtrusive sensor [10]. The reliability of these measurements depends on the quality of the recorded signals so quality assessment is of paramount importance.

Outside the context of clinical monitoring, wearable sensors have gained popularity in the past few years in the wellness and fitness markets. In fact, by 2020 it is projected that 411 million wearable devices will be sold globally [12]. Wearable fitness trackers usually combine a heart rate monitor and accelerometer and derive vital sign parameters which do not, however, need to be accurate to clinical standards [13]. Additionally, their hardware does not need to be certified as a medical device.

1.1.2 Obstacles to Clinical Adoption

Despite the proliferation of wearable health and wellness sensors witnessed in the past few years and the move of the healthcare community towards 24-h continuous monitoring, most of wearable monitoring systems proposed in the literature have still not found their way into widespread clinical use [14]. Although the technology has experienced significant advancements over the last decade, vital sign measurements are still largely unreliable when moving outside controlled clinical environments and into real-life scenarios [6]. Some of the challenges that need to be overcome are the reliability of wireless technology, quality of measurements, power management of wearable monitors, security and patient confidentiality [6]. Patient and clinician acceptance is an important barrier for the widespread adoption of such systems. For patient acceptance to be achieved, unobtrusiveness and ease-of-use are important challenges that need to be met [15]. For clinician acceptance to be achieved, the technology needs to advance to a reliable level. These two requirements are often at odds with each other since the ergonomic requirements of patient comfort often result in a loss of signal quality. Research indicates that monitoring systems which are deemed obtrusive and/or difficult-to-use result in reduced patient compliance [13]. To increase patient compliance, sensors are being progressively miniaturized; "digital patch" devices, smaller than a matchbox, are now being developed which can be worn unobtrusively under clothing for up to a fortnight without needing to be removed. "Smart" garments, where sensors are embedded into clothing or held in special pockets are also being currently developed, but still remain within the domain of research [16]. Ambient and contactless monitoring systems (for example via the iPPG) are also gaining momentum [17]. However, as monitoring systems become progressively smaller and lighter, the quality of the recordings decreases. In general, data in recordings obtained from ambulatory patients are corrupted by motion artefact and are noisier than data obtained from non-mobile patients. For systems where data collection is to take place without clinical supervision, signal quality and reliability of measurements derived from signals obtained via wearable sensors are of utmost importance for patient and clinician acceptance to be achieved. Artefact-corrupted signals lead to erroneous

vital sign measurements. An undesirable effect may be that abnormal vital sign measurements, that are significant predictors of physiological exacerbation and mortality in hospital patients [18, 19], go unnoticed, thus compromising patient care and health outcomes. Another undesirable effect may be the occurrence of a high number of false alerts, leading to the phenomenon of "alarm fatigue", whereby clinical staff become desensitized to, and ultimately ignore, alerts from monitoring systems [6]. "Alarm fatigue" widely recognised as a major medical device technology hazard [20] is partly caused by the need for high sensitivity (i.e., not missing a serious event). Incorporating signal quality considerations into "smart" alarm systems is an obvious remedy for alleviating this undesirable phenomenon [13, 14, 20].

1.1.3 Signal Quality Assessment

It is becoming evident that incorporating signal quality assesment (SQA) strategies in recorded physiological signals is a crucial step towards the successful deployment of continuous health monitoring systems, most importantly as part of telemonitoring applications. While a lot of research activity has been devoted to techniques for suppressing noise to improve the reliability of measurements, sometimes it is not possible to extract a valid measurement from an artefact-corrupted signal; detecting noisy segments is critical and should be addressed first. By detecting and ignoring unreliable signal the reliability of measurements will improve, resulting in improved outcomes. This will, in turn, lead to patient and clinician acceptance and eventually to the widespread adoption of continuous health monitoring systems.

1.2 Design Considerations and Requirements

Before describing approaches for providing signal quality assessment it is important to consider many issues that may impact the overall system design and its potential use.

The aim of quality assessment systems is to identify instances of artefact (or "noise") in a segment of signal such that the vital sign measurements extracted from that segment can be ignored or corrected/enhanced. The terms *quality assessment*, *quality appraisal* and *artefact/noise detection* are considered equivalent thereof. Artefacts are essentially signal interferences with randomly varying amplitudes, frequencies and duration that corrupt physiological signals recorded in clinical settings [21]. They are most commonly the result of motion and can be manifested on the signal as low- or high-frequency noise which corrupts its morphology. Figure 1.1 shows examples of ECG signals with distinct kinds of artefact.

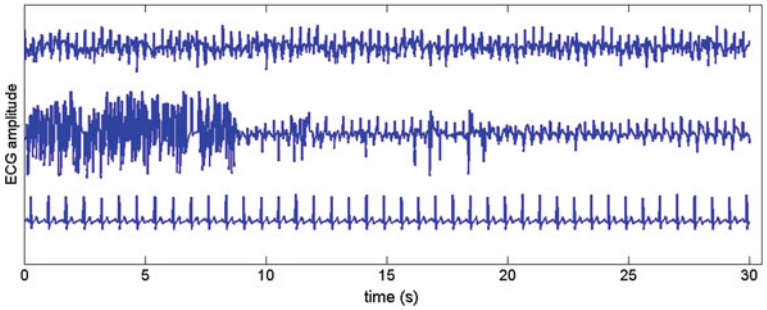

Fig. 1.1 Example ECG signals obtained from wearable sensors. *Bottom* clean ECG signal, *Middle* signal corrupted by artefact that would result in erroneous HR measurement, *Top* signal with low-frequency motion artefact (baseline wander). This is a borderline case since the segment might be suitable for measuring HR within a clinically-acceptable margin of error, but would not be suitable for diagnostic applications

There is no general definition of signal quality since quality requirements are relevant to the application at hand. For example, monitoring systems for high-risk patients, e.g., in critical care wards, have more stringent reliability requirements than fitness/athletic performance monitoring systems. Additionally, quality requirements differ depending on what the signal is intended to assess. For example, if a recorded segment of ECG is only needed for extracting a valid HR measurement, the quality criterion will be more relaxed compared to a diagnostic application intended to determine the presence of coronary heart disease (CHD). That is because for the measurement of HR, precise beat morphology is not important as long as R-peaks can be clearly identified, whereas for diagnosing CHD, beat morphology needs to be clear. Defining the quality requirement for the application at hand is, thus, an important first step towards designing an efficient and effective SQA system. Related to this, is the choice of population demographic and labeling strategy for building the system.

An important consideration when designing SQA systems is the ability to detect and remove artefacts amongst physiological signal variability and pathophysiological changes [22]. When artefacts mimic physiologic data the problem of separating the two becomes complex [23, 24], creating the risk of confusing signal corruption with the existence of a physiological abnormality, e.g., erroneously identifying a segment exhibiting arrhythmia as noise [25]. This is an important consideration when choosing which features to extract from the signal and what their limits of normality are, in terms of signal quality.

SQA systems need to maintain the time resolution required by the clinical application they are intended for [22]. In the context of continuous physiological monitoring, most often, this requirement is for the system to work in real-time. Especially when used in critical care wards, systems need to allow the real-time identification of quickly developing changes in signals which may signify an adverse health event. The required time resolution will not only advice the permitted

computational cost of the system but also the window of signal required for the signal quality assessment. Using shorter time-windows (e.g., 10 s) may limit the type of features that can be used (if using trend-based approaches, for example) and the information that can be extracted (e.g. heart rate variability (HRV) information requires a larger window of signal for reliable extraction) but increases the chance of obtaining at least one reliable HR value at a satisfactory frequency. Longer time-windows (e.g. 60 s) permit the use of trend-based approaches and the extraction of HRV information but may result in increased data retention, since, depending on the quality decision criterion, entire 60 s windows may be rejected, despite possibly containing shorter segments of reliable signal.

In terms of system performance, the ability to reject unreliable signals is preferable to the ability to retain reliable ones; however, a system designed for exhibiting high sensitivity would result in unnecessarily high data retention, which might create large gaps between obtaining reliable measurements. This may be undesirable in high acuity cases where obtaining "mildly erroneous" measurements frequently might be preferable. A system designed for exhibiting high specificity, on the other hand, may result in a high number of false alerts or the possibility of an adverse event being missed which is also undesirable, as explained previously.

When SQA systems are incorporated into wearable monitors (on-board processing), generalizability of the methods to different sensors might not be necessary; if they are incorporated into hospital software (i.e. assessment happens after transmission or off-line) then it is important for methods to be able to generalize to different sensors which often include proprietary software with different "black-box" pre-processing steps.

The decision of on-board, remote, or off-line processing is also important with respect to the computational load of the proposed system. If on-board processing is required, the power consumed by a computationally expensive algorithm may require excessive changing or charging of batteries which may lead to disruptions in recordings and system failures [26]. The computational cost is less important when using remote or off-line processing.

1.3 Basic Design Framework

The common approach for performing quality assessment of a signal is shown in Fig. 1.2. In the first step, appropriate features are extracted from the signal which are likely to differ between clean and artefact-corrupted segments; these may be individual morphological features or trends within a time-window of signal. Pre-processing steps (like filtering to a physiologically probable range) are often applied at this step. In the second step, the features are subjected to one or more decision rules which may be determined heuristically, via trial and error, or may be learnt using large quantities of expertly labeled data.

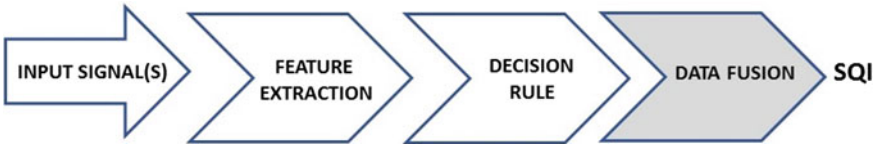

Fig. 1.2 Flowchart of generalized procedure for performing quality assessment and obtaining a Signal Quality Index (SQI). The *shaded step* is optional and is used in approaches utilizing multiple signals or multiple channels of the same signal

Often, SQA systems consist of a cascade of decision steps, based on a variety of features and decision rules starting from simple viability checks (such as the detection of flat lines, signal saturation or obvious outliers) and progressing to more sophisticated checks for making assessments on the more ambiguous signals. The output of this procedure is often called a Signal Quality Index (SQI) which is binary in most cases ("acceptable/reliable" or "unacceptable/unreliable") but can also assign degrees of acceptability to a segment [27]. Multiparameter approaches, combining quality information from multiple channels of the same signal or multiple signals, additionally include a data fusion step where individual quality indices are combined to provide the overall SQI of a signal segment or a vital sign measurement.

1.3.1 Population Demographics and Data Specifications

Before collecting data for building the SQA system, the population demographic needs to be carefully considered. Subjects on specific medications or with pre-existing conditions may have different underlying physiology which may affect signal morphology and trends [28]. Younger subjects will have different physiological characteristics compared to older subjects. Also, the level and type of activity is important; most researchers use protocols with specific voluntary activities (such as hand motions, walking and exercising) whereas others use data collected from subjects going about with their daily activities. While the latter approach will contain more realistic artefact, the risk exists of ending up with very little artefact-corrupted data and a very unbalanced dataset for training an SQA algorithm. Additionally, when data collection follows a specific experimental protocol, labeling will be more reliable and knowledge may be gained on the effect of specific actions on the quality of the recorded signals. Lastly, sampling rate and duration of recording need to be considered, depending on the intended analysis and application.

1.3.2 Establishment of Ground Truth

The problem of automating the assessment of signal quality can be best addressed by considering what the automated system is intended to replace. In the absence of such an automated system, unreliable measurements would be rejected by manual assessment by a human expert (most often a clinician). The criterion for assessment depends on the clinical scenario and the specific requirements of the system it is intended for. Any proposed system needs to be built and validated on labelled data and its performance will be tailored to the labelling criteria. Most validation datasets are labeled by expert annotators based on specific rules provided by the developers. The labelled dataset is then called the "ground truth" and can be used for determining thresholds of decision rules, or building machine learning models, and for testing the performance of the proposed system. The frequent problem of inter-rater variability [29] can be addressed by using multiple annotators and having ambiguous segments reviewed by additional annotators until a consensus is established. The labeling rule needs to be clearly explained to the human annotators such that labeling is consistent. Such a rule can range from "the signal can be used to derive a valid HR" or "the signal is clinically usable" to "the signal can be used for a full diagnostic assessment". Rule-based SQA systems need to be tailored to the quality requirements of the clinical application via the same rule; data-based approaches, incorporating learnt thresholds or machine learning approaches will be automatically tuned to the labeling rule which was used, without the need for further processing. When developing an SQA system, it is important to use datasets which were labeled following the same labeling criteria; failure to do so, may result in decreased performance of the system and the failure to generalize.

1.3.3 Performance Evaluation

The most common way of providing assessments for the performance of a proposed system is by calculating the accuracy of the system, as well as the sensitivity and specificity of the classification output in relation to the ground truth. These statistical indices are derived from the following measurements:

- True Positives (TP), number of unacceptable segments classified correctly.
- True Negatives (TN), number of acceptable segments classified correctly.
- False Negatives (FN), number of unacceptable segments classified incorrectly.
- False Positives (FP), number of acceptable segments classified incorrectly.

The overall accuracy is then given by

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

Accuracy is a good measure of the overall system performance; however, it does not give any insight into the type of error the system produces; assessing the performance of the system separately for the two distinct kinds of error (false positives and false negatives) is important because (i) The database might be imbalanced (ii) One type of error might have more profound consequences compared to the other for the application at hand. In the case of SQA systems, a high number of FP will result in a system where a lot of data is unnecessarily wasted. A result of this type of error is ending up with large gaps between reliable measurements obtained from monitors. This may be undesirable in high acuity clinical environments where measurements need to be recorded in a high frequency. Specificity measures this type of error via the following formula:

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

which essentially measures the percentage of acceptable segments classified correctly (a related measure which may also be used is Negative Predictive Value (NPV) which measures the percentage of segments labeled as acceptable which were correctly classified). A high number of FN, on the other hand, will result in a system where measurements from unacceptable signals are considered reliable and erroneous vital sign measurements are produced. An unwanted result of this type of error is the possibility of missing out on an adverse event (because abnormal vital sign measurements appeared normal) or a high number of false alerts (where normal vital sign measurements appear abnormal). This type of error is measured via the sensitivity measure, given by,

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3}$$

which measures the percentage of unacceptable segments which were correctly classified (as with sensitivity and PPV, a related measure which may also be used is Positive Predictive Value (PPV) which measures the percentage of segments labeled as unacceptable which were correctly classified).

For most clinical scenarios, reducing the number of FNs is more important; that is why most SQA systems aim to maximize system sensitivity instead of specificity. In the case of lifestyle and fitness monitoring, maximizing overall system accuracy may be preferable.

1.4 Challenges and Future Directions

Despite the significant effort that has been made in the field of signal quality assessment for physiological signals, some challenges still need to be overcome before these systems can be incorporated into clinical practice. These are:

- *Distinguishing between signal artefacts from pathophysiological changes*: most SQA systems are trained on data from normal rhythms, without (known) arrhythmias or other pathophysiological abnormalities present. It is, therefore, unknown whether SQA systems can distinguish between signal artefact and the presence of a physiological abnormality. This creates the risk that a signal with a physiological abnormality will be misclassified as noise, preventing the detection of a possibly clinically important event [25]. A recent study on the ability of proposed SQA systems to discriminate between artefact in ECGs and the presence of arrhythmias concluded that different SQA systems performed differently depending on the type of arrhythmia present and that SQA systems developed from normal rhythms may underperform on pathological ECGs [25]. This is an important limitation of proposed systems, especially because the population most likely to need continuous physiological monitoring are the elderly and the chronically ill, who are also very likely to suffer from conditions which may affect their cardiovascular physiology. In the future, large databases with a variety of arrhythmia types need to be used for tuning algorithms to check for robustness in the presence of a variety of conditions in order to improve overall performance [22, 25].
- *Algorithms generalizing to heterogeneous clinical settings and sensors*: most proposed algorithms are highly specific to a clinical setting and require modification for reuse in a different one. This was the conclusion of a recent review on artefact detection in critical care units [21]. Since most algorithms are hard-coded to monitor specific data types and frequencies, their use is usually limited to a specific sensor, population, and clinical setting. This is aggravated by the fact that most commercially available monitors have undisclosed proprietary pre-processing algorithms [14]. Tuning algorithms to databases from heterogeneous populations, monitors and clinical settings will help alleviate this problem in the future, ensuring the flexibility and generalizability of proposed approaches.
- *Reducing data retention*: most proposed systems place more importance in rejecting artefact-corrupted signal than retaining clean signal. This may, however, result in a low frequency of obtaining a valid measurement (with large gaps of no valid measurement) which may be undesirable in high acuity cases where the required frequency of obtaining measurements is high [30]. More flexible classification schemes, assigning different quality ratings to signal segments (instead of binary classification schemes) may give the option to clinicians to obtain "moderately erroneous" vital sign measurements at a higher frequency instead of no measurements at all for extended periods of time.
- *Minimizing computational power*: the requirement of low-computational power, in situations where processing is required on-board, points to the usage of the simplest and most computationally efficient model, rather than more sophisticated but computationally expensive ones. To ensure robustness in quality assessment, this requirement needs to be balanced with the need for a multi-dimensional representation of noise (via the use of many features) such

that the human approach to labeling signals based on diagnostic quality can be approximated as well as possible.

Lastly, an important development in SQA is the emergence of sensor fusion approaches that rely on the fact that noise manifests itself differently on different signals. In the hospital, for example, HR can be obtained from the ECG, ABP and PPG signals, which are measured using different sensors, at different body sites, thus providing independent measurements. Sources of artefact in the different signals are mostly uncorrelated; as a result, sensor fusion approaches can provide improved vital sign measurements by assigning measures of reliability to each signal and either weighing the HR measurements accordingly or just picking the measurement from the most reliable signal at any given point in time [31].

1.5 Summary

This chapter has presented an introduction to the topic of signal quality assessment in physiological monitoring, together with an overview of the considerations for designing such systems. The basic design framework of Signal Quality Assessment (SQA) systems was then briefly introduced before presenting the current technical challenges and future directions. Chapters 2 and 3 will provide an overview of techniques for providing SQA to the two most widely utilized physiological signals, the Electrocardiogram (ECG) and the Photoplethysmogram (PPG). It is hoped that these chapters will provide the reader with the relevant framework for developing SQA systems which can also be extended to a variety of other physiological signals.

References

1. United Nations. (2013). *World population ageing: 1950–2050. Magnitude and speed of population aging*.
2. Clifford, G. D., & Clifton, D. A. (2012). Wireless technology in disease management and medicine. *Annual Reviews in Medicine, 63,* 479–492.
3. Rivera-Ruiz, M., Cajavilca, C., & Varon, J. (2008). Einthoven's string galvanometer: The first electrocardiograph. *Texas Heart Institute Journal, 35*(2), 174–178.
4. Orphanidou, C., Clifton, D. A., Khan, S., Smith, M., Feldmar, J., & Tarassenko, L. (2009). Telemetry-based vital sign monitoring for ambulatory hospital patients. In *Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC 2009*) (pp. 4650–4653).
5. Budinger, T. F. (2003). Biomonitoring with wireless communications. *Annual Reviews in Biomedical Engineering*, *5*, 383–412.
6. Baig, M. M., & Gholamhosseini, H. (2013). Smart health monitoring systems: An overview of design and modelling. *Journal of Medical Systems, 37,* 9898.
7. Ben-Israel, N., Tarasiuk, A., & Zigel, Y. (2010). Nocturnal sound analysis for the diagnosis of obstructive sleep apnea. In *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2010)* (pp. 6146–6149).

8. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering, 57,* 884–889.
9. Nangalia, V., Prytherch, D. R., & Smith, G. B. (2010). Health technology assessment review: Remote monitoring of vital signs-current status and future challenges. *Critical Care, 14,* 233.
10. Meredith, D. J., Clifton, D., Charlton, P., Brooks, J., Pugh, C. W., & Tarassenko, L. (2012). Photoplethysmographic derivation of respiratory rate: A review of relevant physiology. *Journal of Medical Engineering & Technology, 36,* 1–7.
11. Orphanidou, C. (2017). Derivation of respiration rate from ambulatory ECG and PPG using ensemble empirical mode decomposition: Comparison and fusion. *Computers in Biology and Medicine, 81,* 45–54.
12. Lamkin, P. (2017). Forbes, https://www.forbes.com/sites/paullamkin/2016/02/17/wearable-tech-market-to-be-worth-34-billion-by-2020/#6dc6ba003cb5. Accessed August 1, 2017.
13. Bonnici, T., Orphanidou, C., Vallance, D., Darrel, A., & Tarassenko, L. (2012), Testing of wearable monitors in a real-world hospital environment: What lessons can be learnt? In *Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (pp. 79–84).
14. Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., & Tarassenko, L. (2015). Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IIEEE Journal of Biomedical and Health Informatics, 19* (3), 832–838.
15. Zheng, Y.-L., Poon, C. C. Y., Lo, B. P. L., Zhang, H., Zhou, X.-L., Yang, G.-Z., et al. (2014). Unobtrusive sensing and wearable devices for health informatics. *IEEE Transactions on Biomedical Engineering, 61*(5), 1538–1554.
16. Pandian, P. S., Mohanavelu, K., Safeer, K. P., Kotresh, T. M., Shakunthala, D. T., Gopal, P., et al. (2008). Smart vest: Wearable multi-parameter remote physiological monitoring system. *Medical Engineering & Physics, 30*(4), 466–477.
17. Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D. A., & Pugh, C. (2014). Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement, 35,* 807–831.
18. Buist, M., Bernard, S., Nguyen, T. V., Moore, G., & Anderson, J. (2004). Association between clinically abnormal observations and subsequent in-hospital mortality: A prospective study. *Resuscitation*, *62*, 137–141.
19. Andersen, L. W., Kim W. Y., Chase M., et al. (2016). The prevalence and significance of abnormal vital signs prior to in-hospital cardiac arrest. *Resuscitation*, *98*, 112–117.
20. Cvach, M. (2012). Monitor alarm fatigue, an integrative review. *Biomedical Instrumentation and Technology, 2012,* 268–277.
21. Nizami, S., Green, J. R., & McGregor, C. (2013). Implementation of artifact detection in critical care: A methodological review. *IEEE Reviews in Biomedical Engineering, 6,* 127–142.
22. Jakob, S., Korhonen, I., Ruokonen, E., Vintanen, T., Kogran, A., & Takala, J. (2000). Detection of artifacts in monitored trends in intensive care. *Computer Methods and Programs in Biomedicine, 63,* 203–209.
23. Marquez, M. F., Colin, L., Guevara, M., Iturralde, P., & Hermosillo, A. G. (2002). Common electrocardiographic artifacts mimicking arrhythmias in ambulatory monitoring. *American Heart Journal, 144,* 187–197.
24. Chase, C., & Brady, W. J. (2000). Artifactual electrocardiographic change mimicking clinical abnormality on the ECG. *American Journal of Emergency Medicine, 18,* 312–316.
25. Daluwatte, C., Johannesen, L., Galeotti, L., Vicente, J., Strauss, D. G., & Scully, C. G. (2016). Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs. *Physiological Measurement, 37,* 1370–1382.
26. Sweeny, K. T., Ward, T. E., & McLoone, S. F. (2012). Artifact removal in physiological signals: Practices and possibilities. *IEEE Transactions on Information Technology in Biomedicine, 16*(3), 488–500.

27. Sun, X., Yang, P., Zhang, Y. T. (2012). Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach. In *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2012)* (pp. 3456–3459).
28. Clifford, G. D., Azuaje, F., & McSharry, P. E. (2006). *Advanced methods for tools for ECG data analysis*. Norwood, MA: Artech House.
29. Nelson, J. C., & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research, 9*(5), 475–496.
30. Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., & Tarassenko, L. (2015). Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE Journal of Biomedical and Health Informatics, 19* (3), 832–838.
31. Li, Q., & Clifford, G. D. (2012). Signal quality and data fusion for false alarm reduction in the intensive care unit. *Journal of Electrocardiology, 45,* 596–603.

Chapter 2
Quality Assessment for the Electrocardiogram (ECG)

Abstract In this chapter, we review a variety of signal quality assessment (SQA) techniques that robustly generate automated signal quality indices (SQIs) for the Electrocardiogram (ECG). Because of the many kinds of noise that may be present in the ECG, quantifying noise is not straightforward. However, several proposed techniques have been shown to provide efficient, accurate and robust quality assessment. In this chapter, emphasis will be given on currently proposed techniques for dealing with noise resulting from motion artifact focusing on data recorded from wearable sensors. The aim of this chapter is not to review currently proposed systems as complete solutions but rather to provide the reader with a basic understanding of the general framework for designing SQA systems for the ECG, in order to trigger further research.

Keywords ECG · Signal quality assessment · Feature extraction · Decision rules

In this chapter, we review a variety of signal quality assessment (SQA) techniques that robustly generate automated signal quality indices (SQIs) for the ECG. Because of the many kinds of noise that may be present in the ECG, quantifying noise is not straightforward. However, several proposed techniques have been shown to provide efficient, accurate and robust quality assessment. Standard approaches for measuring noise in the ECG assume stationarity in the dynamics of the noise; many of these techniques are covered in [1]. In this chapter, emphasis will be given on currently proposed techniques for dealing with noise resulting from motion artifact focusing on data recorded from wearable sensors. The aim of this chapter is not to review currently proposed systems as complete solutions but rather to provide the reader with a basic understanding of the general framework for designing SQA systems for the ECG, in order to trigger further research.

C. Orphanidou, *Signal Quality Assessment in Physiological Monitoring*,
SpringerBriefs in Bioengineering, http://doi.org/10.1007/978-3-319-68415-4_2

2.1 The Electrocardiogram (ECG)

The ECG is one of the most commonly recorded signals and can be found in even the most basic clinical environments. The distinctive morphology of the ECG signal can provide any clinically-trained observer a lot of information about the condition of the recorded individual's heart; it is the major visual record that cardiologists use for identifying a wide range of different heart disorders.

The ECG records electrical potential differences between prescribed locations on the surface of the body using skin electrodes; these differences in potential occur during the heart cycle, as the heart muscles contract and expand to pump blood around the circulatory system. The standard ECG set-up consists of 12 leads but often 3- and 5-lead setups are used as they are easier to structure. ECG can now also be measured using chest bands, skin patches and monitoring garments.

2.1.1 ECG Morphology

The ECG waveform is composed of three distinct waveforms, the P, QRS and T waves, each one representing an independent event related to the heart cycle. An example is shown in Fig. 2.1 with the three waveforms marked.

The most dominant wave, which should be present in all ECG recordings is the QRS wave, commonly known as the *QRS complex*. The QRS complex occurs as a result of the depolarization activity in the right and left ventricles [2]. Correct identification of the R-peaks in a segment of ECG is sufficient for measuring a reliable HR. The P-wave, which results from the depolarization of the right and left

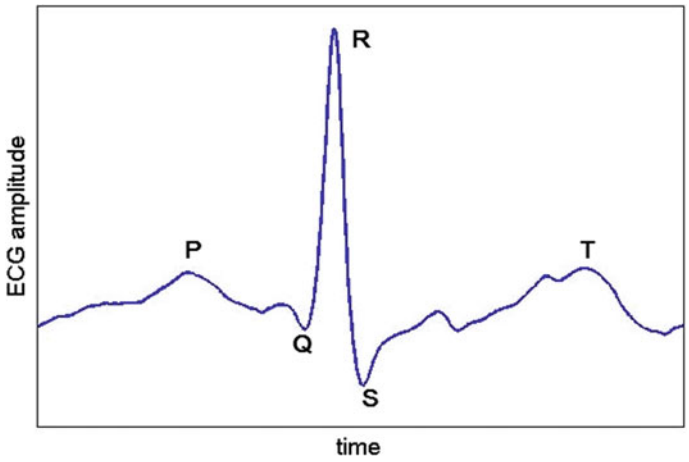

Fig. 2.1 Example PQRST waveform with P, QRS and T waves marked

atria [2] should be present in a healthy ECG recording, however it is absent in the presence of certain conditions, such as atrial fibrillation [3]. The amplitude of the T-wave, which represents the recovery of the ventricles, is also an important indicator or certain heart conditions; T-wave alternans, for example, which is the periodic variation in the peak amplitude and morphology of the T-wave, has been linked to sudden cardiac death [4]. The presence of T-wave abnormalities has also been associated with many heart conditions such as myocardial ischemia [5] and coronary heart disease [6].

When an ECG exhibits clearly the different waveforms and a clean and regular heart beat can be observed without any abnormalities, the heart is said to be in *sinus rhythm* [1]. To correctly identify any deviations from sinus rhythm or any abnormalities in the functioning of the heart, identification of all distinct waveforms present in an ECG recording is important.

In addition to morphological characteristics of single PQRST waveforms, the variation in the time-intervals between successive heart beats (the *RR-intervals*) within a longer ECG segment, known as Heart Rate Variability (HRV) is an important index of the activity of the Autonomic Nervous System (ANS) [7] and should be correctly measured from an ECG segment.

Correctly identifying the different waveforms of the PQRST wave is challenging at the best of times, because of the morphological variability of the signal for different individuals; even for the same individual, different recordings may differ substantially because of electrode type or placement. In the presence of noise, QRS

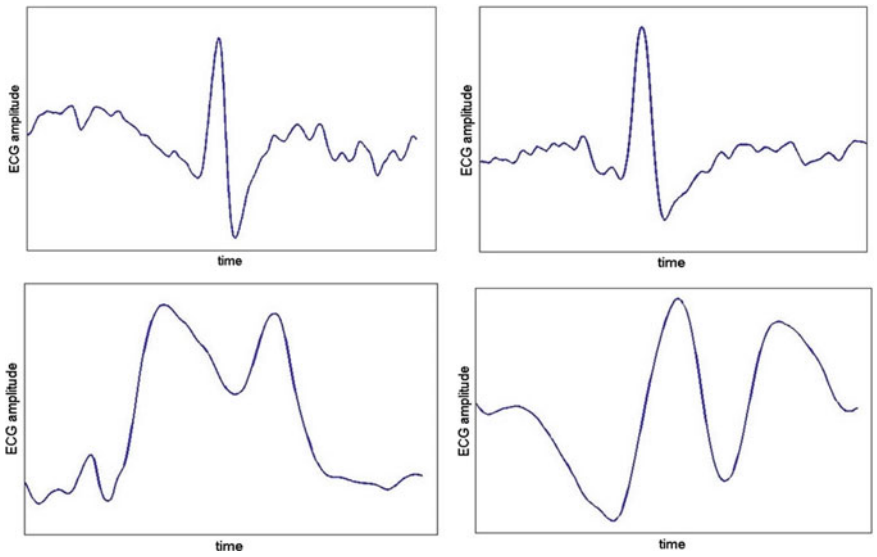

Fig. 2.2 PQRST waves containing noise. The *top* two waveforms contain high-frequency noise. While the R-peak is clean, the P-wave and the T-wave are contaminated with noise and cannot be correctly identified. In the *bottom* two PQRST waves the motion artefact makes determining the R-peak extremely challenging

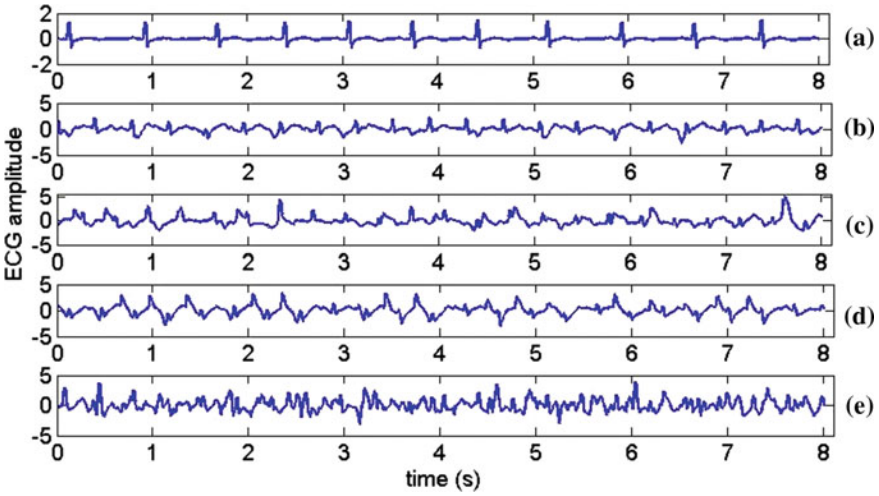

Fig. 2.3 Eight-second ECG segments with distinct kinds of noise. **a** Clean ECG segment. **b** Borderline case; this segment could result in estimating a reliable HR but would be unsuitable for diagnostic purposes. **c**–**e** Progressively noisier ECG segments which would result in unreliable vital sign estimates. As the segments get noisier, the overall morphology becomes more varied

(or R-peak) detection, for example, becomes even more challenging. According to Pan and Tompkins [8], the types of noise commonly found in the ECG which could affect the robustness of QRS detection are "muscle noise, artifacts due to electrode motion, power-line interference, baseline wander, and T-waves with high-frequency characteristics similar to QRS complexes." P-wave and T-wave detection are also similarly affected by noise. The way that noise can interfere with waveform morphology is illustrated in Fig. 2.2 which shows examples of PQRST waveforms containing artefact. As we will see later, a good SQA approach is to identify and utilize morphological features of individual QRS or PQRST waveforms which are present in clean segments of ECG but are altered in the presence of noise. Other promising approaches build on the variability of QRS/PQRST morphology or R-R intervals rather than looking at individual beat waveforms; a clean segment of ECG is expected to have regular morphology in individual beat waveforms whereas a noisy one will be irregular [9]. When setting criteria for regularity/irregularity of QRS complexes in an ECG segment an important consideration is any irregularity which may be attributed to physiology. For example, R-R intervals are expected to show variability because of Respiratory Sinus Arrhythmia (RSA), the cyclic variation of heart rate associated with respiration [10]. The presence of arrhythmias will also add to this variability in the R-R intervals [11]. It is, thus, important for quality assessment approaches to set decision rules which do not result in a high number of false positives because of incorrectly identifying physiological abnormality as noise. Figure 2.3 shows examples of 8-second segments of ECG with different instances of noise showing the variability in beat waveforms which can form the basis for a SQA approach.

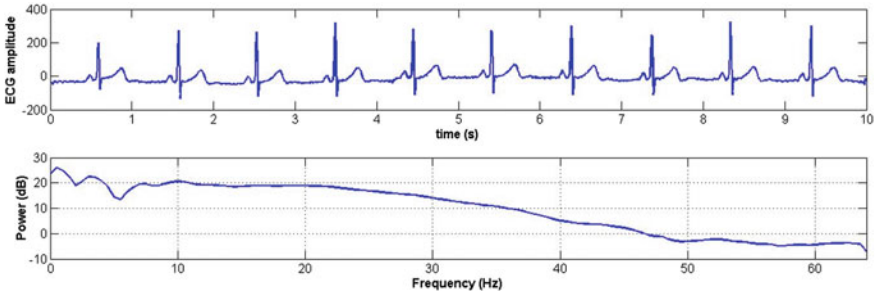

Fig. 2.4 Ten seconds of 128-Hz ECG in sinus rhythm (taken from the MIT-BIH Normal Sinus Rhythm Database) and associated 128-point Welch periodogram calculated with a hamming window and a 64-point overlap. The periodogram calculates the power spectral density (PSD) of the signal, i.e., the power (in dB, vertical axis) contained in each frequency (in Hz, horizontal axis). We note the peak power at approximately 1 Hz which is the HR (which can be verified from the ECG plot) and secondary peaks around 4 Hz (typical frequency of the T-wave), 7 Hz (typical frequency of the P-wave) and 10 Hz (typical frequency of the QRS complex) [1]

2.1.2 Spectral Characteristics of the ECG

The spectrum of a clean ECG in sinus rhythm has a distinct distribution of power; the accepted range of diagnostic ECG is 0.05–100 Hz (even though micro-fluctuations are said to be found in frequencies up to 500 Hz [1]). QRS energy is typically contained in frequencies below 30 Hz with peak power occurring in the range of 4–12 Hz [12]. The presence of notches in the QRS complexes (associated with some conditions) may broaden the power spectrum of the ECG but most power is still contained below 30 Hz. When noise is present in the ECG signal, the distribution of power is altered; an accumulation of power occurs in frequency bands which are either outside the physiologically-attributable range for the ECG or the overall distribution of power which would be considered normal is altered (Fig. 2.4).

Later, in Sect. 2.3.1.4 we will discuss methods for quantifying these alterations in the spectra of noisy signals to extract robust features for differentiating between *acceptable* and *unacceptable* ECG signals.

2.2 ECG Quality Considerations

As discussed in Chap. 1, the definition of signal quality differs depending on the application. For the ECG, we may define signal quality in two ways:

- *Basic quality*: R-peaks are clearly identifiable. In this case, a reliable HR can be extracted as well as some types of arrhythmias and HRV information.

- *Diagnostic quality*: P (if present), QRS and T waveforms are clearly identifiable. In this case, the signal can also be used for clinical diagnosis of more subtle conditions such as myocardial ischemia and coronary heart disease.

The first definition is less strict; as illustrated in Fig. 2.3, in the presence of noise it is often still possible to obtain a reliable HR. Baseline wander, for example, which is often considered to be the result of motion, does not necessarily hinder the measurement of HR. For assessing the quality of data obtained from wearable health and fitness sensors (which would be more likely to contain noise), most often the requirement of *basic quality* is sufficient. Diagnostic applications from the ECG are rarely used when the user is in motion; an exception is wearable systems intended for identifying, for example, episodes of Atrial Fibrillation (AF) [3, 11]. In the case of AF, the absence of the P-wave is a common indicator for the presence of this arrhythmia, so being able to identify it amongst the motion artefact is important.

For the remainder of this chapter, we will take a close look at different approaches proposed in literature for assessing the quality of ECG signals. We present current past and current techniques for feature extraction and for determining associated decision rules. Most approaches focus on obtaining *basic quality* of ECG, rather than *diagnostic quality*. We focus mostly on the former definition since for most diagnostic applications ECG assessment is done while the patient is static so the likelihood of motion artefact occurring is minimal. We separate approaches to single-channel and multi-channel ones. Single-channel approaches rely on extracting morphological or spectral characteristics from a single ECG waveform and applying heuristic, empirically determined, or machine learning-based decision rules. Multi-channel approaches, on the other hand, mostly rely on the level of agreement between features and estimations calculated from multiple simultaneously recorded channels of ECG.

2.3 Single-Channel Approaches

The advantage of single-channel (or single-lead) approaches is that the system is independent of the number of channels recorded [13]. Single-lead approaches are useful for applications requiring the real-time monitoring using wearable sensors and can be used in the most basic clinical environments where the option of multi-channel acquisition is not always available.

2.3.1 ECG Feature Extraction

Most signal quality assessment algorithms propose applying a series of different steps for accepting or rejecting segments of ECG. Often, simple feasibility rules are applied first, to quickly detect common distorting factors such as disconnected

electrodes, poor skin-electrode contact and external electrical interference [14]. These disturbances are successfully tracked via simple rules and result in the quick identification of a significant percentage of unusable segments, often delivering high classification accuracy. More sophisticated approaches are required for questionable segments which pass these initial checks and may produce a physiologically viable measurement of HR which, however, is likely to be erroneous.

We will begin by considering simple feasibility checks and then proceed to discuss ECG features which have been shown to successfully differentiate between "acceptable" and "unacceptable" segments. These features may be based on ECG morphology, spectral characteristics of the ECG or trend-based measures within a window of signal. For most time-domain techniques, the first step is applying a beat detector to the ECG.

2.3.1.1 ECG Beat Detection

Time-domain signal quality assessment algorithms require the use of a robust beat detection algorithm. In fact, most quality assessment schemes are intrinsically linked to the beat detector used since beat detectors themselves have different noise sensitivities. Two popular beat detectors used in many current approaches are the ones proposed by Hamilton and Tompkins [15] and Zong et al. [16]. The former, relies on the application of an elaborate linear and non-linear digital filtering scheme followed by differentiation, squaring, and averaging of the signal before rule-based peak detection, to identify the fiducial point. The latter is based on digital filtering followed by application of the length transform (LT) followed by application of a decision rule. Both methods are well-documented and available in open-source on Physionet [17].

2.3.1.2 Feasibility Checks

Following beat detection, the first step to a quality assessment approach is to apply simple feasibility checks which are often applied in a cascade of decision steps. Even though such rules cannot fully assess signal quality, they can be useful for quickly rejecting segments of low quality, or identifying flat lines and segments of saturated signal.

Flat Line Detection

Real ECG is never constant; even in segments that may appear flat, there is always some low-level noise. Flat segments of ECG, signifying a missing lead [13] are simply identified by tracking the number of consecutive sample points with the same value and setting a threshold to the time-period of unchanged signal value, which is considered to indicate that the signal is unusable [18]. Indicative thresholds

proposed in literature vary from 0.2 s [18] to 1 s [19]. An equivalent technique is searching for a constant derivative of zero where the derivative is obtained by taking differences between consecutive values [20]. A missing lead may not only result in a flat line but also a straight line. Searching for a constant non-zero derivative is also a good approach for identifying such instances of unusable signal.

Amplitude Limits and Variability

Within a given segment of ECG, the range of values that the ECG takes may be used as a simple quality indicator. A small range might indicate that an electrode is not correctly attached rendering the signal useless [18]. A broad range might indicate the presence of high-frequency noise in the form of spikes or low-frequency noise in the form of baseline wander [18]. Common quality indicators include thresholds on the acceptable range (min-max) of ECG values within a segment; Moody [18] proposes a minimum cutoff of 0.2 mV and maximum cutoff of 15 mV. Lead saturation, i.e. amplitude exceeding a threshold indicating the presence of spikes [21] or excessive amplitude which is maintained for a specific time-period (e.g., 200 ms [19]), indicating signal saturation, are also useful quality metrics for identifying unacceptable segments of signal [14, 19], as is the detection of a very small maximum amplitude, steep slopes [19] and sudden amplitude changes [22]. Often, these checks have to satisfy percentage portions of the signal rather than the whole signal for a segment to be rejected.

Baseline Wander

Baseline wander (BW) (or baseline drift), caused by patient movement, can also be quickly identified after the application of a 1 Hz low-pass filter and the assessment of the resulting signal. If the BW exceeds an empirically-set threshold, the signal is classified as unacceptable. As mentioned earlier, the presence of BW does not always render the signal unusable so for many applications the presence of BW is permitted.

Noise-Power Measurements

Standard noise-power measurements may also be applied to the ECG signal which can quantify the artefact present. According to Clifford et al. [1] these include:

- *Root mean square (RMS) power in the isoelectric*[1] *region;*
- *Ratio of the R-peak amplitude to the noise amplitude in the isoelectric region;*
- *Ratio of the peak value of a signal to its RMS value;*
- *Ratio between in-band (5–40 Hz) and out-of-band spectral power;*
- *Power in the residual after a filtering process.*

While these measures have been shown to correctly identify specific instances of noise, they rely on the assumption that the signal is stationary and that the noise we are trying to detect is Gaussian. This is not always the case for the ECG since often noises

[1]The isoelectric region is a range around the isoelectric level which is the region between the P-wave and QRS complex, considered to be the most stable marker for 0 V [Clifford]. It records the short pause between the atrial and ventricular depolarization of the heart.

and artifacts are transient and their onset and duration are unknown. Additionally, some kinds of noise which can occur (e.g. movement artefact) are not Gaussian. The use of these indices is thus limited and they should be used in combination with other indices more suitable for identifying a wider range of noises and artefacts.

Physiological feasibility
The natural limits of cardiovascular physiology which drive the morphology of the ECG signal can be used as signal quality indicators either as individual viability checks or as a cascade of checks which can quickly identify very-low-quality segments and label them as unacceptable. Following beat detection, the following limits have been used to check for physiological feasibility:

- *Average HR*: the average HR measured within a segment of signal needs to satisfy a physiologically valid rate. Proposed limits in literature were 40–180 beats per minute (bpm) (0.67–3 Hz) [9, 23]. If the users are exercising the limits need to be adjusted to allow for increased valid HRs (e.g. up to 300 bpm (5 Hz)). Any estimated HR outside these limits is not physiologically viable and the segment can be automatically rejected.
- *Maximum R-R interval*[2]: based on a minimum viable HR of 40 bpm, the maximum R-R interval within a segment of signal cannot exceed 60/40 bpm = 1.5 s. Allowing for a single missed beat, a maximum R-R interval limit of 3 s has been proposed as a simple feasibility check [23].
- *Maximum R-R interval/Minimum R-R interval*: under normal monitoring conditions, HR is not expected to change rapidly within a short time-segment. As a result, when considering short segments of ECG, R-R intervals are not expected to change substantially. In a short segment of signal (e.g. 10 s), the ratio of the maximum to the minimum R-R interval should not exceed 1.1, since a change in HR of over 10% within 10 s is improbable. Allowing for a single missed beat, a maximum ratio of 2.2 has been proposed as a feasibility rule for acceptable ECG [23].

Particularly for feasibility rules which are based on the distribution of R-R intervals, it is important for thresholds to consider the possible presence of arrhythmias, as they would affect the distribution of R-R intervals. Additionally, the last criterion only applies when considering short segments of signal and must be adapted for larger segments since in larger time-intervals it is possible to have greater changes in the values of instantaneous HR measured via the R-R interval.

2.3.1.3 Trend-Based Approaches

We now discuss trend-based techniques for SQA which search for regularity in a short segment of ECG. In a typical clean ECG recording in sinus rhythm,

[2]The R-R interval is defined as the time difference between successive R-peaks in an ECG segment.

irrespective of the specific QRS morphology of the recording, QRS complexes appear almost periodically and are very similar in morphology. The presence of artefact will distort this regularity. Therefore, the more regular a segment of ECG is, the more reliable it is presumed to be. The techniques we discuss next, search for this regularity using different measures within a short segment of ECG.

High order statistics (HOS)-skewness and kurtosis of the ECG

Skewness and Kurtosis are defined as the third and fourth standardized moments of a probability distribution, respectively; skewness measures the symmetry of the distribution and can take positive or negative values depending on whether the skew is on the right or the left tail of the distribution. Kurtosis, on the other hand, is a measure of how sharp the peak of the distribution is. A distribution with many outliers will have a high (absolute) value of skewness since the outliers will cause asymmetry in the distribution; it is also expected to have a low value of kurtosis since its probability distribution will be flatter and will approach zero slower. A distribution with no outliers will be more symmetric, and will have a sharp peak, thus approach zero faster and, consequently, have high values of skewness and kurtosis. As a result, skewness and kurtosis have been proposed as indices of the presence of outliers in an ECG segment which is equivalent to the presence of noise. The kurtosis of the normal distribution is 3; clean ECG in sinus rhythm has been shown to have a kurtosis larger than 5 [24]. Additionally, Clifford [1] has shown that muscle artifact has a kurtosis of around 5 and baseline wander and power-line interference have kurtosis lower than 5. The measurements of skewness (S) and kurtosis (K) of the distribution of an ECG segment can, thus, be used an index of signal quality [13, 25] and can be calculated for 10 s segments from

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x})^3}{\widehat{\sigma}^3} \tag{2.1}$$

$$K = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x})^4}{\widehat{\sigma}^4} \tag{2.2}$$

where x_i is the discrete signal, $\widehat{\mu}$ and $\widehat{\sigma}$ are the empirical estimates of the mean and standard deviation of the distribution of x_i and N the number of data points in the signal. Li et al. [25] used the kurtosis-based quality index to assign segments with a kurtosis value greater than 5 as *acceptable* and a value of kurtosis lower than 5 as *unacceptable*. Skewness has not been used as a rule-based quality index but both skewness and kurtosis were used by Clifford et al. [13] as inputs into machine-learning-based quality indices. Figure 2.5 shows examples of acceptable and unacceptable ECG segments with their respective distributions and skewness and kurtosis values.

Variability in the R-R Interval

The regularity in the R-R intervals, calculated as the time-intervals between successive R-peaks, after R-peak detection, is another indicator of quality in an ECG segment since heart rhythm is expected to be regular. This measure relies on the reduced

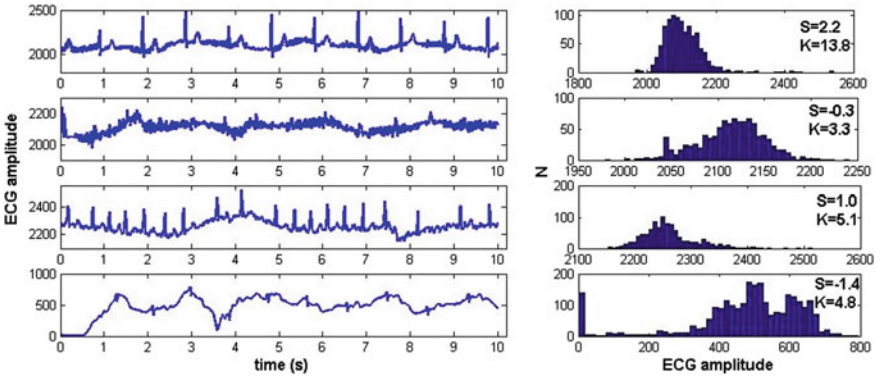

Fig. 2.5 Examples of 10 s ECG segments with corresponding discrete distributions (histograms) and skewness (S) and kurtosis (K) values. The kurtosis values are smaller for artefact-corrupted signals but skewness values are less consistent. For example, the second example from the top contains a large amount of high-frequency noise and the kurtosis value is very low but because of the nature of the noise, the distribution is approximately symmetric giving a low skewness value

performance of the QRS detector in the presence of noise: when artefact is present, the QRS detector underperforms by either missing R-peaks or erroneously identifying noisy peaks as R-peaks. This results in a high degree of variability in the distribution of R-R intervals and a resulting high value in the standard deviation of R-R intervals. Hayn et al. [21], proposed calculating the variability in the R-R intervals of an ECG segment using the coefficient of variation of the R-R intervals, given by

$$C_v = \frac{\widehat{\sigma_{RR}}}{\widehat{\mu_{RR}}}, \tag{2.3}$$

where $\widehat{\sigma_{RR}}$ and $\widehat{\mu_{RR}}$ are the empirical estimates of the mean and standard deviation of the distribution of the R-R intervals within a segment of ECG. The RR-variability quality index was then determined by assigning segments with a value of C_V smaller or equal to 0.64 as *acceptable* and a value of C_V greater than 0.64 as *potentially unacceptable* [21]. The threshold of 0.64 was determined empirically. Figure 2.6 shows examples of *acceptable* and *unacceptable* ECG segments with their respective distributions of R-R intervals and coefficient of variation values.

Template Matching

Another trend-based approach assigning a measure of regularity in an ECG segment relies on the similarities in the morphology of the QRS complexes within a single segment of signal. The template matching approach, used in [23], estimates the average QRS complex in a segment of ECG by performing beat detection and then averaging over all QRS complexes. Each QRS complex is extracted in a window around each detected heartbeat equal to the median R-R interval. Averaging all

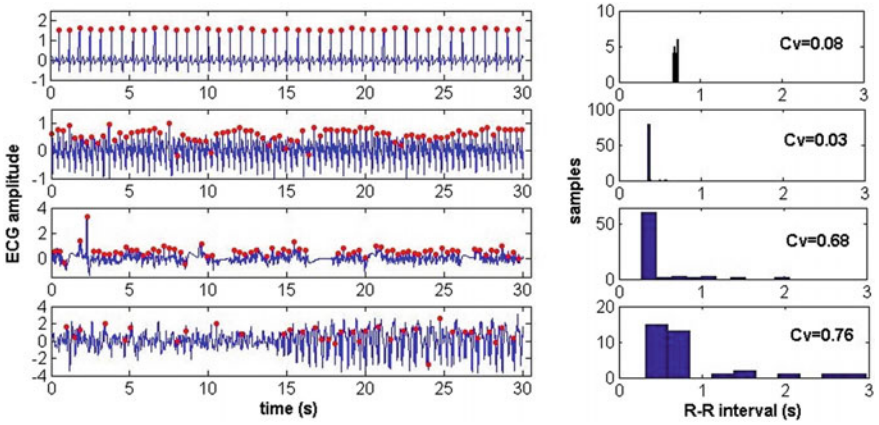

Fig. 2.6 Examples of 30 s ECG segments with *red circles* indicating R-peaks identified by application of the Hamilton and Tompkins QRS detector [15]. On the *right hand side*, we show the discrete distribution of the R-R intervals calculated from the R-peak detection and the corresponding coefficient of variation of the R-R intervals. The *top two figures* show examples of clean ECG in sinus rhythm with very small spread in the distribution of R-R intervals (i.e., no outliers) and a small value of C_v. The third and fourth examples are examples of noisy ECG either with flat regions (third example from the *top*) or an excessive amount of high-frequency noise which results in erroneous QRS detection. For both two cases, the spread in the distribution is large and the corresponding value of C_v is greater than the acceptability cutoff of 0.64. These examples illustrate the dependence of this quality index on the robustness of the QRS detector

QRS complexes provides a QRS template, which is adapted for every different ECG segment. The procedure is illustrated in Fig. 2.7 for a 10 s segment of ECG.

The Pearson's correlation coefficient of the template with each individual QRS complex is then calculated and averaged over all beats in the segment. *Pearson's correlation coefficient*, usually denoted by r for discrete time series analysis, measures the similarity between two time-series x_i and y_i using the following formula:

$$r = \frac{\sum_{i=1}^{N}(x_i - \widehat{x})(y_i - \widehat{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \widehat{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \widehat{y})^2}}, \tag{2.4}$$

where $\widehat{x}$ and $\widehat{y}$ are the empirically determined means of x_i and y_i, respectively, and N is the number of samples in the two time-series.

This approach relies on the fact that in the presence of artefact, some or all QRS complexes will be corrupted; this will distort the QRS template (which is averaged over all QRS complexes) which will then have low correlation with individual QRS complexes. The template-matching quality index is assigned by labeling segments with an average correlation ≥ 0.66 as *acceptable* and a value of average correlation <0.66 as *unacceptable* [23], where the average correlation threshold was determined empirically.

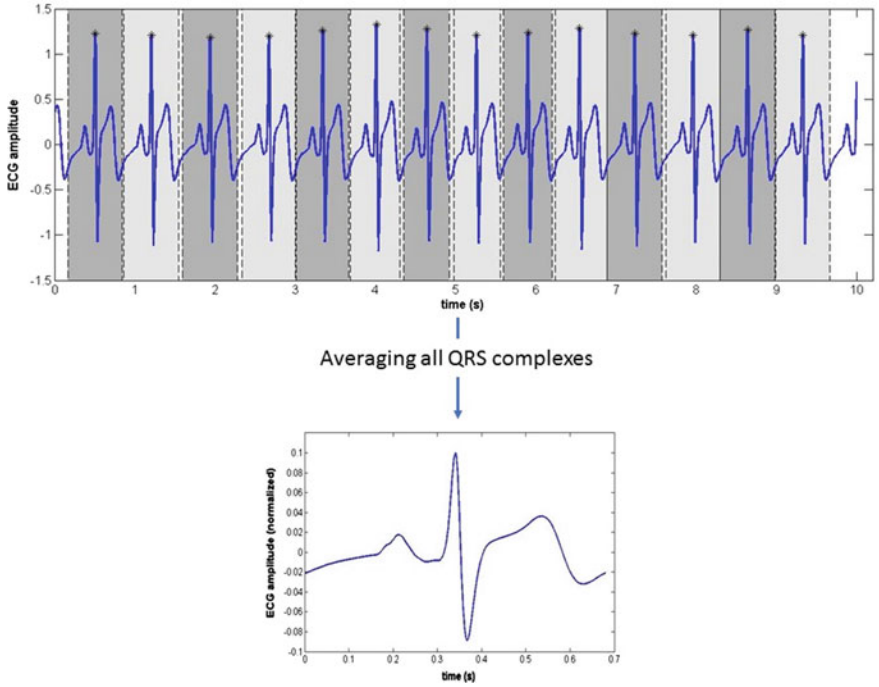

Fig. 2.7 Process for obtaining the average QRS template from a 10 s segment of ECG. The top figure shows individual QRS complexes (in alternate shading for ease of interpretation). The duration of each QRS complex is taken as the median R-R interval in the segment and it is centered around each beat detected (marked in *black stars*). The QRS complexes are then averaged to obtain the QRS template shown in the *bottom plot*

Figure 2.8 shows examples of *acceptable* and *unacceptable* ECG segments with their respective average QRS templates and average correlation coefficient values.

2.3.1.4 Frequency Domain Features

As discussed in Sect. 2.2, noise manifests itself on the spectrum of the ECG as increased power in frequencies outside the physiologically-attributed limits of the ECG or as a change in the distribution of power. We will now discuss some approaches for assessing the quality of the ECG which search for a distortion in the spectrum of the ECG. One approach presented, considers the spectral distribution of the Heart Rate Variability (HRV) signal, a physiologically rich signal which can be estimated from the ECG signal.

Frequency Content Across Different Bandwidths

The distribution of power in different bandwidths of the ECG signal has been investigated as a feature to differentiate between clean and noisy signals initially by

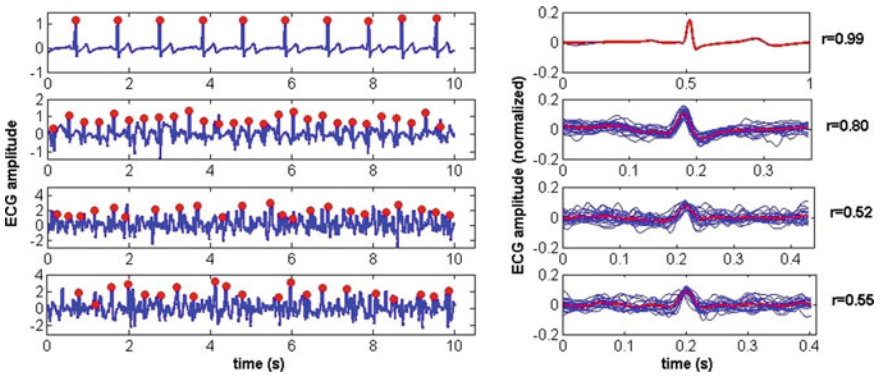

Fig. 2.8 Examples of 10 s ECG segments and extracted QRS templates. The *right-hand plots* show individual QRS complexes (in *blue*) and the QRS template in *red* with corresponding average correlation coefficient. The *top* ECG segment is a clean segment in sinus rhythm. As a result, QRS complexes are almost identical with an average correlation coefficient almost equal to 1. Despite the presence of some noise in the second segment, the robustness of the QRS detector results in a high enough correlation coefficient. The last two segments have a lot of variability in QRS morphology, resulting in an average correlation coefficient below the acceptable threshold of 0.66 and are thus labeled as unacceptable

Allen [26]. Since most physiologically relevant information in the ECG is contained in frequencies below 30 Hz, unusually high power in the frequencies outside this bandwidth will likely signify the presence of noise. The presence of baseline wander (BW) will also distort the distribution of power in the lower frequencies. Allen's seminal work [26] was an investigation on the bandwidths which show the most difference in frequency content between good quality and bad quality ECGs. Using a set of filters, the ECG was split into six different bandwidths such that the signal strength can be measured in each bandwidth. The six bandwidths considered were: low frequency (LF, 0.05–0.25 Hz), lower ECG bandwidth (ECG1, 0.25–10 Hz), higher ECG bandwidth (ECG2, 10–20 Hz), medium frequency (MG, 20–48 Hz), mains powerline noise (50, 48–52 Hz) and high-frequency (HF, 52–100 Hz). The signal strength in each bandwidth was then measured using a root-mean-square (RMS) detector. The system also considers a seventh feature, called the out-of-range event (ORE) feature, defined as the number of times the ECG exceeded a pre-set limit of ± 4 mV (this essentially represents instances of gross movement of the electrodes). The ECG1 and ECG2 bandwidths represented the typical monitoring bandwidth of the ECG and the rest of the bandwidths represented noise [26]. After establishing a baseline of good quality ECG, the authors studied differences in the frequency content features and OREs between good quality ECGs and ECGs collected at night (which were expected to be of lower quality due to uncontrolled user motion). All seven features investigated were found to be higher during the night compared with good quality ECGs with the most significant differences occurring in the LF and ECG1 bandwidths and the OREs feature. These features are thus considered promising for evaluating signal quality.

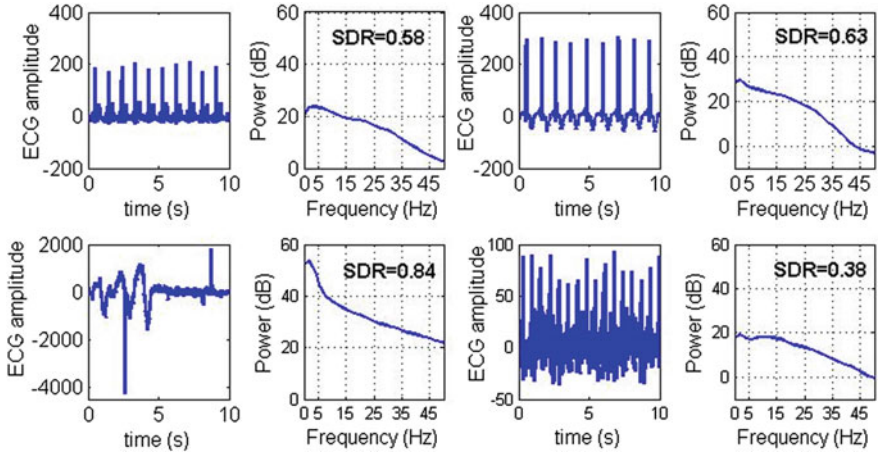

Fig. 2.9 Good quality (*top*) and bad quality (*bottom*) ECG segments with PSD plots and SDR values. The *top* two segments have an SDR value within the limits of acceptable signal whereas the *bottom* two have SDR values outside and are thus labeled unacceptable

Based on the same principle as Allen's investigation, another quality index proposed by Li et al. [25] calculates the ratio between the power spectral density (PSD) (using the Fast Fourier Transform (FFT)) between the 5–14 Hz frequency band and the 5–50 Hz frequency band producing the so-called spectral distribution ratio (SDR) of the ECG. In the presence of high-frequency noise, the SDR is likely to be low. When the SDR is high, it may be the case that excessive QRS-like artifact is present. The SDR-based quality index is then assigned by labeling segments with an SDR value between 0.5 and 0.8 as *acceptable* and signals with SDR values outside this range as *unacceptable* [25], where the thresholds were determined empirically.

Figure 2.9 shows examples of clean and noisy ECG segments with associated PSD plots and SDR values.

Spectral analysis of the HRV signal

It is possible to derive frequency-based features differentiating between *acceptable* and *unacceptable* segments of ECG from the heart rate variability (HRV) signal derived from the ECG. HRV reflects the variation in the time-intervals between consecutive heart beats. HRV is a physiological index, which reflects the interaction of the parasympathetic nervous system with the heart rate, when the heart is in normal sinus rhythm [7]. The classic way to obtain the HRV signal is to firstly calculate the R-R intervals; because the time point of each R-R interval can be either the preceding or succeeding R-peak, and is thus nonperiodic, the next step is to interpolate (using either linear or cubic spline interpolation) in order to obtain a smooth, periodically-sampled signal. HRV has been studied extensively since it reflects the underlying activity of the sympathetic and parasympathetic systems of the Autonomic Nervous System (ANS) and has been found to provide useful

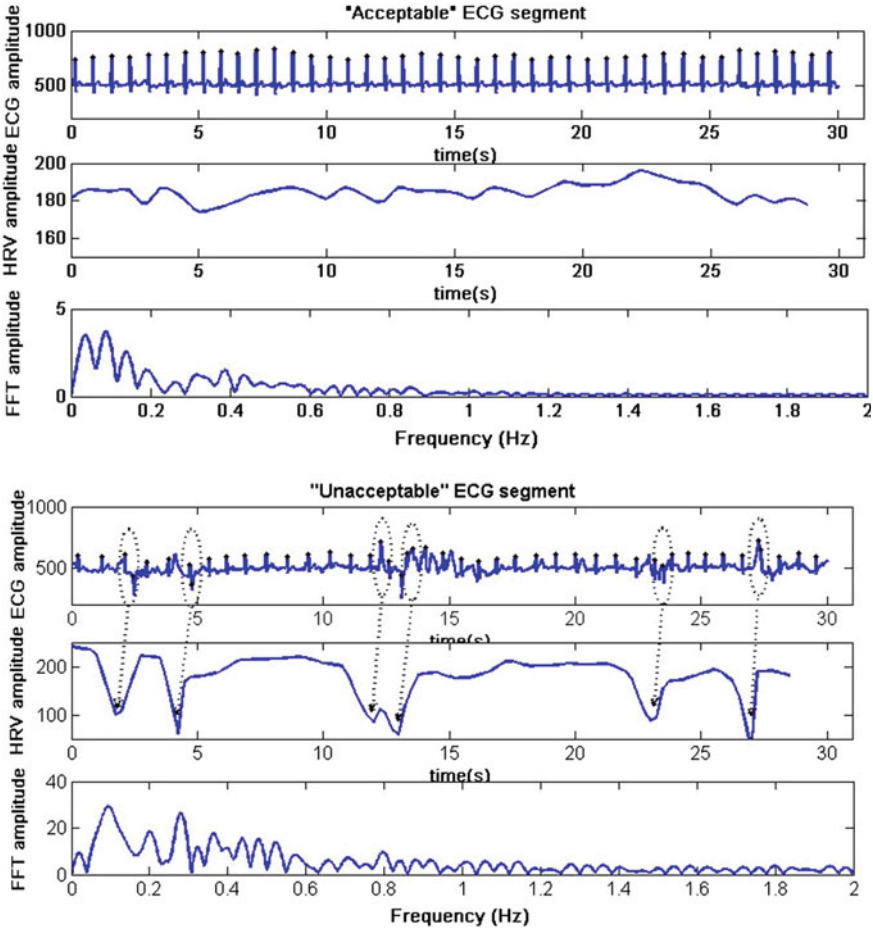

Fig. 2.10 Examples of "acceptable" and "unacceptable" ECG segments with extracted HRV signal and associated FFT spectrum. The HRV signal was derived from the R-R intervals of the ECG using cubic spline interpolation at a frequency of 4 Hz. The presence of noise in the unacceptable segment creates "dips" (*circled*) in the HRV signal which alter the spectrum and increase the strength of frequencies over the physiologically attributable upper limit of 0.4 Hz. Image adapted from [9]

clinical information [9]. The frequency content of the HRV signal is the result of different physiological processes with an upper limit of physiologically-attributed frequencies of 0.4 Hz. Since the calculation of the HRV signal relies on accurate beat detection, in the presence of noise, the HRV signal is distorted. The distorted HRV signal would also have a distorted distribution of energy in the different frequency bands since there would be energy in the non-physiologically-relevant bands, e.g., in frequencies greater than 0.4 Hz [9] (Fig. 2.10).

To extract the energy in different frequency bands wavelet entropy measurements were proposed. Wavelet decomposition allows the representation of the temporal features of a signal at different resolutions [20]. In practice, application of the discrete wavelet transform is the successive application of a two-channel, perfect reconstruction filter bank comprising of a low-pass and a high-pass filter, following by decimation by a factor of 2. The frequency bands associated with each decomposition level depend on the sampling frequency of the signal; at every step of the wavelet decomposition scheme the signal is split into its bottom half bandwidth (because of the application of a high-pass filter) producing the *approximation* wavelet coefficients and its top half bandwidth (as a result of application of a low-pass filter) producing the *detail* wavelet coefficients. The procedure is repeated, dyadically splitting the approximation levels for a specified number of decomposition levels (depending on the characteristics of the signal under study). For the HRV signal, which was interpolated from the R-R intervals time series at a frequency of 4 Hz, 5 levels were used which resulted in 6 sets of wavelet coefficients in the following bandwidths: 2–4 Hz (D1), 1–2 Hz (D2), 0.5–1 Hz (D3), (D4), 0.125–0.25 Hz (D5) and 0–0.125 Hz (A5). The *Shannon entropy* [27] of each decomposition level was then calculated, giving a measure of the disorder at each level. Comparison of the wavelet entropy measurements between acceptable and unacceptable ECG segments showed that all entropy measurements showed statistically significant differences between the two groups and that building a classification scheme using only level D2 (1–2 Hz) as a feature had a high degree of accuracy which was only marginally improved with the addition of D4 (0.25–0.5 Hz) and D5 (0.125–0.25 Hz). This is presumably because the artefact-corrupted segments of ECG had a high frequency content in the 1–2 Hz range whereas the clean segments didn't.

2.3.2 Decision Rules

In this section, we review the types of decision rules which may be used for making assessments of signal quality based on extracted time-domain or frequency-domain features from the ECG. Far better results may be obtained when combinations of features are used. When multiple features are being used together, the decision can be made based on a set of independent rules that need to be satisfied either concurrently or in a cascade of decision steps. In addition, machine learning can provide a "black box" strategy for building robust classifiers using an assortment of different features.

2.3.2.1 Thresholds and Combinations of Rules

The most common approaches for setting thresholds on features extracted from the ECG are based on:

- *Physiological limitations:* in human subjects, there are natural restrictions to the values that HR can have as well as the statistics of the distribution in the R-R intervals. It is, thus, possible to quickly identify signal segments whose derived measurements fall outside these physiologically-attributable restrictions. An important consideration when setting thresholds based on the R-R interval is the frequency and duration of the signal being analyzed, as well as the possible presence of an arrhythmia. A signal with a higher sampling rate will show more variability in the R-R intervals compared to a signal with a lower sampling rate. Furthermore, the "permitted" change in HR or variation in R-R intervals in a 60 s window of signal is greater than in a 10 s window. With respect to variability in R-R intervals an important but tricky task is to account for the variability in R-R intervals which might be the result of the presence of arrhythmias, rather than the presence of noise in the data [28]. This is important so that the system doesn't end up producing false alerts when an arrhythmia occurs.
- *Expected limits of signal characteristics, such as amplitude, rate of change and frequency content, determined heuristically:* these rules are set either heuristically, based on the experience of the expected behavior of the signal itself, rather than the physiology the signal is measuring. In the absence of noise, ECG amplitude, for example, is not expected to change greatly, so many proposed algorithms have set thresholds on the maximum permitted amplitude (measuring it either from the isoelectric region or from the preceding or succeeding minimum within a specific time-window (for example around the QRS complex). Often thresholds are set heuristically, and differ in different approaches and for different datasets. Hayn et al. [21] for example, has set a quality criterion which classifies a signal as "unacceptable" when "*the portion of samples that showed amplitudes of more than* ± 2 mV *was higher than 40% of the analyzed signal*". Moody, on the other hand proposes labelling as "unacceptable" signal segments with a range (maximum-minimum within a 10 s segment) smaller than 0.2 mV, or larger than 1.5 mV [18]. Interestingly, both these cutoffs were set for the same dataset. Most works proposing decision rules based on expected signal characteristics do not contain detailed justifications of the thresholds set and they often seem arbitrary. It is assumed that most authors decide on the thresholds using combinations of experience on expected behavior, knowledge about equipment limits and trial and error using data available for training of the proposed algorithms, as discussed next.
- *Empirical evidence* via *the use of available data:* another way of setting quality cutoffs in extracted features is learning them on datasets available to researchers and developers. The most common approach is using a representative portion of the labelled data as the *training set*, trying different thresholds, and picking the optimal threshold by drawing the Receiver Operating Characteristics (ROC) curve which is a plot of 1-specificity (horizontal axis) against the sensitivity of the system (vertical axis) for different system parameters. If the same weight is given to the two statistical measures then the optimal threshold is the one which minimizes the distance to the point (0,1). When choosing thresholds empirically, the dataset used is important; if the aim is to create a system that can

generalize, parameters need to be set using data from multiple monitors, populations and clinical scenarios [23]. If parameters are tailored to a specific monitor, population and clinical scenario, it is possible that when used on data from a different monitor and in a different clinical setting, the SQA system might underperform (since the type of noise might differ).

As mentioned earlier, most of proposed SQA systems employ combinations of rules based on multiple features which need to be satisfied concurrently or in a cascade of decision steps. The order and combination of steps is mostly chosen either heuristically or to optimize system performance on training data in the same way as explained earlier. Often, the more basic checks such as flat line detection, physiological feasibility checks and checking for extreme signal behaviors are performed first; this way obviously low-quality segments are rejected quickly before more sophisticated and computationally intensive techniques are used for the ambiguous cases.

2.3.2.2 Machine Learning Models

In many cases, machine learning approaches are suitable for use in SQA systems. The more traditional rule-based decision-making approaches we have discussed so far, require prior information about the process to be modelled (i.e., information about how certain features differ in acceptable and unacceptable signal segments). Especially in the more ambiguous cases, the differences between acceptable and unacceptable signals are not clearly understood. Furthermore, the underlying processes which result in an unreliable signal are often very complex and multidimensional. Machine learning favors a *black box* approach: the relationship between features and labels does not need to be fully understood [29]; for a robust signal quality assessment system to be built, the system does not need to understand the underlying labeling process, it just needs to learn how to replicate it. The inclusion of aggregate data in machine learning models may reveal new information that is not seen by the individual. Machine learning models also offer the possibility to model the labeling process in a non-linear fashion which would not be possible, using traditional rule-based approaches. For robust machine learning models to be built, large amounts of labelled training data need to be available. In the way that the human expert, the gold standard of clinical decision-making, gains clinical acumen through experience, via machine learning, the system is taught how to perform a task (in this case classifying signal in terms of quality) using a lot of examples of how it should be done. The more information the system receives, the "experience" grows and the decision making is, thus, improved.

An overview of the use of machine learning in applications using multidimensional clinical data can be found in [29]. In the context of signal quality, Support Vector Machines (SVMs) have been used for building signal quality classifiers [9, 13] as well as Neural Networks via the Multi-Layer Perceptron (MLP) [13]. In Ref.

[9], the quality assessment system used wavelet entropy measurements from the HRV signal, derived from the ECG. A classification model was learnt using expertly annotated data from ambulatory patients, using SVMs with a radial basis function (RBF) kernel. Using only three features, a classification performance of 96% was achieved with a sensitivity of 92.3% and specificity of 99.1%. Despite no prior information being input into the classifier, the performance of the classifier using different combinations of features (which in this case were the entropies in different frequency bands, extracted via wavelet analysis, as explained in Sect. 2.1.3.4) gave an insight into the bandwidths where the differences between the two classes occurred (i.e., where the noise occurred).

In Ref. [13], the authors extracted 7 morphological, spectral and statistical features from 12 channels of simultaneously recorded ECG (see next section for an overview of multi-lead approaches), resulting in 84 features in total. A classifier was then trained using a Support Vector Machine (SVM) also with an RBF kernel and a standard feed-forward Multi-Layer Perceptron Neural Network (MLPNN). Classifiers were firstly tested using all 84 features and then using only the 7 features from a single lead. Like the approach by [9], explained earlier, different combinations of the seven features were tested to find the best. For the single-lead classifiers the results from the MLP and SVMs were almost the same, with the SVM approach slightly outperforming the MLP, with a classification accuracy of 96.5%, Sensitivity of 97.2% and Specificity of 95.8% on the test set using only four features. For the 12-lead case, the best classification accuracy occurred when using five features using the MLP, with an accuracy of 95.9% on the test set.

As discussed in the earlier discussion of empirical determination of thresholds, also in the case of machine learning models, the dataset used for training the system needs to be varied if the aim is for the system to be able to generalize to data from various monitors. Often, quality assessment systems are designed for on-board processing and in that case parameters can be tailored to that specific scenario. If the aim is for the system to be used independently of the monitor used, it needs to be trained on data from a variety of sources.

2.4 Multi-lead Approaches

So far, we have considered approaches extracting features and setting decision rules which can be used for the quality assessment of single-lead ECGs. We will now consider the situation where multiple channels of simultaneously recorded ECG is available. The presence of multiple channels of signal can be exploited for producing more reliable vital sign measurements, by employing methods based on the fusion of information from the various channels. Compared to single-channel approaches, when multiple channels of ECG are available system performance may be improved via a) the extraction of new features derived from the relationship between the different signals b) the use of an increased number of features, compared to single-lead cases, extracted from the different available signals. The 2011

Physionet Challenge addressed this exact problem: the development of an algorithm for assessing the quality of 12-lead ECG recordings collected via a mobile phone in real-time [17, 30]. Researchers had access to 1500 expertly annotated, 10 s in duration 12-lead ECG segments, sampled at 500 Hz, two thirds of which were to be used for training and the rest for testing. Next, we discuss approaches proposed by researchers which exploit the relationship between the multiple leads of ECG to obtain robust quality assessment.

2.4.1 Features Specific to Multi-lead ECGs

Correlation between different leads

According to Mogardo et al. [31] "the 12-lead ECG signals are different projections of the same electrical activation process of the heart". As a result, it is expected that in a clean recording, the 12 leads should look very similar. In a 12-lead ECG contaminated with noise, the similarity between the different leads is presumably reduced (since noise may be manifested differently in the various leads). Correlation, defined in Sect. 2.3.1.3, measures the similarity between two time-series and has been proposed by Kranstein [32] as a quality feature measuring the similarity between different leads. While the way Kranstein measured correlation between the twelve leads is unclear, the potential of the technique may be demonstrated by calculating the correlation coefficient between each possible pair of leads in a 12-lead ECG signal (66 combinations) and taking the mean *absolute* correlation (to account for inverted leads). Such examples for acceptable and unacceptable signals are illustrated in Figs. 2.11 and 2.12, respectively, along with the average inter-lead correlation coefficient. For the acceptable signal, the average absolute correlation coefficient was 0.67, while for the unacceptable signal it was 0.31. In proposed methods, correlation measures were used as inputs into machine learning classifiers, however, it may be possible for the average absolute correlation coefficient to be used as part of a rule-based system, where a threshold of acceptance is determined on available data.

The concept of measuring cross-correlation between the different leads was taken a step further by Morgado [31] who, using the data from the 2011 Physionet Challenge, calculated the covariance matrix of the leads (a matrix in which each element z, k is the covariance value between leads z and k). Like correlation, covariance is also a measure of similarity between two time-series x_i and y_i, given by

$$\text{cov}\ (x, y) = \sum_{i=1}^{N} (x_i - \widehat{x})(y_i - \widehat{y}), \tag{2.5}$$

where $\widehat{x}$ and $\widehat{y}$ are the empirically determined means of x_i and y_i, respectively, and N is the number of samples in the two time-series. (Comparing with the formula for

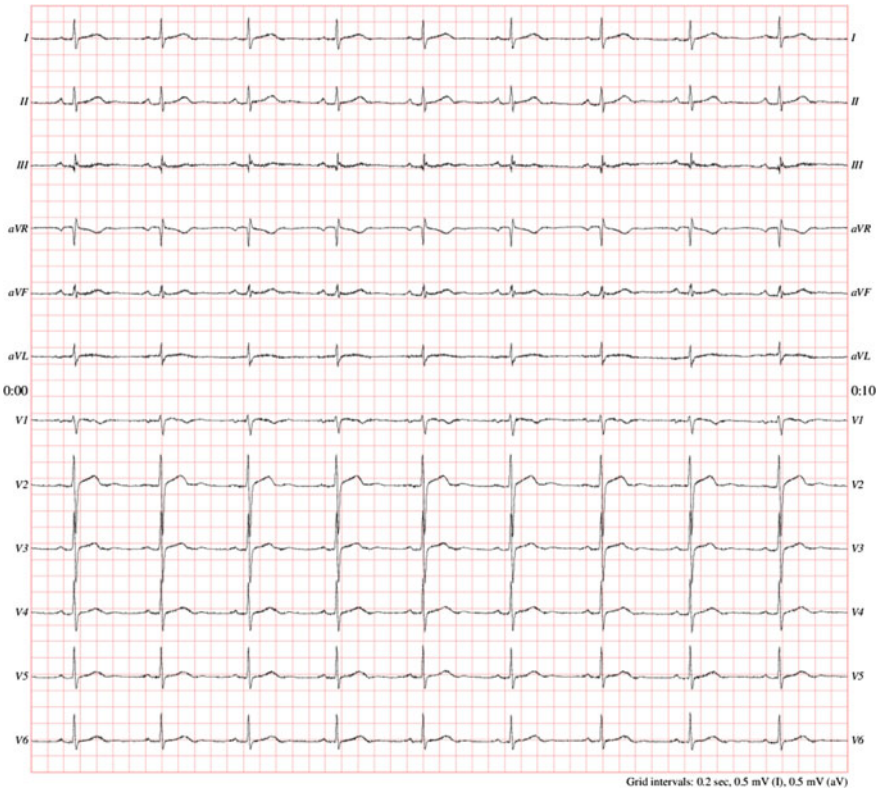

Fig. 2.11 Example of acceptable 12-lead ECG from the 2011 Physionet Challenge (image taken from Physiobank ATM [17]). The average inter-lead correlation coefficient was 0.67

the correlation coefficient (3), we can see that the covariance is the unnormalized equivalent of the correlation coefficient.). The authors used 8 out of the 12 leads of ECG and, thus, extracted an 8×8 covariance matrix which they then analyzed further to extract the eigenvalues. The eigenvalues were then used as features, fed into three binary classifiers, trained on the training data available. The highest accuracy obtained using this novel technique was 92.7% with a sensitivity of 83.1% and a specificity of 95.5%.

Agreement in QRS Detection Between Leads

In a clean 12-lead ECG beats should be detected successfully in all leads at the same time-points. Failure for this to happen signifies the presence of artefact. On this basis, the agreement in QRS detection between leads was proposed as a quality feature by [13] and used as one of the seven features fed into a machine learning-based classifier. This quality index was defined as "the percentage of beats detected on each lead which were detected on all leads" and was measured separately for each lead.

Fig. 2.12 Example of unacceptable 12-lead ECG from the 2011 Physionet Challenge (image taken from Physiobank ATM [17]). The average inter-lead correlation coefficient was 0.31. (Note that one lead is missing)

Lead Crossing Point Detection

This measure, proposed by Hayn et al. [21], was based on the visualization method used by Physionet's ATM [17], which plots all 12 leads in succession. When artefact was present, often one or more of the leads was plotted "over" other leads, obscuring them. Since human annotators used the same view for labeling the ECG records, the authors proposed a quality measure which would penalize records where one or more leads drifted onto another lead. They proposed a measure where the measured the number of crossing points of one lead with any other leads and if the maximum number exceeded 4.9 per second (this was presumably determined

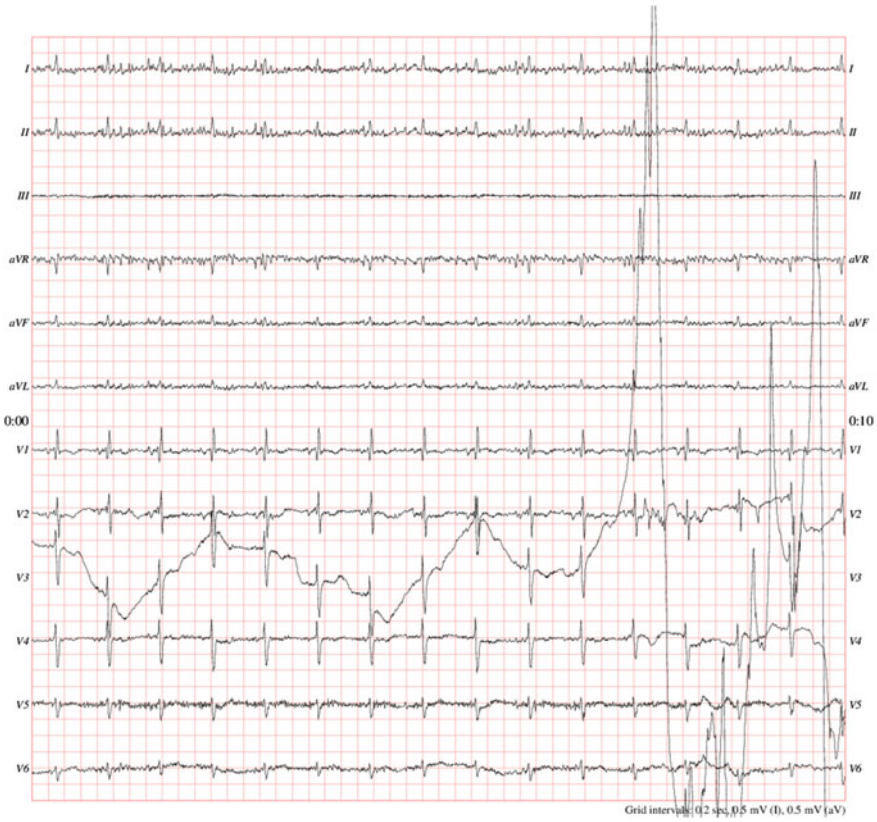

Fig. 2.13 Example of 12-lead ECG from the 2011 Physionet Challenge (image taken from Physiobank ATM [17]) with drifting leads interfering with other leads' plots on the right-hand side

empirically), the signal was classified as "unacceptable". Figure 2.13 shows an example of an ECG record from the Physionet ATM with drifting leads on the right-hand side.

In addition to new features exploiting the relationship between different leads, most researchers proposed techniques which relied on the extraction of features from every lead and using them together either in a rule-based or machine-learning-based scheme for assessing the quality of the ECG records.

2.5 Summary

This chapter introduced the ECG and provided an overview of proposed approaches for feature extraction and of the decision rules employed for designing SQA systems for the ECG. Single-lead approaches have more practical applications but

where available, multi-lead approaches may significantly improve robustness. Not all past and currently-proposed techniques can be presented. Rather, this chapter is intended to introduce the reader to a variety of proposed approaches which have proved successful, and the underlying principles underpinning them, such that further research can be triggered, to improve the robustness of current approaches.

References

1. Clifford, G. D., Azuaje, F., & McSharry, P. E. (2006). *Advanced methods for tools for ECG data analysis*. Norwood, MA: Artech House.
2. Hurst, J. W. (1998). *Naming of the waves in the ECG, with a brief account of their genesis. Circulation*, *98*, 1937–1942 (originally published November 3, 1998).
3. Romero, I., Penders, J., & Kriatselis, C. (2010). P-wave analysis for atrial fibrillation detection in ambulatory recordings. *Journal of Electrocardiology, 43*(6), 647.
4. Monasterio, V., Laguna, P., & Martínez, J. P. (2009). Multilead analysis of T-wave alternans in the ECG using principal component analysis. *IEEE Transactions in Biomedical Engineering, 56*(7), 1880–1890.
5. Channer, K., & Morris, F. (2002). Myocardial ischaemia. *BMJ: British Medical Journal, 324* (7344), 1023–1026.
6. Jacqueline, M., Dekker, J. M., Schouten, E. G., Klootwijk, P., Pool, J., & Kromhout, D. (1995). ST segment and T wave characteristics as indicators of coronary heart disease risk: The Zutphen study. *Journal of the American College of Cardiology, 25*(6), 1321–1326.
7. Pieper, S. J., & Hammill, S. C. (1996). Heart rate variability: Technique and investigational applications in cardiovascular medicine. *Mayo Clinic Proceedings, 70*(10), 354–381.
8. Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering, BME 32*(3), 230–236.
9. Orphanidou, C., & Drobnjak, I. (2017). Quality assessment of ambulatory ECG using wavelet entropy of the HRV signal. *IEEE Journal of Biomedical and Health Informatics*, *21*(5), 1216–1223.
10. Orphanidou, C., Fleming, S., Shah, S. A., & Tarassenko, L. (2013). Data fusion for estimating respiratory rate from a single-lead ECG. *Biomedical Signal Processing and Control, 8*(1), 98–105.
11. Lian, J., Wang, L., & Muessig, D. (2011). A simple method to detect atrial fibrillation using RR intervals. *The American Journal of Cardiology, 107*(10), 1494–1497.
12. Murthy, V. K., Grove, T. M., Harvey, G. A., & Haywood, L. J. (1978). Clinical usefulness of ECG frequency spectrum analysis. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (pp. 610–612), November 1978.
13. Clifford, G. D., Behar, J., Li, Q., & Rezek, I. (2012). Signal quality and data fusion for determining the clinical acceptability of electrocardiograms. *Physiological Measurement, 33,* 1419–1433.
14. Jekova, I., Krasteva, V., Christov, I., & Abacherli, R. (2012). Threshold-based system for noise detection in multilead ECG recordings. *Physiological Measurement, 33,* 1463–1477.
15. Hamilton, P. S., & Tompkins, W. J. (1986). Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE Transactions on Biomedical Engineering, 33* (12), 1157–1165.
16. Zong, W., Moody, G. B., & Jiang, D. (2003). A robust open-source algorithm to detect the onset and duration of QRS complexes. *Computing in Cardiology Conference, 30,* 737–740.
17. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ch, Ivanov P., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220.

18. Moody, B. E. (2011). Rule-based methods for ECG quality control. In *Computing in Cardiology Conference 2011* (Vol. 38, pp. 361–363).
19. Langley, P., Di Marco, L. Y., King, S., Duncan, D., Di Maria, C., Duan, W., et al. (2011). An algorithm for assessment of quality of ECGs acquired via mobile phones. *Computing in Cardiology Conference, 38,* 281–284.
20. Johannesen, L. (2011). Assessment of ECG quality on an android platform. *Computing in Cardiology Conference, 38,* 433–436.
21. Hayn, D., Jammerbund, B., & Schreier, G. (2012). QRS detection based ECG quality assessment. *Physiological Measurement, 33,* 1449–1461.
22. Ho Chee Tat, T., Xiang, X., & Thiam, L. (2011). Physionet challenge 2011: Improving the quality of electrocardiography data collected using real-time QRS-complex and T-wave detection. *Computing in Cardiology, 38,* 441–444.
23. Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., & Tarassenko, L. (2015). Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE Journal of Biomedical and Health Informatics, 19*(3), 832–838.
24. He, T., Clifford, G. D., & Tarassenko, L. (2006). Application of independent component analysis in removing artefacts from the electrocardiogram. *Neural Computing and Applications, 15,* 105–116.
25. Li, Q., Mark, R. G., & Clifford, G. D. (2008). Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiological Measurement, 29*(1), 15–32.
26. Allen, J., & Murray, A. (1996). Assessing ECG signal quality on a coronary care unit. *Physiological Measurement, 17,* 249–258.
27. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423, 623-656.
28. Daluwatte, C., Johannesen, L., Galeotti, L., Vicente, J., Strauss, D. G., & Scully, C. G. (2016). Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs. *Physiological Measurement, 37,* 1370–1382.
29. Orphanidou, C., & Wong, D. (2017). Machine learning models for multidimensional clinical data. In S. U. Khan, A. Y. Zomaya, & A. Assad (Eds.), *Handbook of large-scale distributed computing in smart healthcare, scalable computing and communications* (pp. 177–216). Cham: Springer.
30. Silva, I., Moody, G. B., & Celi, L. (2011). Improving the quality of ECGs collected using mobile phones: The PhysioNet/Computing in Cardiology Challenge 2011. *Computing in Cardiology Conference, 38,* 272–276.
31. Morgado, E., Alonso-Atienza, F., Santiago-Mozos, R., Barquero-Pérez, Ó., Silva, I., Ramos, J., et al. (2015). Quality estimation of the electrocardiogram using cross-correlation among leads. *Biomedical Engineering Online, 14,* 59.
32. Kalkstein, N., Kinar, Y., Na'aman, M., Neumark, N., & Akiva, P. (2011). Using machine learning to detect problems in ECG data collection. *Computing in Cardiology Conference, 38,* 437–440.

Chapter 3
Quality Assessment for the Photoplethysmogram (PPG)

Abstract The Photoplethysmogram (PPG) is fast becoming the most popular monitoring tool because of its ease of measurement, via the pulse oximeter, and because of its ability to provide multiple vital sign measurements from a single signal. However, its high susceptibility to motion artifact limits its reliability for deriving valid vital sign measurements. This chapter introduces the PPG and presents currently proposed approaches for making PPG quality assessments. Rather than presenting proposed techniques as complete solutions, state-of-the-art feature extraction approaches are presented along with widely used decision rules, to provide the reader with a basic understanding of the framework of SQA for the PPG, and encourage further research.

Keywords Photoplethysmogram · PPG · Continuous monitoring · Signal quality assessment · Feature extraction · Decision rules · Data fusion

The Photoplethysmogram (PPG) has long been accepted as a reliable source for measurements of peripheral arterial oxygen saturation (SpO_2) and heart rate (HR). More, recently, much research effort has been put into the addition of reliable breathing rate (BR) [1, 2] and blood pressure (BP) [3] estimations from the PPG, with promising results. As the healthcare community moves into 24-h continuous monitoring for ambulatory patients, the PPG is fast becoming the most popular monitoring tool. It is available in even the most basic clinical environments and it fulfills the desire for small, reliable, low-cost and unobtrusive monitoring, where all basic vital signs (HR, SpO_2, RR and BP) are measured using a single probe. However, the high susceptibility of the PPG to motion artifact limits its reliability for deriving valid vital sign measurements in real-time ambulatory environments. In this chapter, we discuss various methods for assessing the quality of the PPG which may be used for rejecting vital sign measurements which are derived from unreliable segments of signal. Because of the many types of noise that may be present in the PPG and because of its morphological variability, quantifying noise is not straightforward. However, several proposed techniques have been shown to provide efficient, accurate and robust quality assessment and to improve the reliability of vital sign measurements. While the PPG is susceptible to different forms of noise,

C. Orphanidou, *Signal Quality Assessment in Physiological Monitoring*,
SpringerBriefs in Bioengineering, http://doi.org/10.1007/978-3-319-68415-4_3

motion artefact has been the most troublesome; dealing with motion artifact is becoming an even bigger challenge as monitoring devices are being progressively miniaturized and with the development of contact-less PPG [4]. While many techniques are being proposed for eradicating the motion artifact from the PPG, our discussion will only focus on techniques for identifying noise such that artefact-corrupted signal can be ignored. Many important state-of-the-art time-domain and frequency domain approaches will be discussed, accompanied by examples on real-life PPG collected via wearable sensors.

3.1 The Photoplethysmogram (PPG)

Photoplethysmography is an optical measurement technique which measures the volume changes in blood as it moves from the heart towards the periphery and specifically the measurement site, which is most commonly the fingertips [5]. To record a PPG signal, a light emitting diode (LED) illuminates tissue (close to a vessel) with two different wavelengths (red and infrared) and a photodiode on the other side of the tissue measures the intensity of the non-absorbed light at each wavelength. This technology, invented by Takuo Aoyaki, is governed by two physical principles: (a) light absorbance is different for oxygenated and non-oxygenated hemoglobin at the two wavelengths used, and (b) at both wavelengths, the absorbance has a pulsatile (AC) component which reflects the pulsations from the cardiac cycle [6]. The non-pulsatile (DC) component of the PPG signal comprises absorption from the tissue and bones (which are nonchanging), as well as static blood absorption (arterial, venous and in smaller amounts, capillary) [6]. The various components comprising the PPG signal can be seen in Fig. 3.1.

The measurement of oxygenated and non-oxygenated hemoglobin at the red and infrared wavelengths can be used against a standard calibration curve to derive a ratio of the oxygenated-over-total hemoglobin which had been termed SpO_2 and is known as the peripheral oxygen saturation. Since the variation in blood volume reflected in the AC component of the PPG is a result of the pulsation of heart, the PPG signal can be used for deriving the heart rate (HR) (sometimes referred to as pulse rate when derived from the PPG) simply by identifying the peaks in the PPG waveform.

3.1.1 PPG Morphology

Figure 3.2 shows an individual pulse wave from a PPG signal. It comprises of the anacrotic phase (the rising component) which signifies the systole of the heart and the catacrotic phase (the falling component) which signifies the diastole and wave reflections in the periphery [5]. Because of the processes they represent, the two phases are often referred to as the systolic and diastolic waves, respectively. A little

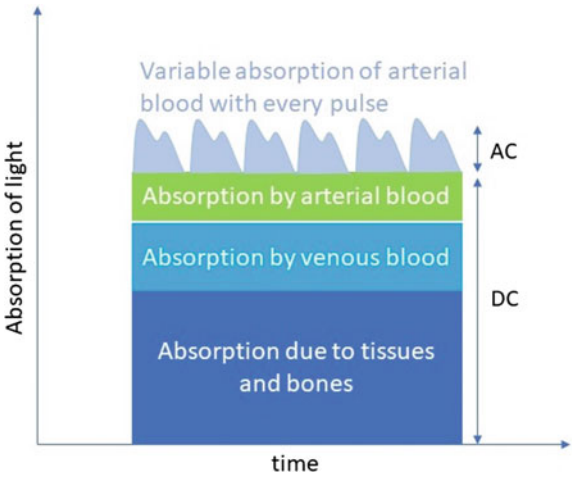

Fig. 3.1 Light absorption which forms the PPG signal. The static (DC) component reflects light absorption from the tissues and bones and static blood and the AC component reflects the pulsations from the cardiac cycle (image adapted from [7])

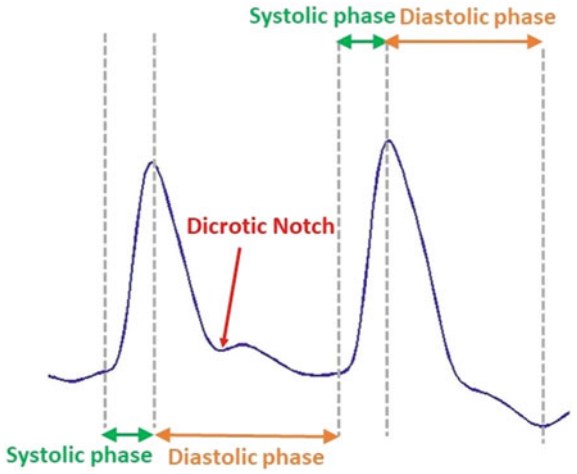

Fig. 3.2 Example of PPG pulse waves with and without dicrotic notch which occur on the same PPG recording. The beginning of the systolic phase signifies the beginning of a pulse and the end of the diastolic phase signifies its end

valley called the dicrotic notch is sometimes seen in the catacronic phase; it signifies the closure of the aortic valve and is usually seen in the PPG of young subjects with healthy artery function [5].

In standard clinical settings, the PPG signal is often displayed on the patient monitors along with the HR measurement extracted. For correct HR measurement from the PPG, identification of the pulse waveform peaks is sufficient. Correct identification of pulse peaks can also aid in the identification of arrhythmias [8].

The variation in the time intervals between successive pulse peaks, called Pulse Rate Variability (PRV) (equivalent to Heart Rate Variability (HRV) for the ECG, provided a sufficiently high enough sampling frequency is used [9]) can also provide useful information for the health and functioning of the Autonomic Nervous System (ANS) and can also be analyzed further for the extraction of Breathing Rate (BR). More specifically, BR information may be extracted from the HF component of the PRV signal [1] (it has also been shown that BR-related information can be extracted from the amplitude and baseline wander modulation of the PPG [2, 10]). While for HR and BR measurements, peak pulse identification is enough, the full diagnostic potential of the PPG requires analysis of its morphology in a more detailed way. Applications exploiting time and frequency-domain characteristics of the PPG for clinical diagnosis include the identification of atherosclerosis [11], artery disease [12], and vascular assessment [5], amongst others.

3.1.2 Spectral Characteristics of the PPG

The power spectrum of a clean PPG from a healthy subject can be seen in Fig. 3.3. Frequencies up to 5 Hz are shown. A typical PPG spectrum will exhibit two peaks; one at the LF end of the spectrum (under 0.5 Hz and most commonly around 0.1 Hz) which represents mostly the activity of the sympathetic nervous system, related to respiration and thermoregulation [13]. The HF component is in the region of 0.5–2 Hz and represents the heart rate. In Fig. 3.3 the HR is at 0.97 Hz (58 bpm). Commonly, any information above 2 Hz concerns only the harmonics of the HR frequency (as can be clearly observed in Fig. 3.3) and has no clinical significance [14]. (In this specific case because the HR frequency was below 1 Hz the first harmonic is observed below 2 Hz.)

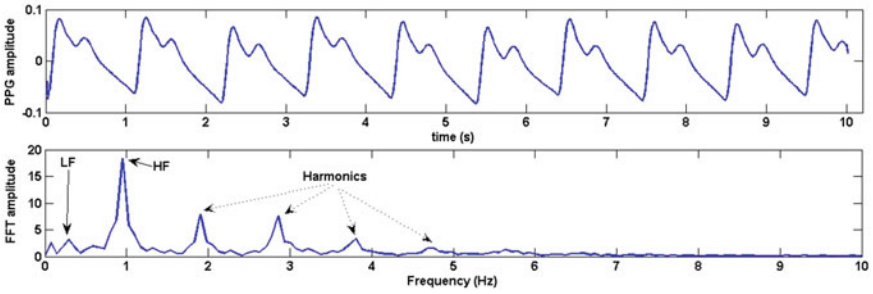

Fig. 3.3 Ten seconds of 500-Hz PPG in sinus rhythm and associated Fast Fourier Transform (FFT) spectrum of the signal. We only show frequencies up to 10 Hz since the remaining components are of no clinical interest. We note the peak power at 0.97 Hz (58 bpm) which is the HR (this can be verified from the PPG plot) and secondary peaks around the LF region (<0.5 Hz) which represent the activity of the sympathetic nervous system. Other peaks in the spectrum represent harmonics of the HR frequency and have no clinical significance

In the presence of noise, it is possible that motion-induced frequencies would coincide with either the LF or HF frequencies, distorting the spectrum of the signal and making interpretation challenging.

3.2 PPG and Noise

Under static recording conditions, identifying the key points on the PPG is straightforward; in the presence of motion artefact, however, the resulting distortion in the PPG waveform, makes its analysis and interpretation challenging. This is even more pronounced when the motion artefact occurs at the same time as incidents of low perfusion (reduced blood flow, often the result of atherosclerosis): when the perfusion is low, the difference between systolic and diastolic pressure is small and the pulse wave is weak; in the presence of noise, the physiological signal may be completely lost amongst the noise. In the clinical environment, motion artefact has been shown to occur because of involuntary activities (such as seizures, shivering or motion during transport) and voluntary activities such as rubbing, scratching, waving and using the hands in order to eat. It has been shown that 70–71% of alarms from pulse oximeters are false positives which occur because of erroneous measurements due to motion [6, 15]. Another recent study showed that apnea-related false desaturation alarm rates were up to 85% [16]. When pulse oximeters are to be used for home-monitoring or for lifestyle/athletic monitoring, the frequency of artifacts significantly increases [17, 18]. Figure 3.4 illustrates the

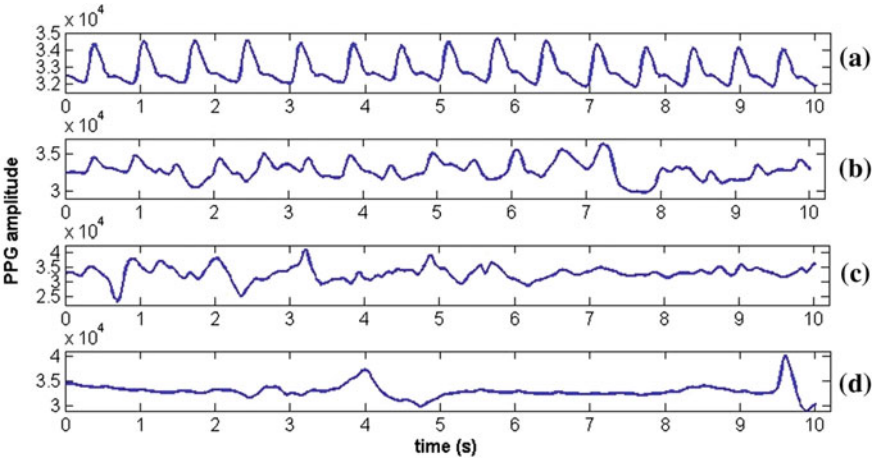

Fig. 3.4 Ten seconds PPG segments with distinct kinds of noise. **a** Clean PPG segment **b** Noisy PPG segment which may result in a HR measurement within an acceptable margin of error but would be diagnostically unusable **c** and **d** Unusable PPG segments. As the segments get noisier, the pulse-wave morphology is distorted or disappears altogether

effect of noise on the PPG using examples of signal recorded from ambulatory patients.

Recenlty, some research effort has been directed towards techniques for analyzing the PPG signal together with simultaneously-recorded accelerometry signals (recorded, for example, in a wrist-watch type monitor). The idea is to identify the main frequencies of motion from the accelerometry signals and remove them from the PPG spectrum in order to calculate a "clean" value of HR [19]. Despite the approach showing promise, relying on the simultaneous measurement of accelerometry is not realistic in many clinical scenarios, since monitoring equipment needs to be kept to a minimum to maximize patient compliance [20] and minimize power consumption.

Additionally, much research effort has been directed towards the suppression of artifact from signals using filtering techniques [21, 22] and signal separation techniques such as Independent Component Analysis [23, 24] and Ensemble Empirical Mode Decomposition [25]. A weakness of these techniques is that they assume that noise occurs in a different domain (time, frequency or statistical) than the physiological signal and that it can be somehow separated and removed [26]. That is often not the case, resulting in the extraction of erroneous vital sign measurements. Identifying signal artefact and rejecting it will result in increased reliability of measurements, even if they are obtained at a lower frequency (due to signal retention).

3.3 PPG Quality Considerations

The definition of signal quality differs depending on the application. We extend the definition of signal quality for the ECG from the previous chapter to the PPG by defining signal quality in two ways:

- *Basic quality*: Pulse peaks are clearly identifiable. In this case, a reliable HR can be extracted as well as some types of arrhythmias and PRV information.
- *Diagnostic quality*: The pulse wave waveforms is clean with systolic and diastolic waves visible. In this case, the signal can also be used for clinical diagnosis.

Unlike the ECG, the diagnostic value of the PPG (based on morphological analysis) is still within the domain of research and its main usage is restricted to the measurement of vital signs. As a result, in the next sections we focus on discussing techniques for assessing the requirement of *basic quality* of PPG.

Even in noise-free PPG recordings from healthy individuals, the signal may be highly non-stationary. Within a short window of signal, pulse waves may vary in height, width and morphology due to physiological changes, sensor type and location. Hardware type and software pre-processing requirements also affect pulse wave morphology, making the extraction of standardized features that can

differentiate between acceptable and unacceptable segments of signal challenging [27]. Additionally, the absence of a widely accepted pulse-peak detector for the PPG introduces an additional challenge to the generalizability of proposed techniques.

For the remainder of this chapter we will take a close look at different approaches proposed in literature for assessing the quality of PPG signals. Proposed approaches rely on extracting morphological, spectral or trend-based characteristics from a single PPG waveform and applying heuristic, empirically determined or machine learning-based decision rules.

3.4 PPG Feature Extraction

Many signal quality assessment algorithms for the PPG propose a series of steps which quickly identify obviously unacceptable segments of signal using simple feasibility rules. These initial rules often detect a significant percentage of unacceptable segments. For the questionable segments and borderline cases more sophisticated techniques are used which extract time-domain or frequency-domain features that can differentiate between acceptable and unacceptable PPG segments, either by comparing with expected pulse wave morphology or by trend-based checks (since a clean segment of signal is expected to be homogeneous in morphology and pulse wave characteristics).

We will begin by considering simple feasibility checks and then proceed to discuss time-domain and frequency-domain PPG features which have been shown to successfully differentiate between *acceptable* and *unacceptable* PPG segments. For the quality assessment of the PPG, most proposed techniques rely on the extraction of morphological features or measures of regularity. Frequency-domain approaches are scarce; we touch upon some of the proposed ones. The features discussed in the next sections have been proposed in literature and have been used in a variety of combinations as part of rule-based or machine learning-based systems for providing assessments as to the acceptability of the PPG for providing reliable vital sign measurements.

3.4.1 PPG Beat Detection

The morphological variability of the PPG signal makes the problem of pulse peak detection challenging. Unlike the ECG, there is no universally accepted approach for pulse peak detection and most proposed works on signal quality assessment which require beat detection as a first step, use custom-made pulse peak detectors. This is a critical issue since systems which rely on robust pulse peak detection are intrinsically linked to the detector used and its robustness to noise: it is impossible to extract and compare morphological features if pulse peaks are not correctly

detected. At the same time, the failure to correctly identify pulse peaks, is usually an indication of the presence of artefact. In conclusion, morphology-based systems which rely on accurate pulse-peak detection, need to be always used in conjunction with the same pulse peak detector. Especially in the case of SQA systems which rely on the use of heuristic decision rules, it is likely that when a different pulse peak detector is used, the system will underperform.

It is worth mentioning that most works on signal quality for the PPG do not provide details on the way that pulse peaks were determined. A pulse-peak detector which has been proved popular and has been used in several relevant publications is an adaptation of [28], an algorithm initially proposed for detecting the onset of Arterial Blood Pressure (ABP) pulses. This algorithm works by firstly calculating the slope sum function (SSF) of the signal and then using adaptive thresholding on its amplitude in combination with a local search function [28]. This algorithm may be adapted to the morphology of the PPG and used for the detection of the onsets of the PPG pulse waves [27].

3.4.2 Time-Domain Features

3.4.2.1 Feasibility Checks

Following beat detection, the first step to an SQA approach is to apply simple viability rules which are often applied in a cascade of decision steps. Even though such rules cannot fully assess signal quality they can be useful for quickly rejecting segments of low quality, or identifying segments of saturated or unusable signal.

Clipping/Signal Saturation Detection
To check for clipping/signal saturation to a minimum or maximum value, a hysteresis threshold was determined by Li and Clifford [27], defining the smallest fluctuation values which should be ignored. The percentage of the pulse wave which did not fall within the fluctuation limits for "clipped signal" was then used as a signal quality measure, identifying low quality pulse beats. Fischer et al. [18] also applied an initial clipping check with upper and lower clipping thresholds determined from the signal channel maximum and minimum.

Physiological Viability
The natural limits of cardiovascular physiology which drive the morphology of the ECG signal, also apply to the PPG, provided a reliable pulse peak detector is used. In the same manner as with the ECG signal, these limits can be used as signal quality indicators either as individual viability checks or as a cascade of checks which can quickly identify very-low-quality segments and label them as "unreliable". Following pulse peak detection, the following limits can be used to check for physiological viability (please note that the same limits have been proposed for the ECG, as discussed in the previous chapter):

- *Average HR*: the average HR measured within a segment of signal needs to satisfy a physiologically valid rate. Proposed limits in literature were 40–180 beats per minute (bpm) (0.67–3 Hz) [29]. In the state of exercise the limits need to be adjusted to allow for increased valid HRs (e.g. up to 300 bpm (5 Hz)). Any estimated HR outside these limits is not physiologically viable and the segment can be automatically rejected.
- *Maximum P-P interval (where P indicates the Pulse Peak and the P-P interval is equivalent to the R-R interval for the ECG)*: based on a minimum viable HR of 40 bpm, the maximum P-P interval within a segment of signal cannot exceed 60/40 bpm = 1.5 s. Allowing for a single missed beat, a maximum R-R interval limit of 3 s has been proposed as a simple feasibility check [29].
- *Maximum P-P interval/Minimum P-P interval*: under normal monitoring conditions, HR is not expected to change rapidly within a short time-segment. As a result, when considering short segments of PPG, P-P intervals are expected to change only within the limits explained by the natural Pulse Rate Variability (PRV-equivalent to the HRV for the ECG[1]). In a short segment of signal (e.g. 10 s), the ratio of the maximum to the minimum R-R interval should not exceed 1.1, since a change in HR of over 10% within 10 s is highly unlikely. Allowing for a single missed beat, a maximum ratio of 2.2 has been proposed as a feasibility rule for acceptable PPG [29].

As discussed in the previous chapter, when feasibility rules are used which are based on the distribution of P-P intervals, it is important for limits to consider the possible presence of arrhythmias, as they would affect the distribution of P-P intervals. Additionally, the last feasibility check is only valid when analyzing short segments of signal and must be adapted for larger segments since in larger time intervals it is possible to have greater changes in the values of instantaneous HR measured via the P-P interval.

3.4.2.2 Morphological Features

Most methods relying on morphological analysis of the pulse wave incorporated pre-processing filtering steps before feature extraction to remove information which was outside the physiological range of HR. Fischer et al. [18] used a 15 Hz low-pass and 0.1 Hz high-pass filter in succession, before proceeding with beat detection and feature extraction. Sukor et al. [31], on the other hand, used a 0.5–5 Hz Butterworth filter, assuming an upper limit of HR of 300 bpm.

Morphological analysis of the PPG, using heuristic thresholds for the various characteristics of the pulse wave have been used extensively in proposed signal quality assessment algorithms for the PPG, either as initial checks for quickly

[1]Studies have shown that provided that a high enough sampling rate is used for the PPG signal and a robust pulse-peak detector is used, PVR is approximately equivalent to the HRV for describing the activity of the Autonomic Nervous System (ANS) [30].

identifying obviously unacceptable beats or as part of multi-feature rule-based systems. While many morphological checks are valid on their own and can identify many disturbed beats, when used in a cascade, they have the advantage of successively identifying more disturbed beats and rejecting them without the danger of false positives, i.e. the incorrect classification of clean beats [18]. Figure 3.4 shows a diagram of a pulse waveform with the different wave characteristics. We base our definitions on the terminology used by Fischer et al. [18], and extend it.

After beat detection, the following morphological characteristics/features were proposed in literature for assessing the quality of PPG pulse waves:

- *Systolic phase duration (SP)/rise time*: this was used by [18] with a permitted range of 0.08 to 0.49 s based on literature reported values.
- *Ratio of systolic to diastolic phase (SP/DP)*: an upper permitted limit of 1.1 was used by [18].
- *Pulse wave duration (PWD)*: a range of 0.27–2.4 s (corresponding to HR values in the range of 25–220 bpm) was used by [18]. Sukor et al. [31] also used this measure as a quality feature using heuristically-determined limits of acceptability.
- *Number of diastolic peaks*: up to 2 diastolic peaks are considered acceptable in the same pulse waveform [18].
- *Variation in pulse wave duration (PWD), systolic phase duration (SP) and pulse wave amplitude (PWA) between successive pulse waves*: PWD and SP variation must be within 33–300% and PWA variation must be within 25–400%. These measures ensure the relative changes are within the permitted limits which were derived from extreme values reported in literature [18].
- *Pulse Wave Amplitude (PWA)*: PWA should not exceed a threshold which was set heuristically [31].
- *Trough depth differences between successive troughs*: the difference between the depth of successive troughs should not exceed a heuristically-determined threshold [31].
- *Relation between pulse wave amplitude (PWA), dicrotic notch amplitude (NA)(if present) and diastolic wave amplitude (DWA)*: [32] used the relationship between these three measures in order to reject pulse waveforms whose patterns were outside some acceptable limits.

The last two features check for similarities between successive pulse waves because in a clean segment of PPG, it is not expected that pulse wave morphology will differ much. In the presence of artefact, pulse wave distortion will result in morphological differences between pulse waves within a short time-window. In the next section, we discuss techniques proposed in literature for assessing the quality of PPG segments which are based on the expectation of regularity and homogeneity in a clean segment of PPG (Fig. 3.5).

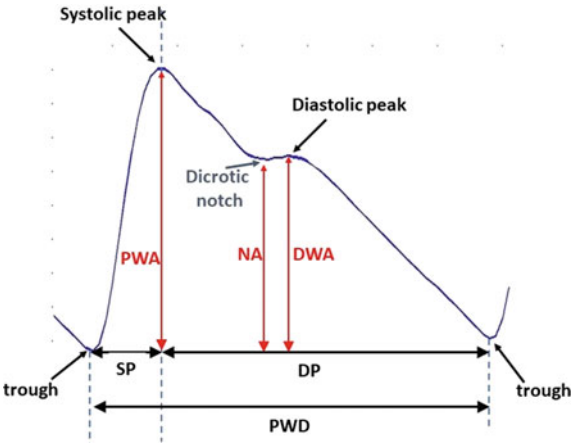

Fig. 3.5 PPG pulse waveform with wave characteristics. *PWD* Pulse wave duration, *SP* systolic phase, *SD* diastolic phase (also termed "rise time" in some publications), *PWA* pulse waveform amplitude, *NA* dicrotic notch amplitude, *DWA* diastolic wave amplitude

3.4.2.3 Trend-Based Approaches

We now discuss trend-based approaches for signal quality assessment which search for regularity in a short segment of PPG. As discussed before, in a clean segment of PPG, pulse waves are expected to be morphologically similar. In the presence of noise, pulse waves become distorted and less similar to their neighboring ones (within a short window of signal). The techniques we discuss next, search for regularity in a segment of PPG, using different proposed measures.

High Order Statistics—Skewness and Kurtosis

Higher order statistics of the distribution of values of a PPG segment give an indication regarding the presence of outliers in the signal which, in turn, indicate the presence of artefact. Skewness and Kurtosis are the third and fourth standardized moments of a probability distribution; skewness measures how symmetric the distribution is and kurtosis measures how sharp the peak of the distribution is. A distribution with many outliers will have low values of skewness and kurtosis since its probability distribution will be asymmetric, flatter and will approach zero slower. A distribution with no outliers will be symmetric, will have a sharp peak and thus approach zero quickly having higher values for skewness and kurtosis. Measurements of skewness and kurtosis have, therefore, been proposed as indicators of the presence of outliers in the PPG segment, which is equivalent to the presence of noise. As defined in the previous chapter, in a discrete time-series, skewness $(\widehat{S})$ and kurtosis $(\widehat{K})$ are calculated from

$$\widehat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x})^3}{\widehat{\sigma}^3} \tag{3.1}$$

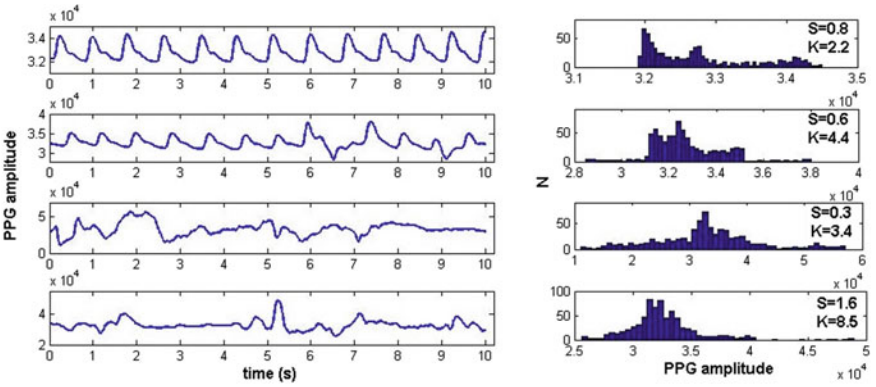

Fig. 3.6 Examples of acceptable and unacceptable PPG segments with associated discrete distributions and skewness (S) and kurtosis (K) values. While skewness and kurtosis are good indicators of signal quality in most cases, in some cases (such as the fourth example from the top) noisy segments may have high values, depending on the type of noise present. As a result, these features may be best used in combination with other quality-related indices

$$\widehat{K} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x})^4}{\widehat{\sigma}^4} \tag{3.2}$$

where x_i is the discrete signal, $\widehat{\mu}$ and $\widehat{\sigma}$ are the empirical estimates of the mean and standard deviation of the distribution of x_i and N the number of data points in the signal. Krishnan et al. [24] used the skewness and kurtosis-based quality index together with other extracted features within a fusion decision scheme. The kurtosis measure was also used by Selvaraj et al. [33] who used ROC analysis to optimize the threshold for acceptable/unacceptable, used within a decision rule scheme with other quality-relevant features. Figure 3.6 shows examples of acceptable and unacceptable PPG segments with their associated distributions and skewness and kurtosis values.

Shannon Entropy

Shannon entropy (SE) is a measure of the information, choice and uncertainty [34] in a system. It has been proposed as a quality-measure for the PPG by Selvaraj et al. [33] on the basis that the presence of noise adds uncertainty to the system. Since, mathematically, SE can be said to describe how much a probability distribution deviates from the uniform distribution [35], it is an indication of non-stationarity and, as an extension, the presence of noise. For a discrete time-series, SE is given by

$$SE = \sum_{i=1}^{N} x_i^2 \log_e\left(x_i^2\right), \tag{3.3}$$

where x_i is the discrete signal and N the number of data points in the signal.

Template Matching
Despite the morphological variability of the PPG pulse wave, template matching approaches have been very popular as measures of regularity (and quality) in a segment of signal. As in the case of the QRS complexes in the ECG, within a clean segment of PPG, it is expected that pulse waveforms will have similar morphology. The morphology of pulses distorted by artifact will be varied, compared to pulses in a clean signal. The template matching approach, has been proposed in various variants by many researchers for the purpose of quality assessment for the PPG signal.

Template Matching Using Euclidean Distance
Sukor et al. [31] extracted all pulses from a PPG segment and aligned them such that their peaks coincide. They then averaged all pulse waveforms to get a reference pulse waveform (the template) and vertically and horizontally aligned all pulses, by offsetting each pulse such that its mean absolute difference from the template was minimized. This was done to ensure that minimal scaling and timing variations did not adversely affect the similarity measurement. Similarity of each pulse with the template was then assessed using two measures:

- The Euclidean distance between each pulse with the reference pulse
- The ratio between the amplitude of the pulse wave under consideration and the template.

Final classification was based on heuristically-determined thresholds.

Template Matching Using Average Correlation Coefficient
Similarly to the approach used for the ECG, described in the previous chapter, in [29] the authors estimated the average pulse waveform in a segment of PPG by performing beat detection and then averaging over all pulse waveforms. Scaling and timing differences were not taken into consideration; each pulse waveform was extracted in a window around each detected heart beat equal to the median P-P interval in the segment. Averaging all pulse waveforms provides the pulse waveform template, which is adapted for every different PPG segment. The procedure is illustrated in Fig. 3.7 for a 10 s segment of PPG.

The *Pearson's correlation coefficient* of the template with each individual pulse waveform was then calculated and averaged over all beats in the segment. *Pearson's correlation coefficient,* usually denoted by r for discrete time series analysis, measures the similarity between two time-series x_i and y_i using the following formula:

$$r = \frac{\sum_{i=1}^{N}(x_i - \widehat{x})(y_i - \widehat{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \widehat{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \widehat{y})^2}}, \tag{3.4}$$

where $\widehat{x}$ and $\widehat{y}$ are the empirically determined means of x_i and y_i, respectively, and N is the number of samples in the two time-series. The correlation coefficient of each pulse waveform with the template within a 10 s segment was then averaged to obtain the average correlation coefficient of the PPG segment. The

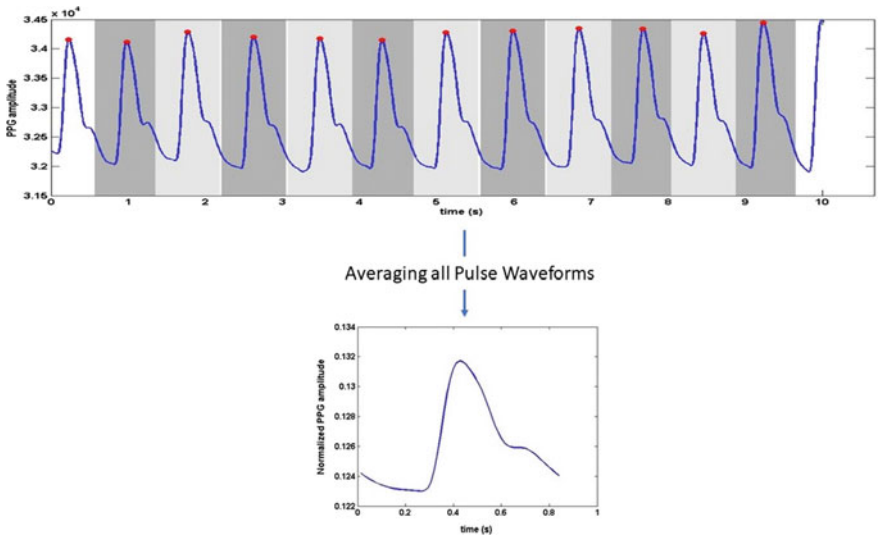

Fig. 3.7 Process for obtaining the average pulse waveform template from a 10 s segment of PPG. The *top figure* shows individual pulse waveforms (in *alternate shading* for ease of interpretation). The duration of each pulse waveform is taken as the median P-P interval in the segment and it is centered around each beat detected (marked in *red circles*). In this example, the first pulse waveform is excluded because the position of its peak does not permit the extraction of a pulse waveform of the required length. The pulse waveforms are then averaged to obtain the pulse waveform template shown in the *bottom plot*

template-matching quality index was assigned by labeling 10 s segments with an *average* correlation greater or equal to 0.86 as *acceptable* and a value of average correlation smaller than 0.86 as *unacceptable* [29], where the average correlation threshold was determined empirically, using data collected from different sensors and clinical environment, to ensure generalizability (Fig. 3.8).

Template Matching Using Correlation with Preceding Pulses

The idea of assessing signal quality by calculating the correlation coefficients between pulses and a reference pulse was also employed by Karlen et al. [15]. Instead of obtaining a template pulse waveform for a specified time window, the authors assessed the quality of each pulse in relation to their similarity to the previous pulses in the same segment of signal (after centering them around their respective peaks). If the correlation coefficient was found to be greater than 0.99, the pulse was accepted as part of the reference pulse set. Up to 10 pulses were used at any time. In case a reference pulse was smaller in duration than the pulse under examination, a correlation coefficient could not be calculated; to avoid this, the authors used first order linear regression to extrapolate the right tail of the reference pulse to the size of the pulse under examination [15].

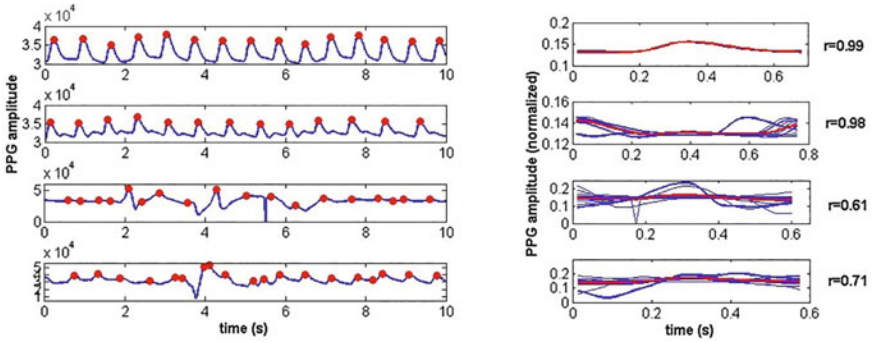

Fig. 3.8 Examples of acceptable and unacceptable PPG segments with their respective average pulse waves (in blue) and pulse wave template (in red). The average correlation coefficient values are also shown. The top two examples which are clean PPGs in sinus rhythm have almost identical pulse waveforms, resulting in really high average correlation coefficients. The bottom two PPG segments contain noise, resulting in average correlation coefficients below to acceptability threshold of 0.86

Template Matching Using Dynamic Time Warping (DTW)

Li and Clifford [27] performed template matching using three different approaches. In each 30 s segment, the autocorrelation was calculated and the distance between the two main peaks was taken as the average period of pulse waveforms, L. All pulse waveforms within the segment were extracted beginning at the pulse beat onset (beginning of systolic phase) until length L and averaged to get the pulse waveform template. The correlation of the template with each pulse waveform was then calculated and any pulse waveform with a correlation value less than 0.8 was excluded from the template which was then recalculated using only the acceptable pulse waveforms. If more than half beats were rejected the template of the previous window was used. The process was continued on a beat by beat basis. Three different correlation-based signal quality indices were calculated:

- Correlation coefficient of each pulse waveform from the beat onset for a length of L with the template.
- Correlation coefficient of each pulse waveform from the beat onset to the end of the pulse with the template, after linear stretching or compressing of the beat to match the length of the template.
- Correlation coefficient of each pulse waveform from the beat onset to the end of the pulse with the template, after using Dynamic Time Warping to match the length of the pulse waveform with the template [27].

Dynamic Time Warping (DTW) is a technique for aligning two time-series of different length, using a non-linear transformation. Given two time series P (the template) and Q (the pulse waveform) of lengths $i = 1...N$ and $j = 1...M$, respectively, a $N \times M$ matrix is constructed where element (i, j) contains the distance of points $d(p_i, q_j)$. After using a piecewise linear approximation algorithm to

transform P and Q to short line sequences, $d(p_i, q_j)$ is calculated as the absolute difference between the slopes in each short line. A cumulative distance measure is then calculated, $c_{i,j}$, defined as

$$c_{i,j} = min \begin{cases} c_{i-1,j} + d(p_i, q_j) l(p_i) \\ c_{i-1,j-1} + d(p_i, q_j)(l(p_i) + l(q_j)) \\ c_{i,j-1} + d(p_i, q_j) l(q_j) \end{cases}$$

where $l(p_i)$ and $l(q_j)$ are the duration of line p_i and q_j [27]. The optimal warping path is chosen as the one which minimizes the cumulative distance and it is then applied to Q as a non-linear transformation to obtain the best alignment with P before the correlation coefficient is calculated.

The cumulative cost (also known as alignment, matching or warping cost) can also be considered a measure of similarity between two time-series in the sense that the more similar they are, the "cheaper" it will be to align them. On this basis, Sun et al. [32] used the alignment cost as a measure of similarity between pulse waveforms in a PPG segment and a template pulse waveform which was extracted as the first clean pulse period of the PPG segment. Unlike [27] who aligned each beat with the template using DTW and then calculated a correlation coefficient, in [32], the authors applied DTW to align each beat with the template and took the alignment cost as the quality index.

3.4.3 Frequency-Domain Features

As discussed in Sect. 3.2, the spectrum of a clean PPG segment should normally have peaks in the LF and HF components (where the LF component represents the parasympathetic activity of the heart and HF the HR) and at various harmonics of the HF frequency. Frequency-domain approaches for quality assessment via the signal spectrum attempt to assess the presence of additional non-physiologically related content in the signal spectrum. It is well-known that noise alters the spectrum of the PPG either by increasing the content of frequencies outside the physiologically attributable limits and, thus, changes the distribution of power in the spectrum. We now discuss two approaches for assessing the quality of the PPG which search for a distortion in its spectrum.

Kurtosis of the Fourier Spectrum

As discussed in Sect. 3.4.2.3, the Kurtosis of a probability distribution measures the sharpness of the distribution; a distribution with no outliers is expected to have a sharp peak. In the presence of outliers the distribution becomes wider and flatter, taking longer for its tails to reach zero. In the same way that a clean PPG signal is expected to have a relatively high value of kurtosis in the time-domain, its Fourier spectrum is also expected to have a sharp peak, since it has a finite number of significant frequency components (LF, HF and harmonics). An artefact-corrupted

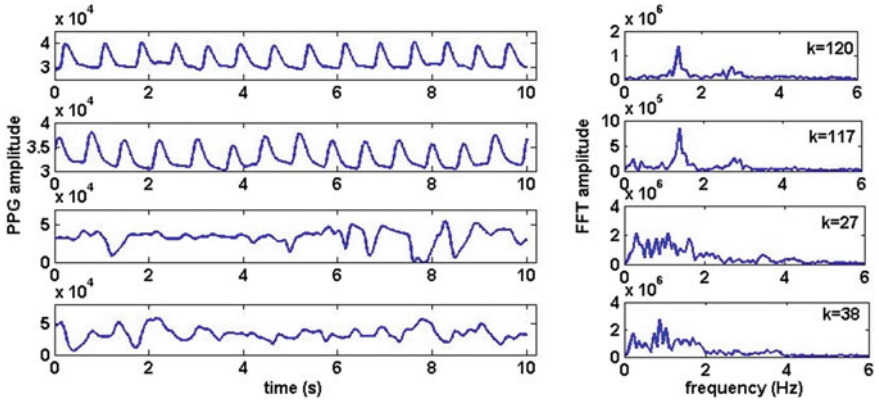

Fig. 3.9 Example acceptable and unacceptable PPG segments with associated Fourier spectra and kurtosis of Fourier spectrum (K). The acceptable segments (*top two*) have fewer frequency components, resulting in sharper distributions and higher values of kurtosis. Artefact-corrupted segments (*bottom two*) have more frequency components and smaller kurtosis values

segment of signal will result in a Fourier spectrum with additional artefact-related frequency components present, altering the shape of the distribution. Based on this idea, Krishnan et al. [24] proposed measuring the kurtosis of the distribution of the Fourier spectrum of a PPG segment and using it as a quality index. Figure 3.9 shows example acceptable and unacceptable PPG segments with their respective Fourier spectrums and associated kurtosis values.

Quadratic Phase Coupling (QPC) Using the Bi-spectrum

The second frequency-domain feature we discuss was also proposed by Krishnan et al. [24]. The idea is to exploit the relationship between the different frequencies present in the spectrum of the PPG to find differences between clean and noisy segments. The spectrum of a signal is a well-known technique for analyzing the decomposition of the power of a signal over frequency. This power is obtained via the second moment of the signal and provides information about the signal content in each frequency band but no phase information. Taking the third moment of the signal (equivalent to the measurement of skewness of a distribution discussed earlier) gives us the bi-spectrum which analyzes the interactions between the frequency components in a signal and is, thus, a function of two frequencies, unlike the power spectrum which is a function of one [36]. In a clean PPG segment when the harmonics interact, phase relations also exist, known as Quadratic Phase Coupling (QPC). Krishnan, thus, proposed calculating the bi-spectrum of PPG segments and measuring the QPC. In clean PPG segments, there is a strong self-coupling between the fundamental components of the frequency spectrum, whereas in artifact-corrupted signals, QPC occurs between random frequency components. By noting down the frequencies which were coupled in each PPG segment bi-spectrum, the segment was assigned as acceptable if strong self-coupling was observed and as unacceptable if self-coupling was absent [24].

3.5 Decision Rules

In this section, we review the types of decision rules which may be used for making assessments of signal quality based on extracted time-domain or frequency-domain features from the PPG. Most proposed algorithms utilize multiple extracted features for making signal quality assessments. When multiple features are being used together, the decision can be made based on a set of independent rules that need to be satisfied either concurrently or in a cascade of decision steps. Data fusion and machine learning approaches have also been proposed and have been shown to provide robust strategies for building classifiers using an assortment of different extracted features.

3.5.1 Physiological Thresholds and Heuristics

The most common approaches for setting thresholds on features extracted from the PPG are based on:

- *Expected limits of signal characteristics, such as pulse waveform duration and amplitude, pulse rate variability, differences between neighboring beats,* etc.: these rules are set based on expected physiological limits or heuristically, based on the experience of the expected behavior of the signal itself. Thresholds set on morphological characteristics are, for example, driven by literature on the physiology of the PPG; for example, pulse waveform duration (PWD) needs to reflect the physiological limits of HR. The same applies for thresholds set on the duration of the systolic phase, or ratio of systolic to diastolic phase as set in [18]. In addition to rules related to the morphology of individual pulse waveforms, expected physiology also drives the thresholds set on the permitted variability in a short segment of signal, or between consecutive beats. Examples of physiologically-driven decision rules rare those set by [29] and [18]. Often thresholds are set heuristically, and differ in different approaches and for different datasets. Sukor et al. [31], used morphological features of the pulse waveform using threshold-based decision rules and determined the thresholds heuristically, using trial and error, to maximize performance. A disadvantage of heuristic rules is the possibility of ending up with a system which is tailored to a specific dataset and would underperform when used with different datasets. As in the case of the ECG, most works proposing decision rules based on expected signal characteristics do not contain detailed justifications of the thresholds set and they are often set arbitrarily. It is assumed that most authors decide on the thresholds using combinations of experience on expected behavior as well as knowledge about equipment limits.
- *Empirical evidence via the use of available data*: many authors optimize thresholds for decision rules using trial and error, utilizing data available to them and choosing thresholds which will maximize performance. The most common

approach is using a representative portion of the labelled data as the *training set*, try different thresholds and pick the optimal threshold by drawing the Receiver Operating Characteristics (ROC) curve which is a plot of 1-specificity (horizontal axis) against and the sensitivity of the system (vertical axis) for different system parameters. If the same weight is given to the two statistical measures then the optimal threshold is the one which minimizes the distance to the point (0,1). When choosing thresholds empirically, the dataset used is important; if the aim is to create a system that can generalize to data from various monitors then, parameters need to be set using data from multiple monitors [29]. If parameters are tailored to a specific monitor and clinical scenario, it is possible that when used on data from a different monitor and in a different clinical setting, it might underperform (since the kind of artifact present might differ).

As in the case of the ECG, most proposed SQA systems for the PPG use combinations of rules based on multiple features which need to be satisfied concurrently or in a cascade of decision steps. The order and combination of steps is mostly chosen either heuristically or to optimize system performance on training data in the same way as explained earlier. Many systems perform the basic checks such as clipping detection and feasibility checks first to quickly reject obviously low-quality segments (which are often a large proportion of the unacceptable segments) before more sophisticated and computationally intensive techniques are used for the more ambiguous cases.

3.5.2 Data Fusion

The combination of multiple features (or sources of information) to obtain a single informative variable is known as data fusion [37]. In the context of signal quality assessment for the PPG, data fusion approaches have emerged in the recent years, combining different extracted features in more sophisticated ways to produce more robust decisions regarding the quality of signals under investigation.

3.5.2.1 Kalman Filters

The Kalman filter (KF) is an optimal state estimation method which calculates estimates for a continuous valued state that evolves over time based on existing observations of the state [38]. The estimations of parameters are made based on conditional probabilities which are obtained from observations of the values of the variables in time. The evolution of the parameter of interest $\boldsymbol{x}(t)$ is, thus, described using an explicit statistical model [37]. Another statistical model is also used to describe the way that the observations, $\boldsymbol{z}(t)$, are related to $\boldsymbol{x}(t)$. Unlike other probabilistic approaches, the estimated value of the variable is calculated as an average and not as a most likely value. The explicit description of the process and

observations, and the consistent use of statistical measures of uncertainty make the Kalman Filter framework ideal for incorporating distinctive features into one basic algorithm. Furthermore, at each point in time, it is possible to evaluate the role each feature plays in the performance of the system, making it an ideal approach for data fusion. The KF framework was initially applied in the problem of signal quality assessment by Li and Clifford [27] for fusing ECG- and ABP-extracted features for improved HR measurements. Sun et al. [32], applied the KF framework to the problem of signal quality assessment for the PPG to inspect four features extracted from the PPG and used the outcome in combination with other rule-based metrics to provide a four-level quality assessment of individual segments.

3.5.2.2 Machine Learning

Rule-based decision-making approaches often require prior information about the process to be modelled or rely on arbitrarily-set thresholds. Very often the relationship between extracted features and signal quality is not clearly understood. Also, the underlying processes which result in an unacceptable signal are multi-dimensional and often very complex. Machine learning favors an approach where the relationship between features and labels does not need to be fully understood [37]; for a robust signal quality assessment system to be built, the system does not need to understand the underlying labeling process, it just needs to learn how to replicate it. The inclusion of aggregate data in machine learning models may reveal additional information that is not seen by the human labeler. Machine learning models also offer the possibility to model the labeling process in a non-linear fashion which would not be possible, using traditional rule-based approaches. For robust machine learning models to be built, large amounts of labelled training data need to be available. The more data is available, from diverse data sources, the better the automatic decision-making will be with good generalization properties.

In the context of signal quality assessment for the PPG, machine learning approaches using the Multi-Layer Perceptron (MLP) have been proposed [27]. The authors fed combinations of four and six features extracted from PPGs to a 3-layer MLP with a sigmoid activation function and optimized the number of hidden nodes on annotated training data. They compared the performance of the signal quality MLP classifier with a heuristic fusion decision-rule using the same features and found that the machine learning-based system outperformed the heuristic fusion decision-rule based system with an accuracy of 95.2% (using six features) on the test set compared to 91.8% on the same set, using the heuristic decision rule.

This illustrates the potential of machine learning applications, especially when a large amount of data is available for training. In the same paper, the authors tested the performance of the classifier after removing each one of the six features and found that the performance was reduced by a range of 0.6–3.8% depending on the feature removed. When the improvement offered by an individual feature is

marginal, it may be optimal to remove it from the system altogether to balance accuracy with computational efficiency, especially when real-time processing is required.

3.6 Summary

In this chapter, we introduced the Photoplethysmogram (PPG) and provided an overview of approaches for feature extraction and for determining decision rules for designing SQA systems for the PPG. The most critical issue for the success of such systems is the identification and extraction of PPG features which can differentiate between acceptable and unacceptable segments of signal. The choice of features and decision rules depend on the application; features and decision rules which are independent of signal characteristics specific to a monitor, patient population and clinical scenario are more likely to have the ability to generalize and find wider application.

References

1. Orphanidou, C. (2017). Derivation of respiration rate from ambulatory ECG and PPG using ensemble empirical mode decomposition: Comparison and fusion. *Computers in Biology and Medicine, 81,* 45–54.
2. Pimentel, M. A. F., Johnson, A. E. W., Charlton, P. H., Birrenkott, D., Watkinson, P. J., Tarassenko, L., et al. (2017). Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering, 64*(8), 1914–1923.
3. Shelley, K. H. (2007). Photoplethysmography: Beyond the calculation of arterial blood pressure and heart rate. *Anesthesia Analgesia, 105,* S31–S36.
4. Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D. A., & Pugh, C. (2014). Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement, 35,* 807–831.
5. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement, 28,* R1–R39.
6. Peterson, M. T., Begnoche, V. I., & Graybeal, J. M. (2007). The effect of motion on pulse oximetry and its clinical significance. *Anesthesia Analgesia, 105,* S78–S84.
7. Young, I. H. (2003). Oximetry. *Australian Prescriber, 26*(6), 132–135.
8. Tang, S. C., Huang, P.-W., Hung, C.-S., Shan, S.-M., Shieh, J.-S., Lai, D.-M., et al. (2017). *Identification of atrial fibrillation by quantitative analyses of fingertip photoplethysmogram*. Scientific Reports 7, Article number: 45644.
9. Choi, A., & Shin, H. (2017). Photoplethysmography sampling frequency: Pilot assessment of how low can we go to analyze pulse rate variability with reliability? *Physiological Measurement, 38,* 586–600.
10. Charlton, P. H., Bonnici, T., Tarassenko, L., Clifton, D. A., Beale, R., & Watkinson, P. J. (2016). An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiological Measurement, 37*, 610–626.

11. Imanaga, I., Hara, H., Koyanagi, S., & Tanaka, K. (1998). Correlation between wave components of the second derivative of plethysmogram and arterial distensibility. *Japanese Heart Journal, 39,* 775–784.
12. Suganthi, L., Manivannan, M., Kunwar, B. K., Joseph, G., & Danda, D. (2015). Morphological analysis of peripheral arterial signals in Takayasu's arteritis. *Journal of Clinical Monitoring and Computing, 29*(1), 87–95.
13. Sviridova, M., & Sakai, K. (2015). Human photoplethysmogram: New insight into chaotic characteristics. *Chaos, Solitons & Fractals, 77,* 53–63.
14. Kamal, A. A. R., Harness, J. B., Irving, G., & Mearns, A. J. (1989). Skin photoplethysmography: A review. *Computer Methods and Programs in Biomedicine, 28,* 257–269.
15. Karlen, W., Kobyashi, K., Ansermino, J. M., & Dumont, G. A. (2012). Signal quality estimation using repeated Gaussian filters and cross-correlation. *Physiological Measurement, 33,* 1617–1629.
16. Monasterio, V., Burgess, F., & Clifford, G. D. (2012). Robust neonatal apnoea-related desaturation classification. *Physiological Measurement, 33,* 1503–1516.
17. Sweeny, K. T., Ward, T. E., & McLoone, S. F. (2012). Artifact removal in physiological signals: Practices and possibilities. *IEEE Transactions on Information Technology in Biomedicine, 16*(3), 488–500.
18. Fischer, C., Dömer, B., Wibmer, T., & Penzel, T. (2017). An algorithm for real-time pulse waveform segmentation and artifact detection in photoplethysmograms. *IEEE Journal of Biomedical and Health Informatics, 21*(2), 372–381.
19. Zhang, Z. (2015). Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE Transactions on Biomedical Engineering, 62*(8), 1902–1910.
20. Bonnici, T., Orphanidou, C., Vallance, D., Darrel, A., & Tarassenko, L. (2012). Testing of wearable monitors in a real-world hospital environment: What lessons can be learnt? In *Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 79–84.
21. Lee, H. W., Lee, J. W., Jung, W. G., & Lee, G. K. (2007). The periodic moving average filter for removing motion artifacts from PPG signals. *International Journal of Control Automation Systems, 5,* 701–706.
22. Graybeal, J. M., & Peterson, M. T. (2004). Adaptive filtering and alternative calculations revolutionizes pulse oximetry sensitivity and specificity during motion and low perfusion. In *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2004)*, pp. 5363–5366.
23. Kim, B. S., & Yoo, S. K. (2006). Motion artifact reduction in photoplethysmography using indendent component analysis. *IEEE Transactions in Biomedical Engineering, 53,* 566–568.
24. Krishnan, R., Natarajan, B., & Warren, S. (2008). Motion artifact reduction in photoplethysmography using magnitude-based frequency domain independent component analysis. In *Proceedings of the 17th International Conference on Computer Communications and Networks (ICCCN 2008)*, pp. 1–5.
25. Pittara, M., Theocharides, T., & Orphanidou, C. (2017). Estimation of pulse rate from ambulatory PPG using ensemble empirical mode decomposition and adaptive thresholding. In *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2017)*.
26. Hayes, M. J., & Smith, P. R. (1998). Artifact reduction in photoplethysmography. *Applied Optics, 37,* 7437–7446.
27. Li, Q., & Clifford, G. D. (2012). Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological Measurement, 33,* 1491–1501.
28. Zong, W., Heldt, T., Moody, G. B., & Mark, R. G. (2003). An open-source algorithm to detect onset of arterial blood pressure pulses. *Computing in Cardiology Conference, 30,* 259–262.

29. Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., & Tarassenko, L. (2015). Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE Journal of Biomedical and Health Informatics, 19*(3), 832–838.
30. Lu, S., Zhao, H., Ju, K., Shin, K., Lee, M., Shelley, K., et al. (2008). Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? *Journal of Clinical Monitoring and Computing, 22*(1), 23–29.
31. Sukor, J. A., Redmond, S. J., & Lovell, N. H. (2011). Signal quality measures for pulse oximetry through waveform morphology analysis. *Physiological Measurement, 32,* 369–384.
32. Sun, X., Yang, P., Zhang, Y.-T. (2012). Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach. In *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2012)*, pp. 3456–3459.
33. Selvaraj, N., Mendelson, Y., Shelley, K. H., Silverman, D. J., & Chon, K. H. (2011). Statistical approach for the detection of noise/artifacts in photoplethysmogram. In *Proceedings of the 33th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011)*, pp. 4972–4975.
34. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379–423, 623–656.
35. Tong, S., Li, Z., Zhu, Y., & Thakor, N. T. (2007). Describing the nonstationarity level of neurological signals based on quantifications of time-frequency representation. *IEEE Transactions on Biomedical Engineering, 54*(10), 1780–1785.
36. Collins, W. B., White, P. R., & Hammond, J. K. (1998). Higher-order spectra: The bispectrum and trispectrum. *Mechanical Systems and Signal Processing, 12*(3), 375–394.
37. Orphanidou, C., & Wong, D. (2017). Machine learning models for multidimensional clinical data. In S. U. Khan, A. Y. Zomaya, & A. Assad (Eds.), *Handbook of large-scale distributed computing in smart healthcare, scalable computing and communications* (pp. 177–216). Cham: Springer.
38. Welch, G., & Bishop, G. (2001). *An introduction to the Kalman filter*. ACM SIC-CRAPH, 2001 Course Notes.