Reference Database and Performance Evaluation of Methods for Extraction of Atrial Fibrillatory Waves in the ECG

Raúl Alcaraz[1]**, Leif Sörnmo**[2]**, José J. Rieta**[3]

[1]Research Group in Electronic, Biomedical and Telecommunication Engineering, University of Castilla-La Mancha, Cuenca, Spain.
[2]Department of Biomedical Engineering and Center of Integrative Electrocardiology, Lund University, Lund, Sweden
[3]BioMIT.org, Electronic Engineering Department, Universitat Politecnica de Valencia, Valencia, Spain

E-mail: `raul.alcaraz@uclm.es`

Abstract. *Objective*: This study proposes a reference database, composed of a large number of simulated ECG signals in atrial fibrillation (AF), for investigating the performance of methods for extraction of atrial fibrillatory waves (f-waves). *Approach*: The simulated signals are produced using a recently published and validated model of 12-lead ECGs in AF. The database is composed of eight signal sets which together account for a wide range of characteristics known to represent major challenges in f-wave extraction, including high heart rates, high morphological QRST variability, and the presence of ventricular premature beats. Each set contains 30 5-min signals with different f-wave amplitudes. The database is used for the purpose of investigating the statistical association between different indices, designed either for use with real or simulated signals. *Main results*: Using the database, which is made available at the Physionet repository of physiological signals, the performance indices *unnormalized ventricular residue* (uVR), designed for real signals, and the *root mean square error*, designed for simulated signals, were found to exhibit the strongest association, leading to the recommendation that uVR should be used when characterizing performance in real signals. *Significance*: The proposed database facilitates the comparison of performance of different f-wave extraction methods and makes it possible to express performance in terms of the error between simulated and extracted f-wave signals.

Submitted to: *Physiol. Meas.*

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia affecting roughly 33.5 million people worldwide (Chugh, Roth, Gillum & Mensah 2014). Since the prevalence of AF is closely related to aging, it is expected to increase significantly in coming years, likely to reach epidemic proportions by the middle of this century (Krijthe, Kunst, Benjamin, Lip, Franco, Hofman, Witteman, Stricker & Heeringa 2013, Colilla, Crow, Petkun, Singer, Simon & Liu 2013). The arrhythmia has a large effect on morbidity and mortality (Fauchier, Villejoubert, Clementy, Bernard, Pierre, Angoulvant, Ivanes, Babuty & Lip 2016) as it may provoke hemodynamic alterations, which in turn predispose the occurrence of cerebrovascular accidents (Sposato, Cipriano, Saposnik, Ruíz Vargas, Riccio & Hachinski 2015).

Since the pathophysiological mechanisms causing and maintaining AF are still not completely understood (Schotten, Dobrev, Platonov, Kottkamp & Hindricks 2016), various types of noninvasive techniques have been developed for better understanding of the mechanisms. With respect to the ECG, being the most well-established noninvasive technique, it is only until fairly recently that attention has been directed towards the atrial fibrillatory waves (f-waves) and their characterization (Sörnmo, Alcaraz, Laguna & Rieta 2018). For example, the f-wave repetition rate, i.e., the dominant atrial frequency (DAF), has proven useful for monitoring drug therapy and predicting spontaneous and drug-induced AF termination (Platonov, Corino, Seifert, Holmqvist & Sörnmo 2014). More advanced approaches to characterizing f-wave morphology, e.g., based on nonlinear analysis and entropy measures (Alcaraz & Rieta 2017), can help select patients who would respond to therapies such as electrical cardioversion, catheter ablation, and Maze surgery. In addition, f-wave representation by image mapping techniques has proven useful for noninvasive localization of focal impulse sources of AF (Cuculich, Wang, Lindsay, Faddis, Schuessler, Damiano, Li & Rudy 2010).

Since the QRST complexes overshadow the f waves, it is highly desirable to extract a connected f-wave signal containing more samples for characterization than what is contained in the TQ intervals, especially in short RR intervals which are rather frequent in AF (Sörnmo, Petrėnas, Laguna & Marozas 2018). Different signal processing principles have been explored for f-wave extraction which may be categorized as follows: average beat subtraction and variants (Stridh & Sörnmo 2001, Lemay, Vesin, van Oosterom, Jacquemet & Kappenberger 2007, Dai, Jiang & Li 2013), principal component analysis (Castells, Mora, Rieta, Moratal-Pérez & Millet 2005, Langley, Stridh, Rieta, Millet, Sörnmo & Murray 2006, Alcaraz & Rieta 2008), interpolation and singular spectral analysis (Sassi, Corino & Mainardi 2009), independent component analysis (Rieta, Castells, Sánchez, Zarzoso & Millet 2004, Phlypo, Zarzoso & Lemahieu 2010), adaptive filtering using an echo state network (Petrėnas, Marozas, Sörnmo & Lukosevicius 2012), extended Kalman filtering (Kheirati Roonizi & Sassi 2017), and diffusion geometry (Malik, Reed, Wang & Wu 2017).

The performance of a method for f-wave extraction is influenced by several factors of which the number of leads is important as more leads should provide better immunity to respiratory-induced modulation of the QRS amplitude; if desired, multiple leads can be used

to produce a global f-wave signal. Certain methods need to learn the signal characteristics before extraction can take place, and, therefore, the minimum signal duration is another important factor. Ventricular premature beats (VPBs) represent a major challenge since such beats often have morphologies that deviate considerably from what is observed in beats originating from the atria. The presence of noise and artifacts is yet another factor which imposes a limitation on performance. In addition, it should be noted that certain methods require that AF detection is first performed since f-waves need to be present if the observed signal is to be properly processed; if sinus rhythm is present, it is often unclear whether the extracted signal will contain P waves.

While performance evaluation of AF detectors can rely on publicly available, annotated ECG databases, this observation does not apply to performance evaluation of methods for f-wave extraction. One important reason to this important difference is that expert annotations cannot be easily related to the accuracy of the extracted signal. As a consequence, a wide variety of proprietary databases have been analyzed to assess the quality of extracted signals, where use is made of an equally wide variety of performance indices. Under these circumstances, it is exceedingly difficult to compare the performance of different extraction methods.

In an attempt to mitigate this limitation, the present study proposes a reference ECG database for evaluating and comparing the performance of f-wave extraction methods, consisting of a large set of simulated signals with widely different characteristics.‡ The database makes it possible to compute performance indices such as the sample-to-sample error between simulated and extracted f-wave signals. Even with the availability of a reference database, other datasets with real ECGs may nonetheless have to be considered to shed light on extraction performance when certain specific signal characteristics are present. To quantify performance in such situations, the reference database can be used to investigate the statistical association between indices designed for real signals and indices designed for simulated signals, with the aim of finding an index for real signals which convey information similar to that conveyed by any of the indices for simulated signals. In the present study, it is hypothesized that such an index is well-suited for performance evaluation based on real ECG databases.

The manuscript is organized as follows. Section 2 describes the composition of the proposed reference database with simulated ECGs, where emphasis is put on the characteristics of the f-waves and the QRST complexes. Section 3 provides a survey of performance indices, designed either for use with simulated or real signals. Using the reference database, Section 4 presents a statistical analysis of the performance indices, leading up to an index recommended for use with real signals. Various aspects of the database and performance indices are discussed in Section 5, and some concluding remarks are provided in Section 6.

‡ The proposed reference database will be made available at the Physionet data repository (Goldberger, Amaral, Glass, Hausdorff, Ivanov, Mark, Mietus, Moody, Peng & Stanley 2000).

2. Materials

A reference database composed of simulated ECG signals is created using the model recently proposed in (Petrėnas, Marozas, Sološenko, Kubilius, Skibarkiene, Oster & Sörnmo 2017), offering a large number of choices such as type of f-wave (synthetic or real), type of QRST complex (synthetic or real), type of RR-interval series (synthetic or real), and type of noise and artifacts. The reference database contains 12-lead signals, however, in the present study, only lead V_1 is considered as it typically exhibits the largest f-wave amplitude. Moreover, all simulated signals are in AF since the aim is to evaluate f-wave extraction performance, not P wave extraction. Consequently, the option to switch between sinus rhythm and AF is disabled in the simulation model, as is also the change in DAF occurring during the first minutes after the onset of AF and the last minute before spontaneous termination (Bollmann, Sonne, Esperer, Toepffer, Langberg & Klein 1999, Ravelli, Masè, Greco, Faes & Disertori 2007, Alcaraz & Rieta 2009).

Real RR-interval series, extracted from ECGs in the Physionet Long-Term Atrial Fibrillation Database, are used to produce signals in which the heart rate ranges from 80 to 150 beats per minute (bpm). Synthetic RR-intervals are not considered.

Synthetic f-waves are generated using an extended version of the sawtooth model (Stridh & Sörnmo 2001), whereas the number of harmonics and the DAF are treated as random variables (the number of harmonics can take the values 1, 2, or 3 with equal probability, and the DAF is uniformly distributed in the interval $[5, 12]$ Hz). Using an echo state network (Petrėnas et al. 2012), *real f-waves* have been extracted from a proprietary database of ECGs obtained from patients clinically diagnosed with either paroxysmal or persistent AF. For both synthetic and real f-waves, the amplitude is expressed as a root mean square (RMS) value, set to either 5, 10, 20, 30, 40, or 50 μV.

Synthetic QRST complexes are generated using an extended version of the single-dipole model described in (Sameni, Clifford, Jutten & Shamsollahi 2007), where morphological variability is introduced by treating the model parameters as uniformly distributed random variables. *Real QRST complexes* have been selected from the PTB Database (Bousseljot, Kreiseler & Schnabel 1995), mostly composed of ECGs in sinus rhythm with moderate morphological variability. However, it is well-known that f-wave extraction becomes increasingly more difficult to perform as the variability becomes increasingly more pronounced; therefore, the simulation model is complemented by a component labeled *real QRST complexes with high morphological variability* (HMV). Since such QRST complexes are often observed just before transitions from sinus rhythm to AF, QRST complexes are selected from such time segments in patients with paroxysmal AF and introduced in the model in the same way as synthetic and real QRST complexes. Figure 1 illustrates simulated signals with different degrees of morphological variability.

Since the morphology of VPBs typically differs considerably from that of beats originating from the atria, the simulation model is complemented with the option to insert *real VPBs*. These beats are selected from episodes of sinus rhythm in patients with paroxysmal AF, and introduced in the model according to the following rule: the length of the coupling interval
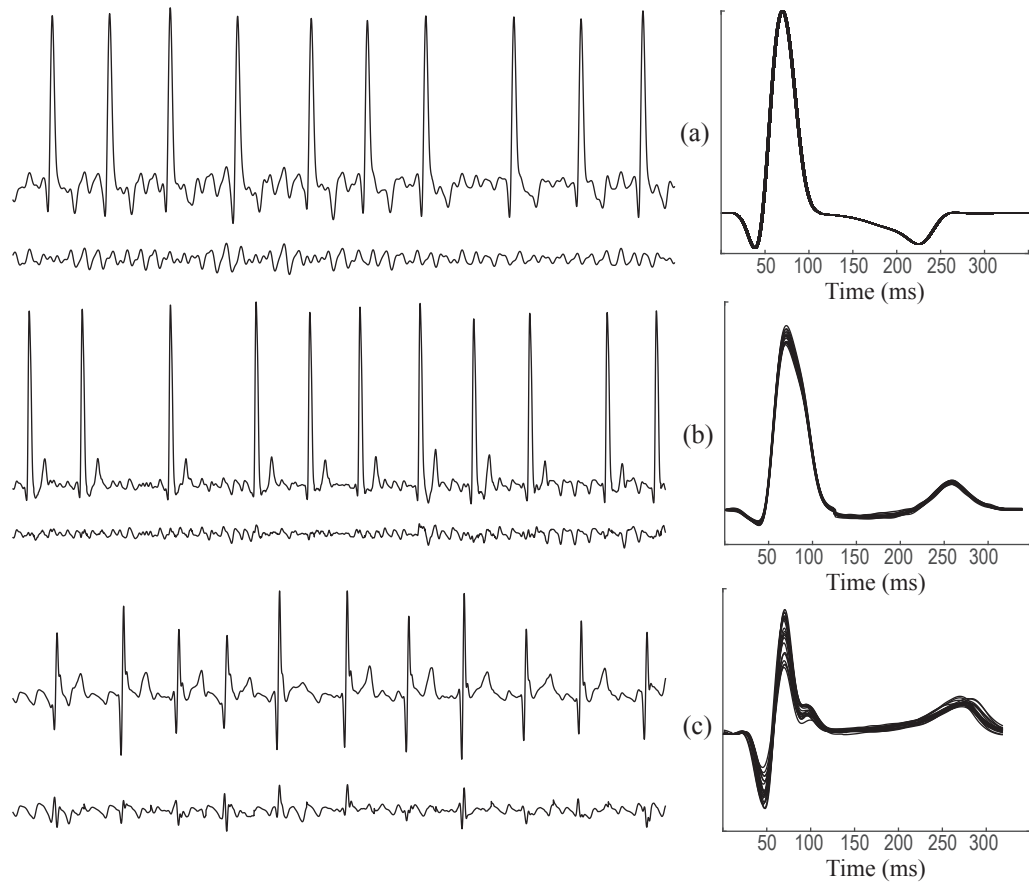
Figure 1. Simulated ECG signals in AF (left column) composed of (a) synthetic f-waves and QRST complexes, (b) real f-waves and synthetic QRST complexes, and (c) real f-waves and real QRST complexes with high morphological variability. The right column displays superimposed QRST complexes.
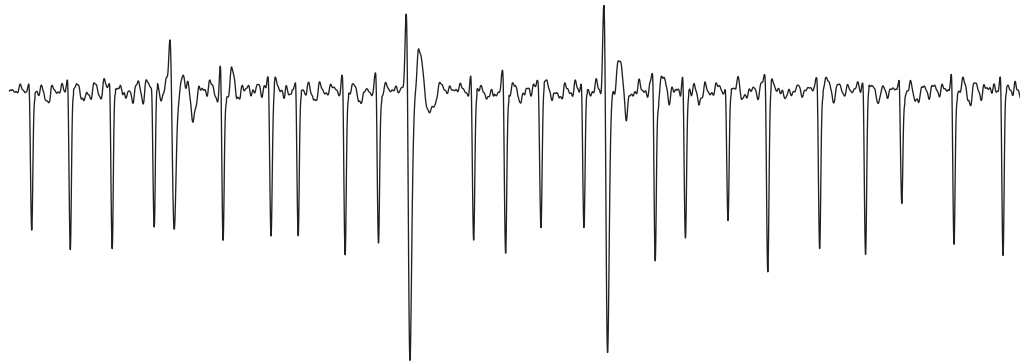
Figure 2. Simulated 10 second-length ECG containing ventricular premature beats.

is 70% of the mean RR interval of the five preceding normal beats, whereas the compensatory pause is about twice the coupling interval. The percentage of VPBs in a simulated signal is set to 5%. Figure 2 illustrates a simulated ECG signal containing VPBs.

Eight sets of signals (labeled S1–S8) are defined by different combinations of the above-

Table 1. Composition of sets with simulated signals used for evaluating f-wave extraction performance.

Set	Real f-waves	Synthetic f-waves	Real QRST complexes	Synthetic QRST complexes	Real QRST complexes with HMV	Real VPBs
S1	×		×			
S2		×	×			
S3	×			×		
S4		×		×		
S5	×				×	
S6		×			×	
S7	×				×	×
S8		×			×	×

mentioned components, see Table 1. In each set, five different 5-min signals are simulated for each of the six f-wave amplitudes, and thus a total of 30 signals. It should noted that the reference database is noise-free, except for the noise which is present in real signals; see Section 5 for a discussion on this matter.

3. Methods

Most indices quantify extraction performance by exploring the properties of the extracted f-wave signal in the QT interval because f-wave distortion is largely confined to this interval. Hence, it is assumed that the observed signal $x(n), n = 0, \ldots, N_{QT} - 1$, whether simulated or real in nature, contains the samples in the interval enclosed by the onset of the Q wave and the end of the T wave. Since wave delineation is extremely difficult to perform accurately in the presence of f-waves, see, e.g., (Dai, Yin & Li 2016), the onset and end of the QT interval are related to the R wave position by fixed distances. The simulated and extracted f-wave signals are denoted $s(n)$ and $\hat{s}(n)$, respectively, and indexed in the same way as $x(n)$. All signals are assumed to be zero-mean, achieved by linear, time-invariant high-pass filtering to remove baseline wander.

For simplicity, beat indexing is omitted in the description below of performance indices. It is straightforward to compute an overall performance figure by averaging the values obtained for all beats of the analyzed signal.

3.1. Performance indices for simulated signals

The *root mean square error* (RMSE) is a well-established performance index for simulated signals, used in diverse biomedical applications of which f-wave extraction is one. This index is defined by (Stridh & Sörnmo 2001, Petrėnas et al. 2012)

$$\mathcal{S}_{RMSE} = \sqrt{\frac{1}{N_{QT}} \sum_{n=0}^{N_{QT}-1} (s(n) - \hat{s}(n))^2}, \tag{1}$$

where $\mathcal{S}$ indicates that the index is designed for simulated signals. The *normalized root mean square error* (nRMSE) is another commonly used index (Castells, Rieta, Millet & Zarzoso 2005, Alcaraz & Rieta 2008, Mateo & Rieta 2013, Kheirati Roonizi & Sassi 2017, Malik et al. 2017), defined by

$$\mathcal{S}_{\text{nRMSE}} = \sqrt{\frac{\frac{1}{N_{\text{QT}}} \sum_{n=0}^{N_{\text{QT}}-1} \left(s(n) - \hat{s}(n)\right)^2}{s_{\text{RMS,QT}}}}, \tag{2}$$

where $s_{\text{RMS,QT}}$ is the RMS value of $s(n)$ in the QT interval.

Yet another well-known performance index is the *sample correlation coefficient* (Rieta & Hornero 2007, Alcaraz & Rieta 2008, Mateo & Rieta 2013), defined by

$$\mathcal{S}_{\text{CC}} = \frac{\frac{1}{N_{\text{QT}}} \sum_{n=0}^{N_{\text{QT}}-1} s(n)\hat{s}(n)}{s_{\text{RMS,QT}}\hat{s}_{\text{RMS,QT}}}, \tag{3}$$

where $|\mathcal{S}_{\text{CC}}| \leq 1$. It should be noted that $\mathcal{S}_{\text{CC}}$ measures morphological similarity, but is invariant to differences in amplitude between $s(n)$ and $\hat{s}(n)$. This property must be taken into account when f-wave amplitude is subject to analysis in applications such as the prediction of AF termination in patients with persistent AF who have undergone catheter ablation (Nault, Lellouche, Matsuo, Knecht, Wright, Lim, Sacher, Platonov, Deplagne, Bordachar, Derval, O'Neill, Klein, Hocini, Jaïs, Clémenty & Haïssaguerre 2009).

3.2. Performance indices for real signals

Real signals call for other types of performance indices than do simulated signals since $s(n)$ is unknown. Real-signal indices, denoted by $\mathcal{R}$, aim to quantify the amount of QRS-related residuals present in $\hat{s}(n)$. The simplest performance index characterizes the local *signal-to-noise ratio* (SNR) by the logarithm of the ratio of the R-wave amplitudes $x(n_{\text{R}})$ and $\hat{s}(n_{\text{R}})$ (Mateo & Rieta 2013),

$$\mathcal{R}_{\text{SNR}} = 10 \cdot \log_{10}\left(\frac{x(n_{\text{R}})}{\hat{s}(n_{\text{R}})}\right), \tag{4}$$

where n_{R} is the location of the R-peak.

The *ventricular residue* (VR) is an often used performance index, originally proposed in (Alcaraz & Rieta 2008). This index can be defined either without normalization, then referred to as the *unnormalized ventricular residue* (uVR),

$$\mathcal{R}_{\text{uVR}} = \hat{s}_{\text{RMS,QRS}} \cdot \max_{n=0,\ldots,N_{\text{QRS}}-1} |\hat{s}(n)|, \tag{5}$$

or with normalization,

$$\mathcal{R}_{\text{VR}} = \frac{\mathcal{R}_{\text{uVR}}}{\hat{s}^2_{\text{RMS,1-min}}}, \tag{6}$$

where $\hat{s}_{\mathrm{RMS,QRS}}$ is the RMS value of $\hat{s}(n)$ in the QRS interval and $\hat{s}^2_{\mathrm{RMS,1\text{-}min}}$ is the RMS value of $\hat{s}(n)$ in a 1-min interval centered around the R-wave peak, respectively. The QRS interval contains N_{QRS} samples.

The *modified ventricular residue* (mVR) index was introduced with the aim of improving the performance evaluation of extracted signals with low-amplitude f-waves (Malik et al. 2017). The index is defined by

$$\mathcal{R}_{\mathrm{mVR}} = \frac{1}{4}\left(\frac{Q_{50}[|\bar{\hat{s}}_{\mathrm{QT}}(n)|]}{Q_{50}[|\bar{x}_{\mathrm{cTQ}}(n)|]} + \frac{Q_{50}[|\bar{x}_{\mathrm{cTQ}}(n)|]}{Q_{50}[|\bar{\hat{s}}_{\mathrm{QT}}(n)|]} \right) \cdot \\ \left(\frac{Q_{95}[|\bar{\hat{s}}_{\mathrm{QT}}(n)|]}{\max[|\bar{x}_{\mathrm{cTQ}}(n)|]} + \frac{\max[|\bar{x}_{\mathrm{cTQ}}(n)|]}{Q_{95}[|\bar{\hat{s}}_{\mathrm{QT}}(n)|]} \right),$$

where $Q_{\alpha}[x(n)]$ denotes the $\alpha\%$ quantile of $x(n)$. The quantiles are computed using the following two median-corrected signals:

$$\bar{\hat{s}}_{\mathrm{QT}}(n) = \hat{s}(n) - Q_{50}[\hat{s}(n)], \quad n = 0, \ldots, N_{\mathrm{QT}} - 1, \tag{7}$$

and

$$\bar{x}_{\mathrm{cTQ}}(n) = x_{\mathrm{cTQ}}(n) - Q_{50}[x_{\mathrm{cTQ}}(n)], \quad n = 0, \ldots, N_{\mathrm{cTQ}} - 1, \tag{8}$$

where $x_{\mathrm{cTQ}}(n)$ is obtained by concatenation of the samples contained in 30 TQ intervals preceding the current beat and 30 TQ intervals following. Thus, $\bar{\hat{s}}_{\mathrm{QT}}(n)$ relates only to the current beat, whereas $\bar{x}_{\mathrm{cTQ}}(n)$ relates to a much longer signal segment whose length N_{cTQ} depends on the prevailing heart rate.

Yet another approach to evaluating extraction performance is to analyze the relevant scales of the time–scale representation $w(s, \tau)$, obtained by computing the continuous wavelet transform of $\hat{s}(n)$; s and τ denote scale and translation, respectively (Langley 2015). In this approach, the scalogram $|w(s, \tau)|^2$ is first normalized across all translations,

$$E(s, \tau) = \frac{|w(s, \tau)|^2}{\int |w(s, \tau)|^2 d\tau}, \tag{9}$$

and, then, for the purpose of characterizing the temporal distribution of $E(s, \tau)$, the *wavelet entropy* $W(s)$ is computed for scale s,

$$W(s) = -\int E(s, \tau) \cdot \ln\left(E(s, \tau)\right) d\tau. \tag{10}$$

Using the fourth-order derivative of the Gaussian function as the mother wavelet, the following three performance indices were proposed, characterizing

- *ventricular activity* ($\mathcal{R}_{\mathrm{V}}$), defined as the minimum $W(s)$ across the scales containing frequencies from 12.5 to 50 Hz.

- *atrial activity* ($\mathcal{R}_{\mathrm{A}}$), defined as the average of $W(s)$ across the scales containing frequencies from 3 to 10 Hz.

- both *ventricular and atrial activity* ($\mathcal{R}_{\mathrm{VA}}$), defined as the average of $W(s)$ across the scales containing both ventricular and atrial activity, i.e., 3 to 50 Hz.

3.3. Comparison of performance indices

The statistical association between indices designed for real and simulated signals is investigated for the purpose of identifying those $\mathcal{R}$-indices which convey information similar to any of the $\mathcal{S}$-indices. The identified indices are deemed to be particularly well-suited for evaluating performance on real signals.

For each pair $(\mathcal{S}, \mathcal{R})$, the distribution of points is fitted, in the least squares sense, using linear, power, exponential, and logarithm functions. The goodness-of-fit is quantified by the degrees-of-freedom adjusted R^2 statistic ($0 < R^2 < 1$), estimating the proportion of the variance in the dependent variable predicted from the independent variable; a better fit is reflected by a larger value of R^2. For these four functions, the largest R^2 is selected to characterize the association between each pair of indices.

A comparison of performance indices requires a method for f-wave extraction. Here, the well-known average beat subtraction (ABS) is chosen since, despite its limitations, it continues to be popular in clinically oriented studies, see, e.g., (Fujiki, Sakabe, Nishida, Mizumaki & Inoue 2003, Shah, Yamane, Choi & Haïssaguerre 2004, Beckers, Anne, Verheyden, van der Dussen de Kestergat, van Herk, Janssens, Willems, Heidbuchel & Aubert 2005, Grubitzsch, Modersohn, Leuthold & Konertz 2008, Sterling, Huang & Ghoraani 2015); the rationale behind this choice is discussed in Section 5.

The performance indices were implemented using the parameter settings provided in the original publications. For normal beats, the onset of the QRST interval started 70 ms before the R wave peak, and the length of the QRST interval was set to the shortest RR interval. For VPBs, the onset was instead set to 90 ms before the R wave peak and the length was established by increasing the shortest RR interval with 20%. It was visually verified that all individual beats, whether normal beats or VPBs, were properly contained in the related window.

4. Results

Table 2 presents the R^2 statistic for pairs of indices, obtained for the real f-waves contained in the sets S1, S3, S5, and S7. Almost all pairs are associated with low values of R^2, and it is only $(\mathcal{S}_{\mathrm{RMSE}}, \mathcal{R}_{\mathrm{uVR}})$ that is close to 1, i.e., $R^2 = 0.94$, thus indicating that these two indices convey similar information. On the other hand, the normalized ventricular residue index $\mathcal{R}_{\mathrm{VR}}$ is not strongly associated with any of the $\mathcal{S}$-indices. This observation becomes even more pronounced for the three indices $\mathcal{R}_{\mathrm{V}}, \mathcal{R}_{\mathrm{A}}$, and $\mathcal{R}_{\mathrm{VA}}$, which are all associated with R^2-values close to zero. It is also noted that $\mathcal{S}_{\mathrm{CC}}$ is not associated with any of the $\mathcal{R}$-indices.

Table 3 presents the R^2 statistic for pairs of indices, now obtained for the synthesized f-waves contained in the sets S2, S4, S6, and S8. Comparing these results with those in Table 2, it is evident that the analysis of indices for real and synthesized f-waves leads to essentially the same result, i.e., $\mathcal{R}_{\mathrm{uVR}}$ again conveys information similar to that of $\mathcal{S}_{\mathrm{RMSE}}$.

The overall results of R^2 statistics, obtained from all the sets S1–S8 merged, differ marginally from the results in Tables 2 and 3 and are therefore not presented.

Table 2. R^2 statistic for pairs of performance indices, obtained for real f-waves in the sets S1, S3, S5, and S7.

	$\mathcal{R}_{\mathrm{SNR}}$	$\mathcal{R}_{\mathrm{VR}}$	$\mathcal{R}_{\mathrm{uVR}}$	$\mathcal{R}_{\mathrm{mVR}}$	$\mathcal{R}_{\mathrm{V}}$	$\mathcal{R}_{\mathrm{A}}$	$\mathcal{R}_{\mathrm{VA}}$
$\mathcal{S}_{\mathrm{RMSE}}$	0.61	0.65	0.94	0.39	0.04	0.08	0.27
$\mathcal{S}_{\mathrm{nRMSE}}$	0.40	0.37	0.38	0.78	0.03	0.10	0.31
$\mathcal{S}_{\mathrm{CC}}$	0.37	0.18	0.28	0.27	0.14	0.06	0.26

Table 3. R^2 statistic for pairs of performance indices, obtained for synthesized f-waves in the sets S2, S4, S6, and S8.

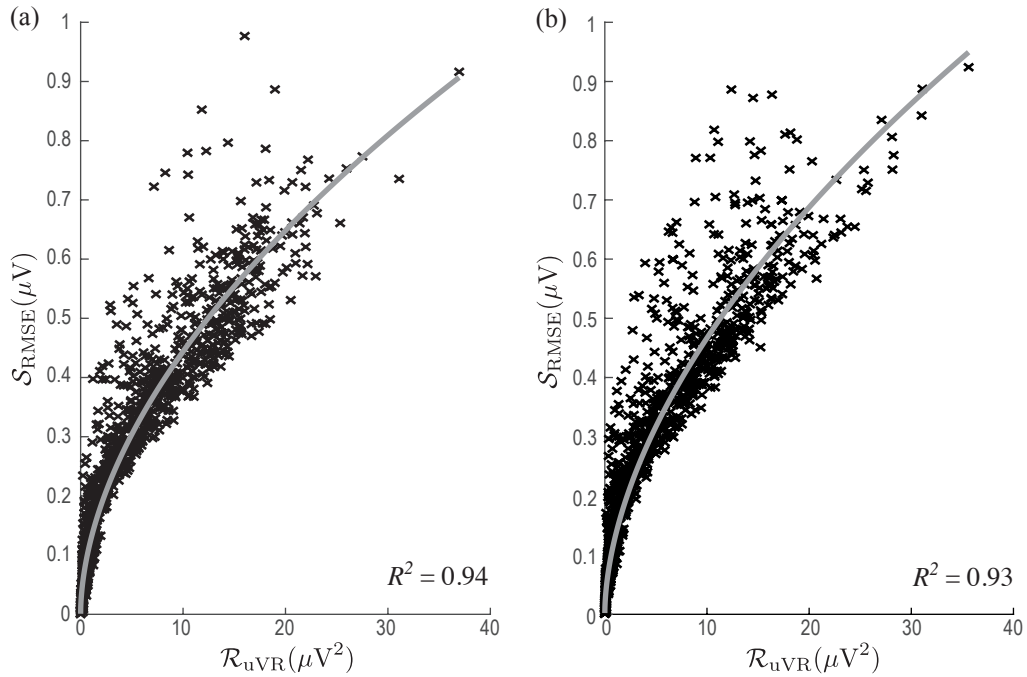|	$\mathcal{R}_{\mathrm{SNR}}$	$\mathcal{R}_{\mathrm{VR}}$	$\mathcal{R}_{\mathrm{uVR}}$	$\mathcal{R}_{\mathrm{mVR}}$	$\mathcal{R}_{\mathrm{V}}$	$\mathcal{R}_{\mathrm{A}}$	$\mathcal{R}_{\mathrm{VA}}$
$\mathcal{S}_{\mathrm{RMSE}}$	0.54	0.71	0.93	0.46	0.17	0.05	0.19
$\mathcal{S}_{\mathrm{nRMSE}}$	0.27	0.36	0.38	0.79	0.04	0.04	0.21
$\mathcal{S}_{\mathrm{CC}}$	0.21	0.15	0.20	0.31	0.12	0.02	0.18

Figure 3. Association between $\mathcal{R}_{\mathrm{uVR}}$ and $\mathcal{S}_{\mathrm{RMSE}}$ for (a) real f-waves and (b) synthesized f-waves.

Figure 3 shows the scatter plots of the pair with the strongest statistical association, i.e., $(\mathcal{R}_{\mathrm{uVR}}, \mathcal{S}_{\mathrm{RMSE}})$, divided into real and synthesized f-waves. Figure 4 shows the pair with the next strongest association, i.e., $(\mathcal{R}_{\mathrm{mVR}}, \mathcal{S}_{\mathrm{nRMSE}})$. Each dot in the scatter plot results from computing the pair of indices from a 5-min signal contained in the analyzed data sets. In all four scatter plots, the power function offers the best fit to the data sets.
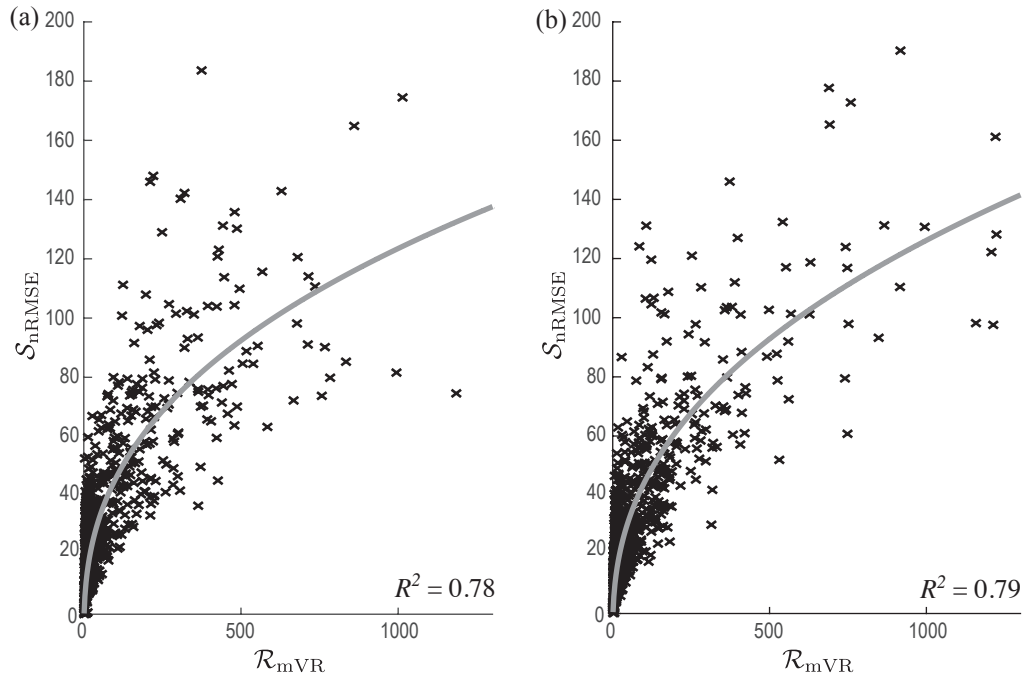
Figure 4. Association between $\mathcal{R}_{\mathrm{mVR}}$ and $\mathcal{S}_{\mathrm{nRMSE}}$ for (a) real f-waves and (b) synthesized f-waves.

5. Discussion

5.1. Database with simulated signals

Simulated ECG signals have been widely used for evaluating the performance of f-wave extraction methods. The signals are usually composed of three components — QRST complexes, f-waves, and noise — extracted either from real or synthesized ECGs. The QRST complexes can be extracted from real ECGs in sinus rhythm following P-wave cancellation (Alcaraz & Rieta 2008, Petrėnas et al. 2012, Dai et al. 2013, Donoso, Figueroa, Lecannelier, Pino & Rojas 2013, Lee, Song, Shin & Lee 2012) or synthesized using a single dipole model (McSharry, Clifford, Tarassenko & Smith 2003, Kheirati Roonizi & Sassi 2017). Similarly, the f-wave signal can be produced by concatenating the TQ intervals of real ECGs in AF (Alcaraz & Rieta 2008, Donoso et al. 2013, Lee et al. 2012) or synthesized using the sawtooth model (Stridh & Sörnmo 2001) or its variants (Petrėnas et al. 2012, Kheirati Roonizi & Sassi 2017, Malik et al. 2017). The noise may be taken from a database as done in (Petrėnas et al. 2017) or synthesized using, e.g., autoregressive modeling (Sameni et al. 2007). In most studies on f-wave extraction, however, the noise component is disregarded.

It is evident that simulated signals can be composed in many different ways, and, therefore, it is not surprising that the characteristics of databases differ considerably from study to study. As a result, a fair comparison of performance is extremely difficult to accomplish, with the exception of those cases where the authors have implemented some other method(s) to compare performance with that of the proposed method; ABS is commonly the method implemented for comparison, see, e.g., (Stridh & Sörnmo 2001, Alcaraz &

Rieta 2008, Lee et al. 2012, Dai et al. 2013).

The proposed reference database of simulated ECG signals not only facilitates the comparison of performance, but it also highlights signal characteristics known to represent major challenges in f-wave extraction. These characteristics include high heart rates, high morphological QRST variability, the presence of VPBs, different f-wave amplitudes and repetition rates, and f-wave morphology involving a stochastic component. Some of these characteristics (i.e., high heart rates, different f-wave amplitudes and repetition rates) are integral to the synthesized signals, whereas other characteristics (i.e., high morphological QRST variability and the presence of VPBs) are featured, either alone or together, in data sets S5–S8, see Table 1. Thus, the strengths and weaknesses of a method can be better pinpointed by presenting the performance results on a set-by-set basis rather than as a result obtained for the whole database.

When QRST complexes are synthesized, the morphological variability may be modest or pronounced depending on how the model parameters are chosen. For most extraction methods, increased morphological variability translates to worsened performance, especially for methods involving ensemble averaging of the QRST complexes. Moreover, the degree of determinism with which f-waves are synthesized may notably influence performance. Indeed, almost perfect extraction can be achieved when the synthesis of f-waves is completely deterministic, i.e., no stochastic component is involved (Petrénas et al. 2017). Obviously, such results say little about the practical value of the method subject to evaluation.

5.2. Performance indices

A method for f-wave extraction needs to be chosen before the statistical association between any two indices $\mathcal{R}$ and $\mathcal{S}$ can be investigated. In the present study, the well-known ABS was chosen because of its simplicity, its earlier use in a study on performance evaluation (Langley 2015), and the fact that the development of extraction methods has had ABS as a natural starting point.

Using the proposed reference database, the second part of the present study addresses the problem of determining which of the considered $\mathcal{R}$-indices provide information statistically associated with any of the $\mathcal{S}$-indices. The results show that the strongest association ($R^2 = 0.94$) exists between the pair $(\mathcal{S}_{\mathrm{RMSE}}, \mathcal{R}_{\mathrm{uVR}})$, whereas no other pair has an association exceeding 0.8. This particular pair offers the advantage of characterizing performance in clinically relevant terms, i.e., in units of μV, and thus avoids the problems associated with interpreting normalized indices such as $\mathcal{S}_{\mathrm{nRMSE}}$ and $\mathcal{S}_{\mathrm{CC}}$. It should be emphasized that a strong association to a certain $\mathcal{S}$-index does not imply that the index is perfect and cannot be bettered.

In addition to the indices described in Section 3.1, a few other $\mathcal{S}$-indices have been considered in the literature. The SNR improvement (Kheirati Roonizi & Sassi 2017) is defined as the ratio between the power of $(x(n) - s(n))$ and the power of $(s(n) - \hat{s}(n))$, and thus differ from the herein described $\mathcal{S}$-indices since it involves the power of $x(n)$. As a consequence, the accuracy of $\hat{s}(n)$ is evaluated correctly when the power of $x(n)$ is held fixed from one realization of the simulated signal to the next. If not, an improved SNR may

be due to changes in the characteristics of the QRST complex (which are unrelated to f-wave extraction). Therefore, this index was not considered in the present study. The mean absolute error of the DAF (Sassi et al. 2009) is an index which indirectly quantifies the performance in simulated signals, being well-suited for situations when the overall aim is to analyze the DAF. However, this index requires that the DAF is estimated from $s(n)$, which thus introduces an unwanted dependence on the technique used for spectral estimation. Since this index does not convey much information on how f-wave morphology is preserved in $\hat{s}(n)$, it was neither considered in the present study.

Likewise, a few $\mathcal{R}$-indices described in the literature were not considered in the present study. The spectral concentration (Castells, Rieta, Millet & Zarzoso 2005, Llinares, Igual & Miró-Borrás 2010, Phlypo et al. 2010, Malik et al. 2017) is defined as the normalized spectral power in a small interval centered around the DAF; a more concentrated spectrum is judged to reflect better performance. However, the usefulness of this index may be questioned since it is well-known that the DAF can vary over time, implying that the dominant spectral peak may be broadened for physiological reasons. Neither does this index provide much information on the magnitude of QRST-related residuals. Another approach to quantifying performance is to compare the f-wave amplitude representative of the QT interval with that representative of the TQ interval (Langley et al. 2006), based on the assumption that good performance is manifested by an insignificant difference in amplitude between the two intervals. A nonparametric statistical test is used to determine whether the amplitudes in the two intervals are significantly different. This test provides a rather coarse characterization of performance, being expressed in terms of statistical significance of differences between the two populations, and, therefore, it is not considered in the present study. A similar reasoning applies to the statistical approach proposed in (Lemay et al. 2007).

The present study considers performance indices quantifying the characteristics of $\hat{s}(n)$ in the QT interval because most f-wave distortion is confined to this interval. However, for certain methods, it may be motivated to quantify the distortion in the TQ interval as well, so that a more complete characterization of extraction performance is provided. Methods which may need both QT- and TQ-related performance indices include independent component analysis (Rieta et al. 2004, Phlypo et al. 2010), autoregressive modeling and prediction error analysis (Wang, Rao, Shepherd & Beggs 2008), and the echo state network (Petrėnas et al. 2012). The indices described in Section 3 can also be used for characterizing the distortion in the TQ interval.

Although the emphasis of the present study is on the use of a single $\mathcal{R}$-index to characterize performance in real ECGs, it may very well be motivated to combine the proposed index $\mathcal{R}_{\text{uVR}}$ with some other index, e.g., designed with reference to the particular application of interest.

Some limitations of the study merit consideration. Extraction performance has been quantified in individual leads, implying that the performance of multi-lead methods, e.g., (Rieta et al. 2004, Donoso et al. 2013), producing a global f-wave signal, cannot be properly evaluated as this signal does not match any of the original leads. Hence, indices need to be developed so that the performance of such methods can be evaluated and compared

as well. Another potential limitation of the study is that noise-free signals have been analyzed, although a certain degree of noise is present in real signals. However, noise-free signals were employed to ensure that any association that may exist between different pairs of indices was not masked by noise. If performance in noise is of interest to investigate, it is straightforward to add noise to the simulated signals of the database. Since f-wave extraction is likely preceded by preprocessing for noise removal, a comparison of performance is rendered difficult due to that the preprocessing implemented usually differ from study to study, causing f-wave morphology to be altered differently.

6. Conclusions

A reference database with simulated ECG signals is proposed for the purpose of evaluating the performance of methods for f-wave extraction. The signals account for a wide range of challenging characteristics such as high heart rates, high morphological QRST variability, and the presence of ventricular premature beats. Using this database, the information conveyed by performance indices designed either for use with simulated or real signals was investigated. The strongest statistical association was found for the index pair (RMSE,uVR), leading to the recommendation that the uVR should be used when characterizing performance in real signals.

Acknowledgments

References

Alcaraz, R. & Rieta, J. J. (2008). Adaptive singular value cancellation of ventricular activity in single-lead atrial fibrillation electrocardiograms, *Physiol. Meas.* **29**(12): 1351–1369.

Alcaraz, R. & Rieta, J. J. (2009). Non-invasive organization variation assessment in the onset and termination of paroxysmal atrial fibrillation, *Comput. Methods Programs Biomed.* **93**(2): 148–154.

Alcaraz, R. & Rieta, J. J. (2017). Applications of nonlinear methods to atrial fibrillation, *in* R. Barbieri, E. P. Scilingo & G. Valenza (eds), *Complexity and Nonlinearity in Cardiovascular Signals*, Springer, chapter 15, pp. 387–426.

Beckers, F., Anne, W., Verheyden, B., van der Dussen de Kestergat, C., van Herk, E., Janssens, L., Willems, R., Heidbuchel, H. & Aubert, A. E. (2005). Determination of atrial fibrillation frequency using QRST-cancellation with QRS-scaling in standard electrocardiogram leads, *Proc. Comput. Cardiol.*, Vol. 32, pp. 339–342.

Bollmann, A., Sonne, K., Esperer, H. D., Toepffer, I., Langberg, J. J. & Klein, H. U. (1999). Non-invasive assessment of fibrillatory activity in patients with paroxysmal and persistent atrial fibrillation using the Holter ECG, *Cardiovasc Res* **44**(1): 60–66.

Bousseljot, R., Kreiseler, D. & Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet, *Biomedizinische Technik* **40**: 317.

Castells, F., Mora, C., Rieta, J. J., Moratal-Pérez, D. & Millet, J. (2005). Estimation of atrial fibrillatory wave from single-lead atrial fibrillation electrocardiograms using principal component analysis concepts, *Med. Biol. Eng. & Comput.* **43**: 557–560.

Castells, F., Rieta, J. J., Millet, J. & Zarzoso, V. (2005). Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias, *IEEE Trans. Biomed. Eng.* **52**(2): 258–267.

Chugh, S. S., Roth, G. A., Gillum, R. F. & Mensah, G. A. (2014). Global burden of atrial fibrillation in developed and developing nations, *Glob. Heart* **9**(1): 113–119.

Colilla, S., Crow, A., Petkun, W., Singer, D. E., Simon, T. & Liu, X. (2013). Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population, *Am. J. Cardiol.* **112**(8): 1142–1147.

Cuculich, P. S., Wang, Y., Lindsay, B. D., Faddis, M. N., Schuessler, R. B., Damiano, Jr, R. J., Li, L. & Rudy, Y. (2010). Noninvasive characterization of epicardial activation in humans with diverse atrial fibrillation patterns, *Circulation* **122**(14): 1364–1372.

Dai, H., Jiang, S. & Li, Y. (2013). Atrial activity extraction from single lead ECG recordings: evaluation of two novel methods, *Comput. Biol. Med.* **43**(3): 176–183.

Dai, H., Yin, L. & Li, Y. (2016). QRS residual removal in atrial activity signals extracted from single-lead: a new perspective based on signal extrapolation, *IET Signal Process.* **10**(9): 1169–1175.

Donoso, F. I., Figueroa, R. L., Lecannelier, E. A., Pino, E. J. & Rojas, A. J. (2013). Atrial activity selection for atrial fibrillation ECG recordings, *Comput. Biol. Med.* **43**(10): 1628–1636.

Fauchier, L., Villejoubert, O., Clementy, N., Bernard, A., Pierre, B., Angoulvant, D., Ivanes, F., Babuty, D. & Lip, G. Y. H. (2016). Causes of death and influencing factors in patients with atrial fibrillation, *Am. J. Med.* **129**(12): 1278–1287.

Fujiki, A., Sakabe, M., Nishida, K., Mizumaki, K. & Inoue, H. (2003). Role of fibrillation cycle length in spontaneous and drug-indcued termination of human atrial fibrillation—Spectral analysis of fibrillation waves from surface electrocardiogram, *Circ. J.* **67**: 391–395.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K. & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, *Circulation* **101**: E215–220.

Grubitzsch, H., Modersohn, D., Leuthold, T. & Konertz, W. (2008). Analysis of atrial fibrillatory activity from high-resolution surface electrocardiograms: Evaluation and application of a new system, *Exp. Clin. Cardiol.* **13**: 29–35.

Kheirati Roonizi, E. & Sassi, R. (2017). An extended Bayesian framework for atrial and ventricular activity separation in atrial fibrillation, *IEEE J. Biomed. Health Inform.* **21**(6): 1573–1580.

Krijthe, B. P., Kunst, A., Benjamin, E. J., Lip, G. Y. H., Franco, O. H., Hofman, A., Witteman, J. C. M., Stricker, B. H. & Heeringa, J. (2013). Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060, *Eur. Heart J.* **34**(35): 2746–2751.

Langley, P. (2015). Wavelet entropy as a measure of ventricular beat suppression from the electrocardiogram in atrial fibrillation, *Entropy* **17**: 6397–6411.

Langley, P., Stridh, M., Rieta, J. J., Millet, J., Sörnmo, L. & Murray, A. (2006). Comparison of atrial signal extraction algorithms in 12-lead ECGs with atrial fibrillation, *IEEE Trans. Biomed. Eng.* **53**: 343–346.

Lee, J., Song, M.-h., Shin, D.-g. & Lee, K.-j. (2012). Event synchronous adaptive filter based atrial activity estimation in single-lead atrial fibrillation electrocardiograms, *Med. Biol. Eng. & Comput.* **50**(8): 801–811.

Lemay, M., Vesin, J.-M., van Oosterom, A., Jacquemet, V. & Kappenberger, L. (2007). Cancellation of ventricular activity in the ECG: evaluation of novel and existing methods, *IEEE Trans. Biomed. Eng.* **54**(3): 542–546.

Llinares, R., Igual, J. & Miró-Borrás, J. (2010). A fixed point algorithm for extracting the atrial activity in the frequency domain, *Comput. Biol. Med.* **40**: 943–949.

Malik, J., Reed, N., Wang, C.-L. & Wu, H.-T. (2017). Single-lead f-wave extraction using diffusion geometry,

Physiol. Meas. **38**(7): 1310–1334.

Mateo, J. & Rieta, J. J. (2013). Radial basis function neural networks applied to efficient QRST cancellation in atrial fibrillation, *Comput. Biol. Med.* **43**(2): 154–163.

McSharry, P. E., Clifford, G. D., Tarassenko, L. & Smith, L. A. (2003). A dynamical model for generating synthetic electrocardiogram signals, *IEEE Trans. Biomed. Eng.* **50**(3): 289–294.

Nault, I., Lellouche, N., Matsuo, S., Knecht, S., Wright, M., Lim, K. T., Sacher, F., Platonov, P., Deplagne, A., Bordachar, P., Derval, N., O'Neill, M. D., Klein, G. J., Hocini, M., Jaïs, P., Clémenty, J. & Haïssaguerre, M. (2009). Clinical value of fibrillatory wave amplitude on surface ECG in patients with persistent atrial fibrillation, *J. Interv. Card. Electrophysiol.* **26**: 11–19.

Petrėnas, A., Marozas, V., Sološenko, A., Kubilius, R., Skibarkiene, J., Oster, J. & Sörnmo, L. (2017). Electrocardiogram modeling during paroxysmal atrial fibrillation: application to the detection of brief episodes, *Physiol. Meas.* **38**(11): 2058–2080.

Petrėnas, A., Marozas, V., Sörnmo, L. & Lukosevicius, A. (2012). An echo state neural network for qrst cancellation during atrial fibrillation, *IEEE Trans. Biomed. Eng.* **59**(10): 2950–2957.

Phlypo, R., Zarzoso, V. & Lemahieu, I. (2010). Atrial activity estimation from atrial fibrillation ecgs by blind source extraction based on a conditional maximum likelihood approach, *Med. Biol. Eng. & Comput.* **48**(5): 483–488.

Platonov, P. G., Corino, V. D. A., Seifert, M., Holmqvist, F. & Sörnmo, L. (2014). Atrial fibrillatory rate in the clinical context: natural course and prediction of intervention outcome, *Europace* **16**: iv110–iv119.

Ravelli, F., Masè, M., Greco, M. D., Faes, L. & Disertori, M. (2007). Deterioration of organization in the first minutes of atrial fibrillation: A beat-to-beat analysis of cycle length and wave similarity, *J. Cardiovasc. Electrophysiol.* **18**: 60–65.

Rieta, J. J., Castells, F., Sánchez, C., Zarzoso, V. & Millet, J. (2004). Atrial activity extraction for atrial fibrillation analysis using blind source separation, *IEEE Trans. Biomed. Eng.* **51**(7): 1176–1186.

Rieta, J. J. & Hornero, F. (2007). Comparative study of methods for ventricular activity cancellation in atrial electrograms of atrial fibrillation, *Physiol. Meas.* **28**(8): 925–936.

Sameni, R., Clifford, G. D., Jutten, C. & Shamsollahi, M. B. (2007). Multichannel ECG and noise modeling: Application to maternal and fetal ECG signals, *J. Adv. Signal Process.* pp. 1–14.

Sassi, R., Corino, V. D. A. & Mainardi, L. T. (2009). Analysis of surface atrial signals: Time series with missing data?, *Ann. Biomed. Eng.* **37**: 2082–2092.

Schotten, U., Dobrev, D., Platonov, P. G., Kottkamp, H. & Hindricks, G. (2016). Current controversies in determining the main mechanisms of atrial fibrillation, *J. Intern. Med.* **279**(5): 428–438.

Shah, D. C., Yamane, T., Choi, K. J. & Haïssaguerre, M. (2004). QRS subtraction and the ECG analysis of atrial ectopics, *Ann. Noninvasive Electrocardiol.* **9**: 389–398.

Sörnmo, L., Alcaraz, R., Laguna, P. & Rieta, J. J. (2018). Characterization of f waves, *in* L. Sörnmo (ed.), *Atrial Fibrillation from an Engineering Perspective*, Springer, chapter 6, pp. 221–279.

Sörnmo, L., Petrėnas, A., Laguna, P. & Marozas, V. (2018). Extraction of f waves, *in* L. Sörnmo (ed.), *Atrial Fibrillation from an Engineering Perspective*, Springer Nature, chapter 4, pp. 137–330.

Sposato, L. A., Cipriano, L. E., Saposnik, G., Ruíz Vargas, E., Riccio, P. M. & Hachinski, V. (2015). Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis, *Lancet Neurol* **14**(4): 377–87.

Sterling, M., Huang, D. T. & Ghoraani, B. (2015). Developing a new computer-aided clinical decision support system for prediction of successful postcardioversion patients with persistent atrial fibrillation, *Comput. Math. Methods Med.* .

Stridh, M. & Sörnmo, L. (2001). Spatiotemporal QRST cancellation techniques for analysis of atrial fibrillation, *IEEE Trans. Biomed. Eng.* **48**(1): 105–111.

Wang, G., Rao, N., Shepherd, S. J. & Beggs, C. B. (2008). Extraction of desired signal based on AR model with its application to atrial activity estimation in atrial fibrillation, *J. Adv. Signal Process.* **8**: 1–9.