IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 February 2023, accepted 26 February 2023, date of publication 9 March 2023, date of current version 16 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253933

APPLIED RESEARCH

Practical R-R Interval Editing for Heart Rate Variability Analysis Using Single-Channel Wearable ECG Devices

KANA EGUCHI[1], (Senior Member, IEEE), AND RYOSUKE AOKI[2]
[1]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi, Kanagawa 243-0198, Japan
[2]NTT Human Informatics Laboratories, Nippon Telegraph and Telephone Corporation, Yokosuka, Kanagawa 239-0847, Japan

Corresponding author: Kana Eguchi (kana.eguchi@ieee.org)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Networked Robot and Gadget Project, NTT Service Evolution Laboratories under Application No. ERP-18-003.

ABSTRACT Recent innovations in wearable devices have expanded the usage opportunities of single-channel electrocardiography (ECG) recordings in a daily life environment and enabled a variety of indirect daily activity monitoring based on heart rate variability (HRV). In general, wearable ECGs rarely undergo visual inspection by medical experts and therefore may contain noise or artifacts. Although noise/artifact-induced changes in ECG waveforms are known to cause misdetection of the QRS complex (i.e., the most distinguishable ECG components comprised of Q wave, R wave, and S wave), its complete suppression might be technically impossible. Since misdetection occurs in the QRS complex unit, we propose reframing the traditional HRV analysis flow by subdividing the R-R interval (RRI) editing into four steps in accordance with the processing detail (i.e., identification and editing) and its target unit (i.e., QRS complex or RRI). In addition, as a dubious QRS complex identification method for practical use, we utilize the amplitude at the detected point assuming the use of a single-channel wearable ECG without a reference. Initial evaluations using pseudo/real ECG datasets including ECGs with noise/artifacts show that the proposed processing/unit-based subdivision is theoretically effective for improving HRV calculation accuracy, and that the dubious QRS complex identification method for practical use also maintains this effect. Our study starting from practical HRV analysis using single-channel wearable ECG devices encourages reexamining each step in HRV analysis through the interdisciplinary research of clinical medicine and engineering/informatics that reveals the relationship of every two adjacent steps from the perspective of theory and practice.

INDEX TERMS Electrocardiography (ECG), heart rate variability (HRV), R-R interval (RRI) editing, single-channel/single-lead, wearable device.

I. INTRODUCTION

Heart rate variability (HRV) quantifies the fluctuations within the sequence of inter-beat intervals [1], [2]. From its inception, HRV has been recognized as one of the most promising quantitative biomarkers of autonomic nervous activity [1], and several recent studies have clarified the relationship between HRV and the target status (e.g., sleep apnea [3], sleep stages [4], [5], or driver drowsiness [6]). HRV is continuing to gain attention as a biomarker that enables a variety of indirect daily activity monitoring systems based on the estimated status.

HRV can easily be obtained from electrocardiography (ECG). In general, it is calculated as the fluctuation of the R-R interval (RRI), which is the interval between two adjacent R waves (Fig. 1); this is because the R wave (or the QRS complex comprised of Q wave, R wave, and S wave) is one of the most distinguishable and robust ECG components corresponding to one heartbeat. Technological innovations from the 2010s onward have made ECG recording possible for

The associate editor coordinating the review of this manuscript and approving it for publication was Sung-Min Park.

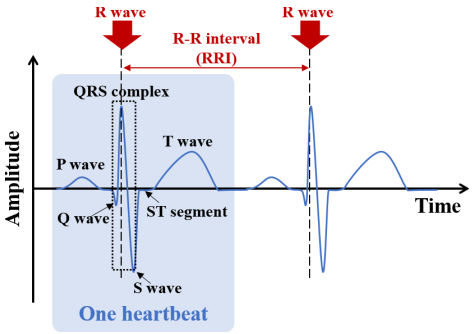


FIGURE 1. ECG components and R-R interval (RRI).

several hundred USD or even around USD 100 by focusing on a single-channel ECG recording [7], [8], [9], [10], [11], [12], [13]. As single-channel wearable ECG devices enabled unaided setup by non-experts while posing minimal disruption to the user's daily life, they have also expanded the usage opportunities of ECG in the daily life environment by reducing the cost of daily monitoring, including operation management. Thanks to these characteristics, several participatory social experiments have been conducted using commercial single-channel wearable ECG devices [12], [13]. Since the technological advances of these easy-to-use single-channel wearable ECG devices have continued to the present day, we expect much of the basic research focused on the relationship between HRV and the target status (e.g., [3], [4], [5], [6]) to move on to participatory social implementation in the near future. The challenge has now become how to calculate HRV as appropriately as possible from single-channel ECGs recorded in the daily life environment without any reference ECGs.

In theory, HRV should be calculated from the normal-to-normal (NN) intervals comprising two adjacent normal QRS complexes stemming from pure sinus node depolarization [1]. However, noise and artifacts in the recorded ECG degrade the accuracy of the R wave (or QRS complex) detection by changing the apparent morphology of the ECG waveform including the R wave (or QRS complex) [14], finally resulting in HRV miscalculation. Although accurate ECG recording would be the logical solution, it is realistically impossible for single-channel wearable ECGs, especially for non-clinical healthcare services that are typically recorded during daily life activities. In the first place, it is easy for the ECGs to be contaminated with noise/artifacts due to external factors such as body movements, respiration, and perspiration [15], [16]. In addition, theoretically, the morphology of an ECG waveform can be changed depending on the location of the electrode itself (i.e., ECG recording lead; see detail in Appendix. A) [17], [18], which can be influenced by how the device is worn along with the person's physique. Moreover, when using single-channel shirt-type wearable ECG devices, the ECGs are also affected by the inherent characteristics of clothing: for example, a shirt including embedded electrodes will deform along with body movement, which may cause displacements between the electrodes and the skin surface and thereby result in impedance fluctuations [15] that can be observed as noise/artifacts. Since it is not realistic to ask users to manually keep the same condition all the time or to ask well-trained medical experts to visually inspect all ECGs, it has become a challenge to calculate HRV as appropriately as possible in the data processing of HRV analysis, not in the ECG recording, while suppressing the effect of noise, artifacts, and morphological changes in the ECG waveform.

As the first step of HRV analysis, QRS complex detection has been a major research topic for the past several decades [19], [20], [21], [22], [23], [24], [25]. However, complete suppression of misdetection has not yet been accomplished and might be technically impossible: since the frequency characteristics of artifacts and the real QRS complex are so similar, one might be mistakenly detected as the other [24]. As the second step of HRV analysis, RRI editing has been another major research topic [2], [16], [26], [27], [28], [29], [30], [31], [32]. Although many approaches have been proposed, the majority of them cannot be applied to our assumed situation. Historically, RRI editing has focused on the misdetection of physiologically inadequate beats (e.g., arrhythmic beats) from ''clean'' ECG without any noise/artifacts, so its dubious RRI identification step generally uses duration information. As such, when the duration is physiologically reasonable, these methods might overlook an RRI including one or two misdetected QRS complexes, which are derived from the processing failure of a QRS complex detection algorithm. To avoid this issue, another conventional RRI editing method [28] considers the morphological similitude of the QRS complex in addition to the duration of the RRI. However, this method is not sufficient for our assumed scenario either: the morphology of an ECG waveform can change depending on the ECG recording lead as well as noise/artifacts, so the dubious RRI identification method for our situation cannot postulate a similar morphology in the QRS complex. For a more accurate HRV calculation using single-channel wearable ECG devices under the daily life environment, we need to develop a new RRI editing method for misdetected QRS complexes while utilizing information other than the morphological similitude.

In this paper, we propose reframing the traditional HRV analysis flow [1] by subdividing the RRI editing into four steps in accordance with the processing detail (i.e., identification and editing) and its target unit (i.e., QRS complex or RRI): the dubious *QRS complex* identification step, the dubious *QRS complex* editing step, the dubious *RRI* identification step, and the dubious *RRI* editing step. The first combination targeting the dubious *QRS complexes* extracts only the sequence of possibly accurately detected points, and the second combination targeting the dubious *RRIs* then extracts only the sequence of possible NN intervals. In addition, as a dubious QRS complex identification method for practical use, we utilize the amplitude at each detected point assuming the use of a single-channel wearable ECG without a reference. Since misdetection occurs mainly with

ECGs containing noise/artifacts, our method detects dubious QRS complexes by considering the signal quality of ECG at the detected point as an indirect indicator of misdetection possibility.

The original concept of identifying misdetected QRS complexes by QRS complex unit for improving RRI editing was proposed in our previous work [32], [33], where we excluded dubious RRIs comprising misdetected QRS complexes by RRI unit. In this study, we go one step further by implementing the processing/unit-based subdivision in consideration of the QRS complex misdetection-induced cascading effects over the HRV analysis processing flow. The first contribution of this study is to clarify the importance of processing/unit-based subdivision in the HRV analysis processing flow (Fig. 9) and confirm its validity through two experiments: one using pseudo ECG data created from open data and the other using real ECG data recorded by a commercial single-channel shirt-type wearable ECG device. The second contribution is to clarify the ideal step structure of the HRV analysis flow (Fig. 21) in consideration of the validation experiment results with its performance at the theoretical/practical level, inter-step relationships, and the original definition of each HRV feature calculation.

II. POTENTIAL ISSUES INHERENT IN TRADITIONAL HRV ANALYSIS FLOW UNDER DAILY LIFE ENVIRONMENT

Since the traditional HRV analysis flow comprises several steps [1], we need to consider both tacit understanding in each step and the inter-step relationships between every two adjacent steps when proposing an RRI editing method, which is an intermediate step of HRV analysis. In addition, potential issues to address should be clarified from the perspective of both theory and practice. Most of the conventional studies from the engineering perspective seem to unintentionally neglect or overlook the medical prerequisites for each step in HRV analysis, especially when these prerequisites are regarded as tacit understanding. To make matters worse, most conventional studies targeting HRV analysis, even including review papers, may lack the whole picture; in general, they tend to focus on the step of interest alone or at most the inter-step relationships between the two adjacent steps including the step of interest. This means that researchers may potentially overlook issues if they already possess the tacit understanding required. In fact, on the basis of the source codes, we confirmed that several open-source libraries for HRV analysis at present might not be able to calculate HRV features accurately in some cases because they seem to overlook both tacit understanding and inter-step relationships. In other words, researchers using these open-source libraries may unintentionally make a mistake in HRV analysis.

In this section, we highlight potential issues induced by overlooking tacit understanding and inter-step relationships to clarify the potential issues stemming from the combination of the traditional HRV analysis flow and ECGs with noise/artifacts. We first present an overview of the traditional HRV analysis flow [1] and then discuss the processing details step-by-step from start to end. To highlight potentially overlooked issues in each step under a daily life environment, our discussion includes the perspectives of both theory and practice. Finally, to identify which issues to address, we summarize the recommended policies for each step along with potentially remaining issues.

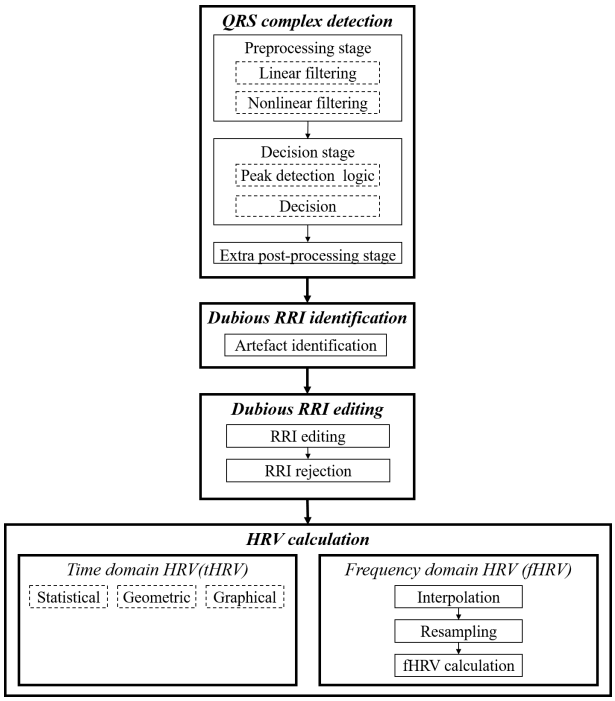


FIGURE 2. Reframed traditional heart rate variability (HRV) analysis flow for clarifying potential issues inherent in traditional HRV analysis flow.

A. OVERVIEW OF TRADITIONAL HRV ANALYSIS FLOW

To clarify potential issues inherent in the traditional HRV analysis flow, we reframe the traditional HRV analysis flow defined in the guidelines [1] as the following four steps (Fig. 2): (i) QRS complex detection, (ii) dubious RRI identification (defined as *artefact identification* in the traditional flow), (iii) dubious RRI editing (the step comprising *RRI editing* and *RRI rejection* in the traditional flow), and (iv) HRV feature calculation. Although several conventional methods focusing on *RRI editing* feature one integrated step comprising what we call the dubious RRI identification step and dubious RRI editing step, we explain their characteristics separately here, as we feel the former is more important than the latter. Specifically, if we cannot identify dubious RRIs at all, the dubious RRI editing step might be useless regardless of the RRI editing method used.

Since HRV features should be calculated from the NN intervals [1], we define the aims of each step as follows.

(i) QRS complex detection: detects as many accurate QRS complexes as possible
(ii) Dubious RRI identification: identifies as many dubious RRIs as possible to exclude non-NN intervals in the dubious RRI editing step

(iii) Dubious RRI editing: edits the sequence of dubious RRIs to obtain the sequence of NN intervals
(iv) HRV feature calculation: calculates HRV features from the sequence of NN intervals

In this paper, we use the term *non-NN interval* to mean an RRI including one or two non-normal QRS complexes and the term *dubious RRI* to mean an RRI suspected of being a non-NN interval.

Issues in HRV analysis inevitably pile up along with the progression of steps. Theoretically, the traditional ECG processing flow [1] assumes only one-way processing, which means the later steps suffer from all the remaining issues of the preceding steps. This makes it difficult for the following steps to recover from the issues that should ideally have been solved in the preceding steps.

For a practical situation, we need to keep in mind that the QRS complex detection step is firstly influenced by the quality of the recorded ECGs. If we cannot completely suppress noise/artifacts at the level of the ECG signal, the remaining noise/artifacts will hamper QRS complex detection and cause detection errors [14]. Since the RRI used in HRV analysis is generally calculated as the time interval between two adjacent points detected as R waves (or QRS complexes), the influence of detection errors is not limited to the QRS complex detection step but might spread to the subsequent steps. Potential issues raised in HRV analysis under the daily life environment should therefore be unraveled in consideration of the theoretical role, processing detail, step-specific issues, and cascading effects on the following steps.

B. SUMMARY OF PROCESSING DETAILS IN EACH STEP AND POTENTIAL ISSUES

To clarify which issues to address, this subsection briefly summarizes the processing details in each step while highlighting tacit understandings together with the potential issues raised in the daily life environment.

1) QRS COMPLEX DETECTION

Normal ECG waveforms corresponding to at least one heartbeat comprise a P wave, the QRS complex (comprised of a Q wave, R wave, and S wave), an ST segment, and a T wave (Fig. 1) [17]. Among these, the QRS complex is the most striking component [19]. For the purpose of calculating the inter-beat interval, RRI is therefore generally used. In this sense, the QRS complex detection step aims to detect as many accurate QRS complexes (or just R waves) from the recorded ECG as possible.

The general scheme for QRS complex detection is built on a two-stage structure [19]: a preprocessing stage comprising linear/nonlinear filtering sub-steps for denoising or feature extraction, and a decision stage comprising a peak detection logic sub-step and a decision sub-step for selecting applicable QRS complex candidates alone. Because the decision stage acts as an ''output filter'' of the preprocessing stage, ideally, the preprocessing stage should line up all the applicable QRS complex candidates while suppressing detection error. Towards this ultimate goal, various preprocessing approaches have been developed, including those using digital filters [20], [21], wavelet transform [22], [23], [24], or convolutional neural networks (CNN) [25]. As for the decision stage, the commonly used approach has been threshold determination on amplitude regardless of the approach applied in the preprocessing stage. Several recent studies including [24] have utilized a determination rule other than the threshold on amplitude.

TABLE 1. Confusion matrix.

Data class	*Classified as positive*	*Classified as negative*
Labeled as positive	True positive (TP)	False negative (FN)
Labeled as negative	False positive (FP)	True negative (TN)

To clarify cascading effects on the subsequent steps, we would like to explicitly explain the errors induced in the QRS complex detection step. In this paper, we define three types of errors in consideration of the confusion matrix on QRS complex detection (Table 1) [34], [35]: misrecognition, detection artefact,[1] and detection errors. Here, we use the term ''artefact'' (the middle character of the word is ''*e*'') to indicate the errors that occurred in the QRS complex detection step, whereas the term ''artifact'' (the middle character of the word is ''*i*'') to indicate the signal quality of the recorded ECG (see III-B for detail). Misrecognition and detection artefact correspond to the ''true positive (TP)'' class, where an algorithm correctly detects the QRS complex, and detection error corresponds to either the ''false positive (FP)'' or ''false negative (FN)'' class, where an algorithm cannot correctly detect the QRS complex. FP means the situation where an algorithm detects an irrelevant point as a QRS complex (i.e., misdetected QRS complex), and FN means the situation where an algorithm overlooks a QRS complex and it remains undetected (i.e., overlooked QRS complex). The ''true negative (TN)'' class does not exist from the perspective of QRS complex detection; the algorithm only targets ''positive'' detection (i.e., QRS complex) without ''negative'' detection and there are no ''negative'' labels in the reference either (i.e., all reference QRS complexes are defined as ''positive'').

The detailed definition of misrecognition, detection artefact, and detection errors are as follows. We define misrecognition as the situation where an algorithm detects a physiologically non-normal beat (i.e., an arrhythmic beat) as TP. The main cause of this error is improperly defined labels. For example, the label of a QRS complex during supraventricular arrhythmias might not necessarily be ''non-normal''

[1] In recognition of the work by Malik et. al. [36], we use the word ''artefact'' to indicate the errors that occurred in the QRS complex detection step. What Malik et. al. call a ''recognition artefact'' is the combination of what we call a ''detection artefact'' and ''detection error'' in this paper.

but rather ''normal'' because the theoretical shape of the QRS complex is quite similar to or even the same as that of a normal beat. Complete suppression of misrecognition requires both appropriate labelling and more accurate handling of arrhythmias, which can be accomplished by utilizing the dubious RRI identification step in conjunction with the dubious RRI editing step.

In contrast to misrecognition, detection artefact and detection errors are simple technical errors caused by the algorithms. We define detection artefact as a slight time gap between the detected point and the exact observation time of the R peak. In practical terms, a detection artefact is considered to be less than the duration of a normal QRS complex in a healthy participant (i.e., 0.10 s) [17]. As error compensation for detection artefact, we may additionally implement an extra post-processing stage for precisely determining the temporal location of the assumed QRS candidate after the decision stage [19], [37].

By defining the detection artefact as above, we can regard the detected point as a detection error whose time gap from its corresponding QRS complex label exceeds 0.10 s. In contrast to detection artefacts, we cannot completely suppress detection errors: considering the two-stage structure of QRS complex detection, both FNs and FPs might be caused by a mismatch of feature extraction in the preprocessing stage as well as by inappropriate selection in the decision stage. Technically speaking, suppression of noise can be accomplished by improving the performance of the preprocessing stage including filtering because its frequency characteristics might be quite different from the QRS complex. However, it is nearly impossible to completely suppress the influence of artifacts in both the preprocessing and decision stages because their frequency characteristics and morphology are quite similar to those of the QRS complex. For this reason, detection errors due to artifacts would have to be dealt with in the subsequent steps of HRV analysis.

For appropriate HRV analysis, the steps subsequent to the QRS complex detection step should be able to handle detection errors (i.e., both FNs and FPs) that may be recorded in ECG with artifacts regardless of the algorithm. This is crucial because the RRI sequence including FNs or FPs does not appropriately reflect actual heart activity.

2) DUBIOUS RRI IDENTIFICATION

Since HRV features should be calculated from the NN intervals in principle [1], the aim of the dubious RRI identification step is to identify dubious RRIs.

Dubious RRIs are classified into two types on the basis of their origin: physiological disturbances in the heart (e.g., the RRI during arrhythmia) and technological disturbances in the HRV analysis (e.g., an RRI comprising one or two detection errors). Although the origins are different, both disturbances may cause changes to the ECG waveform or inter-beat interval.

When ECG devices are capable of recording clean ECG without any noise/artifacts, the target of dubious RRI identification is only the physiological disturbances in the heart, namely, the RRIs during arrhythmia. As several arrhythmias represented as a premature atrial contraction (PAC) or premature ventricular contraction (PVC) can commonly occur in healthy individuals [38], dubious RRI identification typically refers to the RRI duration [26], [27]. However, considering the practical situation in a daily life environment, it is nearly impossible to record clean ECGs completely free from noise/artifacts. The dubious RRI identification methods used in a practical situation should therefore target the dubious RRIs coming from both physiological disturbances in the heart (e.g., arrhythmic beats) and technological disturbances in HRV analysis (e.g., dubious RRI comprising FNs or FPs).

Technically, we can divert the conventional duration-based dubious RRI identification method just as it is to dubious RRI comprising FNs alone. Specifically, the duration of a dubious RRI comprising FNs can be longer than a real RRI (i.e., prolonged RRI), and its value is thus the sum of two or more real RRIs. However, we cannot divert these methods to dubious RRIs comprising FPs. Since the duration of these RRIs is not necessarily changed nor is their variability consistent, duration-based dubious RRI identification methods may overlook an FP by regarding it as ''normal'' when its duration seems to be physiologically reasonable. For these reasons, we cannot simply declare that prolonged RRIs should be the dubious RRIs comprising FNs under the practical situation, which means that we should subdivide the dubious RRI identification in accordance with its editing target unit (i.e., QRS complex or RRI).

Liu et al. proposed an alternative dubious RRI identification method that considers the morphological similitude of the QRS complex in addition to the duration of the RRI, and confirmed its suitability for the four types of dubious RRIs by QRS complex unit (i.e., FN, FP, PAC, or PVC) [28]. When we consider ECGs with noise/artifacts, however, utilizing the morphological similitude of QRS complexes is still not sufficient to identify dubious RRIs comprising FPs. In ECG with noise, we can often observe apparent changes in the QRS complex due to the superposition principle, which makes it seems as though the QRS complex was recorded from different ECG recording leads, even if the actual ECG recording lead is unchanged (described in detail in III-B). Because this method may misidentify a QRS complex as an FP when the ECG waveform is apparently different from the others due to the superposition principle, we need to develop a new dubious RRI identification method for dubious RRIs comprising FPs.

3) DUBIOUS RRI EDITING

Using the results of the dubious RRI identification step as a basis, the dubious RRI editing step aims to extract the NN intervals alone by editing dubious RRIs that do not come from pure sinus node depolarization.

Regardless of the identification unit (QRS complex or RRI) in the dubious RRI identification step, conventional dubious RRI editing methods try to edit dubious RRIs by RRI unit to obtain the applicable RRI sequence. These methods can be classified into RRI deletion [2], [29], replacement [2], [4], [26], [27], [29], [30], and RRI interpolation [2], [16], [31], [32]. Note that RRI deletion here includes *RRI rejection* in the traditional HRV flow [1], namely, rejecting the impulse in the RRI sequence is theoretically equivalent to deleting the RRIs that make the ''impulse.'' These RRI editing approaches are non-exclusive and can be used in combination with each other.

Before moving on to the potential issues raised in this step, let us briefly revisit the definition of RRI. Although this definition can be traditionally regarded as a tacit understanding, it seems to be unintentionally neglected in most studies targeting RRI editing. In principle, RRI is calculated as the time interval between two adjacent points detected as R waves (or QRS complexes) in such a way that the ideal RRI sequence satisfies

$$RRI_{n+1}(x) = RRI_n(x) + RRI_n(y), \tag{1}$$

where n is a natural number and $n \geq 2$. $RRI_n(x)$ stands for the observation time of the n-th RRI, whereas $RRI_n(y)$ stands for the value of the n-th RRI. In actual use, the first RRI is calculated from the second and first detected points so that $RRI_1(x)$ becomes the observation time of the second detected point while $RRI_1(y)$ becomes an absolute value of the difference in the observation time between the first and second detected points. In an ideal RRI sequence, from start to finish, every two RRI elements should satisfy the relationship expressed as (1) without any missing value.

Considering this RRI definition and the aforementioned RRI editing methods, there are potentially two types of incorrect RRI editing consequences corresponding to their mathematical approach: the deletion approach (i.e., RRI deletion) may cause missing values (hereafter, ''missing RRI''), whereas the insertion approach (i.e., RRI interpolation) may cause a value difference between the plausible RRI and the theoretically insertable RRI (hereafter, ''type-A RRI gap''). Namely, type-A RRI gap means the situation in which a plausible RRI obtained from an arbitrary method cannot be ensured (1). Mathematically, the RRI replacement approach may combine deletion and insertion, so the RRI replacement would potentially cause both of these two.

Both the missing RRI and the type-A RRI gap are undoubtedly inappropriate at the level of the RRI sequence. Although missing RRI would ensure the original RRI definition expressed as (1), there would be a value missing as its name stands for. Meanwhile, the type-A RRI gap may even collapse the original RRI definition expressed as (1). We should also point out that avoiding the type-A RRI gap is not necessarily easy. A simple resolution of the type-A RRI gap by value compensation using (1) often leads to another issue, namely, the value difference between the originally obtained plausible RRI and the actually inserted RRI (hereafter, ''type-B RRI gap''). In other words, we are forced to encounter either a type-A or type-B RRI gap unless the three values (plausible RRI, theoretically insertable RRI, and actually inserted RRI) are coincident with each other.

To make matters worse, the missing RRI, type-A RRI gap, and type-B RRI gap may even decrease the accuracy of the HRV features in consideration of the attention point and calculation flow (described in detail in II-B4.). To come up with a reasonable dubious RRI editing method for target HRV features, we should therefore also take the HRV feature calculation step into account.

4) HRV FEATURE CALCULATION

HRV features are one of the most promising quantitative markers of autonomic activity [1], and a number of indices have been proposed [1], [39], [40]. HRV features are roughly divided into time domain HRV features (tHRVs) and frequency domain HRV features (fHRVs), examples of which are respectively shown in Tables 2 and 3.

The major tHRVs are divided into statistical measures [1] and geometric measures [1], [39]. Statistical measures can be further divided in accordance with their point of focus: one type focuses on the characteristics of whole target NN intervals using simple statistics (e.g., mean NN intervals and SDNN), and the other on the characteristics of topical NN intervals expressed by defined formulae (e.g., RMSSD, SDSD, and pNN50). For calculating the latter statistical measures focusing on topical NN intervals, in general, we first calculate the differences between every two adjacent NN intervals and then apply defined formulae. Geometric measures, on the other hand, focus on the geometric and/or graphic properties of the resulting pattern of NN intervals (e.g., TINN, CVI, and CSI), so the point of focus will vary depending on each feature.

As for fHRVs, they focus on the frequency characteristics inherent in the tachogram of NN intervals that can be measured as specific frequency components in their power spectral density (PSD) [1]: low frequency (LF) components (0.04–0.15 Hz) and high-frequency (HF) components (0.15–0.40 Hz). We therefore need three preprocessing substeps before fHRV calculation: data interpolation, data resampling, and spectral analysis. Since RRIs are, in principle, unequally spatial data derived from individual heartbeats, all target RRIs should be equally spaced by data interpolation and data resampling before spectral analysis is performed.

The accuracy of HRV features is influenced by both physiological disturbances and technological disturbances. The overlooking of ectopic beats is known to induce fHRV miscalculation due to the miscalculation of PSD [29]. Regarding technological disturbances, both detection artefact and detection errors may disrupt accurate tHRV calculation [36]. Since detection artefact can be corrected by extra post-processing in the QRS complex detection step, we focus here on appropriately dealing with detection errors (i.e., FPs and FNs). In this context, as emerging issues under the practical

TABLE 2. Examples of time domain HRV features.

Type	*Feature*	*Unit*	*Focus point*	*Description*	*Evaluation target*
Statistical	Mean RRI	ms	Volume	Average of RRIs.	✓
Statistical	SDNN	ms	Variability	Standard deviation of all NN intervals.	✓
Statistical	SDSD	ms	Difference, adjacency	Standard deviation of differences between adjacent NN intervals. Mathematically equivalent to RMSSD [41].	
Statistical	RMSSD	ms	Difference, adjacency	Square root of the mean of the sum of the squares of differences between adjacent NN intervals. Mathematically equivalent to SDSD [41].	✓
Statistical	pNN50	%	Difference, adjacency	Proportion derived by dividing NN50 by the total number of all NN intervals.	✓
Geometric	TINN	ms	Volume, variability	Baseline width of the minimum square difference triangular interpolation of the highest peak of the histogram of all NN intervals.	
Geometric	Cardiac vagal index (CVI)	N/A	Volume, difference, variability, adjacency	Index of cardiac vagal function that is not affected by sympathetic activity. According to the original definition, the measure is calculated by log10(L×T) [39]. This study obtains L as SD1 and T as SD2 based on [42].	✓
Geometric	Cardiac sympathetic index (CSI)	N/A	Volume, difference, variability, adjacency	Index of cardiac sympathetic function except in the resting supine condition, which is not affected by vagal activity. The measure is calculated by the ratio L/T [39]. In the same manner as CVI calculation, this study obtains L as SD1 and T as SD2 based on [42].	✓

NN: normal-to-normal
NN50: the number of interval differences of successive NN intervals greater than 50 ms
L: longitudinal axis of Lorenz plot
T: transverse axis of Lorenz plot
SD1: standard deviation of the Poincaré cross-wise
SD2: standard deviation of the Poincaré lengthwise

TABLE 3. Examples of frequency domain HRV features.

Feature	*Unit*	*Description*	*Evaluation target*
LF	ms^2	Power in low-frequency range (0.04–0.15 Hz).	✓
HF	ms^2	Power in high-frequency range (0.15–0.40 Hz).	✓
LF/HF	N/A	Ratio of LF/HF.	✓

situation, we should emphasize that even the missing RRI or RRI gap induced in the dubious RRI editing step might unintentionally influence the calculation of both tHRVs and fHRVs. In contrast to the laboratory setting where these HRV features are originally developed, it is basically impossible to be completely free from all technical disturbances.

Regarding tHRVs, both missing RRI and RRI gaps may decrease the accuracy depending on the situation. Although simple deletion of dubious RRIs would be more effective than doing nothing in theory, missing RRI might lead to a miscalculation in the statistical tHRVs focusing on the whole target NN intervals (e.g., mean NN intervals and SDNN) due to the ''presence bias'' of the remaining RRIs. This would be more problematic, especially when targeting ECGs recorded during/after exercise using wearable ECG devices. Considering physiological heart activity, heart rate may increase/decrease during/after exercise. If we were to delete lots of dubious RRIs (e.g., FNs and FPs induced by noise/artifacts during exercise activity), it would ultimately result in calculating inappropriate tHRVs from only a few remaining RRIs reflecting ''presence bias'' rather than the whole target RRIs (Fig. 3). Aside from this, geometric tHRVs (e.g., CVI, and CSI) and statistical tHRVs focusing on every two adjacent RRIs (e.g., RMSSD, SDSD, and pNN50) are also influenced by missing RRIs depending on their observation time management. Let us introduce an example of CVI and CSI calculation to highlight potentially overlooked issues in HRV calculation. The Poincaré plot used for CVI and CSI calculation is depicted as a scatter graph whose x-axis is the value of the n-th RRI and y-axis is the value of the $n+1$-th RRI. If the tHRV calculator only utilizes the value of the RRI sequence without its observation time, technically, it might miscalculate CVI and CSI due to an incorrectly depicted Poincaré plot (Fig. 4). Specifically, without the observation time of each RRI, the tHRV calculator might not notice the presence of missing RRI and instead depict dots even when

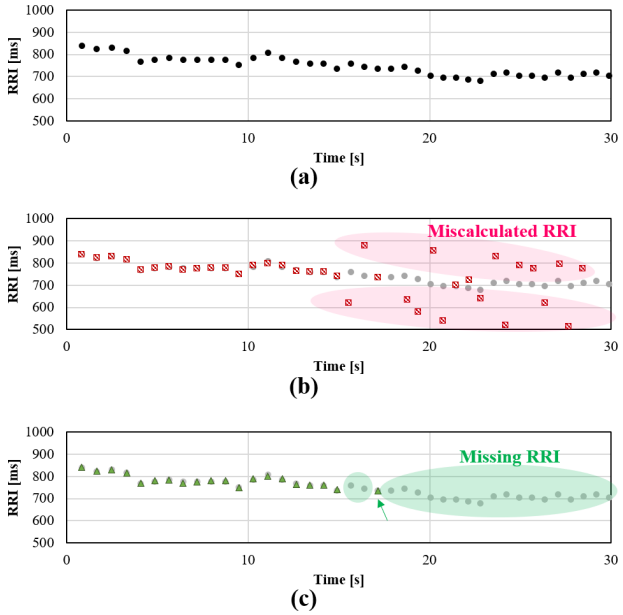


FIGURE 3. Examples of RRI miscalculation and RRI presence bias caused by deleting miscalculated RRIs. (a) Reference RRIs. (b) Example of miscalculated RRIs including FPs and FNs. (c) Example of edited RRIs by deletion. In (b) and (c), we assume the signal quality of ECG in the latter part (after 15 s) is poorer than that in the former part (before 15 s), which causes the RRI miscalculation. Gray circles in (b) and (c) indicate correct RRIs. In (c), RRIs have presence bias in the former part: the number of RRIs here may be higher than in the latter part due to missing RRIs. Green arrow in (c) indicates non-adjacent RRIs that may cause incorrect depiction of Poincaré plot without managing its observation time (see details in Fig. 4).

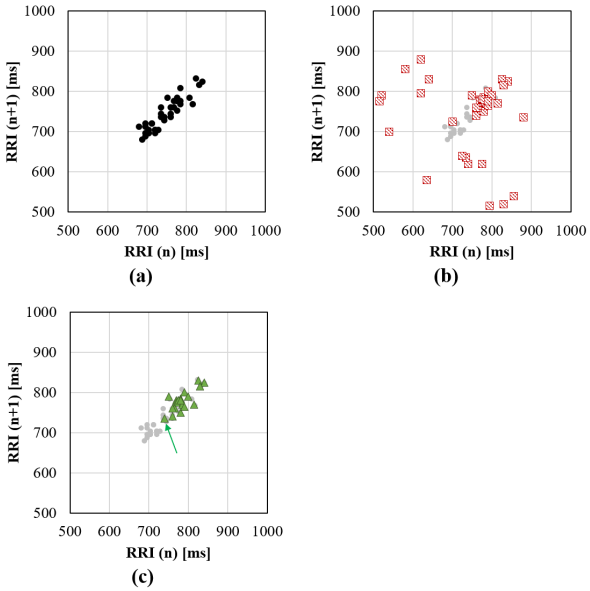


FIGURE 4. Example of incorrectly depicted Poincaré plot using RRIs from Fig. 3. (a) Reference RRIs. (b) Example of miscalculated RRIs including FPs and FNs. (c) Example of edited RRIs by deletion. Gray circles in (b) and (c) indicate correct RRIs. Comparing (b) and (c), (c) is able to obtain the Poincaré plot similar to (a). Without managing the observation time of each RRI, however, the dot indicated by the green arrow in (c) is incorrectly depicted.

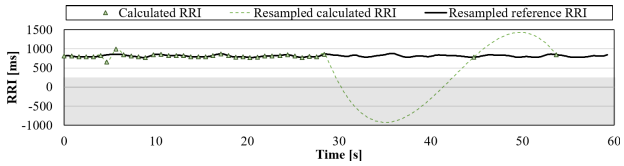


FIGURE 5. Example of RRI outliers caused by spline interpolation. Several resampled calculated RRIs using the spline interpolation function are lower than the lower limit of RRI in a healthy participant (i.e., 250 ms indicated by gray shaded area).

it is unreasonable (e.g., when a missing RRI is observed between two RRIs, neither of which are adjacent). A similar miscalculation due to adjacency confirmation failure might occur in the calculation of statistical tHRVs focusing on every two adjacent RRIs (e.g., RMSSD, SDSD, and pNN50). On the basis of the source codes, we confirmed that currently available HRV analysis techniques using open-source libraries also suffer from this adjacency confirmation failure. Regarding the RRI gap, a type-B RRI gap may also cause a miscalculation in the statistical tHRVs focusing on every two adjacent RRIs (e.g., RMSSD, SDSD, and pNN50) or several graphical tHRVs (e.g., CVI and CSI) because it focuses on every two adjacent RRIs. Although a type-A RRI gap may evade this issue, we cannot determine whether to permit the presence of a type-A gap because it cannot ensure the original definition of RRI shown in (1).

FHRVs may also be under the influence of both the missing RRI and RRI gaps. Missing RRI stemming from RRI deletion may cause another kind of RRI outlier in the data interpolation preprocessing sub-step: when using a cubic spline function for interpolating missing RRIs, we may observe out-of-range values for the RRI in a healthy participant, or even impossible ones in consideration of the heart activity due to the oscillation of the function [16] (Fig. 5). However, this issue might also be overlooked unless we check the result of cubic spline interpolation in detail; since this cubic spline interpolation itself is mathematically correct, no processing error would occur. Since the data resampled from RRI outliers does not appropriately reflect real heart activity, fHRV miscalculation is bound to occur. Although this issue could be suppressed by performing pre-interpolation before the data interpolation sub-step, this ad-hoc solution might cause another fHRV miscalculation due to the RRI gap. Unless the pre-interpolation results match plausible values, the type-B RRI gap would cause fHRV miscalculation by false RRI fluctuation. As we mentioned in the discussion on tHRVs, although the type-A RRI gap may evade this issue, we cannot determine whether to permit the presence of a type-A gap because it cannot ensure the original definition of RRI shown in (1). We should emphasize that these issues occurring at the level of the RRI sequence could also have a cascading effect on the PSD calculation: unless the calculated RRI sequence matches the real RRI sequence, we may miscalculate PSD and ultimately miscalculate fHRV.

To sum up, inappropriate dubious RRI editing may decrease the accuracy of both tHRVs and fHRVs. Under the practical situation, we need to decide how to suppress this unintentional influence induced by missing RRIs and RRI

gap over target HRV features along with the dubious RRI editing step in consideration of the characteristics of the target HRV features.

C. RECOMMENDED POLICIES IN TRADITIONAL HRV ANALYSIS FLOW UNDER DAILY LIFE ENVIRONMENT AND POTENTIALLY REMAINING ISSUES

To clarify the current technical limitations and potential requirements for considering the proposed method, this subsection highlights recommended policies in the traditional HRV analysis flow step-by-step. As HRV features should be calculated from the NN interval sequence in principle, as a whole, we ultimately need to aim for all dubious RRIs stemming from physiological disturbances and technological disturbances. Supposing a daily life environment where HRV analysis mainly targets non-clinical healthcare services, however, it would be better to prioritize dubious RRIs coming from technological disturbances (i.e., FPs and FNs). This is because ECGs recorded under that situation would comprise noise/artifacts with higher probability than transient arrhythmic beats in healthy participants, and detection errors are totally irrelevant to the real heart activity.

Regarding the QRS complex detection step, we should at least consider how to obtain QRS complexes of sufficient quality and quantity for calculating the target HRV features. Recall that the number of detection errors and their type (i.e., FNs or FPs) depend in large part on the performance of the applied algorithm. Since only the QRS complex detection step has the unique capability of detecting QRS complexes from recorded ECGs, we should use certain algorithms that can detect as many plausible QRS complexes as possible while suppressing FNs. In other words, we should use an algorithm that is capable of following the apparent waveform change in the QRS complex to suppress noise-induced FPs and FNs. Regardless of the algorithm, we should then implement an extra post-processing stage to compensate for detection artefact so that the quality of TPs is ensured. However, complete suppression of artifact-induced FPs and FNs only by the QRS complex detection step is nearly impossible: both the frequency and morphology characteristics of the artifacts are similar to those of the QRS complexes. Hence, we appropriately compensate for the remaining FPs and FNs in the following steps.

Regarding the dubious RRI identification step, we propose subdividing the dubious RRI identification step in accordance with the editing target unit (i.e., QRS complex or RRI), as FPs and FNs lead to different consequences: FPs cause miscalculated RRI, whereas FNs cause prolonged RRI. Furthermore, considering practical RRI calculation, FPs and FNs may be differently observed. The RRI comprising FPs, of course, comprises one or two FPs, but the RRI only comprising FNs may be observed as simply prolonged RRIs comprising two TPs (i.e., the FN class corresponds to the ''defined label of QRS complex'' overlooked in the QRS complex detection step, not to ''detected point as QRS complex''). Considering these points, the appropriate identification target unit for FPs and FNs should be different: FPs should be identified as a QRS complex unit, whereas FNs as an RRI unit. Therefore, we need to develop a dubious RRI identification method for dubious RRI comprising FPs. Assuming the practical situation, we should clarify here that the potential minimum requirement may become ''not using morphological similitude of QRS complex.'' Due to the superposition principle, an apparent QRS complex in ECG with noise might be changed just as if the QRS complex is recorded from different ECG recording leads, which means that utilizing the morphological similitude of QRS complexes might cause the misdetection of FPs. Meanwhile, one way of identifying the RRIs comprising FNs would be to adopt conventional dubious RRI identification methods that focus on the duration of RRIs.

Regarding the dubious RRI editing step, again, the RRI editing-induced influence over each HRV feature would be different depending on the consequences of the incorrect RRI editing (i.e., missing RRI or RRI gap) and even the target HRV features. We should therefore come up with a plausible RRI editing method for target HRV features in consideration of the potential issues emerging under the assumed environment. In other words, we should newly consider appropriate countermeasures depending on the identification target unit, which might not be accomplished by simply using the conventional RRI editing approach (i.e., deletion, interpolation, and replacement). For example, FP rejection should include consequent RRI re-calculation using non-FPs (i.e., TPs) identified by the QRS complex unit, whereas RRI re-calculation for FNs should include plausible value calculation by the RRI unit (generally, this plausible value might be smaller than the initially calculated prolonged RRI). As the influence over each HRV feature varies, we may have to decide whether to prioritize the accuracy at the level of RRI sequence satisfying (1) or at the level of HRV features. For example, if we need HRV features for status prediction utilizing the relationship between HRV features and target status, we may prioritize the accuracy at the level of HRV features.

To sum up, we need to newly consider both the dubious RRI identification step and the dubious RRI editing step in addition to the potential influence derived from FPs and FNs. We also need to determine the appropriate execution sequence for these two steps along with their subdivision: since the QRS complex is theoretically a smaller unit than RRI, the identification and editing for FPs might need to be done before that for FNs. Considering the current technical achievements in each step together with step-specific issues and their cascading effects, we will first work on dubious RRI identification for FPs. This is because the identification of FPs has not yet been accomplished even though it is a necessity for its editing, nor can we utilize any of the identification techniques from conventional studies.

III. ECG CHARACTERISTICS RECORDED UNDER DAILY LIFE ENVIRONMENT

In the use case of practical HRV monitoring under a daily life environment, FPs are inevitably under the influence of

the ECG characteristics recorded by wearable ECG devices. Unlike in-hospital ECG devices used for clinical purposes, however, these ECGs rarely get a visual inspection by experts, especially for non-clinical purposes. Since the traditional ECG processing flow [1] assumes one-way processing only, the signal quality of ECG in principle will influence the accuracy of the HRV features.

In this section, we unravel the ECG characteristics and limitations unique to wearable ECG devices at the level of ECG signals. We first introduce the single-channel wearable ECG devices assumed in this study, followed by the potential issues induced by their use in a daily life environment. Here, we aim to clarify potentially overlooked issues inherent in wearable ECG devices, which can be a major premise for our study. We then define the terms ''noise'' and ''artifacts'' used in this paper while showing the apparent changes they can create. We finish this section by summarizing the requirements for the proposed method.

In this study, we compare wearable ECGs to the standard 12-lead ECG to clarify the ECG characteristics from the perspectives of theory and practice. Since engineers can easily start ECG analysis without studying the background represented as Kaggle, we introduce the essential principles of ECG including the standard 12-lead ECG in Appendix A. We assume this section should be read with the basic knowledge described there.

A. ECG CHARACTERISTICS RECORDED BY SINGLE-CHANNEL WEARABLE ECG DEVICE AND POTENTIAL ISSUES UNDER DAILY LIFE ENVIRONMENT

1) TYPES OF SINGLE-CHANNEL WEARABLE ECG DEVICES AND THEIR RECORDING DURATION

To meet the growing demand for more appropriate diagnosis or estimation related to heart activity under the daily life environment, a variety of wearable ECG devices have been developed to increase the scope of ECG recording opportunities [7], [8], [9], [10], [11], [12], [13]. Although usability and signal quality are in an inevitable trade-off relationship, each device seems to improve usability for non-experts within a permissible range for the intended usage.

Depending on where the electrodes are placed, currently available wearable ECG devices can be classified into patch-type [7], [13], shirt-type [8], [9], [10], [12], and wristband-type [11]. In many cases, these devices record a single-channel ECG by means of bipolar measurement (see Appendix A for details), so the duration of a continuous ECG recording depends on the duration of electrode-to-skin contact. For HRV analysis during daily life activity, the patch-type or shirt-type is therefore theoretically preferable. Since these two types feature two electrodes that can be placed on the skin surface around the chest without disturbing daily life activities, long-term ECG recording suitable for HRV analysis is possible. A wristband-type is less suitable because it only records ECGs when the user is touching one electrode with the left hand and the other with the right (or vice versa), which can be fundamentally disruptive to daily activity.

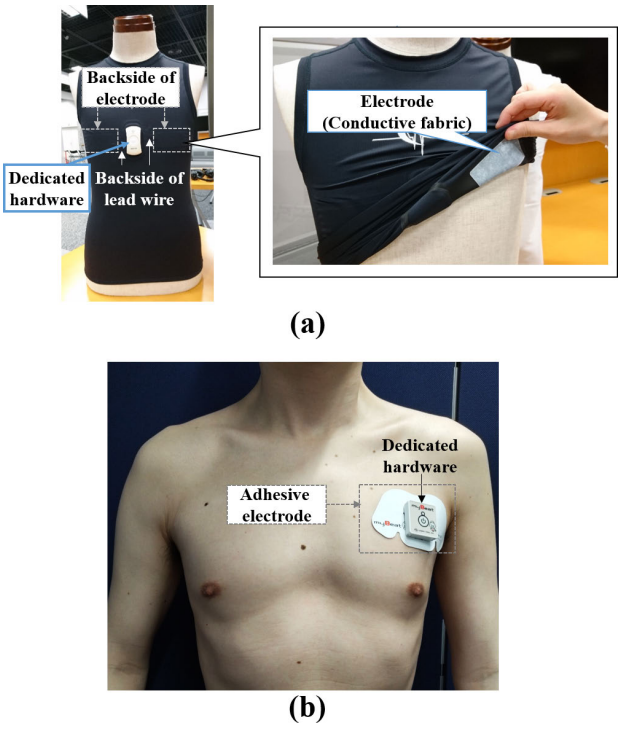


FIGURE 6. Examples of commercial single-channel wearable ECG devices that have been used for participatory social experiments in Japan. (a) Shirt-type comprising a specially designed shirt (C3fit IN-pulse, Goldwin, Inc., Tokyo, Japan) and an attachable dedicated wearable ECG device (hitoe®transmitter 01, NTT DOCOMO, INC., Tokyo, Japan). (b) Patch-type (myBeat, Union Tool Co., Tokyo, Japan).

2) TYPICAL SETTINGS AND CHARACTERISTICS IN THE USE OF COMMERCIAL SINGLE-CHANNEL WEARABLE ECG DEVICES UNDER DAILY LIFE ENVIRONMENT

Fig. 6 shows examples of commercial single-channel wearable ECG devices that have been used for participatory social experiments [12], [13]. In the shirt-type device shown in Fig. 6(a), a special shirt has electrodes/lead wires embedded inside it, so the setup for ECG recording only requires directly wearing the shirt on the body and attaching a small dedicated device for the ECG recording. In the patch-type device shown in Fig. 6(b), the setup requires users to affix electrodes and attach a small dedicated device for the ECG recording. Comparing the two types, the shirt-type one tends to be used more for non-clinical purposes (e.g., bus driver fatigue management), as it only costs around USD 100, which is significantly cheaper than the patch-type, and its electrode is reusable by machine-washing. Note that the ECG data recording method is different depending on device-to-device and even system-to-system; for example, whether it is directly recorded in a small dedicated device itself or transferred to other devices (e.g., a smartphone, personal computer, or even database via relay devices such as a smartphone or receiver module). Regardless of the ECG data recording method, users can normally check the recorded ECG in an application.

The signal quality of ECGs recorded by the aforementioned single-channel wearable ECG devices is not constant: often they are simultaneously under the influence of ''active'' ECG recording conditions over time (e.g., electrode-to-skin contact conditions and electrode conditions) and ''static'' device specifications of the hardware/software that remain unchanged during ECG recording (e.g., analog-to-digital (A/D) converter and built-in filters).

As one of the ''active'' ECG recording conditions, physical skin-to-electrode contact can easily change during daily life activity. For example, friction can occur between an electrode and the skin surface, the back side of the electrode may bump into some other object, or a part of the electrode may not be in direct contact with the skin surface. Electrode conditions can change along with daily activity as well: for example, humidity (e.g., perspiration) can cause changes in electrodes, or they can be deformed by physical issues such as flexion or expansion. All of these condition changes can potentially lead to displacements between an electrode and the skin surface, which might ultimately result in impedance fluctuations [15] observed as noise/artifacts in an ECG recording.

We should emphasize again here that an ECG recording lead should be regarded as an ''active'' ECG recording condition when using wearable ECG devices. Unlike in-hospital ECG devices that are set up by experts, ECG recording leads are not necessarily used as intended when a non-expert sets them up unaided. When targeting daily life activities, as stated earlier, an ECG recording lead can be changed along with body movements depending on the device type. Shirt-type wearable ECG devices with embedded electrodes have a pronounced tendency to do this (e.g., the electrode can easily slip off). Since the ECG waveform is principally affected by ECG recording leads (see Appendix A.2 for details), this represents a significant influence over the ECG waveform.

Since so many factors can influence an ECG recording and they are rarely exactly the same every time, identical noise/artifacts are not repeatedly observed even when the same user performs the same movements. This makes it nearly impossible to record long-term ECG without any noise/artifacts during daily life activity. Moreover, despite the potential instability in the quality of ECG recordings, wearable ECG devices utilized for non-clinical healthcare services are rarely inspected visually by well-trained medical experts. Since it is not realistic to ask users to manually unify all ''active'' ECG recording conditions, we need to suppress the influence of noise/artifacts in the data processing of HRV analysis, not in the ECG recording.

B. APPARENT CHANGES IN ECG WAVEFORM CAUSED BY NOISE AND ARTIFACTS

Moving forward, we separately use the term ''noise'' or ''artifacts'' depending on the distinguishability of the QRS complex only on the basis of the waveform.

In this paper, ''noise'' means an ECG with a low-frequency component in which we can visually recognize a QRS complex only on the basis of the waveform (defined as ''baseline wander (BW)'' in the MIT-BIH Noise Stress Test Database (NSTDB) [43], [44], [45]). Although apparent waveform changes in the recorded ECG are primarily caused by the ECG recording lead, similar apparent changes in the QRS complex are often observed in ECGs with noise due to the superposition principle (Fig. 7). Specifically, the visible part of a QRS complex buried in noise can be observed as if the QRS complex was recorded from a different chest lead, even

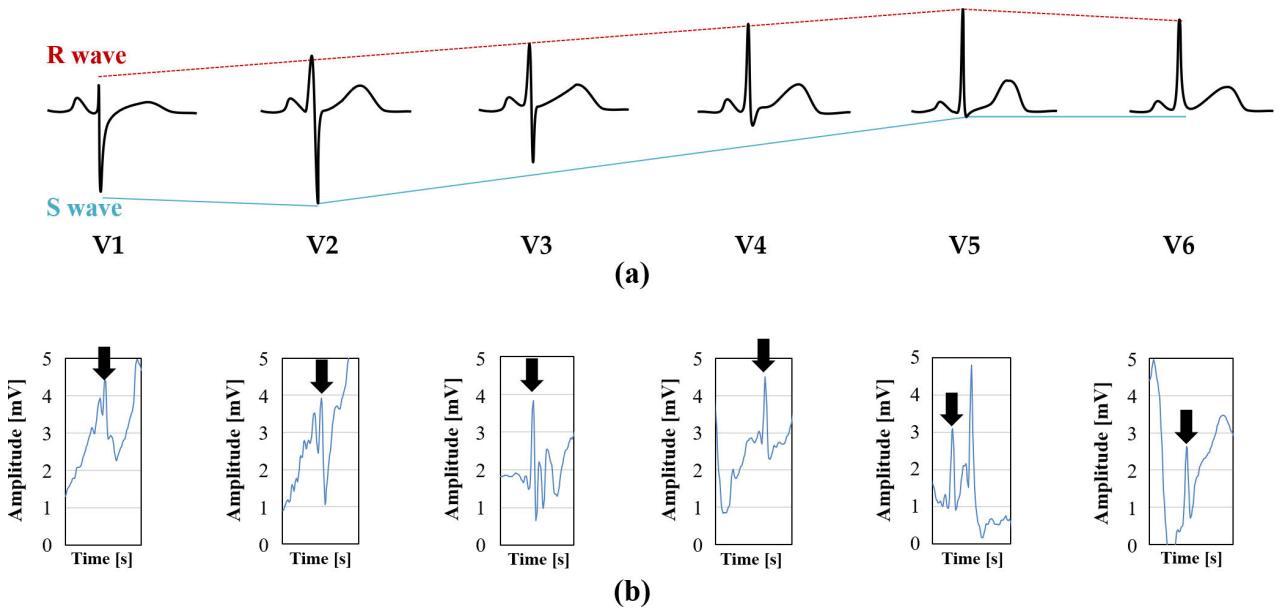


FIGURE 7. Example of ECG waveform change in QRS complex. (a) Variations in QRS complex patterns of chest leads from V1 to V6 (this set of ECG waveforms originally appeared in [18], on which we drew additional lines indicating R wave and S wave). (b) Apparent waveform change observed in wearable ECGs contaminated by noise/artifacts. In (b), black arrows indicate the position of QRS complexes and the duration of each window is 0.5 s. Previous research [24] has demonstrated that the Pan-Tompkins algorithm [20] could not detect any of the QRS complexes shown in (b), whereas SCWF [24] was able to do so.

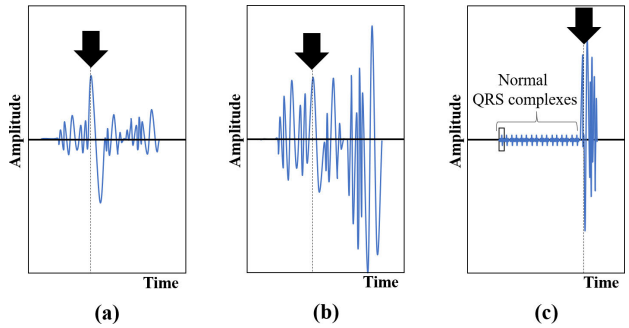


FIGURE 8. Conceptual diagram of the relationship between QRS complex in artifacts and its identifiability. (a) Example of an identifiable QRS complex. (b) Example of an unidentifiable QRS complex buried in artifacts at the same level. (c) Example of an unidentifiable QRS complex buried in artifacts at the greater level. Black arrows indicate the location of the QRS complex, which is exactly the same in both (a) and (b). In (a) to (c), the dotted line connected to the black arrow indicates the location of the local maxima of the R wave.

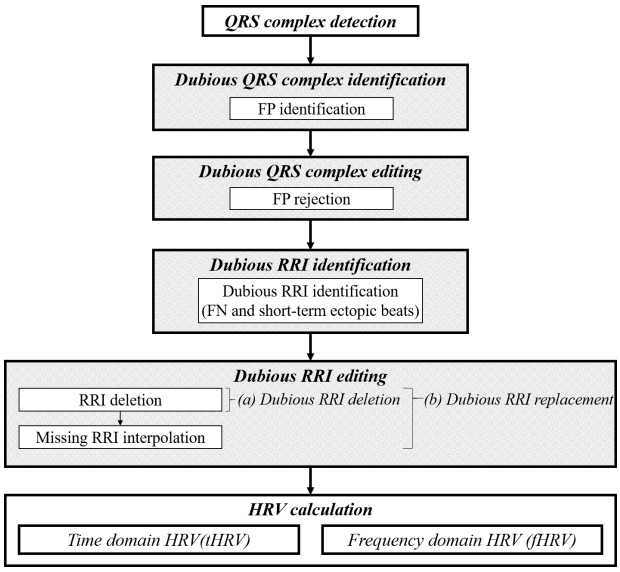


FIGURE 9. Overview of proposed heart rate variability (HRV) analysis flow for the suppression of miscalculation induced by technological disturbances (i.e., FP- and FN-induced miscalculation). The shaded area indicates our proposal. Fig. 9 modified Fig. 2 by subdividing the editing target identification and the actual editing into four steps in accordance with the target unit (i.e., QRS complex or RRI). Note that we omit several sub-steps in the conventional HRV analysis flow here (i.e., QRS complex detection step and HRV calculation step) that are irrelevant to the proposal of this study.

though the actual ECG recording lead seems to be constant. When targeting an ECG that may be suffering from noise, we should remain cognizant of these apparent waveform changes regardless of the ECG recording lead.

Meanwhile, ''artifacts'' in this paper means an ECG with a high-frequency component in which we cannot recognize a QRS complex only on the basis of the waveform (defined as ''electrode motion artifact (EM)'' or ''muscle artifact (MA)'' in NSTDB [43], [44], [45]). Unlike ECGs with noise, whether we can distinguish a QRS complex recorded in an ECG with artifacts depends totally on the amplitude of artifacts. Specifically, we can distinguish it if the amplitude of artifacts is sufficiently smaller than the actual QRS complex, but cannot distinguish it at all when the amplitude of artifacts is approximately the same as or larger than the actual QRS complex (Fig. 8) [16].

C. SUMMARY OF REQUIREMENTS FOR PROPOSED METHOD

The main issue caused by the presence of noise/artifacts is the miscalculation of HRV features induced by detection errors. Considering the one-way-only processing in the traditional ECG processing flow, along with the technical limitations in improving the performance of the QRS complex detection step (discussed in II-B1 and II-C), we need to come up with a new dubious RRI identification step as well as a dubious RRI editing step targeting the RRI comprising FPs and FNs. Ideally, these steps can be used in practical situations under the daily life environment regardless of the algorithm applied for QRS complex detection.

Assuming a practical situation targeting daily activity monitoring, ECGs for HRV analysis are recorded only by one ECG device. When using a single-channel wearable ECG device for this purpose, we cannot distinguish the class of the detected QRS complexes (i.e., TPs, FNs, FPs) in a conventional way. In principle, these classes should be determined on the basis of the confusion matrix requiring a ''reference'' for evaluation. To enable more appropriate HRV analysis in a daily life environment, we need to develop a method that distinguishes the class of detected QRS complexes by utilizing only the information available in one single-channel ECG device rather than the morphological similitude of the QRS complex.

IV. METHODS

One of the key challenges when it comes to achieving a more appropriate HRV analysis using a single-channel wearable ECG device under the daily life environment is how to suppress the technological disturbances that induce miscalculation (i.e., miscalculation induced by FPs and FNs) at the level of RRI or HRV features. With the aim of suppressing of FP-induced miscalculation even when using a single-channel wearable ECG device, we propose reframing the conventional dubious RRI identification step and dubious RRI editing step by subdividing them in accordance with the target unit (i.e., QRS complex or RRI).

Fig. 9 depicts an overview of the modified HRV analysis flow including the four proposed steps: dubious QRS complex identification, dubious QRS complex editing, dubious RRI identification, and dubious RRI editing. We adopt this step-by-step solution to deal with the practical situation in which FPs may occur in combination with FNs. In the dubious QRS complex identification step and dubious QRS complex editing step, we first aim to extract the sequence

of possible TPs from the sequence of detected points. Then, in the dubious RRI identification step and dubious RRI editing step, we aim to extract the possible NN interval sequence from the sequence of possible TPs. In addition, as a dubious QRS complex identification method for practical use assuming a single-channel wearable ECG without a reference, we utilize the amplitude at the detected point as an indirect indicator of misdetection possibility.

In this section, we introduce the processing details step-by-step in the order of execution.

A. IDENTIFICATION AND EDITING FOR QRS COMPLEX UNIT

The aim of the dubious QRS complex identification step and dubious QRS complex editing step is to extract a possible TP sequence alone by editing all dubious QRS complexes from the sequence of detected points obtained in the QRS complex detection step.

As the initial target of dubious QRS complexes assuming wearable ECGs, we herein aim for FPs that might be observed during ECGs with noise/artifacts. Considering the origin and cascading effects of FPs, the most plausible editing method for FPs is, in theory, uniquely determined to reject FPs before the RRI calculation.

1) FP IDENTIFICATION

This step aims to identify possible FPs utilizing only the information available in one single-channel ECG device regardless of ''active'' ECG recording condition. This is because the waveform changes in a QRS complex due to the superposition principle as well as the unobvious/unstable ECG recording lead, both of which can be changed along with body movements.

Since wearable ECG devices vary depending on the participant and the environmental conditions, and the number of FPs varies depending on the algorithm used for QRS complex detection, it is extremely difficult to create a training dataset for FP classification from scratch. Moreover, as there is currently no gold standard algorithm, it would be unrealistic to make a new training dataset for every algorithm with appropriate classification labels. In this situation, as the very first attempt to identify FPs from wearable ECGs, we therefore use an unsupervised classification based on predetermined heuristic rules that do not require any training dataset.

In practical terms, FPs are mainly caused by inappropriate processing of ECG with noise/artifacts in the QRS complex detection algorithm. Considering the general performance of the currently available algorithms, this indicates that the detected points from ECG with noise/artifacts are probably FPs depending on the algorithm, whereas the detected points from clean ECG without any noise/artifacts are probably TPs regardless of the algorithm. In other words, we can consider the quality of the ECG at the detected point to be an indirect indicator of the possibility of FP. We therefore try to discriminate FPs based on the recording status of the ECG at the detected point.

To this end, we set three possible ECG recording statuses: artifacts, noise, and clean. Regardless of the algorithm performance, we presume the detected points from ECG with artifacts without references are FPs. According to our interview with experts, the detected points from ECG with artifacts without references cannot be regarded as TPs even if that detected point looks like an actual QRS complex through visual inspection. Conversely, there is no unified determination for the detected points from ECG with noise without references. Because the robustness to noise varies in each algorithm even in terms of its detecting point, we need to determine how to deal with the detected points from ECG with noise without references by considering the algorithm performance against ECG with noise. In fact, we confirmed that the algorithm developed by Shimauchi et al. [24] was able to accurately detect QRS complexes from ECG with noise, whereas the Pan-Tompkins algorithm (PTA) [20] could not (Fig. 7(b)). For FP identification based on the recording status of ECG, we should therefore independently handle artifacts and noise to determine the appropriate processing for each algorithm.

Considering the difficulties in QRS complex detection, the likelihood of FPs decreases in the order of artifacts, noise, and clean ECG regardless of the algorithm. To determine the ECG quality at each detected point, the proposed method utilizes a three-stage evaluation based on two criteria: one for artifacts and the other for noise. Specifically, we first evaluate the ECG quality using the criterion for artifacts, and if it does not meet this criterion, we evaluate it again using the criterion for noise. Finally, the ECG is regarded as clean when it does not meet both criteria. In other words, the proposed method classifies each detected point as artifacts, noise, or clean corresponding to the worst conceivable quality at that point.

We first define the criterion for clean ECG on the basis of practical confirmation. In our previous study [24], we confirmed that even a conventional benchmark QRS complex detection algorithm such as PTA is able to accurately detect QRS complexes without any detection error from ECGs with slight noise (i.e., ECG with BW alone). For this reason, we regard apparent ECG waveform changes within the range of chest leads (i.e., from V1 to V6) as ''clean,'' and excessive apparent ECG waveform changes that pass over that range as ''noise'' or ''artifact.'' The criteria for discriminating ''noise'' and ''artifacts'' are as follows.

For the artifact criterion, we utilize the amplitude of the QRS complex (hereafter, QRS amplitude) obtained as the difference between the R wave peak and S wave depth (Fig. 10) that can absorb the difference in the height of the R wave among chest leads. Assuming HRV analysis using a single-channel ECG device aimed at non-clinical healthcare services, the main targets are healthy participants whose ECGs are unlikely to correspond to morbid criteria. In other words, we can consider the detected point to be an artifact when it

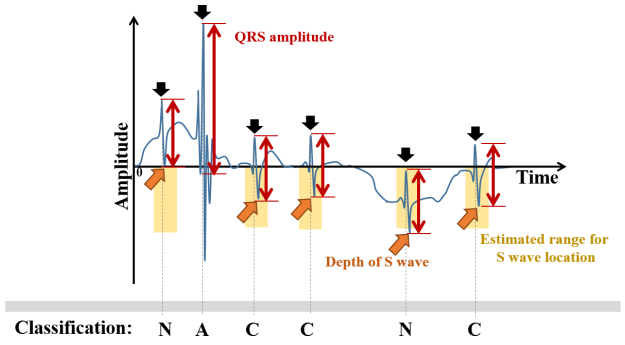


FIGURE 10. Conceptual diagram of FP identification based on the amplitude of detected point. Classification of each detected point is shown by the initial character of corresponding class: A, artifacts; N, noise; C, clean.

exceeds the morbid criteria in the medical field. We therefore focus on the QRS amplitude criterion of left ventricular hypertrophy (LVH), which is a major heart disease affecting QRS amplitude. Here, we utilize the Sokolow-Lyon criterion [46], RV5+SV1>3.5 mV, where RV5 is the R wave in V5 and SV1 is the S wave in V1. In principle, an R wave and S wave recorded by a single-channel wearable ECG device cannot simultaneously become RV5 and SV2 (see Appendix A for details), so our method evaluates the detected point by the QRS amplitude using that criterion. Note that we consider the subtle differences in R waves due to thoracic respiration to be sufficiently small and would thus be negligible.

For the noise criterion, we also utilize the amplitude at the detected point: namely, apparent ECG waveform changes from within the range of chest leads (i.e., from V1 to V6) are considered "clean." To discriminate "clean" ECG regardless of the apparent difference in the shape of the QRS complex, we focus on the theoretical sequential apparent difference in an S wave. As discussed in Appendix A.2, the depth of an S wave reaches the maximum at V2 and is shallower at V5 and V6, and it will be approximately the same as the height of an R wave in the transitional zone around V3 or V4. Since the S wave depth in a "clean" ECG is between V2 and V6, we set two criteria for noise discrimination and then regard the detected point as "noise" when its S wave depth is out of that range, namely, the range between two criteria representing the possible location ranges of the S wave depth between V2 and V6. The value of the S wave depth in chest leads is nearly 0 in V6 and becomes half that of the QRS amplitude in the transitional zone and more than half in V2. We therefore calculate the S wave depth in V2 using (2) and in V6 using (3).

$$S_{depth}\,(V2) = -\frac{QRSamplitude}{2} - a \times \frac{QRSamplitude}{2} \quad (2)$$

$$S_{depth}\,(V6) = -\frac{QRSamplitude}{2} + a \times \frac{QRSamplitude}{2} \quad (3)$$

Here, a is a positive real number used as a coefficient. Considering the S wave depth in V6, a should be set to around 1.

2) FP EDITING

This step aims to extract a TP sequence from the sequence of detected points by rejecting all possible FPs based on the results of the FP identification step.

As mentioned in IV-A1, we regard the detected points from ECG with artifacts without references as FPs, so we should undoubtedly reject all the detected points classified as "artifacts" regardless of the algorithm. In contrast, the detected points classified as "clean" should be regarded as TPs that must not be rejected.

Unlike the detected points classified as "artifacts" or "clean," whether to reject the detected points classified as "noise" should be determined in consideration of the algorithm performance applied for the QRS complex detection. For this determination, we require algorithm-specific validation tests when targeting several HRV features focused on different characteristics.

B. IDENTIFICATION AND EDITING FOR RRI UNIT

The aim of the dubious RRI identification step and dubious RRI editing step is to extract a possible NN sequence alone by editing all the dubious RRIs from the RRI sequence calculated after the dubious QRS complex editing step.

1) PREREQUISITES FOR CONSIDERING DUBIOUS RRI IDENTIFICATION AND DUBIOUS RRI EDITING

As an initial attempt to divide traditional dubious RRI identification and dubious RRI editing in accordance with the target unit, we propose a basic dubious RRI identification and dubious RRI editing using simple mathematics. Our intention is that these can be used as a benchmark in future studies.

As an initial target of dubious RRI identification assuming wearable ECGs, we aim for two kinds of dubious RRI that we can identify mainly on the basis of the RRI duration: FNs that might be observed during/after ECGs with noise/artifacts, and possible arrhythmic beats whose duration significantly differs from the majority of the remaining RRIs (i.e., single or short-term ectopic beats represented as PAC).

With this dubious RRI editing, note that we will likely face HRV feature miscalculation due to the RRI gaps or missing RRIs unless all RRI elements meet the real value. In other words, it is important that the dubious RRI editing method be able to obtain accurate target HRV features rather than doing nothing while allowing the RRI gaps or missing RRIs. However, unlike FPs that theoretically should be rejected before RRI calculation, there are currently no appropriate dubious RRI editing methods or corresponding execution order, even in conventional studies. Considering the traditional calculation methods for each HRV feature along with its theoretical focus point, plausible RRI editing methods may be different depending on the target HRV features. For tHRVs focusing on the topical NN intervals, we need to consider RRI deletion suitable for suppressing the issues stemming

from the RRI gap; for tHRVs focusing on the whole NN intervals, either RRI deletion or RRI replacement would be suitable depending on which one reflects the distribution of the ideal RRIs; for fHRVs, the RRI replacement should prevent oscillation of the interpolation function in the sub-steps for fHRV calculation (i.e., data interpolation and data resampling). Here, we consider the most plausible approach for fHRVs through risk comparison. RRI replacement for fHRVs is inevitably prone to risk, as it potentially has cascading effects on the PSD calculation and might lead to fHRV miscalculation. At the same time, the severity of the oscillation of the interpolation function might be higher than that of the PSD miscalculation, so that we should prioritize the suppression of the former. In this sense, RRI replacement is potentially a better countermeasure than RRI deletion alone. Note that our considerations of all the potential RRI editing methods are theoretical only. Validation tests for each target HRV feature type will be required to determine the best way to edit dubious RRIs.

As an initial validation of dubious RRI editing in consideration of traditional calculation methods for each HRV feature along with its theoretical focusing point, we propose taking a different editing approach depending on the target HRV feature. For example, RRI deletion alone is presumably suitable for tHRVs focusing on topical NN intervals, whereas RRI replacement may be better for fHRVs. We need to investigate how RRI deletion or RRI replacement affects the calculation of tHRVs focusing on the whole NN intervals. In our method, to improve the accuracy of the target HRV features, each dubious RRI editing is allowed to use several different RRI editing approaches together with its corresponding dubious RRI identification: as discussed in II-B3, each conventional RRI editing approach can be used in combination with others. Here, we conduct dubious RRI identification and dubious RRI editing as a pair, and implement several pairs in a sequential manner (starting with the first pair and then moving on to the next until finishing with the last pair).

Since this paired processing is performed sequentially, we describe the dubious RRI identification together with dubious RRI editing by its RRI editing approach: deletion and replacement. Since RRI replacement technically uses exactly the same RRI deletion as in RRI deletion alone, we discuss only deletion and interpolation below.

2) DUBIOUS RRI IDENTIFICATION AND DUBIOUS RRI EDITING BY DELETION APPROACH

In this approach, the dubious RRI identification identifies dubious RRIs that should be excluded from the HRV calculation, then the dubious RRI editing deletes them. Since the sequence of detected points only includes TPs after the dubious QRS complex editing step, we presume that dubious RRIs can be identified simply on the basis of the RRI duration.

As a target of this deletion, we set two types of dubious RRI: ''out of the range of the normal'' and ''of the majority.'' Here, out of the range of the normal refers to RRIs shorter/longer than the normal RRI in a healthy participant (i.e., shorter than 250 ms or longer than 1500 ms), and of the majority means RRIs whose duration deviates from the majority of RRIs (i.e., out of the range defined by mean $\pm n \times$ standard deviation, where n is an integer). Since dubious RRI deletion targeting RRIs out of the range of the normal alone is not sufficient for deleting possible arrhythmic beats, we further propose a two-step deletion: first, delete abnormal RRIs by conducting a pair of dubious RRI identification and dubious RRI deletion targeting the dubious RRIs whose duration is out of the range of the normal, and second, delete ''outliers'' in the remaining RRIs by conducting another pair of them targeting the dubious RRIs whose duration is out of the range of the majority. This sequential order is determined on the basis of the theoretical fact that out of the range of the majority can be influenced by the presence and the volume of ''outliers.''

Through this two-step dubious RRI editing by deletion, we expect to obtain more accurate HRV features than when doing nothing, especially when calculating tHRVs focusing on the topical NN intervals.

3) DUBIOUS RRI IDENTIFICATION AND DUBIOUS RRI EDITING BY INTERPOLATION APPROACH

In this approach, the dubious RRI identification identifies dubious RRIs missing from the consecutive RRI sequence (i.e., missing RRIs), then the dubious RRI editing interpolates them. These steps can be rephrased as ''missing RRI identification'' and ''missing RRI interpolation.'' Our objective with this interpolation is to improve the accuracy of the target HRV features, which cannot be accomplished by the deletion approach alone. For tHRVs focusing on the whole NN intervals, we should suppress the presence bias of the remaining RRIs. Meanwhile, for fHRVs, we should avoid the oscillation of the spline function in the data interpolation sub-step as far as possible. In this sense, we try to avoid reproducing RRI outliers that may decrease the accuracy of the target HRV features.

As an initial attempt to replace dubious RRIs in combination with dubious RRI deletion, we utilize the same interpolation for calculating tHRVs focusing on the whole NN intervals and fHRVs. Because the interpolation method itself potentially influences the missing RRI identification method (i.e., affecting how long the missing RRI should be interpolated), we first consider the interpolation method used in dubious RRI editing by the interpolation approach, and then consider which dubious RRI identification method can avoid reproducing RRI outliers in combination with the interpolation method.

Among the interpolation methods listed in Peltola's review article [2], degree zero (i.e., direct current (DC) components), degree one (i.e., linear), and spline use simple mathematics and thus can be easily utilized for our purposes. We decided against using spline interpolation due to the possible issue with the oscillation of the interpolation function (mentioned in II-B4). Among the remaining two methods, DC interpolation itself does not comprise the frequency components

of the fHRV we focus on (i.e., LF: 0.04–0.15 Hz; HF: 0.15–0.40 Hz), and theoretically, the linear interpolation may result in a miscalculation of fHRV due to overestimation of low-frequency components depending on the duration of the interpolation (i.e., the duration of missing RRIs) [32]. The influence of DC interpolation on the target frequency components is limited to two parts: immediately before and immediately after the interpolation. Compared to the long-term influence induced by linear interpolation, this might be more suitable for suppressing PSD miscalculation. For these reasons, we opted to use DC components (i.e., the average of the remaining RRIs after dubious RRI editing by deletion) for interpolating the missing RRIs.

On the basis of the original definition of RRI described in II-B3, we interpolate missing RRIs while satisfying (1). Since the endpoint of a missing RRI is a TP that cannot be rejected, the missing RRI interpolation requires at the very least that we recalculate RRI to satisfy (1) while allowing the type-B RRI gap. Although a missing RRI itself can theoretically be identified on the basis of the observation time of RRI, we need to consider the possible interpolation-induced RRI outliers for determining the target duration of missing RRI identification. When a missing RRI is shorter than the interpolation value, we cannot interpolate it, and when a missing RRI is not sufficiently longer than the interpolation value, the very last interpolated value might result in an RRI outlier. This outlier could ultimately lead to an inappropriate fHRV calculation due to the oscillation of the spline function in the data interpolation sub-step for fHRV calculation. For these reasons, we only interpolate a missing RRI when it exceeds a predetermined threshold that is the same as the deletion criterion on the maximum normal RRI in a healthy participant.

This missing RRI editing by interpolation should enable us to obtain more accurate HRV features than doing nothing, especially when calculating tHRVs focusing on the whole NN intervals as well as fHRVs.

V. EVALUATION

We performed an initial validation to determine whether the proposed method can improve the accuracy of HRV features compared to the conventional method regardless of the algorithm used for QRS complex detection.

In this section, we first present an overview of the two experiments we conducted and then describe the experimental conditions of each in detail. We then report the results of each experiment in independent sub-sections. In both experiments, we used the same ECG dataset as our previous study [24], but we here briefly summarize the experimental conditions before moving on to the results so as to ensure clarity and reproducibility.

A. EXPERIMENT OVERVIEW

1) EXPERIMENT DESIGN

We conducted the two experiments shown as follows.

- *Experiment 1*: Targeting the sequence of detected points obtained from a *pseudo ECG dataset*. This dataset comprises eight ECGs created from two different open data sources (the MIT-BIH Arrhythmia Database (MITDB) [43], [47], [48] and NSTDB [43], [44], [45]) assuming ECGs recorded by single-channel shirt-type wearable ECG devices.
- *Experiment 2*: Targeting the sequence of detected points obtained from a *real ECG dataset*. This dataset comprises three ECGs recorded by a single-channel shirt-type wearable ECG device during exercise activity.

In each experiment, the sequence of detected points obtained by the target algorithm (i.e., PTA or SCWF) was used as input data for the proposed method and evaluated separately for each algorithm. We used the code by Sedghamiz [49], a PTA program code that works on the MATLAB$^{\delta}$ environment (The MathWorks, Inc., Natick, MA, USA). As in our previous study [24], each experiment used the reference of RRIs for comparative evaluation. The details of the ECG dataset along with its references are described for each experiment.

All experiments were conducted after preparing the target data in the storage medium of a personal computer (CPU, Intel ®Core™i7-7700 3.60 GHz; RAM, 32.0 GB; OS, Windows 10). All analyses were conducted offline post-experiment.

2) TARGET ALGORITHMS

As discussed in II-C, we should use an algorithm with high performance, or else we could end up with inaccurate HRV features. However, it has not yet been clarified whether and to what degree the dubious RRI identification/editing steps contribute to improving the accuracy of HRV calculation. Since our point of interest is the inter-step relationships within HRV analysis, we need to clarify whether our proposed method can improve the accuracy of HRV calculation by compensating for the difference in the performance of the QRS complex detection step. Without any intermediate processing between the QRS complex detection step and the HRV calculation step, in theory, the number of FPs and FNs would directly degrade the accuracy of the HRV features: the greater the number of FPs or FNs, the lower the accuracy of the HRV features. Since the detailed breakdown of FPs and FNs (e.g., their quantity, ratio, and generation timing) varies in each algorithm even when targeting exactly the same ECGs, validation tests targeting several different algorithms should give us a better idea of the effectiveness of the proposed method in a practical use case.

To clarify these interests through our experiments, we target two algorithms having different performances with ECGs contaminated by noise/artifacts: PTA [20], which is commonly used as a benchmark algorithm for QRS complex detection and is also utilized in gold-standard HRV analysis software [50], and single complex wavelet filtering together with morphology-based peak selection (hereafter,

SCWF) [24], which we developed in an earlier study. In our previous work [24], we confirmed that the sequence of detected points derived from both algorithms contained FPs and FNs as expected when the target ECGs were contaminated. SCWF performed better than PTA in terms of reducing the number of both FPs and FNs during and after the contamination of noise/artifacts. In this study, we regard SCWF as an algorithm that performs well in the presence of noise/artifacts and PTA as an algorithm that performs worse.

B. EXPERIMENTAL CONDITIONS IN EACH EXPERIMENT

1) OVERVIEW OF EVALUATION PERSPECTIVES

As discussed in Section IV, our proposal includes two independent perspectives: subdividing the editing target identification and the actual editing into four steps in accordance with the target unit (i.e., QRS complex or RRI), and dubious QRS complex identification utilizing the amplitude at the detected point. We clarify the effect of each by evaluating from the following two perspectives.

- *Theoretical (i.e., ideal):* Corresponding to the evaluation of subdividing the editing target identification and the actual editing into four steps in accordance with the target unit (i.e., QRS complex or RRI). Evaluating the effectiveness of dubious QRS complex editing and dubious RRI editing based on its theoretical performance.
- *Practical (i.e., real):* Corresponding to the evaluation of dubious QRS complex identification utilizing the amplitude at the detected point. Evaluating the effectiveness of dubious QRS complex editing based on its practical performance.

In this study, we set 12 experimental conditions on each evaluation target algorithm: six for the theoretical performance evaluation and six for the practical performance evaluation. The details of each condition are described in the following V-B2 and V-B3.

Since there are two target algorithms, we investigate the results of the 12 evaluation conditions targeting the sequence of detected points obtained by PTA and by SCWF.

2) CONDITIONS FOR THEORETICAL PERFORMANCE EVALUATION

We first introduce the basic experimental conditions with regard to the theoretical performance evaluation.

Among the four steps comprising the proposed method, only two are responsible for the *editing* that might change the accuracy of the HRV features: the dubious QRS complex editing step and the dubious RRI editing step. To determine whether we should execute these two steps, we set the four evaluation conditions shown in Table 4.

However, we should also clarify which dubious RRI editing approach (deletion or replacement) performs better as a sub-evaluation of the theoretical performance evaluation. As mentioned in IV-B1, the effectiveness of the dubious RRI editing step will be different depending on the editing approach (i.e., deletion or replacement) and target HRV features. Thus, we need to subdivide conditions β and δ accordingly to evaluate the performance of each approach.

TABLE 4. Combination of evaluation conditions.

		Dubious QRS complex editing	
		Don't do	*Do*
Dubious RRI editing	*Don't do*	α	γ
	Do	β	δ

For these reasons, we set the following six conditions for each target algorithm as the basic experimental conditions.

- *Condition* α: Do not use any intermediate processing between the QRS complex detection step and the HRV calculation step.
- *Condition* β*-1*: Do not use the dubious QRS complex editing step, but do use the dubious RRI editing step featuring the deletion approach (described in IV-B2).
- *Condition* β*-2*: Do not use the dubious QRS complex editing step, but do use the dubious RRI editing step featuring the replacement approach implemented by the combination of deletion (described in IV-B2) and interpolation (described in IV-B3).
- *Condition theoretical-*γ (*t-*γ hereafter): Do use the dubious QRS complex editing step based on its theoretical performance, but do not use any dubious RRI editing step.
- *Condition theoretical-*δ*-1* (*t-*δ*-1* hereafter): Do use the dubious QRS complex editing step based on its theoretical performance, and then use the dubious RRI editing step featuring the deletion approach (described in IV-B2).
- *Condition theoretical-*δ*-2* (*t-*δ*-2* hereafter): Do use the dubious QRS complex editing step based on its theoretical performance, and then use the dubious RRI editing step featuring the replacement approach implemented by the combination of deletion (described in IV-B2) and interpolation (described in IV-B3).

Conditions *t-*γ, *t-*δ*-1*, and *t-*δ*-2* apply the dubious QRS complex editing step based on its *theoretical* performance while making use of the reference in each dataset. In other words, they use the information of *erroneously detected QRS complexes*. We obtained these erroneously detected QRS complexes through a comparative evaluation between the sequence of detected points and that of the reference QRS complexes.

3) CONDITIONS FOR PRACTICAL PERFORMANCE EVALUATION

We introduce the following extra experimental conditions with regard to the practical performance evaluation.

As discussed in IV-A2, whether to reject the detected points classified as "noise" should be determined on the basis of validation tests for each algorithm while targeting several HRV features focused on different HRV characteristics.

For this reason, we should clarify whether or not we use the detected points regarded as ''noise'' as a sub-evaluation of the practical performance evaluation.

Since there are three-tuple conditions using the dubious QRS complex editing step (i.e., conditions γ, δ-*1*, and δ-*2*), the number of extra conditions becomes six: the product of the three-tuple condition and the two different approaches with regard to the use of the detected points regarded as ''noise.'' One is a *fail-safe* approach that does not use the detected points regarded as ''noise'' (i.e., conditions *practical1*-γ, *practical1*-δ-*1*, and *practical1*-δ-*2*), and the other is a *fail-soft* approach that does use the detected points regarded as ''noise'' (i.e., conditions *practical2*-γ, *practical2*-δ-*1*, and *practical2*-δ-*2*).

- *Condition practical1*-γ (*p1*-γ hereafter): Do use the *practical* dubious QRS complex editing step (described in IV-A), but do not use any dubious RRI editing step. Here, the practical dubious QRS complex editing step takes the *fail-safe approach* in which only the detected points classified as ''clean'' are used for the following analyses while those classified as ''noise'' or ''artifacts'' are rejected.
- *Condition practical1*-δ-*1* (*p1*-δ-*1* hereafter): Do use the *practical* dubious QRS complex editing step (described in IV-A) taking the *fail-safe approach*, and then use the dubious RRI editing step taking the deletion approach (described in IV-B2).
- *Condition practical1*-δ-*2* (*p1*-δ-*2* hereafter): Do use the dubious QRS complex editing step (described in IV-A) taking the *fail-safe approach*, and then use the dubious RRI editing step taking the replacement approach implemented by the combination of deletion (described in IV-B2) and interpolation (described in IV-B3).
- *Condition practical2*-γ (*p2*-γ hereafter): Do use the dubious QRS complex editing step (described in IV-A), but do not use any dubious RRI editing step. Here, the practical dubious QRS complex editing step takes the *fail-soft approach* in which the detected points classified as ''clean'' and ''noise'' are used for the following analyses while those classified as ''artifacts'' are rejected.
- *Condition practical2*-δ-*1* (*p2*-δ-*1* hereafter): Do use the dubious QRS complex editing step (described in IV-A) taking the *fail-soft* approach, and then use the dubious RRI editing step taking the deletion approach (described in IV-B2).
- *Condition practical2*-δ-*2* (*p2*-δ-*2* hereafter): Do use the dubious QRS complex editing step (described in IV-A) taking the *fail-soft* approach, and then use the dubious RRI editing step taking the replacement approach implemented by the combination of deletion (described in IV-B2) and interpolation (described in IV-B3).

Here, we set a in (2) and (3) for noise classification to 1.5 to prevent undervaluation/overvaluation of classifying ''noise.''

C. EXPERIMENT 1: EVALUATION TARGETING SEQUENCE OF DETECTED POINTS OBTAINED FROM PSEUDO ECG DATASET CREATED BY OPEN DATA

In experiment 1, we first evaluated the performance of the proposed method targeting the sequence of detected points obtained from a *pseudo ECG dataset* that assumes ECGs recorded by a single-channel shirt-type wearable ECG device under a daily life environment. This evaluation mainly targets the accuracy of the HRV features, as the proposed method was originally developed for improving the accuracy of HRV features even when targeting single-channel ECGs contaminated by noise/artifacts.

1) EVALUATION CONDITIONS IN EXPERIMENT 1

Through the evaluation of the dubious QRS complex identification step targeting the *pseudo ECG dataset* (described in Appendix E.1), we confirmed that no detected points were classified as ''artifacts.'' Hence, experiment 1 targets only nine experimental conditions: α, β-*1*, β-*2*, *t*-γ, *t*-δ-*1*, *t*-δ-*2*, *p1*-γ, *p1*-δ-*1*, and *p1*-δ-*2*. The three *p1* conditions belong to the *fail-safe* approach that only uses the ones classified as ''clean'' for the following analyses.

2) OVERVIEW OF PSEUDO ECG DATASET

We utilized an artificially generated pseudo ECG dataset created by mixing noise/artifacts into the ECG while fixing the ECG recording lead. As the ECG, we used the first 60 s of V5-derived ECGs belonging to ID no. 100 in MITDB [43], [47], [48], as this was considered easy enough for both algorithms to accurately detect R waves (or QRS complexes). Assuming the situation in which certain noise/artifacts suddenly occur, ECGs of 30 to 40 s were replaced with pseudo ECGs derived by

$$targetECG = ECG + n \times irregular\ wave, \tag{4}$$

where n is an integer.

Prior work [16] has shown that noise/artifacts observed in ECGs recorded by a single-channel shirt-type wearable ECG device may be similar to the combination of three irregularities defined in NSTDB [43], [44], [45]: BW, EM, and MA. To calculate (4), the previous study [24] therefore regarded V5-derived ECGs of ID no. 100 in MITDB [43], [47], [48] as the *ECG* and *noise 1* of BW, EM, and MA provided in NSTDB [43], [44], [45] as the *irregular wave*. On the basis of the combination of three types of *irregular wave*, the pseudo ECG dataset comprises the following eight ECGs in total: (i) RAW (without any noise/artifacts), (ii) BW, (iii) EM, (iv) MA, (v) BW+EM, (vi) BW+MA, (vii) EM+MA, and (viii) BW+EM+MA. When mixing two or more irregular waves, the summed value of both was used. Assuming that the ECGs were recorded by a single-channel shirt-type wearable ECG device, integer n in (4) was set to 3 on the basis of a visual comparison of the QRS complex between the pseudo ECG generated by (4) and the real ECG recorded by a commercial single-channel shirt-type wearable ECG device [16].

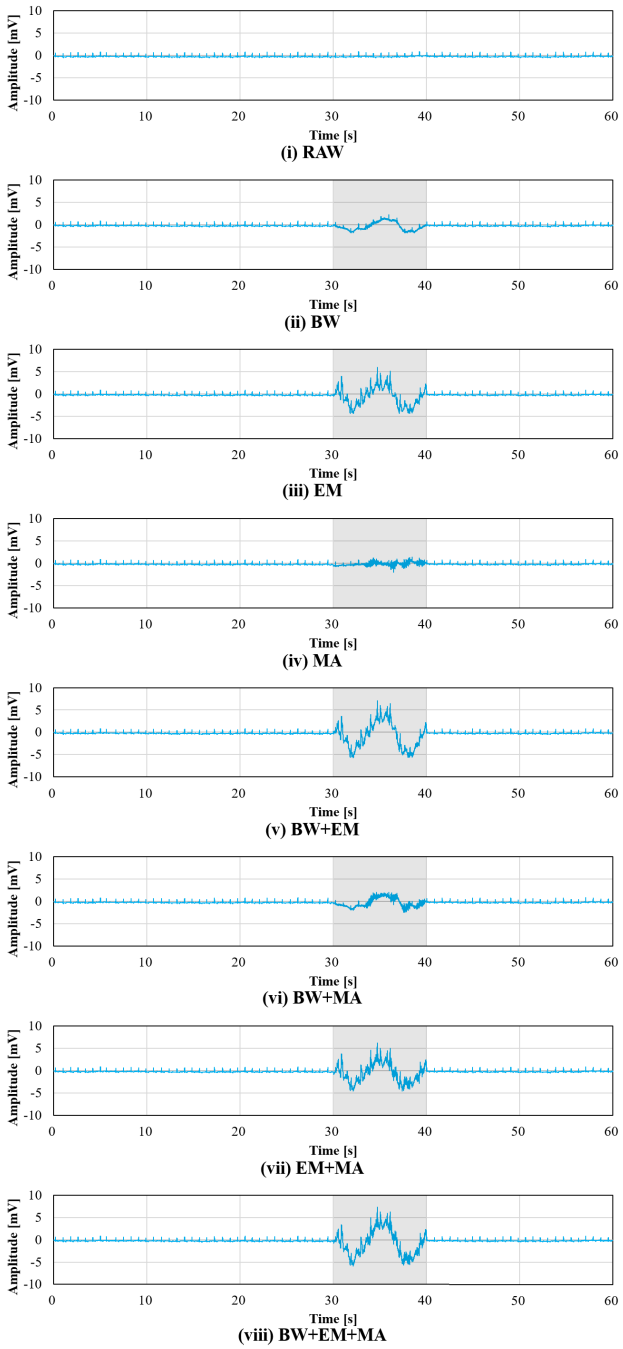


FIGURE 11. Target pseudo ECGs in experiment 1. Black shaded area in each ECG indicates period when the target pseudo ECG undergoes noise/artifacts created by (4).

Fig. 11 shows the target pseudo ECGs created in accordance with the above procedures.

ECG and *irregular wave* were synchronized with the same sample number. Since the complex wavelet utilized in SCWF assumes the ECG sampled at 200 Hz, all the target ECG data were down-sampled after the additive synthesis using (4). No preprocessing was applied because we wanted to obtain the detected points by the original performance of each QRS complex detection algorithm.

The signal quality of this dataset measured by the signal-to-NOISE ratio (SNR) defined in Appendix B.1 was 0.566 ± 1.02. Note that we use the capitalized term "NOISE" to mean the components other than heart activity; this is calculated from independent data of each *irregular wave* and that of the ECG other than P-QRS-T. The detailed SNR for each target ECG is shown in Appendix C.

3) REFERENCE QRS COMPLEXES

In experiment 1, we used the "annotations" of MITDB ID no. 100 in the first 60 s as the reference from (i) to (viii) regardless of the algorithm. This is because noise/artifacts only cause apparent changes in the ECG waveform without changing the heart activity (i.e., the actual positions of QRS complexes).

A total of 75 points were annotated as the reference QRS complexes in the first 60 s. On the basis of the guidelines provided by the PhysioBank annotation [51], we did not use the very first annotation "+" observed at 0.050 s because it is not related to an actual QRS complex. However, we used the annotation "A" observed at 5.678 s as a reference because the premature atrial contraction represented as "A" does not cause changes to the shape of a QRS complex, according to the clinical definition [17]. Overall, 74 points from 0.214 to 59.508 s were used as the references of the QRS complexes. As in our previous study [24], we moved each annotation to the local maxima of the corresponding R wave.

4) RULES FOR CLASSIFYING DETECTED POINTS IN THEORETICAL CONDITIONS

In conditions t-γ to t-δ-2, we used the classification results of detected points obtained manually through comparative evaluation to the reference QRS complexes.

To obtain TPs as accurately as possible while suppressing overvaluation or undervaluation, we utilized a specific time range for classification: a detected point observed within 0.10 s of the observation time of the corresponding reference QRS complex is classified as TP, and otherwise, as FP. Here, 0.10 s is the normal duration of one QRS complex among healthy people, according to the clinical definition [17].

5) TIME ADJUSTMENT AND QRS AMPLITUDE CALCULATION FOR DETECTED POINTS IN PRACTICAL DUBIOUS QRS COMPLEX IDENTIFICATION

Assuming that a detected point obtained from each QRS complex detection algorithm might be close to an exact QRS complex (e.g., within the normal range from Q wave to S wave), we first adjusted the observation time of each detected point from its original value to the local maxima observed before and after 0.10 s from each detected point. Here, 0.10 s is the normal duration of one QRS complex among healthy people, according to the clinical definition [17]. We then regarded the local minima observed after 0.10 s from each detected point as S wave, and calculated the QRS amplitude of each detected point as the difference in amplitude between the local maxima and the local minima.

6) TARGET HRV FEATURES

As frequently used HRV features, we targeted the six tHRVs shown in Table 2 and three fHRVs shown in Table 3 marked as *evaluation target*.

The target tHRVs were calculated in accordance with the description in Table 2. Before the calculation of tHRVs focusing on the characteristics of topical NN intervals (i.e., RMSSD, pNN50, CVI, and CSI), we independently checked the adjacency using (1), and only RRIs that were confirmed to be adjacent to the RRIs immediately before and after were utilized for the calculation.

For the fHRV calculation, we implemented three conventional preprocessing sub-steps: data interpolation, data resampling, and spectral analysis. Since the sequence of RRIs calculated from (1) is not sampled at a constant frequency, we resampled at 8 Hz using the linear function. In the spectral analysis, we first windowed the target RRIs with a Hann window and then calculated PSD by using an autoregressive model at the sixteenth order, where PSDs at low and high frequencies become stable regardless of respiration.

7) EVALUATION METHODS FOR HRV FEATURES

All HRV features theoretically reflect the characteristics of the target RRI sequence, so we conducted a pre-evaluation to investigate the RRI sequence obtained under each condition. As the evaluation points in this experiment mainly focus on the accuracy of the target HRV features, this pre-evaluation investigates the physiological RRI changes and provides a brief summary at the level of the RRI sequence originating from each experimental condition in the form of an RRI tachogram.

For the performance evaluation of target HRV features, we considered the root mean squared error (RMSE) using the HRV features calculated from the reference RRIs as an ideal value. To assess the influence of the combination of the dubious QRS complex editing step and dubious RRI editing step on the target HRV calculation, on each HRV feature, we calculate RMSEs between the ideal value and the calculated value obtained by each condition and compare them among the nine experimental conditions. Here, we assess the performance of each condition using a box plot of RMSEs.

8) PRE-EVALUATION OF RRIS

Fig. 12 and Fig. 13 show the RRI tachograms obtained by each algorithm. As a whole, the physiological RRI changes in the reference RRIs were confirmed to be stable. The only exception was a temporary RRI change due to PAC at 5.675 s (shortened) and 6.670 s (lengthened).

Before investigating each experimental condition, we briefly discuss the original performance of the two target algorithms under condition α. Overall, PTA obtained FPs and FNs when noise/artifacts were present (from 30 to 40 s), and even after the noise/artifacts (after 40 s). In contrast, SCWF obtained FPs and FNs during noise/artifacts (from 30 to 40 s) and was free of them after noise/artifacts (after 40 s).

Comparing the six experimental conditions from α to *t-δ-2* targeting *theoretical* performance, overall, the only execution of the dubious QRS complex editing step (i.e., condition *t-γ*) resulted in obtaining rather inappropriate prolonged RRIs due to FP rejection regardless of the algorithm. When ignoring missing RRIs, the dubious QRS complex editing step together with the dubious RRI editing step taking the deletion approach (i.e., condition *t-δ-1*) was able to obtain the most accurate RRI sequence. A similar tendency was confirmed in the *practical* dubious QRS complex editing step (i.e., conditions *p1-γ* to *p1-δ-2*): among these three conditions, *p1-γ* performed the worst, whereas *p1-δ-1* performed the best when ignoring missing RRIs.

In summary, overall, executing both the dubious QRS complex editing step and dubious RRI editing step will presumably improve the accuracy at the level of RRIs. The plausible RRI editing approach should be clarified through the evaluation targeting HRV features.

9) RESULTS OF HRV FEATURES

Fig. 14 and Fig. 15 show the RMSEs of target HRV features calculated under each experimental condition targeting the detected points obtained from PTA and SCWF. Steel-Dwass testing revealed a significant difference between several pairs of experimental conditions in the mean RRI obtained by SCWF. Here, one of each pair was consistently the condition *p1-δ-2:* between the condition α ($p = 0.003$), *β-1* ($p = 0.003$), *β-2* ($p = 0.003$), *t-γ* ($p = 0.001$), *t-δ-1* ($p = 0.001$), and *t-δ-2* ($p = 0.003$). Steel-Dwass tests also revealed that there was no significant difference between any pairs of the experimental conditions in any of the remaining HRV features regardless of the algorithm.

Comparing the six experimental conditions α to *t-δ-2* targeting *theoretical* performance, as indicated at the level of RRI sequence, either condition *t-δ-1* or *t-δ-2* was able to achieve the highest accuracy in most of the target HRV features regardless of the algorithm. Regarding PTA, based on the median, *t-δ-1* was the best condition for pNN50, RMSSD, CVI, and HF, whereas *t-δ-2* was best for mean RRI, SDNN, LF, and LF/HF. Excluding these, condition *t-γ* performed best for CSI. Regarding SCWF, based on the median, *t-δ-1* was the best condition for SDNN, pNN50, RMSSD, and CVI, whereas *t-δ-2* was best for mean RRI, CSI, LF, and HF. Excluding these, conditions α to *β-2* performed best for LF/HF. In a broad sense, these results indicate that executing both the dubious QRS complex editing step and dubious RRI editing step would be effective to improve the accuracy of HRV features regardless of the algorithm, but the suitable dubious RRI editing method may vary depending on the target HRV feature.

Next, we investigated whether the best-performing condition for each target HRV feature was consistent with the theory in consideration of the original focus point of each HRV feature and physiological RRI changes. Regarding tHRVs, we investigated target tHRVs in conditions α to *t-δ-2* targeting *theoretical* performance from three categories

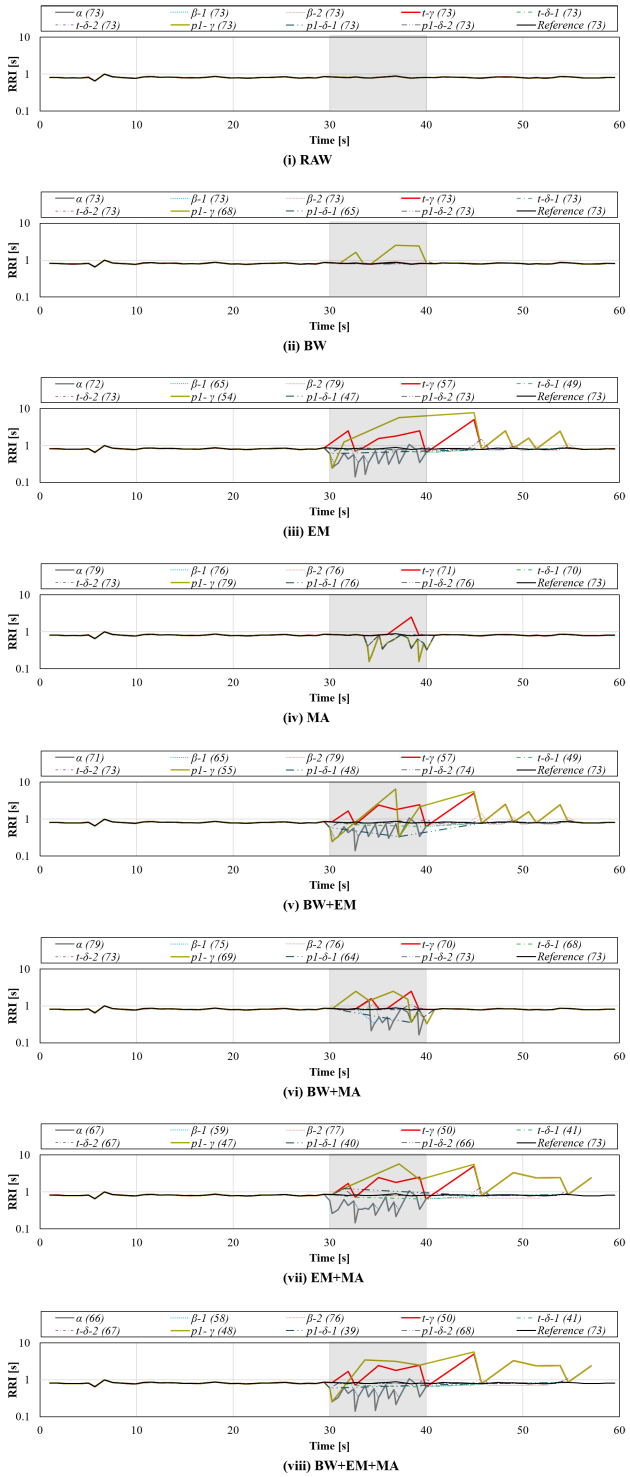


FIGURE 12. RRI tachograms calculated from QRS complexes obtained by PTA. Black shaded areas indicate period when the target pseudo ECG undergoes noise/artifacts created by (4). Regardless of the target ECG, only the rejection of QRS complex (conditions *t* -γ and *p1*-δ-*2*) caused inappropriately prolonged RRIs.

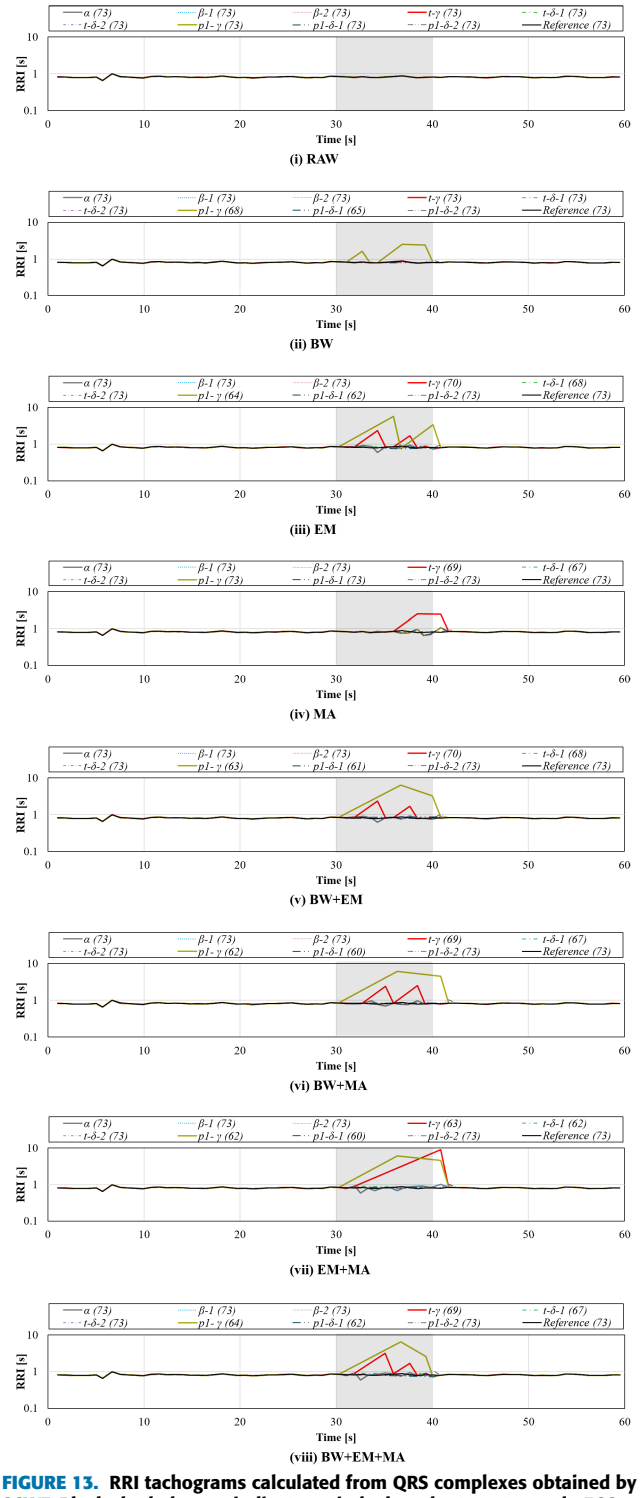


FIGURE 13. RRI tachograms calculated from QRS complexes obtained by SCWF. Black shaded areas indicate period when the target pseudo ECG undergoes noise/artifacts created by (4). Regardless of the target ECG, only the rejection of QRS complex (conditions *t* -γ and *p1*-δ-*2*) caused inappropriately prolonged RRIs.

corresponding to their original focusing point: volume (i.e., mean RRI), topical characteristics between every two adjacent RRIs (i.e., pNN50, RMSSD, CVI, and CSI), and variability (i.e., SDNN). Regarding mean RRI, we confirmed that the replacement approach (i.e., condition *t*-δ-*2*) would effectively improve the accuracy. This result is consistent with the theoretical prediction in consideration of the physiologically stable RRI changes in this experiment. In other words, when physiological RRI changes are stable, replacing

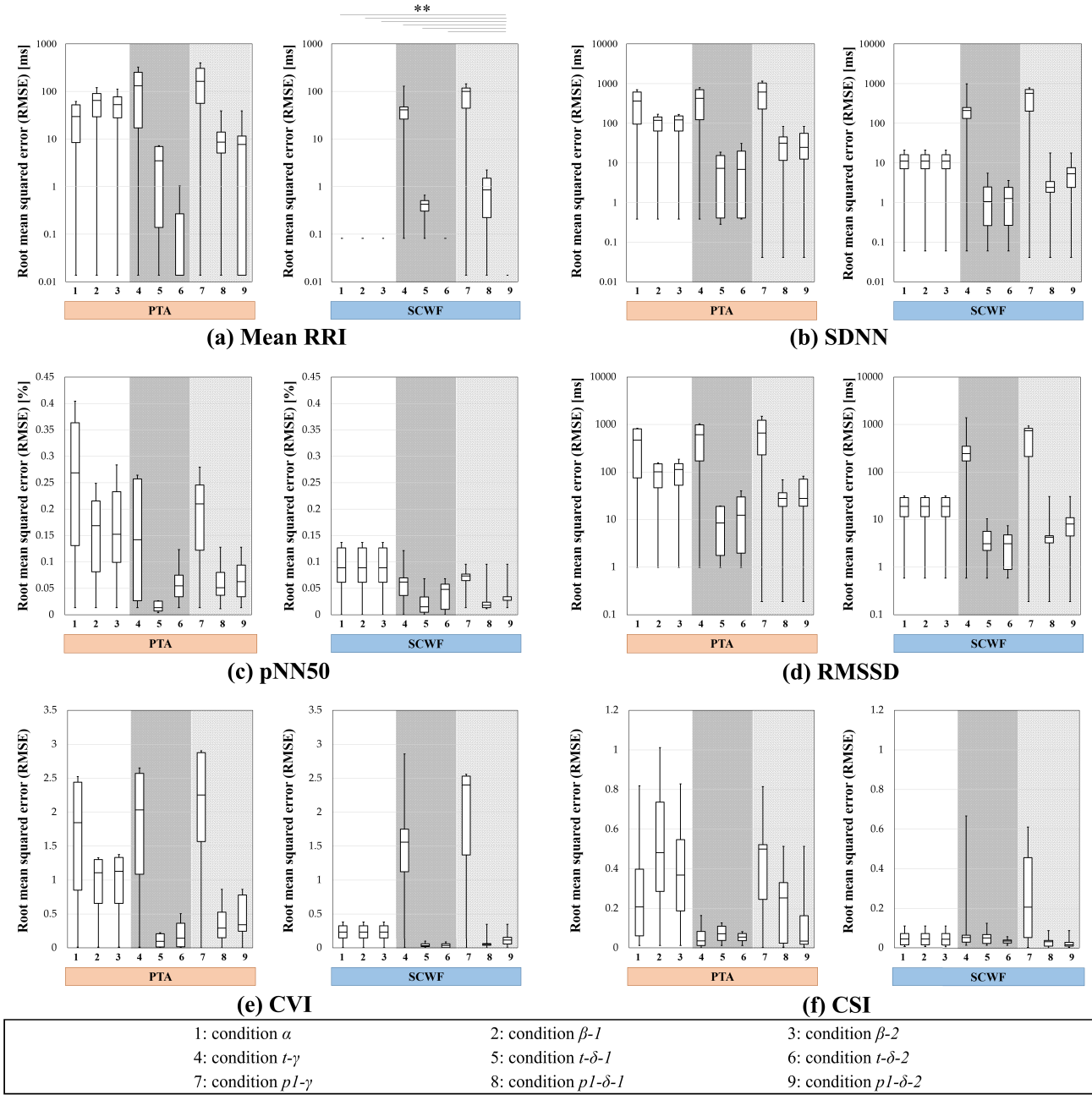


FIGURE 14. tHRVs calculated under each condition. The shaded areas in each graph indicate the set of conditions with regard to QRS complex rejection: no shade, without QRS complex rejection; darker shade, with QRS complex rejection by *theoretical* performance; lighter shade, with QRS complex rejection by *fail-safe practical* performance. Conditions α, β-1, β-2, t-δ-2, and p1-δ-2 in (a) obtained by SCWF are depicted as points instead of boxes because the RMSEs calculated from all eight target ECGs were the same. Here, their value was also the same in conditions α, β-1, β-2, and t-δ-2.

dubious RRIs with the average value of RRIs would be effective. Regarding tHRV features focusing on every two adjacent RRIs (i.e., pNN50, RMSSD, CVI, and CSI), the dubious RRI deletion would perform better unless the dubious replacement accomplishes perfect interpolation. Since these four tHRV features require adjacency to be ensured between every two RRIs at the minimum, interpolating imperfect RRIs cannot satisfy this initial requirement. As expected, the deletion approach (i.e., condition *t-δ-1*) performed the best in pNN50, RMSSD, and CVI: specifically, it performed well in pNN50 because it mainly focuses on whether the difference exceeds 50 ms. In this experiment, however, the best condition for CSI was the replacement approach (i.e., condition *t-δ-2*). Considering the original definition, CVI and CSI would be theoretically under the influence of volume, difference, and variability between every two adjacent RRIs. It is true that the miscalculation of geometric features using the Poincaré plot may fail depending on the remaining RRI, but the

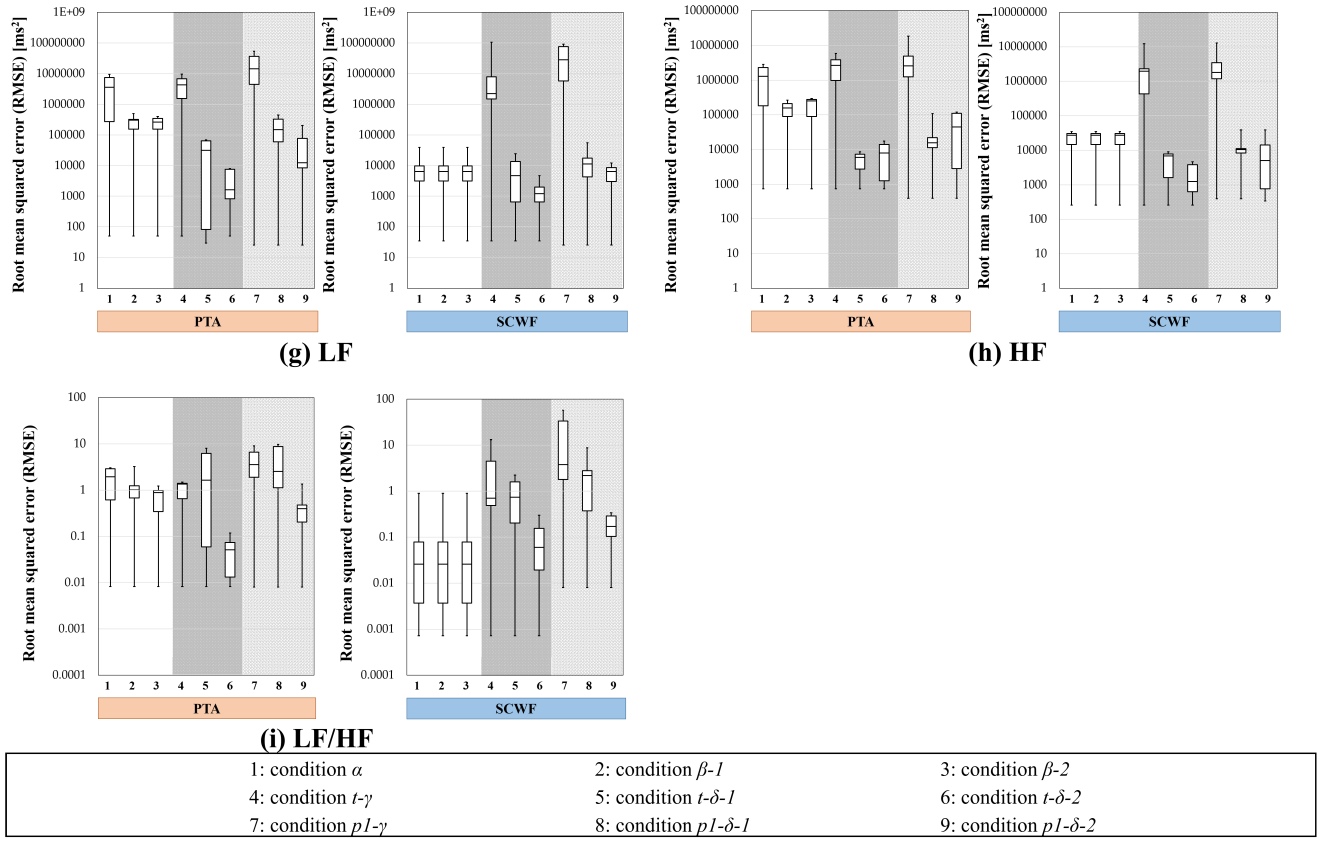


FIGURE 15. fHRVs calculated under each condition. The shaded areas in each graph indicate the set of conditions with regard to QRS complex rejection: no shade, without QRS complex rejection; darker shade, with QRS complex rejection by *theoretical* performance; lighter shade, with QRS complex rejection by *fail-safe practical* performance.

dubious RRI deletion would be better than the replacement in terms of ensuring adjacency. Unlike the aforementioned tHRV features, we cannot theoretically determine whether the deletion or replacement approach would be better for tHRVs focusing on variability (i.e., SDNN). From the perspective of suppressing the overvaluation of false fluctuation due to FPs, we should at least delete dubious RRIs. However, the suitability of the replacement would theoretically depend on its interpolation performance. In this experiment, we cannot confirm a consistent advantage in either the deletion approach or the replacement approach.

Regarding fHRVs, the replacement approach (i.e., condition *t-δ-2*) would be suitable from the perspective of suppressing the overvaluation of low-frequency components. Focusing on the results of LF, as consistent with the theory, the replacement approach (i.e., condition *t-δ-2*) performed better than the deletion approach (i.e., condition *t-δ-1*) regardless of the algorithm. This difference might originate from the preprocessing of the fHRV calculation. Considering the experimental conditions in this experiment, in theory, missing RRIs at the level of RRI sequence would be altered to the resampled RRIs by using the linear function in the preprocessing of the fHRV calculation. Since the deletion approach generates more missing RRIs for a longer period of time than the replacement approach, this might cause an overvaluation of low-frequency components. The replacement approach, on the other hand, interpolates missing RRIs at the level of the RRI sequence by the average value of the RRIs classified as "clean." These might act as direct current components and enable the overvaluation of low-frequency components to be suppressed. The difference in the RMSE of PTA and SCWF under the *t-δ-1* condition supports this hypothesis: PTA is more vulnerable to this issue than SCWF due to the longer missing RRI. However, regarding HF and LF/HF, we cannot confirm a consistent advantage in either the deletion approach or the replacement approach. Since LF/HF is the ratio of LF and HF, the order of magnitude in both components influences the accuracy. In other words, simply suppressing the overvaluation of LF does not necessarily contribute to obtaining a more accurate LF/HF depending on the RRI sequence. As a result of this ordering issue, the RMSE of SCWF under the *t-δ-2* condition could not reach the level of conditions *α* to *β-2*.

As a temporal summary of the comparative evaluation of the six experimental conditions from *α* to *t-δ-2* targeting *theoretical* performance, we conclude that the execution of both the dubious QRS complex editing step and dubious RRI editing step (conditions *t-δ-1* and *t-δ-2*) are, *in theory*,

potentially effective to obtain more accurate RRIs and HRV features. The execution of the dubious QRS complex editing step alone (condition *t-γ*) resulted in obtaining an inappropriate RRI sequence and HRV features compared to the conventional dubious RRI editing methods alone (conditions *β-1* and *β-2*). Regarding the dubious RRI editing method, the deletion approach (condition *t-δ-1*) is theoretically suitable for pNN50, RMSSD, CVI, and CSI and the experimental results supported this hypothesis in pNN50, RMSSD, and CVI. Meanwhile, the replacement approach (condition *t-δ-2*) is better for mean RRI targeting physiologically stable RRI sequences and LF.

Regarding the *practical* dubious QRS complex editing step (conditions *p1-γ* to *p1-δ-2*), the same tendency was confirmed regardless of the algorithm. Among these three conditions, either *p1-δ-1* or *p1-δ-2* performed the best. Specifically, the deletion approach (condition *p1-δ-1*) performed better for pNN50, RMSSD, and CVI, whereas the replacement approach (condition *p1-δ-2*) performed better for mean RRI and fHRVs. Of note, in the mean RRI obtained by SCWF, condition *p1-δ-2* performed significantly better than the *theoretical* dubious QRS complex editing step. The results of the RRI sequence suggest that this improvement was caused accidentally by the overvaluation of FPs together with dubious RRI replacement: condition *p1-δ-2* rejected more RRIs than actual FPs and replaced missing RRIs with values that were closer to the average. Overall, the combination of the *practical* dubious QRS complex editing step and dubious RRI editing step enabled us to obtain more accurate HRV features than the conventional dubious RRI editing methods (conditions *β-1* and *β-2*), even though the accuracy of several target HRV features did not reach the level of the *theoretical* performance (conditions *t-δ-1* and *t-δ-2*).

Comparing condition *α* between PTA and SCWF, as we confirmed at the level of RRI sequence, SCWF performed better. Comparing the same condition between PTA and SCWF, in general, the difference in the original algorithm performance was directly linked to the accuracy of the target HRV features. Although the combination of the dubious QRS complex editing step and dubious RRI editing step was able to improve the accuracy of HRV features, this proposed combination cannot turn over the difference in the original algorithm performance on the QRS complex detection. However, applying the proposed combination enabled PTA to improve the accuracy of the target HRV features at the same level as SCWF without the proposed combination.

In summary, the results of experiment 1 demonstrate that applying both the dubious QRS complex editing step and dubious RRI editing step is potentially effective for improving the target HRV features regardless of the algorithm, even by its *practical* performance. Regarding the dubious RRI editing, the deletion approach is presumably more suitable for pNN50, RMSSD, CVI, and CSI whereas the replacement approach is better for mean RRI and LF. Applying the proposed combination enables the QRS complex detection algorithm with poor performance to improve the accuracy of the target HRV features at the same level as the one with better performance that does not apply the proposed combination.

TABLE 5. 13 workouts comprising radio exercise no. 1 and approximate duration of each.

No.	*Workout*	*Duration [s]*
1	Stretching the body	15*
2	Swinging the arms and bending/stretching the legs	15
3	Rotating the arms	15
4	Spreading the chest	15
5	Side bending of the body	15
6	Bending the body back and forth	15
7	Body twisting	15
8	Stretching the arms up and down	15
9	Bending the body diagonally downwards and spreading the chest	15
10	Rotating the body	15
11	Jumping with both legs	10
12	Swinging the arms and bending/stretching the legs (the same as no. 2)	15
13	Deep breaths	25

*Including 8.5 s for "pause."

D. EXPERIMENT 2: EVALUATION TARGETING SEQUENCE OF DETECTED POINTS OBTAINED FROM REAL ECG DATASET RECORDED BY SINGLE-CHANNEL SHIRT-TYPE WEARABLE ECG DEVICE

In this experiment, we evaluated the performance of the proposed method targeting the sequence of detected points obtained from a *real* ECG dataset recorded during an exercise activity by a single-channel shirt-type wearable ECG device. As in experiment 1, the accuracy of HRV features was the focus of evaluation.

1) EVALUATION CONDITIONS IN EXPERIMENT 2

Through the evaluation of the dubious QRS complex identification step targeting the *real ECG dataset* (described in Appendix E.2), we confirmed that the detected points were classified as "artifacts," "noise," or "clean." Hence, we target the same 12 experimental conditions described in V-B: *α*, *β-1*, *β-2*, *t-γ*, *t-δ-1*, *t-δ-2*, *p1-γ*, *p1-δ-1*, *p1-δ-2*, *p2-γ*, *p2-δ-1*, and *p2-δ-2*.

2) TARGET EXERCISE

Under the assumption of a daily life environment, the previous study [24] targeted ECGs recorded while the participants performed "radio exercise no. 1 [52], [53]." This exercise was originally designed to help ordinary people improve their physical fitness, so it comprises exercises for all parts of the body, including jumping and twisting/stretching/bending of the trunk. The 13 short workouts in this exercise are listed in Table 5 and can be easily performed by ordinary people from children to the elderly. The detailed movements of each workout are provided in Appendix B of the previous study [24]. Each workout is set to music with short commands on the next workout, and the entire set takes approximately three min from start to finish [52], [53].

Before the experiment, all participants watched the video and practiced the movements. In case a participant forgot any of the movements, we also played the exact same video during the experiment.

3) DEVICE SETUP FOR CREATING REAL ECG DATASET

To evaluate RRI and HRV regardless of the signal quality of the target ECG, the previous study [24] used the sequence of detected points obtained by two independent wearable ECG devices at the same time: a commercial single-channel shirt-type device that records the target ECGs for the performance evaluation, and a three-channel patch-type device with medical approval that records the reference ECGs for the comparative evaluation.

In the experiment, a participant first wore the three-channel patch-type wearable ECG device for the reference ECG and then wore the single-channel shirt-type wearable ECG device for the target ECG. After wearing both devices, we confirmed that each was able to record ECGs without interfering with the other device.

4) OVERVIEW OF REAL ECG DATASET

The real ECG dataset comprises three ECGs recorded from three healthy male participants (age: 30.0 $\pm$ 1.633) while performing "radio exercise no. 1." All participants provided written informed consent. In accordance with the approximate duration of each workout in the video, we took the same approach as the previous study [24] and targeted the ECGs recorded for 200 s from the start of the designated music.

Fig. 16 shows the three target ECGs. The signal quality measured by the signal-to-ARTIFACTS ratio (SAR) defined in Appendix B.2 was 1.70 $\pm$ 3.46. Note that we use the capitalized term "ARTIFACTS" to mean the possible components other than heart activity. The details of the SAR per target ECG and its relevance to the movement are discussed in Appendix D. As shown in Fig. 16 and Table 7, as a whole, the ECG obtained from participant 3 (Fig. 16(iii)) was comparatively clean regardless of the movements. For participants 1 (Fig. 16(i)) and 2 (Fig. 16(ii)), on the other hand, several parts of the recorded ECGs were contaminated with noise/artifacts.

Since we used two individual devices for recording the target ECG and reference ECG, we paired both per participant in accordance with the method described in V-D8. Note that the SNR and the SAR of the reference ECG were not available because the wearable ECG device used for obtaining it cannot output raw ECG data.

5) THREE-CHANNEL WEARABLE ECG DEVICE USED FOR RECORDING REFERENCE ECGS

The reference ECGs were recorded by a three-channel patch-type Holter ECG monitoring device (Cardy 303 pico+, SUZUKEN CO., LTD., Aichi, Japan) with medical approval. This device can record three-channel bipolar ECGs simultaneously (i.e., NASA, CM5, and an auxiliary lead that is similar to half of CC5) at a sampling rate of 125 Hz. This three-channel ECG recording enables the suppression of obtaining FPs or FNs due to the processing failure of the recorded ECGs. More detailed information on this device can be found in the previous study [24].

6) REFERENCE QRS COMPLEXES OBTAINED FROM REFERENCE ECGS

Fig. 17 shows the reference RRIs obtained from each participant by using the Holter ECG device. Regarding physiological changes in the RRI sequence, overall, the RRIs were shortened along with the progress of elapsed time regardless of the participants. Since experiment 2 targets the ECGs recorded during exercise, this RRI shortening is physiologically reasonable.

On the basis of the previous study [24], there were no RRI calculation failures in the reference ECG of participant 1. For participants 2 and 3, on the other hand, several RRI calculation failures were observed by the annotations: specifically, the device failed to obtain RRIs observed at 170.28 s and 170.56 s for participant 2, and at 85.856 s and 86.144 s for participant 3. As we did in the previous study, we excluded those four RRIs and used the rest of them as the reference QRS complexes for the comparative evaluation of the detected points from the target ECGs. In total, the following points were set as the reference QRS complexes in this study: 274 points for participant 1, 332 points for participant 2, and 321 points for participant 3.

7) SINGLE-CHANNEL WEARABLE ECG DEVICE USED FOR RECORDING TARGET ECGS

The target ECGs were recorded by a commercial single-channel shirt-type wearable ECG device comprising a specially designed shirt (Toray Industries, Inc., Tokyo, Japan) and an attachable wearable ECG device (hitoe®transmitter 01, NTT DOCOMO, INC., Tokyo, Japan). Lead wires and electrodes made of the functional material hitoe®(Toray Industries, Inc., Tokyo, Japan) were embedded inside the specially designed shirt. Note that hitoe®is an electroconductive textile fabric made of nano-fiber yarn (fiber diameter 700 nm; polyester) coated with a PEDOT-PSS polymer thermobonding composition [9], [54].

The combination of the specially designed shirt with embedded electrodes and an attachable wearable ECG device is capable of recording a single-channel ECG whose measurement lead is similar to a bipolar chest lead CC5 [55] at a sampling rate of 200 Hz. More detailed information on this device can be found in the previous study [24].

After finishing the target ECG recording, all ECG data were transferred to a personal computer, and the designated QRS complex detection algorithm was applied offline.

8) RULES FOR CLASSIFYING DETECTED POINTS

To unify the time counting systems of the two independent devices, the previous study used the time elapsed since the observation time of the first RRI.

To classify the detected points as accurately as possible under the situation in which the sampling rates of two ECGs were different (i.e., 125 Hz for the reference ECGs and 200 Hz for the target ECGs) when raw ECG data are not necessarily available, the previous study only classified a detected point as TP when it was observed within 0.10 s from the observation time of the corresponding reference RRI, and otherwise classified it as FP. The previous study assumed that these rules were also capable of resolving sampling rate differences in the same QRS complex: a sampling rate difference induced by a theoretical time gap (i.e., 0.003 s) was comparatively smaller than the time gap between the reference RRI and the detected point (i.e., less than 0.10 s).

9) TARGET HRV FEATURES

We calculated all the HRV features targeting the sequence of the detected points obtained by each algorithm from the target ECGs (i.e., about 200 s from the start of the designated music).

Since this experiment targets the ECGs recorded during exercise activity, we only target the six tHRVs shown in Table 2 marked as *evaluation targets*. As in experiment 1, we independently checked the adjacency of two adjacent RRIs by (1) before calculating tHRVs focusing on the characteristics of topical NN intervals (i.e., RMSSD, pNN50, CVI, and CSI).

10) EVALUATION METHODS FOR HRV FEATURES

As in experiment 1, we conducted a two-step evaluation: one for RRIs and the other for HRV features. As the pre-evaluation of the target RRIs, we first confirmed the physiological RRI changes by the tachograms of the reference RRIs. We then investigated a brief summary at the level of the RRI sequence originating from each experimental condition based on the tachogram of the calculated RRIs.

For the performance evaluation of the target tHRVs, we used the RMSE with the HRV features calculated from the reference RRIs as an ideal value. For each HRV feature, we calculated the RMSE between the ideal value and the calculated value obtained under each condition and then compared them among the 12 experimental conditions. We then assessed the performance of each condition using the box plot of RMSEs.

11) PRE-EVALUATION OF RRIS

Fig. 18 and Fig. 19 show the RRI tachograms obtained under each experimental condition. Regarding the original performance of the target algorithms shown as condition α in each target ECG, PTA and SCWF performed differently: PTA resulted in obtaining FPs and FNs during/after noise/artifacts, whereas SCWF resulted in obtaining FPs and FNs during noise/artifacts and was able to recover after the contamination. Regardless of the algorithm, the number of FPs and FNs seems to be directly proportional to the signal quality of the target ECG: both algorithms seldom resulted in obtaining FPs or FNs in the ECG with better quality (i.e., participant 3), whereas both frequently resulted in obtaining FPs or FNs in the ECG with worse quality (i.e., participants 1 and 2).

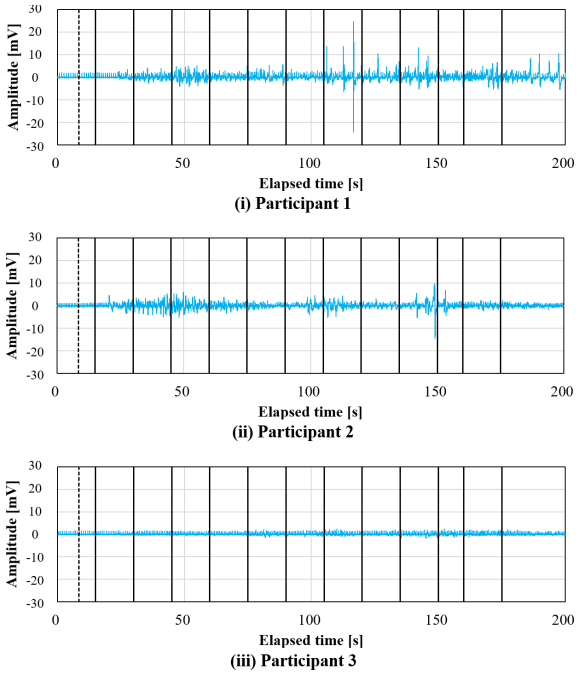


FIGURE 16. Target real ECGs in experiment 2. All ECGs were recorded by a commercial single-channel shirt-type wearable ECG device during the target exercise. The solid vertical lines indicate the end of each workout (i.e., switching point of the movements), whereas the dotted one indicates the end of "pause" in workout no. 1.

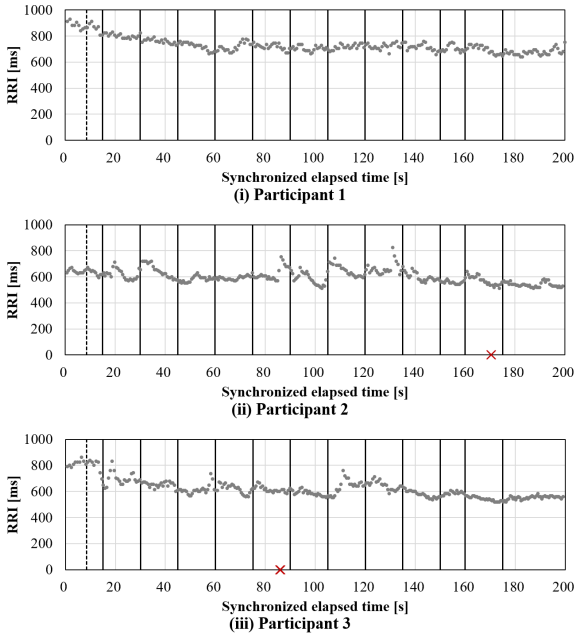


FIGURE 17. Reference RRIs obtained from each participant by a series of three-channel wearable Holter ECG monitoring devices. The solid vertical lines indicate the end of each workout (i.e., switching point of the movements), whereas the dotted one indicates the end of "pause" in workout no. 1. The red crosses at 170.28 s and 170.56 s for (ii) Participant 2 and at 85.856 s and 86.144 s for (iii) Participant 3 indicate erroneous RRIs according to the annotation (i.e., RRIs excluded from the reference RRIs).

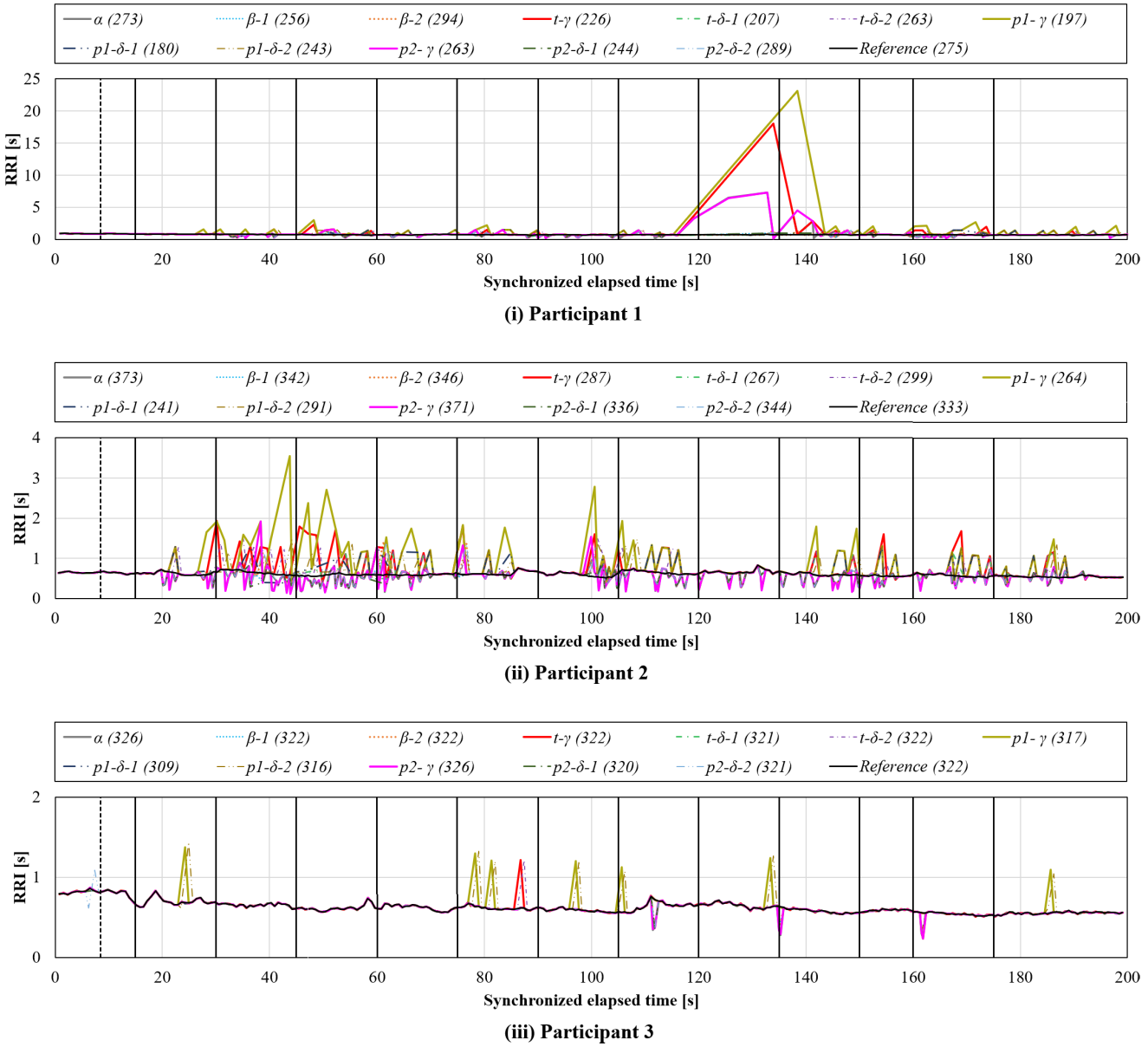


FIGURE 18. RRI tachograms calculated from QRS complexes obtained by PTA. The solid vertical lines indicate the end of each workout (i.e., switching point of the movements), whereas the dotted one indicates the end of "pause" in workout no. 1.

As confirmed in experiment 1, among the six experimental conditions from α to *t-δ-2* targeting *theoretical* performance, the execution of only the dubious QRS complex editing step (i.e., condition *t-γ*) resulted in obtaining inappropriately prolonged RRIs regardless of the algorithm. When ignoring missing RRIs, the dubious QRS complex editing step together with the dubious RRI editing step taking the deletion approach (condition *t-δ-1*) was able to obtain the most accurate RRI sequence. Regarding the *practical* dubious QRS complex editing step (conditions *p1-γ* to *p2-δ-2*), a similar tendency was consistently confirmed: among these six conditions, *p1-γ* and *p2-γ* performed worse due to prolonged RRIs, whereas *p1-δ-1* and *p2-δ-1* performed better if we ignore missing RRIs.

Comparing the two *practical* approaches (conditions *p1-γ* to *p1-δ-2* corresponding to the fail-safe approach and conditions *p2-γ* to *p2-δ-2* corresponding to the fail-soft approach), each performed differently: the fail-soft approach ensured a greater number of RRIs than the fail-safe approach when ignoring several misidentifications of FPs, whereas the fail-safe approach ensured a better quality of RRIs when ignoring the number of RRIs.

In conditions *t-γ*, *p1-γ*, and *p2-γ* of participants 2 and 3, there were a few FPs caused by the absence of the reference RRI. As we mentioned in V-D6, the RRIs observed at 170.28 s

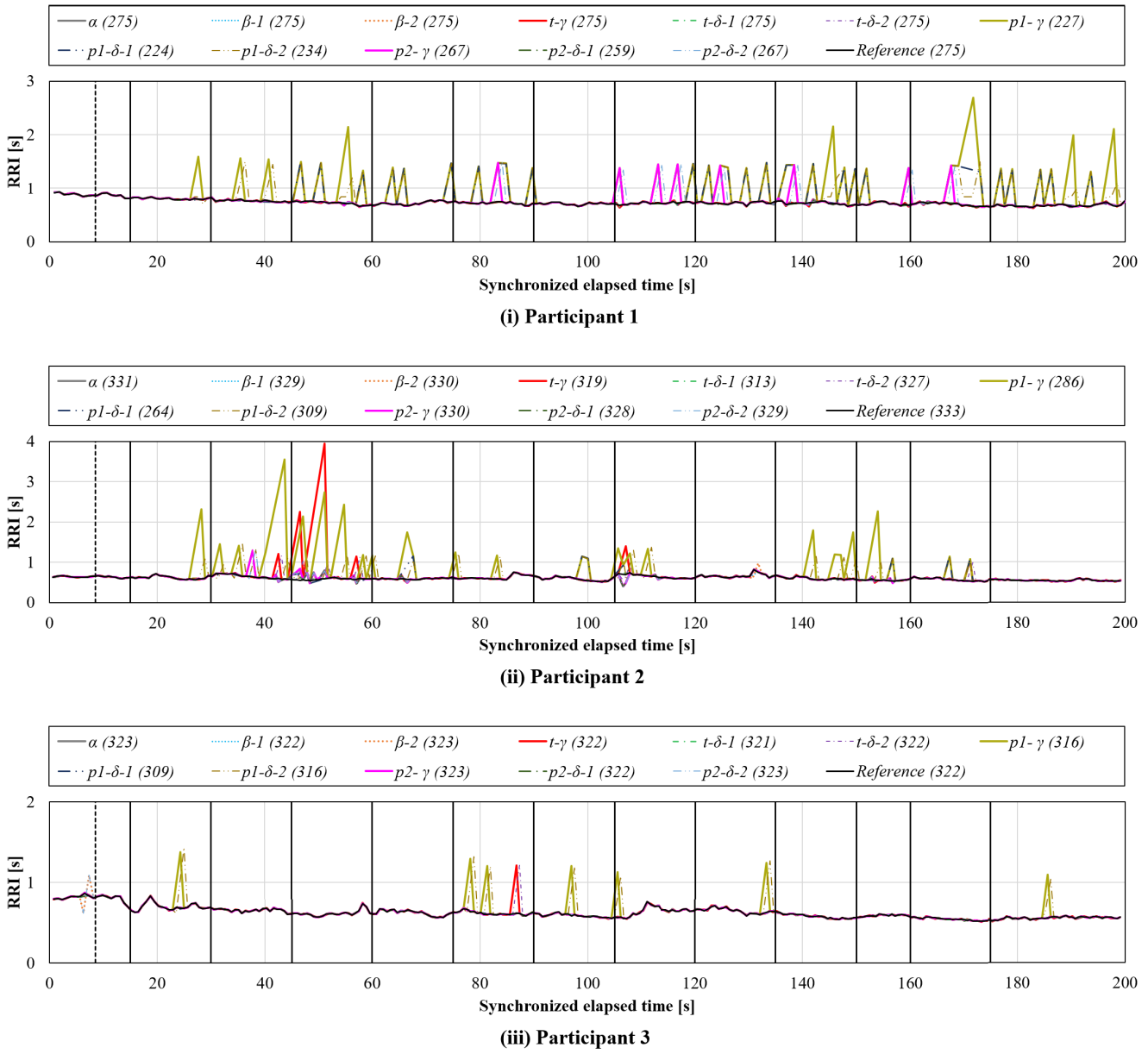


FIGURE 19. RRI tachograms calculated from QRS complexes obtained by SCWF. The solid vertical lines indicate the end of each workout (i.e., switching point of the movements), whereas the dotted one indicates the end of "pause" in workout no. 1.

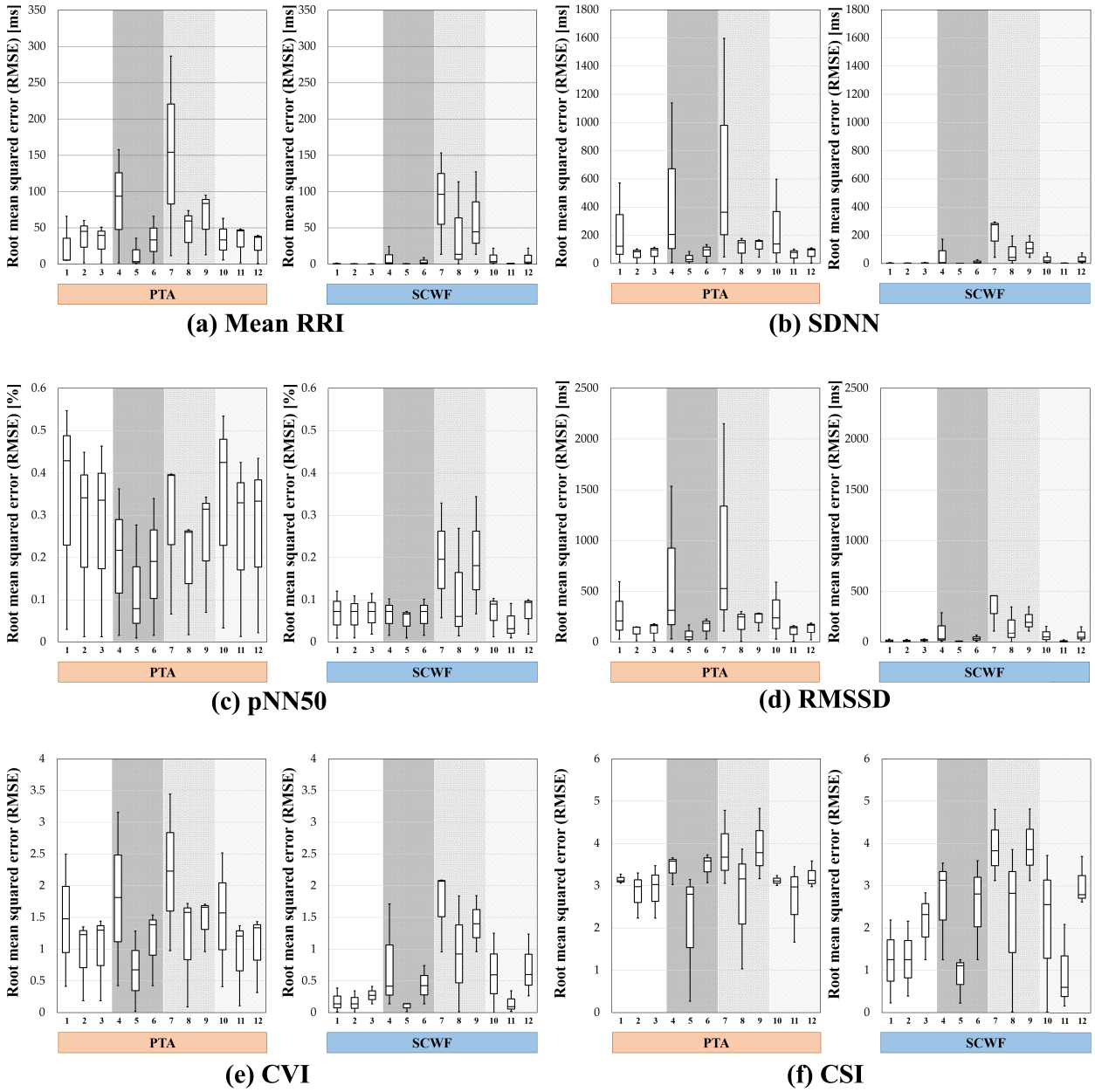
and 170.56 s for participant 2, and at 85.856 s and 86.144 s for participant 3, were excluded, so these RRIs would be unfairly rejected even if they were correct. Considering the original performance of SCWF, the only prolonged RRI observed in condition *t-γ* of participant 3 might have suffered from this issue.

In summary, overall, executing both the dubious QRS complex editing step and dubious RRI editing step seems likely to improve the accuracy at the level of RRIs. We will investigate the plausible RRI editing approach through the evaluation targeting HRV features.

12) RESULTS OF HRV FEATURES

Fig. 20 shows the RMSEs of the target tHRV features calculated under each experimental condition targeting the detected points obtained by PTA and SCWF. Steel-Dwass tests revealed no significant difference between any pairs of the experimental conditions in any of the target HRV features regardless of the algorithm.

Comparing the six experimental conditions from *α* to *t-δ-2* targeting *theoretical* performance, based on the median, the deletion approach (i.e., condition *t-δ-1*) consistently enabled the most accurate HRV features to be obtained among all the target HRV features regardless of the algorithm. We herein



1: condition α	2: condition β-*1*	3: condition β-*2*
4: condition *t*-γ	5: condition *t*-δ-*1*	6: condition *t*-δ-*2*
7: condition *p1*-γ	8: condition *p1*-δ-*1*	9: condition *p1*-δ-*2*
10: condition *p2*-γ	11: condition *p2*-δ-*1*	12: condition *p2*-δ-*2*

FIGURE 20. **tHRVs calculated under each condition. The shading in each graph indicates the set of conditions with regard to QRS complex rejection: no shade, without QRS complex rejection; the darkest shade, with QRS complex rejection by *theoretical* performance; middle gray shade, with QRS complex rejection by *fail-safe practical* performance; the lightest shade, with QRS complex rejection by *fail-soft practical* performance.**

investigate whether the best performing condition for each target HRV feature was consistent with theory in consideration of the original focus point of each HRV feature and the physiological RRI changes. In experiment 2, unlike in experiment 1, the deletion approach (condition *t*-δ-*1*) performed better for mean RRI calculation than the replacement approach (condition *t*-δ-*2*). Considering the physiological RRI shortening in experiment 2, this seems to be reasonable. When RRIs are consistently changing, the replacement approach using the average value of RRI would not necessarily be effective for mean RRI calculation: the average value itself would be under the influence of the presence bias of the RRIs, so replacement using this value might cause false changes. Specifically, real ECGs were confirmed to be

clean at the beginning of the exercise but contaminated with noise/artifacts in the middle (see Appendix D for details), so the average value might strongly reflect the RRI value at the beginning of the exercise. This false change might lead to the miscalculation of the mean RRI. Although the deletion approach is not a perfect solution, it would perform better unless it causes inappropriate results due to the presence bias of RRIs. Regarding tHRVs focusing on every two adjacent RRIs (i.e., pNN50, RMSSD, CVI, and CSI), the deletion approach (condition *t-δ-1*) consistently performed better for the same reasons described in experiment 1: in short, it was able to better suppress false RRI fluctuations compared to the replacement approach (condition *t-δ-2*). As we mentioned, this result would be reasonable in consideration of physiological RRI shortening, the possible value used for the replacement approach, and the possible false variability between the remaining RRI and the replaced RRI. Regarding tHRVs focusing on the variability of the whole target RRIs (i.e., SDNN), we cannot theoretically assert whether the deletion or replacement approach would be better; it might depend on whether the false RRI variability fits with the real RRI variability.

A temporal summary of the *theoretical* performance evaluation (conditions α to *t-δ-2*) is consistent with experiment 1. Specifically, executing both the dubious QRS complex editing step and dubious RRI editing step would be, *in theory*, potentially effective for improving the accuracy of HRV features. On the basis of physiological RRI shortening due to exercise, the deletion approach (condition *t-δ-1*) would be preferable with regard to the dubious RRI editing step.

The performance of the *practical* dubious QRS complex editing step (conditions *p1-γ* to *p2-δ-2*) was, in a broad sense, consistent with the results of the *theoretical* performance evaluation. Based on the median, condition *p2-δ-1* performed the best in most HRV features regardless of the algorithm except for two HRV features by PTA: condition *p2-γ* for mean RRI and condition *p1-δ-1* for pNN50. Although the performance of these best *practical* conditions did not reach the level of the *theoretical* performance, they were able to obtain more accurate HRV features than the conventional dubious RRI editing methods (conditions α to β*-2*) in most cases; the exceptions were mean RRI by PTA and SCWF as well as SDNN by SCWF. As mentioned in the *theoretical* performance evaluation, we presume these exceptions were mainly caused by the presence bias of RRIs.

Focusing on the performance of the two *practical* approaches (i.e., the fail-safe approach in conditions *p1-γ* to *p1-δ-2* and the fail-soft approach in conditions *p2-γ* to *p2-δ-2*), the fail-soft approach consistently performed better for most of the target HRV features regardless of the algorithm. Considering the RRI sequence and the RMSEs, however, the reason for this would be different depending on each algorithm. The main reason for SCWF was the quality and quantity of RRIs: on the basis of the RRI sequence, SCWF seemed able to detect QRS complexes from ECG with noise, so it does not need to reject the detected points classified as "noise." In other words, we can actively select the fail-soft approach. This hypothesis is supported by the results of participant 3: as we mentioned in V-D11, the only prolonged RRI observed in condition *t-γ* might be caused by the lack of the reference RRI, so SCWF in condition *p2-δ-1* was able to obtain the RRI sequence that might be closest to the actual one. Regarding PTA, on the other hand, the main reason might be the number of RRIs and their presence bias. Specifically, in participant 1, missing RRIs generated by the fail-safe approach became extremely large (i.e., 23.16 s), so deleting this prolonged RRI may cause an inappropriate calculation due to the presence bias of RRIs: in theory, the mean RRI might be under the influence of this issue. In other words, the fail-soft approach was passively selected to avoid other issues that might have had a larger impact. This passive selection of the fail-soft approach might not necessarily work well depending on the target HRV features: for example, because the fail-safe approach can strictly ensure the quality of RRIs, it worked better for pNN50, which focuses on the local difference between two adjacent RRIs.

Comparing the same condition between PTA and SCWF, as we confirmed in experiment 1, SCWF consistently performed better than PTA. The difference from experiment 1 was the impact of the performance of QRS complex detection on the accuracy of HRV features. Since the target *real* ECG dataset suffered from intermittent noise/artifacts more frequently than the *pseudo* ECG dataset, applying the proposed combination could not allow PTA to obtain the target HRV features at the same level as SCWF without the proposed combination.

In summary, the results of experiment 2 demonstrate that, in a broad sense, applying both the dubious QRS complex editing step and dubious RRI editing step taking the deletion approach is potentially effective for improving the target HRV features regardless of the algorithm. We should mention again here that the accuracy improvement by the proposed intermediate processing might be limited by the performance of the algorithm used for QRS complex detection. Regarding the *practical* dubious QRS complex editing step, the *fail-soft* approach would enable fair processing, but for the pNN50 calculation by PTA, the *fail-safe* approach performed better. Since PTA could not accurately detect QRS complexes from noise, the *fail-safe* approach that only uses the detected points from "clean" can ensure more accurate calculation than the *fail-soft* approach that uses the detected points from "clean" and "noise."

VI. DISCUSSION

A. EFFECTIVENESS AND ADVANTAGES OF OUR PROPOSAL

In this study, we proposed subdividing the dubious RRI identification and dubious RRI editing steps in accordance with the target unit (i.e., QRS complex or RRI). The results of two experiments showed that applying both the dubious QRS complex editing step and dubious RRI editing step is *theoretically* effective for improving the target HRV

features regardless of the algorithm used. Although the proposed framework mainly focuses on technological disturbances in HRV analysis, it may still work well when there are only a few arrhythmic beats. Arrhythmic beats generally accompany inter-beat interval changes, so the influence derived from relatively few arrhythmias can be recovered by a combination of the dubious RRI identification step and dubious RRI editing step. For this reason, the proposed framework shows good potential for non-clinical healthcare services utilizing HRV features when the main target is healthy users. We can therefore conclude that our proposed reframing would be *theoretically* suitable for HRV analysis.

As an initial method for *practical* dubious QRS complex identification assuming a single-channel wearable ECG without a reference, we also proposed utilizing the amplitude at the detected point in consideration of the characteristics of ECGs themselves and wearable ECG devices, which we summarized in Sections II and III. Although the performance of the *practical* dubious QRS complex identification was not perfect, it performed better than the combination of conventional dubious RRI identification/editing steps only targeting the RRI unit.

The three unique advantages in our proposed *practical* dubious QRS complex identification are its applicability, real-timeness, and labor-effectiveness. As shown in our experiments, it can be applied for HRV analysis regardless of the QRS complex detection algorithm because it only uses the ECGs before/after 0.10 s of each detected point. For the same reason, it can be applied in real-time (i.e., 0.10 s immediately after the detection) even though it is post-hoc processing for QRS complex detection. Before introducing the final advantage, high labor-effectiveness, we should reiterate that the calculation required for applying our proposed method is simple, involving just the absolute value of the peak and the gap between the absolute value and the local minima of the peak. As we confirmed in experiment 1, applying the combination of our proposed *practical* dubious QRS complex editing and the dubious RRI editing enables the algorithm with poor performance (i.e., PTA) to improve the accuracy of HRV features at the same level as the one with high performance (i.e., SCWF) without our proposed intermediate processing. Since only the QRS complex detection step has the unique capability of detecting QRS complexes from recorded ECGs, it is clear that developing algorithms with better performance is crucial; as a major premise in HRV analysis, using an algorithm that is tolerant to noise/artifacts is undoubtedly essential, and applying the proposed combination cannot turn over the difference in the original algorithm performance on the QRS complex detection. Considering the possible efforts required for improving the performance of QRS complex detection, however, applying the proposed combination would be the best choice in terms of ensuring labor-effectiveness.

B. BEST APPROACH WITHIN PROPOSED SUBDIVIDED STEPS FOR ACCURATE HRV ANALYSIS REVEALED THROUGH EXPERIMENTS

The results of our experiments also clarified that the best approach for QRS complex editing and RRI editing will vary depending on several factors: the performance of the QRS complex detection algorithm, the target HRV features, and even the physiological RRI changes. Fig. 21 summarizes our findings gleaned through the experiments. For future reference, we highlight insights on our proposed intermediate processing in accordance with the order of HRV analysis.

Regarding the dubious QRS complex identification step and the dubious QRS complex editing step, ideally, all FPs should be identified and rejected without overlooks. Although we validated two approaches in this study (*fail-safe* and *fail-soft*) as *practical* QRS complex editing methods, the experimental results indicated that the appropriate approach might change depending on the performance of the algorithm used for QRS complex detection, the number of RRIs, and the target HRV features. Our findings clarified that the *fail-soft* approach was appropriate for an algorithm with good performance (e.g., SCWF) regardless of the target HRV features. Regarding an algorithm with poor performance (e.g., PTA), the *fail-safe* approach would perform better for tHRVs focusing on every two adjacent RRIs (e.g., pNN50) because it can strictly ensure the quality of RRIs; otherwise, the plausible approach would change depending on the target HRV features.

Regarding the dubious RRI identification step and the dubious RRI editing step, they should at least identify and delete the prolonged RRIs caused by FP rejection in the dubious QRS complex editing step. Although the ideal replacement using complete interpolation for missing RRIs might improve the accuracy of all HRV features, we should choose either deletion or incomplete replacement with a limited performance under the practical situation. Our findings indicated that the following approaches would be suitable from theory and practice: for tHRVs focusing on every two adjacent RRIs (e.g., pNN50, RMSSD, CVI, and CSI), deletion approach; and for fHRVs, replacement approach using direct current components (e.g., average). Regarding the tHRVs focusing on the volume of the whole RRIs (e.g., mean RRI), the best approach might be different depending on the physiological RRI changes. Specifically, the replacement approach using the average of RRIs performed better than the deletion approach when targeting stable RRIs, but it caused rather inappropriate results when targeting changing RRIs (e.g., shortening or lengthening). Regarding the tHRVs focusing on the variability of the whole target RRIs (e.g., SDNN), we cannot theoretically assert whether deletion or replacement would be better; although deletion meets the minimum requirement, replacement might be better depending on whether the false RRI variability matches the real RRI variability.

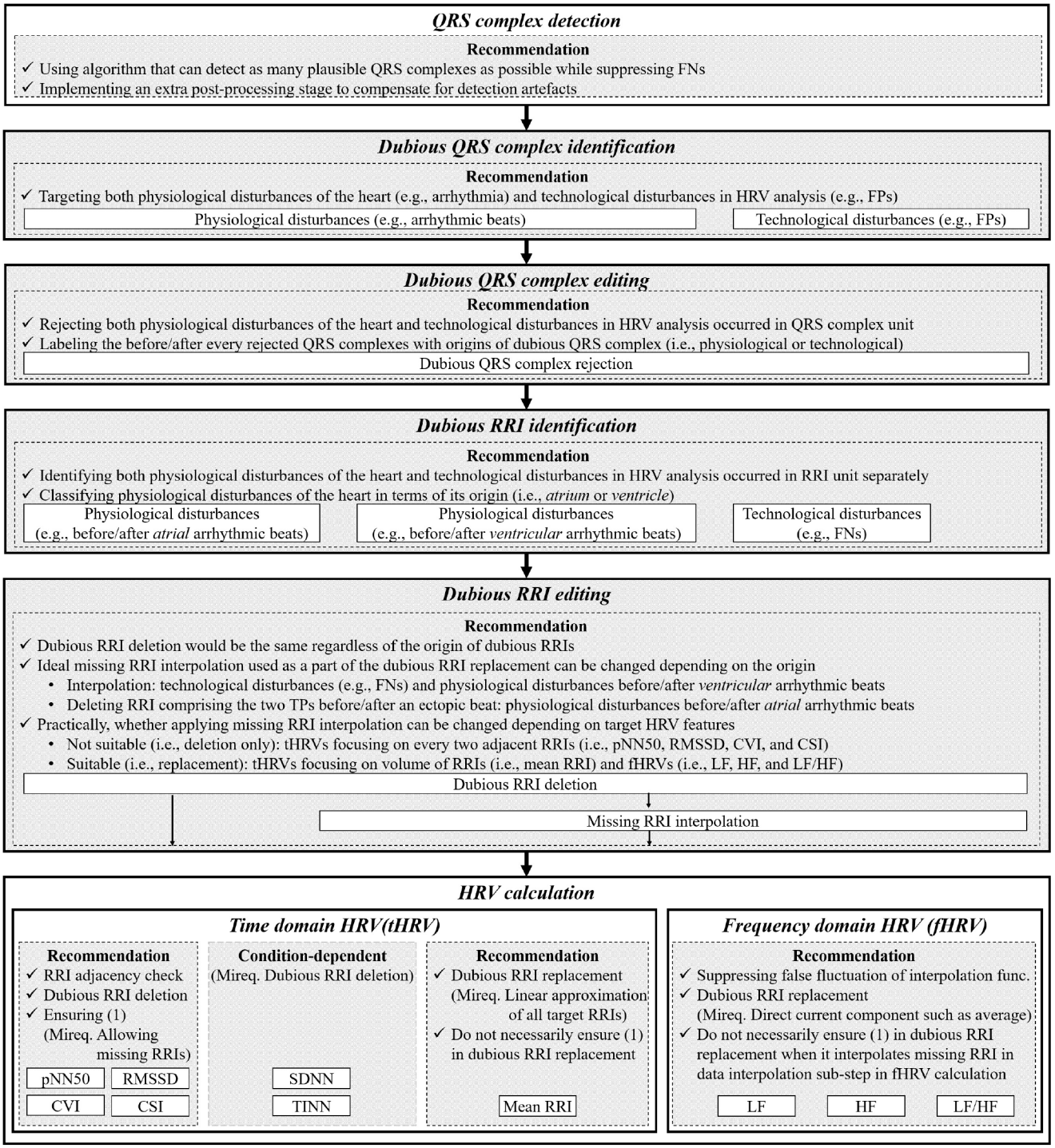


FIGURE 21. Overview of ideal heart rate variability (HRV) analysis flow considering the origin of dubious QRS complex and dubious RRI, physiological characteristics of arrhythmias, and the definition of each HRV feature. Shaded area indicates our proposal. Fig. 21 modified Fig. 9 by subdividing dubious QRS complex identification and dubious RRI identification in accordance with its origin (i.e., physiological disturbances of the heart and technological disturbances in HRV analysis). Note that we omit several sub-steps in the conventional HRV analysis flow here (i.e., QRS complex detection step and fHRVs calculation step) that are irrelevant to the proposal of this study. Definition of abbreviations: Mireq, minimum requirements; interpolation func., interpolation function.

C. IDEAL FRAMEWORK FOR HRV ANALYSIS

Since HRV features should be calculated from the NN interval sequence in theory [1], we ultimately need to ensure that all TPs come from pure sinus node depolarization without any false inter-beat interval changes. In other words, we should rephrase the aim of the intermediate processing steps between

the QRS complex detection step and the HRV calculation step as suppressing the influences derived from ECG disturbances of physiological origins and technological origins to guarantee the accuracy of the RRI sequence and HRV features. In this sense, we need to subdivide the TP class based on whether it comes from pure sinus node depolarization. The TP class in this study simply means that the corresponding beat comes from the heart activity, so a single ectopic beat such as PAC might be regarded as TP and left as it is. Thus, ideally, both the dubious QRS complex identification step and dubious RRI identification step should extend their target to the physiological disturbances of the heart (e.g., arrhythmia) in addition to the technological disturbances in HRV analysis that we targeted in this study (i.e., FPs and FNs).

Fig. 21 shows the ideal framework for the HRV analysis we consider. The ideal dubious QRS complex identification step should aim for *false changes occurring in a heartbeat unit* from the perspective of the technological disturbances (e.g., FPs) and the physiological disturbances (e.g., single ectopic beats such as PVC/PAC or short-run atrial/ventricular tachycardia). Since a false QRS complex (e.g., FP or ectopic beat) should in theory be rejected regardless of its origin, the dubious QRS complex editing step should be uniquely accomplished by rejection. Simultaneously, the ideal dubious RRI identification step should aim for *false changes occurring in an inter-beat interval unit* from the perspective of the technological disturbances (e.g., FNs) and the physiological disturbances (e.g., prolonged inter-beat interval after rejecting arrhythmic beats).

Along with this extension of the identification target, the dubious RRI editing step should change its processing in accordance with the origin of dubious RRIs. Considering the cascading effects of arrhythmia, dubious RRIs coming from physiological disturbances of the *atrium* should be deleted. On the other hand, dubious RRIs coming from technological disturbances in HRV analysis, as well as dubious RRIs coming from physiological disturbances of the *ventricle*, should be replaced with ideal RRIs by combining the dubious RRI deletion and missing RRI interpolation. Physiologically, an *atrial* ectopic beat may reset the sinus rhythm such that the sinus rhythm before and after an *atrial* ectopic beat is different: in other words, there is no ideal beat substitute for an ectopic beat. As we mentioned, an ectopic beat itself should be rejected in theory, so all we can do in this situation is delete the RRI comprising the two TPs before and after an ectopic beat. Regarding dubious RRIs coming from technological disturbances in HRV analysis, as well as dubious RRIs coming from physiological disturbances of the *ventricle*, on the other hand, the sinus rhythm itself is constant: the ultimate goal of the dubious RRI editing step targeting FNs should therefore be to replace dubious RRIs with ideal RRIs.

D. POSSIBLE FUTURE WORK TOWARDS IDEAL FRAMEWORK

As part of the dubious QRS complex identification step assuming the use of wearable ECG devices such as a single-channel shirt-type, we proposed an FP identification based on heuristic criteria that can be used regardless of the ECG recording lead. Although it can identify FPs caused by noise/artifacts with high amplitude, it is still difficult to identify FPs caused by artifacts with approximately the same amplitude as that of the QRS complex (e.g., MA) (see details in Appendix E). Since our experimental results showed that FP rejection would be *theoretically* effective for improving HRV feature calculation, improving the performance of *practical* FP identification is the next challenge.

To improve the FP identification performance and expand its target to include physiological disturbances, we cast an eye to the visual inspection of ECGs. Specifically, with adequate expert knowledge, it is possible for humans to empirically infer whether either physiological or technological disturbances have occurred in each heartbeat by sight while making use of amplitude, timing, and morphology. For example, when we observe an inter-beat interval change or a morphological change of the P-QRS-T part without any noise/artifacts, we can infer that it might be caused by physiological disturbances represented as arrhythmia. Conversely, when we confirm a morphological change of ECGs not limited to the P-QRS-T part, such as a plateaued baseline or overall amplitude change without any health condition change, we can infer that it might be caused by technological disturbances such as noise (e.g., overall amplitude change with a low frequency such as BW), artifacts (e.g., overall amplitude change with a high frequency such as MA or EM), or even detached electrodes (e.g., plateaued baseline). For these reasons, in a broad sense, the dubious QRS complex identification step can be divided into two sub-steps, as proposed in our previous study [56]: first, assessing the presence of a dubious QRS complex or signal quality in a segmented ECG unit, and second, applying the results of the segmented ECG to a corresponding detection point. Approaches for performing the first sub-step are not limited to a specific method but can be borrowed from those in conventional studies: for example, making use of morphological similitude [28], statistical ECG characteristics [56], or even CNN focusing on the graphical characteristics of the heartbeat observed by recurrence plot [57]. Another viable approach would be to apply supervised learning utilizing the dubious QRS complex classification results in this study.

We should emphasize that the conditions required for executing the dubious QRS complex identification step (e.g., data length for input and execution timing) are directly linked to real-time processing: a method that can be implemented directly after the detection of each point (such as the one used in this study) enables real-time processing, otherwise a pause time for processing the dubious QRS complex identification is required. Since there may be trade-offs between real-time processing and the accuracy of dubious QRS complex identification in general, the point of compromise should be determined in consideration of the purpose of the HRV analysis, the quality requirements for dubious QRS complex

identification (e.g., what degree the target of dubious QRS complex identification would be from FPs at the minimum to TPs coming from pure sinus node depolarization at the maximum), and the ECG recording conditions (e.g., ECG recording lead). In this sense, the proposed FP identification based on the QRS amplitude can function as a benchmark of FP identification: specifically, we can use it immediately after detecting each point regardless of the QRS complex detection algorithm and can obtain more accurate HRV features even though the performance may not be perfect.

Regarding the dubious RRI identification step, we might have to change the identification method along with expanding the identification target. As mentioned, the dubious RRI editing step in the ideal framework might edit dubious RRIs differently in accordance with their origin. In this sense, using the duration of RRIs (as done in this study) would not be suitable because, in principle, it cannot distinguish FNs from arrhythmias. For appropriate processing in the dubious RRI editing step, the dubious RRI identification step should at least label the before/after every rejected QRS complex while separating two origins. This labeling might be accomplished together with the dubious QRS complex editing step.

Regarding the dubious RRI editing step, the deletion approach would be the same regardless of the origin of dubious RRIs. We should emphasize several insights here related to missing RRI interpolation used as a part of the replacement approach for dubious RRIs originating from technological disturbances in the HRV analysis. Considering the calculation process of HRV features together with their point of focus, the requirements for interpolating missing RRIs can be changed depending on the timing. Specifically, when interpolating missing RRIs *at the timing of RRI calculation*, the interpolated RRIs should satisfy (1). In contrast, when interpolating missing RRIs *at the timing of HRV calculation*, the interpolated RRI does not necessarily have to satisfy (1). A typical example of the latter is fHRV calculation: the RRI sequence would lose (1) by the data interpolation sub-step in theory, so missing RRI interpolation at this timing may be free from having to satisfy (1). The previous study [16] utilized this theoretical characteristic and proposed interpolating missing RRIs during the preprocessing of fHRVs. Another example of the latter is mean RRI: since mean RRI does not focus on each RRI value in detail, both type-A and type-B RRI gaps stemming from (1) not being satisfied would be negligible. Conversely, missing RRIs should be avoided: the accuracy of mean RRIs may be decreased due to the presence bias of RRIs. For these reasons, even if (1) is not satisfied, interpolation by linear approximation of all the remaining RRIs can increase the accuracy of mean RRI while suppressing the presence bias of RRIs as well as reflecting the overall trend of the RRI sequence. As discussed above, depending on the target HRV features, the missing RRI interpolation approach suitable for improving the accuracy may be different depending on the method, timing, or satisfying (1). Although the use of several different missing RRI interpolation methods depending on target HRV features is technically viable, the point of compromise should be determined in consideration of the characteristics of the target HRV features as well as the assumed situation: extremely complicated missing RRI interpolation may lead to unexpected consequences.

E. LIMITATION

The major limitation of this study is the lack of a comprehensive evaluation: all the aforementioned insights are based solely on the results of experiments obtained from a few studies or desk studies related to the mechanism of the heart. As we have shown, a variety of factors affect each step of HRV analysis in theory, including physiological heart activity changes (e.g., RRIs that are physiologically stable/shortened/lengthened or the number of arrhythmic beats), ''static'' device specifications of hardware/software, ''active'' ECG recording conditions, and even the performance of the algorithm used for QRS complex detection. Although the two experiments in this study provide useful insights for HRV analysis, the findings are limited to two conditions: the situation in which the PTA or SCWF targeting the ECGs recorded by one specific commercial single-channel shirt-type wearable ECG device feature RRIs that are either stable (experiment 1) or shortened (experiments 2).

For a firmer HRV analysis theory construction, it will be necessary to accumulate real data and corresponding evaluation results utilizing several QRS complex detection algorithms. This will enable us to determine which approaches for each step have the highest affinity in consideration of the pros and cons inherent in each method along with their relevance to the accuracy of the HRV calculation. For achieving practical healthcare services using HRV features under the daily life environment, an effective data collection approach will be indispensable because the environmental conditions are constantly changing. We need to validate the results step-by-step and investigate the cascading effects over the estimation target status.

VII. CONCLUSION

Towards more appropriate HRV analysis, in this paper, we proposed reframing the traditional HRV analysis flow [1] by subdividing the RRI editing into four steps (Fig. 9) in accordance with the processing detail (i.e., identification and editing) and its target unit (i.e., QRS complex or RRI). In addition, as a dubious QRS complex identification method for practical use, we utilize the amplitude at the detected point assuming the use of a single-channel wearable ECG without a reference. Our experimental results showed that this processing/unit-based subdivision is theoretically effective for improving the target HRV features, and the dubious QRS complex identification method for practical use also maintains this effect. Furthermore, we confirmed that the plausible dubious RRI editing method would be different depending on the target HRV features and physiological RRI changes. Our experimental results revealed the following indications

with regard to the dubious RRI editing: the deletion approach would be suitable for tHRVs focusing on every two adjacent RRIs (e.g., pNN50); the replacement approach using a linear approximation of all the remaining RRIs would be suitable for tHRVs focusing on the volume (e.g., mean RRI); and the replacement approach using direct current components would be suitable for fHRVs.

Since HRV features should be calculated from the NN interval sequence in theory, we ultimately need to ensure that all TPs come from pure sinus node depolarization with regard to both the heartbeat unit and inter-beat interval unit. The ideal framework for HRV analysis (Fig. 21) should expand its identification target to dubious RRIs coming from the physiological disturbances of the heart (e.g., arrhythmias). As the construction of an ideal framework will necessitate an interdisciplinary research approach, it is indispensable to combine the knowledge of clinical medicine and engineering/informatics from the perspectives of both theory and practice.

APPENDIX

A. ESSENTIAL PRINCIPLES OF ECG

1) PRINCIPLE OF ECG RECORDING AND ITS RELEVANCE TO RECORDING DURATION

As with other electrophysiological signals, ECGs can be recorded by either a bipolar or unipolar measurement. In a bipolar ECG recording, the potential difference between a pair of electrodes is amplified by one amplifier channel. In a unipolar ECG recording, in contrast, the potential difference between one electrode and the reference is amplified. The reference in a unipolar ECG measurement is a potential calculated from several signals recorded by several electrodes [17], [18].

In principle, the duration of continuous ECG recording depends on the duration of electrode-to-skin contact, regardless of the measurement used. In a bipolar measurement, ECG recording is effective so long as two electrodes are placed on the skin, but it will fail and result in recording nothing if either of the electrodes comes off. In a unipolar measurement, the ECG recording is effective when all electrodes are placed on the skin, but again it will fail if any of the electrodes come off. In contrast to a bipolar measurement, the consequences of a failed unipolar measurement are different depending on which electrode detaches from the skin. Specifically, when the electrode comes off, the ECG recording will fail, and when the electrode comprising the reference comes off, the ECG recording itself might continue but the result will be different due to the change in the reference calculation.

2) ECG WAVEFORM AND WAVEFORM CHANGES CORRESPONDING TO ECG RECORDING LEADS

The typical shapes of ECG waveforms are shown in Fig. 1. These shapes are not constant but rather change depending on the ECG recording position. Usually, we interpret them in accordance with the similarity to the standard 12-lead ECG.

The 12-lead ECG is broadly divided into two types based on the ECG recording plane: one comprises chest leads focused on the heart in the horizontal plane, and the other comprises limb leads focused on the heart in the vertical plane. All the chest leads (i.e., V1 to V6) are unipolar ECGs, whereas the limb leads are comprised of three unipolar ECGs (i.e., aV_R, aV_L, aV_F) and three bipolar ECGs (i.e., I, II, III). Since each lead of the 12-lead ECG focuses on the heart from a different direction, ECG waveforms recorded from different leads appear quite different from each other even if the heartbeat they focus on is exactly the same.

Regarding the chest leads (i.e., V1 to V6), this apparent difference is easily observable, especially in the shape of the QRS complexes (Fig. 7(a)). In general, an R wave reaches the maximum at V5, whereas an S wave reaches the maximum at V2. Because these apparent waveform changes are caused by the difference in the electrode placement position in the horizontal plane, the changes in both the R wave and S wave corresponding to the chest leads seem sequential. The height of an R wave and the depth of an S wave become approximately the same in the transitional zone, which is normally located around V3 to V4.

B. MEASURES FOR SIGNAL QUALITY ASSESSMENT

1) SIGNAL-TO-NOISE RATIO (SNR)

We used the signal-to-NOISE ratio (SNR) as the reference index to assess signal quality in experiment 1, in which we have independent data of the signal (i.e., ECG), noise, and artifacts. In this SNR calculation, we regard the QRS complex as the signal. Note that we use the capitalized term "NOISE" to mean components other than heart activity; as described in the manuscript, the lowercase "noise" means the ECG with a low-frequency component (i.e., "baseline wander" in NSTDB [43], [44], [45]) in which we can visually recognize a QRS complex only on the basis of the waveform.

SNR is calculated as the ratio of the amplitude of the signal to the one of NOISE, as

$$SNR = \frac{V_{max}(QRS)}{V_{max}\,(NOISE) - V_{max}(NOISE)}, \tag{B1}$$

where *Vmax(QRS)* stands for the maximum voltage amplitude of the QRS complex among all the QRS complexes obtained from *ECG* (i.e., MITDB [43], [47], [48]), and *Vmax(NOISE)* and *Vmin(NOISE)* stand for the maximum and minimum voltage amplitude of the NOISE, respectively. Hence, the denominator means the difference between the maximum and minimum voltage amplitude of NOISE.

Since NOISE refers to components other than heart activity, we obtain it as

$$NOISE = ECG\,(ISO) + k \times irregularWave. \tag{B2}$$

Comparing (4) to (B2), the only difference is the first term, *ECG(ISO)*, which is calculated here by subtracting the ECGs corresponding to P-QRS-T from the original *ECG* used in (4). On the basis of the clinically known normal duration range in healthy participants [17], we subtracted 0.25 s before the R

wave as the P-R duration and 0.40 s after the R wave as the R-T duration, while assuming that the R wave is located at the center of the QRS complex.

2) SIGNAL-TO-ARTIFACTS RATIO (SAR)

We used the signal-to-ARTIFACTS ratio (SAR) as the index to assess signal quality in experiment 2, in which we cannot necessarily isolate noise and artifacts from the recorded ECGs. In this SAR calculation, the QRS complex is regarded as the signal. Note that we use the capitalized term ''ARTIFACTS'' to mean the possible components other than heart activity; as described in the manuscript, the lowercase ''artifacts'' means the ECG with a high-frequency component in which we cannot recognize a QRS complex only on the basis of the waveform (defined as ''electrode motion artifact'' or ''muscle artifact'' in NSTDB [43], [44], [45]).

As in the study by [58], we calculate SAR as the ratio of the amplitude of the signal to one of ARTIFACTS, as

$$SAR = \frac{V_{max}(TPs)}{V_{max}\,(ARTIFACTS) - V_{max}(ARTIFACTS)}, \quad \text{(B3)}$$

where *Vmax(TPs)* stands for the maximum voltage amplitude of a detected point among all the detected points considered as TPs obtained from *targetECG*. Because FPs are not related to heart activity in theory, using only *Vmax(TPs)* instead of *Vmax(detected points)* can prevent the overestimation of the numerator as well as the underestimation of the denominator. *Vmax(ARTIFACTS)* and *Vmin(ARTIFACTS)* stand for the maximum and minimum voltage amplitude of ARTIFACTS, respectively. Hence, the denominator means the difference between the maximum and minimum voltage amplitude of ARTIFACTS.

Since ARTIFACTS means the possible components other than heart activity indicated as the sequence of TPs, we obtain it as

$$ARTIFACTS = targetECG\,(ISO[TPs])\,. \quad \text{(B4)}$$

TargetECG(ISO[TPs]) is calculated by subtracting the ECGs corresponding to P-QRS-T of the detected points regarded as TPs from the original *targetECG*. The rules for subtracting the P-QRS-T part are the same as those used in the SNR calculation.

In theory, the accuracy of SAR depends on the performance of the algorithm used for the QRS complex detection as well as the determination criteria of TPs. To calculate SAR as appropriately as possible, we used the detected points obtained by SCWF [24], which was confirmed to better suppress FPs due to noise/artifacts compared to PTA [20]. Regarding the determination criteria of TPs, as we did in the manuscript, only detected points classified as ''clean'' are regarded as TPs.

C. SAR VALIDATION TARGETING PSEUDO ECG DATASET

As an initial validation of SAR, we calculated both the SNR and SAR targeting the pseudo ECG dataset used in experiment 1. Table 6 shows all the calculated target indices.

TABLE 6. SNR and SAR calculated in each condition of experiment 1.

	SNR	*SAR*
(i) RAW	3.24	3.24
(ii) BW	0.328	0.376
(iii) EM	0.109	0.124
(iv) MA	0.330	0.594
(v) BW+EM	0.089	0.111
(vi) BW+MA	0.244	0.315
(vii) EM+MA	0.103	0.110
(viii) BW+EM+MA	0.087	0.114

Steel-Dwass tests revealed no significant difference between the two groups ($p = 0.512$).

Before discussing our use of SAR for the signal quality assessment of real ECGs, we would like to clarify its potential issues. Although no significant difference was observed between SNR and SAR, SAR sometimes resulted in obtaining a different value from SNR (e.g., condition (iv) MA). To calculate the SAR more accurately, we need to come up with a new determination criterion for ''artifact'' that can appropriately evaluate muscle artifacts, which do not necessarily have a larger voltage amplitude than the QRS complex. However, ''radio exercise no. 1 [52], [53]'' used in experiment 2 rarely suffered from this issue in consideration of the workout movement: it does not include any workout that involves only isometric contraction of the trunk part. We therefore use the SAR defined in Appendix. B.2 for the signal quality assessment of real ECGs used in experiment 2.

D. SAR OF REAL ECG DATASET

Table 7 shows all the SAR values calculated from the segmented real ECGs. Steel-Dwass tests revealed there were significant differences between participants 1 and 3 ($p = 0.001$) and participants 2 and 3 ($p = 0.001$), and no significant difference between participants 1 and 2 ($p = 0.9$). Among the three, participant 3 was the most stable from start to finish. Conversely, participant 1 was the most unstable: although he had the best signal quality (SAR = 20.6) in workout no. 1, he also had the worst one (SAR = 0.057) in workout no. 8.

As for the SAR in each workout, Friedman tests revealed no significant difference among the 13 workouts ($p = 0.283$). However, the average SAR was lower than 1.00 in workout nos. 6, 7, 8, 10, and 11, which indicates that artifacts were larger than the signal (i.e., the voltage amplitude of the QRS complex). The electrodes of the single-channel shirt-type wearable ECG device we used were located around the pit of the stomach, which means it was probably susceptible to the trunk movement: e.g., bending back and forth (workout no. 6), twisting (workout nos. 7 and 10), or movement that may induce slipping up of the shirt (workout nos. 8 and 11).

Since the signal quality of this real ECG dataset varied in each participant as well as in each workout, we were able to

TABLE 7. SAR during radio exercise no. 1.

Workout no.	*Participant 1*	*Participant 2*	*Participant 3*	*Average*	*Standard deviation*
1	20.6	8.00	1.21	9.94	8.03
2	0.733	0.410	3.46	1.53	1.37
3	0.408	0.217	4.21	1.61	1.84
4	0.276	0.195	2.62	1.03	1.12
5	0.322	0.393	5.00	1.91	2.19
6	0.432	0.537	0.778	0.582	0.145
7	0.781	0.269	0.987	0.679	0.302
8	0.057	0.224	1.34	0.540	0.570
9	0.271	0.458	2.44	1.056	0.981
10	0.186	0.087	0.682	0.918	0.260
11	0.381	0.195	1.02	0.532	0.353
12	0.319	0.569	2.12	1.00	0.797
13	0.180	0.972	2.98	1.378	1.18
Average	1.92	0.963	2.22		
Standard deviation	5.38	2.04	1.34		

validate the performance of the proposed method assuming a situation targeting an ECG recorded by a wearable ECG device under the daily life environment.

E. AUXILIARY EVALUATION ON FP IDENTIFICATION

As we confirmed in the experimental results of the RRI sequence and HRV features, the performance of the *practical* FP identification proposed in this study could not reach the same level as its *theoretical* performance. Although the FP identification performance evaluation was not our main focus, we briefly summarize its current performance here.

From the perspective of FP identification (i.e., misdetected QRS complex identification), the correspondence of each class used in each formula becomes as follows: TP, correctly identified misdetected QRS complex; TN, correctly overlooked QRS complex (i.e., accurately detected QRS complex); FP, falsely misidentified point as misdetected QRS complex; and FN, falsely overlooked misdetected QRS complex.

To evaluate the performance of FP identification in terms of misdetection and overlook regardless of the number of target FPs, this evaluation uses the following five measures: F1 score, precision, recall, and specificity [34], [35], as well as Matthew's correlation coefficient (MCC) [59]. Each measure is calculated from the following formula.

- *Precision (aka positive prediction value)*: the measure that quantifies how well the evaluation target (e.g., model, algorithm) avoids false positives.

$$Precision = \frac{TP}{TP + FP}$$

- *Recall (aka positive rate, sensitivity)*: the measure that quantifies how well the evaluation target (e.g., model, algorithm) avoids false negatives.

$$Recall = \frac{TP}{TP + FN}$$

- *F1 score*: Harmonic mean of the precision and recall measures into a single score [34], [35].

$$F1score = \frac{2 \times Recall \times Precision}{Recall + Precision} = \frac{2TP}{2TP + FN + FP}$$

- *Specificity (aka true negative rate)*: the measure that quantifies how well the evaluation target (e.g., model, algorithm) correctly retains true negatives as they are while avoiding false positives.

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

- *MCC*: the measure for binary classification that generates a high score only if the binary predictor is able to correctly predict the majority of positive data instances and the majority of negative data instances. Unlike other target measures, the value range of MCC is from –1 to 1, where –1 indicates complete mismatch, 1 indicates complete match, and 0 indicates similar to at random.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

1) EXPERIMENT 1: PSEUDO ECG DATASET

Tables 8 and 9 list the results obtained from PTA and from SCWF, respectively. In both tables, N/A indicates target measures that could not be calculated mainly because of the FP overlook. Regardless of the algorithm, we confirmed that there were no detected points classified as ''artifacts.'' All five evaluation measures thus indicate the performance of the dubious QRS complex identification by the fail-safe approach.

Overall, the target measures indicate that the proposed practical FP identification performed better in PTA than in SCWF, possibly because PTA had a greater number of target FPs. Regardless of the algorithm, the proposed practical FP

TABLE 8. FP identification performance in each condition of experiment 1 (PTA).

	Target FP	*F1 score*	*Precision*	*Recall*	*Specificity*	*MCC*
(i) RAW	0	N/A	N/A	N/A	1.00	N/A
(ii) BW	0	N/A	0	N/A	0.932	N/A
(iii) EM	15	0.667	0.611	0.733	0.879	0.574
(iv) MA	8	N/A	N/A	0	1.00	N/A
(v) BW+EM	14	0.800	0.750	0.857	0.931	0.750
(vi) BW+MA	9	0.526	0.500	0.556	0.930	0.464
(vii) EM+MA	17	0.811	0.750	0.882	0.902	0.745
(viii) BW+EM+MA	16	0.824	0.778	0.875	0.922	0.766

TABLE 9. FP identification performance in each condition of experiment 1 (SCWF).

	Target FP	*F1 score*	*Precision*	*Recall*	*Specificity*	*MCC*
(i) RAW	0	N/A	N/A	N/A	1.00	N/A
(ii) BW	0	N/A	0	N/A	0.932	N/A
(iii) EM	3	0.333	0.222	0.667	0.901	0.343
(iv) MA	4	N/A	N/A	0	1.00	N/A
(v) BW+EM	3	0.308	0.200	0.667	0.887	0.320
(vi) BW+MA	4	0.545	0.429	0.750	0.943	0.535
(vii) EM+MA	10	0.857	0.818	0.900	0.97	0.835
(viii) BW+EM+MA	4	0.462	0.333	0.750	0.914	0.460

TABLE 10. FP identification performance in each condition of experiment 2 (PTA).

	Target FP	*F1 score*		*Precision*		*Recall*		*Specificity*		*MCC*	
		p1	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*
Participant 1	47	0.634	0.211	0.513	0.600	0.830	0.128	0.837	0.982	0.561	0.221
Participant 2	86	0.708	N/A	0.633	0	0.802	0	0.861	0.993	0.614	–0.040
Participant 3	4	0.308	N/A	0.222	N/A	0.500	0	0.978	1.00	0.321	N/A

TABLE 11. FP identification performance in each condition of experiment 2 (SCWF).

	Target FP	*F1 score*		*Precision*		*Recall*		*Specificity*		*MCC*	
		p1	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*	*p1*	*p2*
Participant 1	0	N/A	N/A	0	0	N/A	N/A	0.826	0.971	N/A	N/A
Participant 2	12	0.246	N/A	0.156	0	0.583	0	0.881	0.997	0.253	–0.01
Participant 3	1	N/A	N/A	0	N/A	0	0	0.978	1.00	–0.008	N/A

identification could not identify FPs in condition (iv) MA: since MA does not necessarily cause a change in amplitude, our proposed method utilizing the amplitude of each detected point might overlook FPs.

2) EXPERIMENT 2: REAL ECG DATASET

Tables 10 and 11 list the results obtained from PTA and from SCWF, respectively. In both tables, N/A indicates target measures that could not be calculated mainly because of the FP overlook. In contrast to when the pseudo ECG dataset was targeted, we confirmed here that several detected points were classified as ''artifacts.'' All five evaluation measures are thus independently calculated from the dubious QRS complex identification by the fail-safe approach (*p1*) and by the fail-soft approach (*p2*).

As we confirmed in experiment 1, overall, the target measures indicate that the proposed practical FP identification performed better in PTA than in SCWF, presumably for the same reason as in experiment 1 (i.e., the number of target FPs). Comparing the two practical FP identification methods (i.e., condition *p1* by the fail-safe approach and condition *p2* by the fail-soft approach) based on specificity, regardless of the algorithm, the fail-soft approach performed better than the fail-safe approach.

REFERENCES

[1] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, ''Heart rate variability: Standards of measurement, physiological interpretation, and clinical use,'' *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, Mar. 1996, doi: 10.1093/oxfordjournals.eurheartj.a014868.

[2] M. A. Peltola, ''Role of editing of R–R intervals in the analysis of heart rate variability,'' *Frontiers Physiol.*, vol. 3, pp. 1–10, May 2012, doi: 10.3389/fphys.2012.00148.

[3] A. Iwasaki, C. Nakayama, K. Fujiwara, Y. Sumi, M. Matsuo, M. Kano, and H. Kadotani, ''Screening of sleep apnea based on heart rate variability and long short-term memory,'' *Sleep Breathing*, vol. 25, no. 4, pp. 1821–1829, Jan. 2021, doi: 10.1007/s11325-020-02249-0.

[4] M. Xiao, H. Yan, J. Song, Y. Yang, and X. Yang, ''Sleep stages classification based on heart rate variability and random forest,'' *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 624–633, 2013, doi: 10.1016/j.bspc.2013.06.001.

[5] T. Takeda, O. Mizuno, and T. Tanaka, "Time-dependent sleep stage transition model based on heart rate variability," in *Proc. EMBC*, Milan, Italy, Aug. 2015, pp. 2343–2346, doi: 10.1109/EMBC.2015.7318863.

[6] K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, M. Matsuo, and H. Kadotani, "Heart rate variability-based driver drowsiness detection and its validation with EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1769–1778, Jun. 2019, doi: 10.1109/TBME.2018.2879346.

[7] S. Hamada, K. Sasaki, H. Kito, Y. Tooyama, K. Ihara, E. Aoyagi, N. Ichimura, S. Tohda, and T. Sasano, "Effect of the recording condition on the quality of a single-lead electrocardiogram," *Heart Vessels*, vol. 37, no. 6, pp. 1010–1026, Jun. 2022, doi: 10.1007/s00380-021-01991-z.

[8] T. Takagahara, K. Ono, N. Oda, and T. Teshigawara, "'hitoe'—A wearable sensor developed through cross-industrial collaboration," *NTT Tech. Rev.*, vol. 12, no. 9, pp. 1–5, 2014. [Online]. Available: https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201409ra1.html

[9] Y. T. Tsukada, M. Tokita, H. Murata, Y. Hirasawa, K. Yodogawa, Y.-K. Iwasaki, K. Asai, W. Shimizu, N. Kasai, H. Nakashima, and S. Tsukada, "Validation of wearable textile electrodes for ECG monitoring," *Heart Vessels*, vol. 34, no. 7, pp. 1203–1211, Jan. 2019, doi: 10.1007/s00380-019-01347-8.

[10] N. Shiozawa, J. Lee, A. Okuno, and M. Makikawa, "Novel under wear 'smart-wear' with stretchable and flexible electrodes enables insensible monitoring electrocardiograph," in *Proc. World Eng. Conf. Conv.*, vol. 3, Kyoto, Japan, Nov./Dec. 2015, pp. 1–2.

[11] J. M. Raja, C. Elsakr, S. Roman, B. Cave, I. Pour-Ghaz, A. Nanda, M. Maturana, and R. N. Khouzam, "Apple watch, wearables, and heart rhythm: Where do we stand?" *Ann. Transl. Med.*, vol. 7, no. 17, p. 417, Sep. 2019, doi: 10.21037/atm.2019.06.79.

[12] T. Kondo, Y. Yamato, M. Nakayama, A. Chiba, K. Sakaguchi, T. Nishiguchi, T. Masuda, and T. Yoshida, "Natural Sensing with 'hitoe' functional material and initiatives towards its applications," *NTT Tech. Rev.*, vol. 15, no. 9, pp. 1–8, Sep. 2017. Accessed: Feb. 7, 2023. [Online]. Available: https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201709fa3.html

[13] Y. Okubo, T. Tokuyama, S. Okamura, Y. Ikeguchi, S. Miyauchi, and Y. Nakano, "Evaluation of the feasibility and efficacy of a novel device for screening silent atrial fibrillation (MYBEAT trial)," *Circulat. J.*, vol. 86, no. 2, pp. 182–188, Jan. 2022, doi: 10.1253/circj.CJ-20-1061.

[14] G. M. Friesen, T. C. Jannett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 1, pp. 85–98, Jan. 1990, doi: 10.1109/10.43620.

[15] T. Alkhidir, A. Sluzek, and M. K. Yapici, "Simple method for adaptive filtering of motion artifacts in E-textile wearable ECG sensors," in *Proc. EMBC*, Milan, Italy, Aug. 2015, pp. 3807–3810, doi: 10.1109/EMBC.2015.7319223.

[16] K. Eguchi, R. Aoki, S. Shimauchi, K. Yoshida, and T. Yamada, "R-R interval outlier processing for heart rate variability analysis using wearable ECG devices," *Adv. Biomed. Eng.*, vol. 7, pp. 28–38, Feb. 2018, doi: 10.14326/abe.7.28.

[17] S. Watanabe and I. Yamaguchi, Eds., *Shinden-Zu No Yomikata Perfect Manual [ECG Perfect Manual]*, (in Japanese). Tokyo, Japan: Yodosha Co., Ltd., 2006.

[18] J. Okude, *Korenara Wakaru! Kantan Piont Shinden-Zu [Easy-to-Understand Tutorial On Electrocardiogram]*, (in Japanese), 2nd ed. Tokyo, Japan: Igaku-Shoin, 2011.

[19] B.-U. Kohler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 1, pp. 42–57, Jan./Feb. 2002, doi: 10.1109/51.993193.

[20] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985, doi: 10.1109/TBME.1985.325532.

[21] M. Elgendi, B. Eskofier, S. Dokos, and D. Abbott, "Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e84018, doi: 10.1371/journal.pone.0084018.

[22] P. S. Addison, "Wavelet transforms and the ECG: A review," *Physiol. Meas.*, vol. 26, no. 5, pp. R155–R199, Oct. 2005, doi: 10.1088/0967-3334/26/5/R01.

[23] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ECG delineator: Evaluation on standard databases," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 570–581, Apr. 2004, doi: 10.1109/TBME.2003.821031.

[24] S. Shimauchi, K. Eguchi, R. Aoki, M. Fukui, and N. Harada, "R-R interval estimation for wearable electrocardiogram based on single complex wavelet filtering and morphology-based peak selection," *IEEE Access*, vol. 9, pp. 60802–60827, 2021, doi: 10.1109/ACCESS.2021.3070604.

[25] M. Jia, F. Li, J. Wu, Z. Chen, and Y. Pu, "Robust QRS detection using high-resolution wavelet packet decomposition and time-attention convolutional neural network," *IEEE Access*, vol. 8, pp. 16979–16988, 2020, doi: 10.1109/ACCESS.2020.2967775.

[26] J. McNames, T. Thong, and M. Aboy, "Impulse rejection filter for artifact removal in spectral analysis of biomedical signals," in *Proc. IEEE EMBC*, San Francisco, CA, USA, Sep. 2004, pp. 145–148, doi: 10.1109/IEMBS.2004.1403112.

[27] S. Miyatani, K. Fujiwara, and M. Kano, "Deniosing autoencoder-based modification of RRI data with premature ventricular contraction for precise heart rate variability analysis," in *Proc. IEEE EMBC*, Honolulu, HI, USA, Jul. 2018, pp. 5018–5021, doi: 10.1109/EMBC.2018.8513218.

[28] C. Liu, L. Li, L. Zhao, D. Zheng, P. Li, and C. Liu, "A combination method of improved impulse rejection filter and template matching for identification of anomalous intervals in RR sequences," *J. Med. Biol. Eng.*, vol. 32, no. 4, pp. 245–250, 2012, doi: 10.5405/jmbe.1006.

[29] N. Lippman, K. M. Stein, and B. B. Lerman, "Comparison of methods for removal of ectopy in measurement of heart rate variability," *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 267, no. 1, pp. H411–H418, Jul. 1994, doi: 10.1152/ajpheart.1994.267.1.H411.

[30] G. D. Clifford, "Signal processing methods for heart rate variability," Ph.D. dissertation, Dept. Eng. Sci., St. Cross College, Univ. Oxford, Oxford, U.K., 2002.

[31] K. Kamata, K. Kinoshita, and M. Kano, "Missing RRI interpolation algorithm based on locally weighted partial least squares for precise heart rate variability analysis," *Sensors*, vol. 18, no. 11, p. 3870, Nov. 2018, doi: 10.3390/s18113870.

[32] R. Aoki, K. Eguchi, S. Shimauchi, K. Yoshida, and T. Yamada, "Consideration of calculation process assuming heart rate variability analysis using wearable ECG devices," in *Proc. IEEE EMBC*, Honolulu, HI, USA, Jul. 2018, pp. 5693–5696, doi: 10.1109/EMBC.2018.8513449.

[33] K. Eguchi, R. Aoki, K. Yoshida, and T. Yamada, "Reliability evaluation of R-R interval measurement status for time domain heart rate variability analysis with wearable ECG devices," in *Proc. IEEE EMBC*, Seogwipo, South Korea, Jul. 2017, pp. 1307–1311, doi: 10.1109/EMBC.2017.8037072.

[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[35] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*. Newton, MA, USA: O'Reilly Media, 2017.

[36] M. Malik, R. Xia, O. Odemuyiwa, A. Staunton, J. Poloniecki, and A. J. Camm, "Influence of the recognition artefact in automatic analysis of long-term electrocardiograms on time-domain measurement of heart rate variability," *Med. Biol. Eng. Comput.*, vol. 31, no. 5, pp. 539–544, Sep. 1993, doi: 10.1007/BF02441992.

[37] R. J. Ellis, B. Zhu, J. Koenig, J. F. Thayer, and Y. Wang, "A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability," *Physiol. Meas.*, vol. 36, no. 9, pp. 1827–1852, Sep. 2015, doi: 10.1088/0967-3334/36/9/1827.

[38] F. C. Messineo, "Ventricular ectopic activity: Prevalence and risk," *Amer. J. Cardiol.*, vol. 64, no. 20, pp. J53–J56, Dec. 1989, doi: 10.1016/0002-9149(89)91200-9.

[39] M. Toichi, T. Sugiura, T. Murai, and A. Sengoku, "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R–R interval," *J Auton Nerv Syst.*, vol. 62, nos. 1–2, pp. 79–84, Jan. 1997, doi: 10.1016/s0165-1838(96)00112-9.

[40] G. E. Billman, "Heart rate variability—A historical perspective," *Frontiers Physiol.*, vol. 2, no. 86, pp. 1–13, Nov. 2011, doi: 10.3389/fphys.2011.00086.

[41] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability?" *IEEE Trans. Biomed. Eng.*, vol. 48, no. 11, pp. 1342–1347, Nov. 2001, doi: 10.1109/10.959330.

[42] V. Oliveira, W. von Rosenberg, P. Montaldo, T. Adjei, J. Mendoza, V. Shivamurthappa, D. Mandic, and S. Thayyil, "Early postnatal heart rate variability in healthy newborn infants," *Frontiers Physiol.*, vol. 10, pp. 1–12, Aug. 2019, doi: 10.3389/fphys.2019.00922.

[43] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.
[44] G. B. Moody, W. K. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Comput. Cardiol.*, vol. 11, no. 3, pp. 381–384, 1984. [Online]. Available: http://ecg.mit.edu/george/publications/nst-cinc-1984.pdf
[45] PhysioNet. *The MIT-BIH Noise Stress Test Database*. Accessed: May 9, 2022. [Online]. Available: https://archive.physionet.org/physiobank/database/nstdb/
[46] M. Sokolow and T. P. Lyon, "The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads," *Amer. Heart J.*, vol. 37, no. 2, pp. 161–186, Feb. 1949, doi: 10.1016/0002-8703(49)90562-1.
[47] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May/Jun. 2001, doi: 10.1109/51.932724.
[48] PhysioNet. *MIT-BIH Arrhythmia Database*. Accessed: May 9, 2022. [Online]. Available: https://archive.physionet.org/physiobank/database/mitdb/
[49] H. Sedghamiz, "MATLAB implementation of Pan Tompkins ECG QRS detector," 2014. Accessed: Mar. 10, 2023, doi: 10.13140/RG.2.2.14202.59841.
[50] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-aho, and P. A. Karjalainen, "Kubios HRV—Heart rate variability analysis software," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 210–220, Jan. 2014, doi: 10.1016/j.cmpb.2013.07.024.
[51] PhysioNet. *PhysioBank Annotations*. Accessed: May 9, 2022. [Online]. Available: https://archive.physionet.org/physiobank/annotations.shtml
[52] Japan Post Insurance. *Radio Exercise No. 1 (in Japanese)*. Accessed: May 9, 2022. [Online]. Available: https://www.jp-life.japanpost.jp/radio/instruction/radio_first.html
[53] Japan Post Insurance. *Radio Exercise No. 1 (in Japanese)*. Accessed: May 9, 2022. [Online]. Available: https://www.youtube.com/watch?v=_YZZfaMGEOU
[54] S. Tsukada, H. Nakashima, and K. Torimitsu, "Conductive polymer combined silk fiber bundle for bioelectrical signal recording," *PLoS ONE*, vol. 7, no. 4, Apr. 2012, Art. no. e33689, doi: 10.1371/journal.pone.0033689.
[55] H. Blackburn, H. L. Taylor, C. L. Vasquez, and T. C. Puchner, "The electrocardiogram during exercise: Findings in bipolar chest leads of 1,449 middle-aged men, at moderate work levels," *Circulation*, vol. 34, no. 6, pp. 1034–1043, Dec. 1966, doi: 10.1161/01.cir.34.6.1034.
[56] K. Eguchi, R. Aoki, K. Yoshida, and T. Yamada, "R-R interval outlier exclusion method based on statistical ECG values targeting HRV analysis using wearable ECG devices," in *Proc. IEEE EMBC*, Honolulu, HI, USA, Jul. 2018, pp. 5689–5692, doi: 10.1109/EMBC.2018.8513452.
[57] B. M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M. F. Abbod, and J.-S. Shieh, "ECG arrhythmia classification by using a recurrence plot and convolutional neural network," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102262, doi: 10.1016/j.bspc.2020.102262.
[58] N. Meziane, S. Yang, M. Shokoueinejad, J. G. Webster, M. Attari, and H. Eren, "Simultaneous comparison of 1 gel with 4 dry electrode types for electrocardiography," *Physiol. Meas.*, vol. 36, no. 3, pp. 513–529, Mar. 2015, doi: 10.1088/0967-3334/36/3/513.
[59] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

KANA EGUCHI (Senior Member, IEEE) received the B.E. degree from Kyoto Institute of Technology, Kyoto, Japan, in 2010, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, in 2012 and 2020, respectively.

She joined Nippon Telegraph and Telephone (NTT) Corporation, Tokyo, Japan, in 2012. Since July 2022, she has been a Program-Specific Assistant Professor with the Department of Real World Data Research and Development, Graduate School of Medicine, Kyoto University. Her current research interests include biosignal processing, wearable sensing systems, medical engineering, and medical informatics.

Dr. Eguchi is a member of the Institute of Electronics, Information, and Communication Engineers of Japan (IEICE), the IEEE Engineering in Medicine and Biology Society (EMBS), the Japanese Society for Medical and Biological Engineering (JSMBE), the Japanese Society of Sleep Research (JSSR), the Japan Association for Medical Informatics (JAMI), and the Human Interface Society in Japan. Her awards include the Telecom System Technology Award for Students (Honorable Mention) from the Telecommunications Advancement Foundation, in 2019.

RYOSUKE AOKI received the B.E., M.S., and Ph.D. degrees in information sciences from Tohoku University, Miyagi, Japan, in 2005, 2007, and 2014, respectively.

He joined Nippon Telegraph and Telephone (NTT) Corporation, Tokyo, Japan, in 2007. His current research interests include human–computer interaction, interaction design, research through design, medical engineering, human augmentation, and wearable/ubiquitous computing.

Dr. Aoki is a member of the Information Processing Society of Japan (IPSJ).