IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **38** (2017) 1730–1745 https://doi.org/10.1088/1361-6579/aa6e9f

Performance of an open-source heart sound segmentation algorithm on eight independent databases

Chengyu Liu[1], David Springer[2] and Gari D Clifford[1,3]

[1] Department of Biomedical Informatics, Emory University, Atlanta, United States of America
[2] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[3] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, United States of America

E-mail: gari@gatech.edu

Received 22 January 2017, revised 10 April 2017
Accepted for publication 21 April 2017
Published 1 August 2017

Abstract

Objective: Heart sound segmentation is a prerequisite step for the automatic analysis of heart sound signals, facilitating the subsequent identification and classification of pathological events. Recently, hidden Markov model-based algorithms have received increased interest due to their robustness in processing noisy recordings. In this study we aim to evaluate the performance of the recently published logistic regression based hidden semi-Markov model (HSMM) heart sound segmentation method, by using a wider variety of independently acquired data of varying quality. *Approach*: Firstly, we constructed a systematic evaluation scheme based on a new collection of heart sound databases, which we assembled for the PhysioNet/CinC Challenge 2016. This collection includes a total of more than 120 000 s of heart sounds recorded from 1297 subjects (including both healthy subjects and cardiovascular patients) and comprises eight independent heart sound databases sourced from multiple independent research groups around the world. Then, the HSMM-based segmentation method was evaluated using the assembled eight databases. The common evaluation metrics of sensitivity, specificity, accuracy, as well as the F_1 measure were used. In addition, the effect of varying the tolerance window for determining a correct segmentation was evaluated. *Main results*: The results confirm the high accuracy of the HSMM-based algorithm on a separate test dataset comprised of 102 306 heart sounds. An average F_1 score of 98.5% for segmenting S1 and systole intervals and 97.2% for segmenting S2 and diastole intervals were observed. The F_1 score was shown to increases with an increases in the tolerance window size, as expected. *Significance*: The high segmentation accuracy of the HSMM-

based algorithm on a large database confirmed the algorithm's effectiveness. The described evaluation framework, combined with the largest collection of open access heart sound data, provides essential resources for evaluators who need to test their algorithms with realistic data and share reproducible results.

Keywords: heart sound, heart sound segmentation, hidden Markov model (HMM), hidden semi-Markov model (HSMM), PhysioNet/CinC Challenge

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

1. Introduction

Cardiac auscultation is an essential part of physical examination in clinical practice and may reveal many pathologic cardiac conditions such as arrhythmias, valve disease, heart failure, and more. Heart sounds provide important initial clues in disease evaluation, serve as a guide for further diagnostic examination, and thus play an important role in the early detection of cardiovascular diseases (Leatham 1975, Wang *et al* 2007, Springer 2015b). Identification of the exact positions of the fundamental heart sounds (FHSs), i.e. heart sound segmentation, is generally a first step in the automatic analysis of heart sound recordings. FHSs includes the first (S1) and second (S2) heart sounds (Leatham 1975). S1 occurs at the beginning of isovolumetric ventricular contraction, when the closed mitral and tricuspid valves suddenly reach their elastic limit due to the rapid increase in pressure within the ventricles. S2 occurs at the beginning of diastole with the closure of the aortic and pulmonic valves (Tilkian and Conover 2001). Accurate localization of S1 and S2 is a prerequisite for locating the systolic or diastolic regions, allowing the subsequent classification of pathological events within these periods (Springer 2015b, Springer *et al* 2016).

While S1 and S2 are the most recognizable sounds of the heart cycle, there are also other audible sounds, such as the third heart sound (S3), fourth heart sound (S4), systolic ejection click (EC), mid-systolic click (MC), diastolic sound or opening snap (OS), as well as heart murmurs caused by the turbulent, high-velocity flow of blood (Springer 2015b). These components could complicate the identification of S1 and S2 sounds, especially for signals recorded in a noisy environment, making the task of accurate heart sound segmentation challenging.

Numerous heart sound segmentation methods have been studied over the past few decades. For a detailed review of the existing segmentation methods, as well as the size of the database used in each study and the corresponding numerical results, readers can refer to Springer (2015b) and Liu *et al* (2016). Here, we give a brief summary for heart sound segmentation methods, which can be generally classified by four approaches:

- Type 1 covers the envelope-based methods, which use a variety of techniques to construct the envelopes of a heart sound signal and thus to perform heart sound segmentation. Typical methods for constructing the envelope include the Shannon energy (Liang *et al* 1997, Moukadem *et al* 2013), the Hilbert transform (Sun *et al* 2014), extraction of the cardiac sound characteristic waveform (Jiang and Choi 2006, Yan *et al* 2010), and the squared-energy (Ari *et al* 2008).
- Type 2 involves feature-based methods, which require the calculation of heart sound features followed by a decision making process for segmentation. Typical features include amplitude features (Naseri and Homaeinezhad 2013), frequency features (Kumar *et al* 2006), phase features (Varghees and Ramachandran 2014), periodic component features (Pedrosa *et al* 2014), complexity-based features (Nigam and Priemer 2005), multi-level

wavelet coefficient features (Vepa *et al* 2008) and higher order statistics such as the kurtosis (Papadaniil and Hadjileontiadis 2014).

- Type 3 includes machine learning approaches, which employ neural networks or other nonlinear classifiers which leverage the extracted features. Examples include time-delay neural networks (Oskiper and Watrous 2002) multi-layer perceptron neural networks (Sepehri *et al* 2010)), K-means clustering (Gupta *et al* 2007, Chen *et al* 2009), dynamic clustering (Tang *et al* 2012), and unsupervised learning approaches (Rajan *et al* 2006).
- Type 4 are the hidden Markov model (HMM)-based methods, originally proposed by Gamero and Watrous (2003) in the context of heart sounds. The approach was subsequently developed by Ricke *et al* (2005). Gill *et al* then incorporated timing duration (Gill *et al* 2005) and Sedighian *et al* incorporated homomorphic filtering (Sedighian *et al* 2014) within the HMM method to improve segmentation accuracy. Schmidt *et al* generalized HMM to the hidden semi-Markov model (HSMM) by modelling the expected duration of heart sounds (Schmidt *et al* 2010a, 2010b). Springer *et al* then developed the HSMM method further by employing additional input features, logistic regression for emission probability estimation and a modified Viterbi algorithm that addresses boundary conditions in order improve the overall segmentation of short duration real-world heart sound recordings (Springer 2015b, Springer *et al* 2016).

When tested on a database of 10 172 s of heart sounds, Springer's HSMM method (hereafter identified as the HSMM-based method) achieved an average F_1 score of 95.63% on a separate test dataset, significantly improving upon the highest F_1 score of 86.28% achieved by the other reported methods in the literature when evaluated on the same test data (Springer *et al* 2016). This method is generally regarded as the state-of-the-art and was provided as an open source segmentation algorithm in the PhysioNet/Computing in Cardiology (Liu *et al* 2016). In previously published studies, performance evaluations of heart sound segmentation algorithms were mainly based on a single database, which were limited by the recording number, duration, as well as the physiological/pathological conditions (Springer 2015b, Springer *et al* 2016). In addition, the heart sounds from single database were generally recorded with special equipment using fixed microphones in a low noise environment, ensuring good recording quality for testing the segmentation algorithms. Therefore, the comparisons between algorithms and the subsequent confirmation were hindered by the lack of large, multi-environment, and standardized databases of heart sound signals obtained from a variety of healthy subjects and subjects with pathological conditions. Although the HSMM-based method was evaluated on an independent test set of a database with relatively large number of heart beats, its performance still needs to be confirmed with more data. Therefore, in response to this, we assembled a wide collection of heart sound recordings for the evaluation of algorithms for the PhysioNet/Computing in Cardiology Challenge 2016 (2016) (Clifford *et al* 2016, Liu *et al* 2016, PhysioNet/CinC Challenge 2016 2016). This resulted in a total of more than 120 000 s of heart sounds recorded from 1297 healthy subjects/ pathological patients, representing more than ten times the amount of data compared to any previous study, including that used in Springer *et al* (2016). For this Challenge we manually annotated the four major (normal) heart sound states, i.e. S1, systole, S2 and diastole, for each recording (Clifford *et al* 2016, Liu *et al* 2016). This resulted in a standard reference database for training and evaluating automatic segmentation algorithms.

The purposes of this current study were: (1) to construct a systematic scheme for evaluating different segmentation algorithms based on the PhysioNet/CinC Challenge 2016 data and the manually annotated segmentation annotations, and (2) to evaluate the open-source benchmark HSMM-based segmentation algorithm across multiple databases released for the PhysioNet/CinC Challenge 2016.

2. Methods

2.1. Heart sound data

Heart sound recordings from eight independent databases sourced from multiple research groups around the world released in the PhysioNet/CinC Challenge 2016 were employed in this study. These are documented in detail in Liu *et al* (2016). Among the eight available databases, four databases were divided into training and test sets with a 70–30 training–test split, with two databases assigned to the training set and two other databases assigned to the test set for the Challenge. Thus, the Challenge training set includes data from six databases (with file names prefixed alphabetically, *a* through *f*) and the Challenge test set included data from six databases (named *b* through *e*, plus *g* and *i*) (Clifford *et al* 2016, Liu *et al* 2016). The collected data included not only clean heart sounds but also very noisy recordings. They were recorded from both healthy subjects and cardiovascular patients, and from both children and adults. The recordings from the same patient did not appear in both the training and test datasets. The data were also recorded from different locations, depending on the individual protocols used for each database. However, they were generally recorded at the four common recording locations of aortic area, pulmonic area, tricuspid area and mitral area. All recordings were resampled to 2000 Hz using an anti-alias filter and provided in a standard uncompressed (wav) format.

Each recording contains a single channel of heart sound activity with the exception of training set *a*, which also contains a simultaneously recorded ECG (PhysioNet/CinC Challenge 2016 2016, Liu *et al* 2016). In this study, the ECG data provided in training set *a* were used for training the HSMM-based method (as detailed in Springer *et al* (2016)). The other training and test databases were then used as test data.

In clinical practice, the criteria adopted by the cardiologist to annotate the beginning and the ending of S1 and S2 sounds were defined as follows: the beginning of S1 is the start of the high frequency vibration due to mitral closure, the beginning of S2 is the start of the high frequency vibration due to aortic closure, and the endings of S1 and S2 are annotated by the end of the high frequency vibrations (Moukadem *et al* 2013). According to this criteria, manual annotations for the four heart sound states (i.e. S1, systole, S2 and diastole) for each beat for the PhysioNet/CinC Challenge 2016 data were provided by the authors (Liu *et al* 2016).

Some recordings, or some segments of heart sound recordings, were visually noisy or acoustically muffled and hard to interpret, making the manual segmentation of the four heart sound states impossible. These recordings and episodes were manually annotated as 'noisy' and were excluded for the algorithm evaluation. Table 1 summarizes the total numbers of subjects, recordings, as well as the excluded recordings. The numbers of the manually annotated beats from the resultant recordings, as well as the maximum (max), median and minimum (min) values of recording length, are also reported. As shown in table 1, a total of 392 (409 recordings minus the 17 excluded recordings) heart sound recordings (totaling 14 559 beats) were used for training the HSMM-based algorithm and the other independent 3622 heart sound recordings from 951 subjects (totaling 102 306 beats) were used for testing.

2.2. Evaluation approach

Let $\{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ denote the manually annotated onset positions for one of the four heart sound states. A tolerance parameter δ is used for determining the true positive (TP), false positive (FP) and false negative (FN) segmentation for the evaluated heart sound segmentation algorithms. For the ith manually annotated onset position, x_i, we counted the numbers of the state onsets from the automatic segmentation results in two time regions:

Table 1. Summary of data used in this study. Specifically, training set *a* in the PhysioNet/CinC Challenge 2016 was used for training the HSMM-based method since this database contains simultaneously recorded ECG signals. The other Challenge training and test databases were used as test data.

Data type	Database	# Subjects	# Recordings	# Excluded recordings	Recording length (s)			# Beats (manual annotation)			
					Min	Median	Max	Min	Median	Max	Total
Training	Training-a	121	409	17	9.3	35.6	36.5	12	37	78	14559
Test	Training-b	106	490	122	5.3	8	8	4	9	15	3353
	Training-c	31	31	4	9.6	44.4	122.0	15	67	143	1808
	Training-d	38	55	3	6.6	12.3	48.5	6	14	72	853
	Training-e	356	2054	126	8.1	21.1	101.7	4	27	174	59593
	Training-f	112	114	5	29.4	31.7	59.6	7	39	75	4259
	Test-b	45	205	73	6.3	8	8	6	9	16	1269
	Test-c	14	14	1	19.3	54.4	86.9	32	57	107	853
	Test-d	17	24	2	6.1	11.4	17.1	7	11	24	260
	Test-e	153	883	61	8.1	21.8	103.6	3	28	169	26724
	Test-g	44	116	0	15	15	15	9	18	29	2048
	Test-i	35	35	2	15.0	31.7	68.8	18	36	59	1286
	Total	**951**	**4021**	**399**	—	—	—	—	—	—	**102306**

$x_i - \delta \leqslant$ state onsets $\leqslant x_i + \delta$ and $x_i + \delta <$ state onsets $< x_{i+1} - \delta$. Let N_1 and N_2 denote the counted numbers in these two time regions respectively.

For the current heart sound state, the automatically segmented onset was expected to appear in the time region $x_i - \delta \leqslant$ state onsets $\leqslant x_i + \delta$ and should not appear in the other time interval, $x_i + \delta <$ state onsets $< x_{i+1} - \delta$. The TP, FP and FN for each manually annotated heart beat cycle were then defined as:

- TP: if $N_1 > 0$, TP = TP + 1, means that there is an expected state onset in the expected time region.
- FP: (1) if $N_1 > 1$, FP = FP + $N_1 - 1$, means that there are more than one segmented state onsets in the expected time region; (2) if $N_2 > 0$, FP = FP + N_2, means there is/are false segmented state onset/onsets in the unexpected time region.
- FN: if $N_1 = 0$, FN = FN + 1, means that there is a missing segmented state onset in the expected time region.

The metrics of sensitivity (Se, or *recall*), positive predictivity (P_+, or *precision*), accuracy (Acc) and F_1 measure are defined as Springer *et al* (2016):

$$\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{1}$$

$$P_+ = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{2}$$

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \times 100\% \tag{3}$$

$$F_1 = \frac{2 \times \text{Se} \times P_+}{\text{Se} + P_+} \times 100\%. \tag{4}$$

The tolerance parameter δ (defining how close in time a detection and an annotation can be in order to count as a match) has a key effect on the evaluation metrics. For ECG QRS detection evaluation, the ANSI/AAMI EC57 standard recommends a tolerance of 150 ms for identifying a coincident automatic and reference ECG R-peak annotation (American National Standards Institute 2012). In the PhysioNet/CinC Challenge 2013, focused on fetal ECG beat detection, a tolerance of 100 ms was used for determining the true/false matching between reference and detected annotations (Silva *et al* 2013, Clifford *et al* 2014). For heart sound signals, there are four state onsets within a single heart cycle. The tolerance therefore must be shorter. Schmidt *et al* (2010a) used a tolerance of 60 ms and a heart sound was determined as a TP if the middle of the segmented sound state (S1 and S2) was closer than this tolerance to the middle of the manually annotated state. In Springer's HSMM approach, the references were ECG features and thus a tolerance of 100 ms was used, i.e. a TP S1 onset was found to be within this tolerance of the R-peak, and a TP S2 onset was found if the center of the automatically segmented S2 sound was within this tolerance of the corresponding end-T-wave (Springer *et al* 2016). For this study, we evaluated tolerances of $\delta = 20$, 40, 60, 80 and 100 ms to test the effect of this parameter on evaluation metrics and identify a consistent yet logical interval to be used.

2.3. ECG-derived heart sound labelling

The HSMM-based method needs to be trained with reference heart sound labelling to obtain the model parameters. Once the model training is completed, the model can be used for segmenting heart sound recordings directly without any other input information (other than the heart sound recording and the features derived from the recording). In this study, the HSMM model was trained using the training-a database (see table 1) since it includes both heart sound and ECG signals and can provide accurate labels for the S1 and S2 sounds.

Firstly, the locations of R-peak and end-T-wave in ECG signals were obtained as the reference positions for S1 and S2 sounds. R-peaks were detected and confirmed using the combination of four detectors: 'gQRS' (Goldberger *et al* 2000), 'jQRS' (Behar *et al* 2014), a parabolic fitting method (Manriquez and Zhang 2007) and wavelet-based method (Martinez *et al* 2004). T-wave end points were also detected and confirmed using the combination of four detectors: 'ecgpuwave' (Laguna *et al* 1994, Goldberger *et al* 2000), a sliding window area method (Zhang *et al* 2006), a wavelet-based method (Martinez *et al* 2004) and the trapezium area method (Vazquez-Seisdedos *et al* 2011). The agreement between the R-peak and end-T-wave detectors was assessed to derive the ECG signal quality index. First, the agreement between all four R-peak detectors was measured as an F_1 score using the 'bxb' algorithm, available from PhysioNet (Goldberger *et al* 2000). Then, the R-peak detector with the lowest F_1 score was excluded. Over a 4 s window, the ECG signal quality was labeled as the F_1 score of agreement between the remaining three R-peak detectors. In windows with 100% F_1 score, ECG episodes were determined as good signal quality if all three R-peak detectors were within the 100 ms tolerance. For detected end-T-wave positions, the annotation furthest from the median of the four annotations was excluded. Then ECG episodes were determined as good signal quality if the remaining three annotations were all within the 100 ms tolerance of each other. ECG episodes corresponding to poor signal quality were excluded for training the HSMM model.

Subsequently, the four heart sound states were labeled using the reference locations of R-peak and end-T-wave as shown in figure 1. The period from each detected R-peak plus the mean S1 duration was labelled as an S1 sound. The maximum value of the Hilbert envelope of heart sound within a given window centered on the end-T-wave was marked as center of

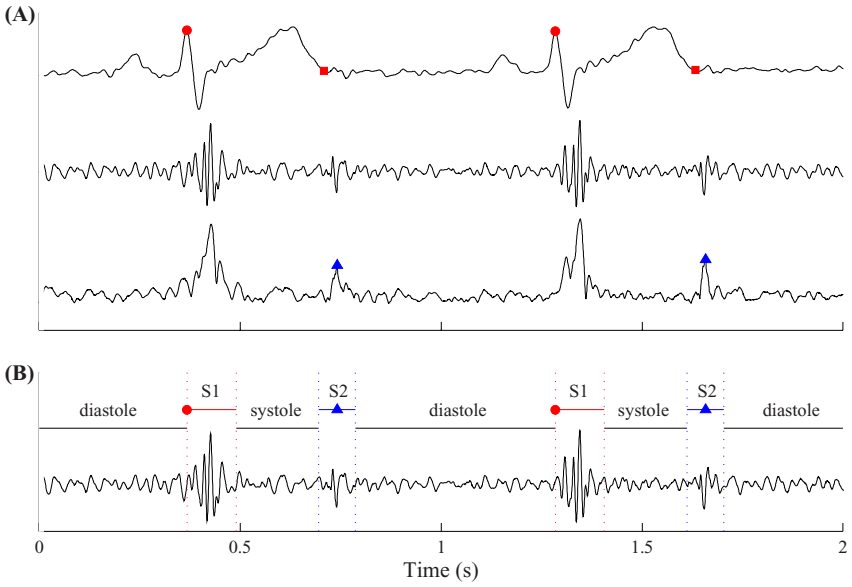

Figure 1. Demonstration of how to label the four heart sound states. (A) From top to bottom, ECG, simultaneously acquired heart sound recording, and the Hilbert envelope of heart sound were shown. The detected R-peak (circle '●') and end-T-wave (square '■') in ECG signal, as well as the locations of the maximum value from the Hilbert envelope of heart sound (triangle '▲'), were also given. (B) Heart sound states were labelled using the reference locations of R-peak (circle '●') and the maximum value from the Hilbert envelope of heart sound (triangle '▲') using the threshold parameters reported in the Schmidt's HSMM model (Schmidt *et al* 2010a).

S2 sound. The period equal to mean S2 duration, centered on this maximum position, was labelled as an S2 sound. The period between S1 and S2 was labelled as systole, and the period between S2 and S1 in next beat cycle was labelled as diastole. The mean S1 duration was set to 122 ms, the mean S2 duration was set to 92 ms, the special window was set as the mean S2 duration plus the standard deviation of S2, i.e. 114 ms. All the parameter values were reported from the Schmidt's HSMM model (Schmidt *et al* 2010a).

2.4. HSMM-based heart sound segmentation algorithm

The HSMM-based model used in this study was developed from Schmidt's HSMM approach, which is a standard HMM (i.e. $\lambda = (A, B, \pi)$) plus a probability model of the time remaining in each heart sound state, i.e. $\lambda = (A, B, \pi, p)$, where A is the transmission matrix of the four heart sound states, B is the observation distribution matrix, π is the initial heart sound state distribution and p is the probability density function of the time expected to remain in each heart sound state. Since p was added to the iteration process of B, only the state transition matrix A is Markovian. Therefore the model is referred to as a HSSM. Springer's HSMM model improved Schmidt's HSMM model in three ways: (1) it uses a logistic regression derived observation function for B matrix to replace the Gaussian distribution function; (2) it extends the Viterbi algorithm to predict the possible state durations beyond the beginning and end of the heart sound signal, to give the state durations at the boundary points; and (3) it uses a combination of four envelope features of the heart sounds for the model inputs,

Table 2. Results of evaluation metrics for the four heart sound states on the training data (training-a).

State	Database	TP	FN	FP	Se	P_+	Acc	F_1
S1	Training-a	14 277	282	307	98.1	97.9	96.0	**98.0**
Systole	Training-a	14 323	283	310	98.1	97.9	96.0	**98.0**
S2	Training-a	14 148	390	412	97.3	97.2	94.6	**97.2**
Diastole	Training-a	14 278	429	462	97.1	96.9	94.1	**97.0**

Note: tolerance parameter δ is set as 100 ms.

including the homomorphic envelope, Hilbert envelope, wavelet envelope and power spectral density envelope (Springer *et al* 2016).

The parameter settings were summarized as follows. The transmission probabilities in matrix A are initialized to 0, except for the possible transitions between successive states (i.e. S1 $\rightarrow$ systole, systole $\rightarrow$ S2, S2 $\rightarrow$ diastole and diastole $\rightarrow$ S1), which were set to 1. The initial state distribution probabilities in matrix π were set to be equal to 0.25 for all four states. The matrix B and p were trained by running the modified Viterbi algorithm over the training data. The four envelopes mentioned above were calculated and were normalized on a per-recording basis by subtracting the mean and dividing by the standard deviation of each recording. After normalization, the four envelope feature vectors are down-sampled to 50 Hz using a poly-phase anti-aliasing filter to increase the speed of computation. Their envelope feature vectors, as well as the corresponding reference annotation of the four heart sound states, were inputted into the HSMM model for training.

The trained HSMM-based model was evaluated on 11 databases in the PhysioNet/CinC Challenge 2016, i.e. training-b, training-c, training-d, training-e, training-f, test-b, test-c, test-d, test-e, test-g and test-i databases. In addition, since the reference annotations used in the training process were from ECG features and were not directly from the manually annotated heart sound states, evaluation metrics from the training data were also reported.

3. Results

3.1. Evaluation results on training data

The performance results of the training data are presented in table 2. This table illustrates all evaluation metrics for the four heart sound states when using a tolerance parameter $\delta = 100$ ms. Compared to the manually annotated results, an automatic HSMM-based algorithm with a 100 ms tolerance achieved 96.0% Acc for onset segmentations of both S1 and systole (i.e. end of S1) sound states, as well as achieved 98.0% of F_1 for both S1 and systole sound states. Meanwhile, it achieved 94.6% and 94.1% Acc for onset segmentations of S2 and diastole (i.e. end of S2) sound states, as well as achieved 97.2% and 97.0% of F_1 for these two states respectively.

3.2. Evaluation results on test data

The results on the independent test data are presented in table 3. This table illustrates the detailed evaluation metrics for the four heart sound states for each test database. A tolerance parameter $\delta = 100$ ms was also used. For onset segmentations of S1 and systole sound states, the numbers of FN beats were 2129 and 2165 respectively whereas the numbers of FP beats were much smaller, with 873 and 904 respectively. However, for onset segmentations of S2 and diastole sound states, the numbers of both FN and FP beats were at relatively high levels, with

Table 3. Results of evaluation metrics for the four heart sound states on the test data (training-b, c, d, e and f and test-b, c, d, e, g and i).

State	Database	TP	FN	FP	Se	P_+	Acc	F_1
S1	Training-b	3325	28	7	99.2	99.8	99.0	**99.5**
	Training-c	1719	89	36	95.1	97.9	93.2	**96.5**
	Training-d	822	31	20	96.4	97.6	94.2	**97.0**
	Training-e	58406	1187	453	98.0	99.2	97.3	**98.6**
	Training-f	4044	215	213	95.0	95.0	90.4	**95.0**
	Test-b	1269	0	0	100	100	100	**100**
	Test-c	805	48	19	94.4	97.7	92.3	**96.0**
	Test-d	250	10	4	96.2	98.4	94.7	**97.3**
	Test-e	26207	517	118	98.1	99.6	97.6	**98.8**
	Test-g	2048	0	0	100	100	100	**100**
	Test-i	1282	4	3	99.7	99.8	99.5	**99.7**
	Total	**100177**	**2129**	**873**	**97.9**	**99.1**	**97.1**	**98.5**
Systole	Training-b	3349	29	7	99.1	99.8	98.9	**99.5**
	Training-c	1713	95	42	94.7	97.6	92.6	**96.2**
	Training-d	818	33	22	96.1	97.4	93.7	**96.7**
	Training-e	58546	1196	459	98.0	99.2	97.3	**98.6**
	Training-f	4050	227	223	94.7	94.8	90.0	**94.7**
	Test-b	1275	0	0	100	100	100	**100**
	Test-c	803	51	22	94.0	97.3	91.7	**95.7**
	Test-d	249	10	4	96.1	98.4	94.7	**97.3**
	Test-e	26286	519	121	98.1	99.5	97.6	**98.8**
	Test-g	2063	0	0	100	100	100	**100**
	Test-i	1280	5	4	99.6	99.7	99.3	**99.6**
	Total	**100432**	**2165**	**904**	**97.9**	**99.1**	**97.0**	**98.5**
S2	Training-b	3268	71	47	97.9	98.6	96.5	**98.2**
	Training-c	1589	218	163	87.9	90.7	80.7	**89.3**
	Training-d	796	45	32	94.6	96.1	91.2	**95.4**
	Training-e	57566	1922	1187	96.8	98.0	94.9	**97.4**
	Training-f	3993	274	269	93.6	93.7	88.0	**93.6**
	Test-b	1258	0	0	100	100	100	**100**
	Test-c	752	101	71	88.2	91.4	81.4	**89.7**
	Test-d	243	16	9	93.8	96.4	90.7	**95.1**
	Test-e	25843	830	425	96.9	98.4	95.4	**97.6**
	Test-g	2041	0	0	100	100	100	**100**
	Test-i	1249	34	32	97.3	97.5	95.0	**97.4**
	Total	**98598**	**3511**	**2235**	**96.6**	**97.8**	**94.5**	**97.2**
Diastole	Training-b	3415	77	50	97.8	98.6	96.4	**98.2**
	Training-c	1603	226	173	87.6	90.3	80.1	**88.9**
	Training-d	820	50	38	94.3	95.6	90.3	**94.9**
	Training-e	58310	1919	1190	96.8	98.0	94.9	**97.4**
	Training-f	4012	274	271	93.6	93.7	88.0	**93.6**
	Test-b	1314	0	0	100	100	100	**100**
	Test-c	765	98	69	88.6	91.7	82.1	**90.2**
	Test-d	256	18	11	93.4	95.9	89.8	**94.6**
	Test-e	26174	824	423	96.9	98.4	95.5	**97.7**
	Test-g	2094	0	0	100	100	100	**100**
	Test-i	1265	34	32	97.4	97.5	95.0	**97.5**
	Total	**100028**	**3520**	**2257**	**96.6**	**97.8**	**94.5**	**97.2**

Note: tolerance parameter δ is set as 100 ms.

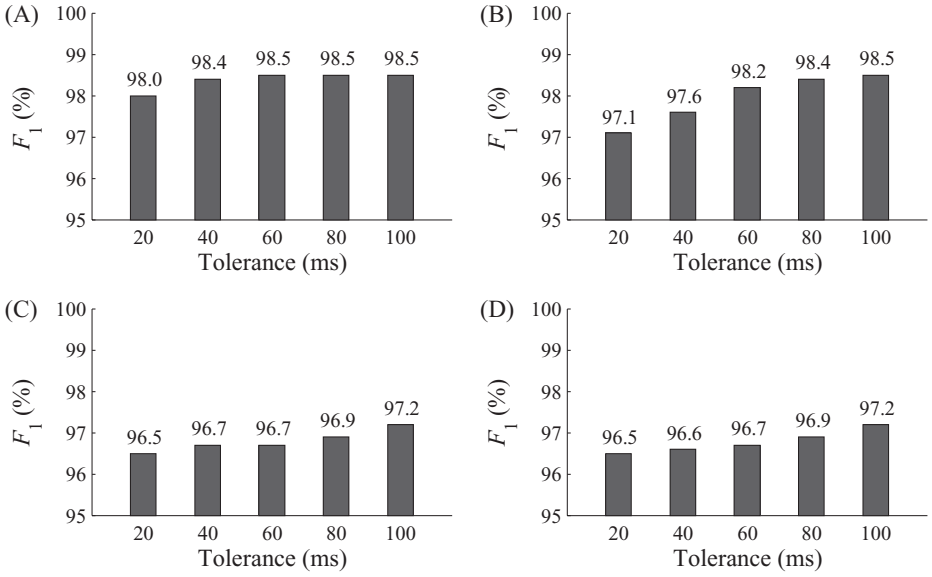

Figure 2. Effect of tolerance parameter δ on the evaluation metric of F_1 from the test data. Four sub-figures show the results for segmenting each of the four heart sound states: (A) S1, (B) systole, (C) S2 and (D) diastole.

3511 and 3520 for FN respectively and with 2235 and 2257 for FP respectively. The HSMM-based algorithm achieved an Acc of 97.1%, 97.0%, 94.5% and 94.5% for S1, systole, S2 and diastole respectively, and achieved an F_1 of 98.5%, 98.5%, 97.2% and 97.2% respectively.

3.3. Effect of tolerance on evaluation metric of F_1

Figure 2 shows the effect of tolerance parameter δ on the evaluation metric of F_1 for the test data. As expected, with the increases of the tolerance δ, the F_1 score increases. When using 20, 40, 60, 80 and 100 ms tolerance respectively, the F_1 scores were found to be 98.0%, 98.4%, 98.5%, 98.5% and 98.5% for S1 onset segmentation, 97.1%, 97.6%, 98.2%, 98.4% and 98.5% for systole onset segmentation, 96.5%, 96.7%, 96.7%, 96.9% and 97.2% for S2 onset segmentation and 96.5%, 96.6%, 96.7%, 96.9% and 97.2% for diastole onset segmentation respectively. The F_1 score is higher than 96% for any choice of tolerance, δ.

4. Discussion

Heart sound segmentation is a prerequisite for the accurate automatic analysis of heart sound recordings. In this study, we evaluated the performance of an open-source HSMM-based segmentation algorithm on a large collection of heart sound data, both made freely available for the PhysioNet/CinC Challenge 2016. The results showed that it has a high accuracy for segmenting heart sound data over multiple data sets, as demonstrated by the high average F_1 scores of 98.5% for segmenting S1 and systole intervals, and 97.2% for segmenting S2 and diastole intervals, using a tolerance of 100 ms. The effect of tolerance parameter on evaluation metrics was evaluated and the F_1 score was shown to increase with longer values of the tolerance as expected.

The traditional time-domain or frequency-domain segmentation methods face challenging for processing signals from multiple sources (Liu *et al* 2016). Spectral properties of heart sounds, as well as possible noise sources, have been well described in Leatham (1975). The typical frequency ranges of different components are: S1 for 10–140 Hz (energy concentration usually in low frequencies of 25–45 Hz), S2 for 10–200 Hz (energy concentration usually in low frequencies of 55–75), S3 and S4 for 20–70 Hz, murmurs can be as high as 600 Hz, respiration for 200–700 Hz (Tilkian and Conover 2001, Liu *et al* 2016). The FHS components that need to be segmented, i.e. S1 and S2, overlap with many noise sources in the frequency domain, which leads to difficulty in separation between heart sounds from abnormal sounds or artifacts using traditional frequency-domain analysis. Moreover, the morphological similarity of the noise to the FHSs makes identification of the latter also extremely difficult using time-domain techniques.

Unlike the traditional time-domain or frequency-domain analysis, the HSMM-based algorithm utilized the inherent probabilistic estimation of the duration-dependent Markov model, as well as combined the logistic regression for emission probability estimation and a modified Viterbi algorithm (Springer 2015a, Springer *et al* 2016), to achieve a strong generalization ability and robustness for segmenting both clean and noisy heart sounds. In clinical practice, the recorded heart sounds could be from both healthy subjects and patients with valvular heart disease, arrhythmia, pulmonary diseases and obesity, and the recordings can be contaminated with both endogenous physiological and exogenous background noise. This is particularly important in real world clinical situations. The probability estimation-based HSMM algorithm was a better method for heart sound segmentation than the traditional time-domain or frequency-domain methods, especially in processing noisy recordings. Figure 3 shows several successful examples of automatic heart sound segmentation. Sub-figure 3(A) shows an example of segmenting a clean heart sound. Sub-figure 3(B) shows the HSMM-based algorithm can ignore large-amplitude artifacts, indicated by the arrows. Sub-figure 3(C) shows the accurately segmenting in an example of loud background noises. Sub-figure 3(D) shows the successful segmentation of a recording with both large-amplitude artifacts and loud background noise. The HSMM-based method is not only able to accurately segment the clean signals, but can also accurately deal with the signals with large-amplitude artifacts (as shown in 3(B)) or with loud noise (as shown in 3(C)), or with both (as shown in 3(D)). In addition, the trained model could be applied directly to the heart sound for segmentation without any need of prior information, such as ECG-derived timings.

Although the HSMM-based algorithm was able to successfully segment the majority of evaluated heart sound recordings, it still fails to segment accurately in some cases. Figure 4 shows typical errors. Sub-figure 4(A) shows an example of FN segmentation, where the algorithm only detected one of the S1 and S2 components in each heart cycle. In this case, there was large variation in the heart rate over a short period. Sub-figure 4(B) shows another example of inaccurate segmentation, where only one FN beat appeared, as indicated by the arrows. Sub-figure 4(C) shows an example of FP segmentation, where a long heart cycle (indicated by the arrow) appeared after an extremely short heart cycle and the algorithm incorrectly detected a beat during this long heart cycle. This is a typical heart sound waveform with arrhythmia. It is clear that the waveform in the extremely short heart cycle has an obvious morphological change due to the ectopic beat. Sub-figure 4(D) shows another example of a FP segmentation indicated by the arrow. Although the FP segmentation is also due to the relatively long heart cycle, there is no obvious morphological change in heart sound waveform. This relatively long heart cycle is due to the irregular sinus rhythm. We identify the improvement for the HSMM-based method to successfully segment the cases reported in figure 4 as our future work.

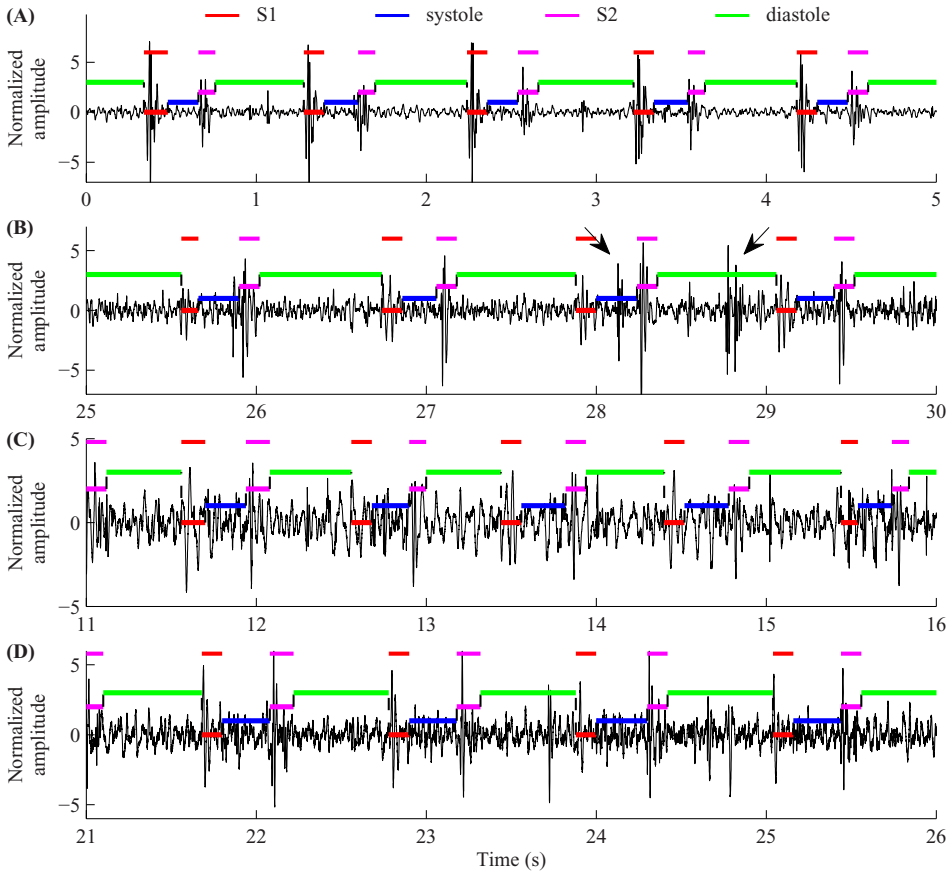

Figure 3. Heart sound signals for the recordings (A) a0109, (B) a0353, (C) a0073 and (D) a0044 taken from the PhysioNet/CinC Challenge 2016, together with successful segmentations using the HSMM-based segmentation algorithm under evaluation. The four automatically segmented heart sound states (S1, systole, S2 and diastole) are shown on the heart sound waveforms as a series of steps. The manually annotated S1 and S2 intervals are shown at the top of each sub-figure (with systole and diastole omitted for clarity).

The high accuracy of the HSMM-based algorithm can be attributed to several factors. First and most importantly, as discussed before, it uses the inherent probabilistic estimation of the duration-dependent Markov model. Additionally, the algorithm combines a logistic regression for emission probability estimation and a modified Viterbi algorithm to achieve strong generalization ability and robustness for segmenting both clean and noisy heart sounds (Springer 2015a, Springer *et al* 2016). Although the algorithm was trained using reference of ECG features, it can still achieve high segmentation accuracy when compared with the manually annotated heart sound onsets (see table 2), indicating that the inherent physiological model built into the method holds true across the large set of independent databases used in this work. Secondly, in the current study only the heart sound data which passed the signal quality assessment were used. The noisy recordings and noisy episodes, for which it was impossible to manually annotate the locations of the heart sounds, were excluded from the evaluation. This reduces the risk of false segmentations. However, this does not imply that the algorithm

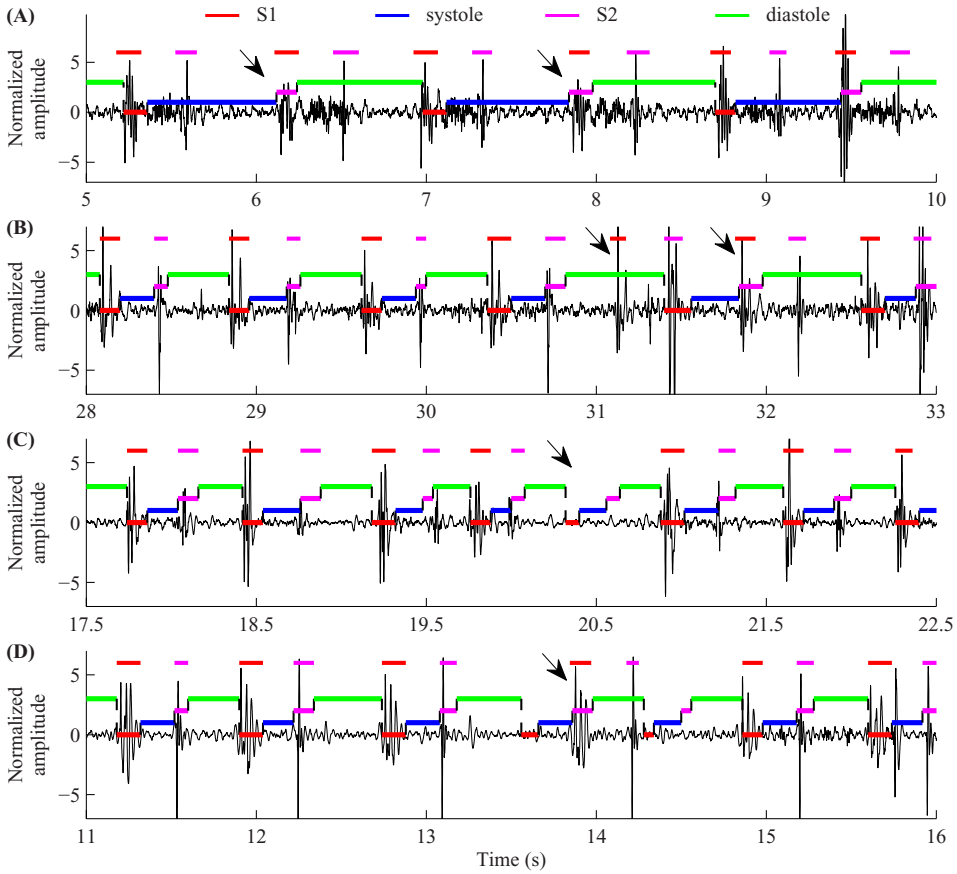

Figure 4. Examples of segmentation failures for the HSMM-based algorithm (indicated by arrows). Heart sound signals for the recordings (A) a0031, (B) a0112, (C) a0284 and (D) a0352 taken from the PhysioNet/CinC Challenge 2016, together with successful segmentations using the HSMM-based segmentation algorithm under evaluation. The four automatically segmented heart sound states (S1, systole, S2 and diastole) are shown on the heart sound waveforms as a series of steps. The manually annotated S1 and S2 intervals are shown at the top of each sub-figure (with systole and diastole omitted for clarity).

needs good quality signals to perform well. In fact, figure 3 has demonstrated that the algorithm can also work well for the noisy recordings, as long as the heard sound signal can be visually segmented into the four states. Thirdly, the heart sound data used in this study were collected from multiple independent databases, for which the qualities and numbers of recordings varied highly. As shown in table 3, both training-e (59 593 beats) and test-e (26 724 beats) databases were from the same data contributor, and they accounted for about 85% of total tested data (102 597 beats). However, the majority of signals in these two databases exhibited little noise and the fundamental S1 and S2 components were clearly identifiable, which significantly contributes to the high segmentation accuracy. Lastly, the average accuracy in test data as shown in table 3 was even higher than that in training data as shown in table 2, illustrating that the HSMM-based model was not over-trained on the training data (training-a). This again highlights that many of the recordings in the training-e and test-e databases exhibited

good signal quality, but confirms its strong generalization ability. It is also worth noting that the F_1 metric rather than the Acc metric was used for the final evaluation. This is because the Acc metric does not provide an adequate representation of the results since it does not account for true negatives.

Limitations exist in the current study. First, some recordings were excluded as 'noisy' recordings. The majority of them are due to poor transducer contact with the skin, or the presence of large amplitude noise. However, some of them are due to the significant murmurs. The reason for exclusion is that both the large amplitude noise and significant murmurs make it impossible to perform manual annotation for the four heart sound states, using either visual inspection or auscultation. Second, in the current study we focused on the evaluation of the HSMM-based algorithm using the large collection of multiple source heart sound data from the PhysioNet/CinC Challenge 2016. Since we designed this challenge and only the HSMM-based algorithm has ever been run on all the data, we have no comparison with other algorithms. We leave this enormous task to other authors, providing the framework herein. Third, as can be seen from figure 4, detection errors exist for the HSMM-based algorithm, especially in the situations of long heart cycles and irregular sinus rhythm. These two key issues are left as future improvements of the HSMM-based segmentation algorithm.

In summary, in the current study we constructed a systematic framework for evaluating heart sound segmentation methods based on the PhysioNet/CinC Challenge 2016 data and evaluated the performance of the open-source HSMM-based algorithm. (For more information about this algorithm, see the released software package on PhysioToolkit: https://physionet.org/physiotools/hss/.) The high segmentation accuracy on the large test database confirmed its effectiveness. The described systematic framework, combined with a significant amount of open data, facilitates a fair comparison between researchers and industry alike, and provides essential resources for entities who need to test their algorithms with realistic data and share reproducible results.

Acknowledgments

We wish to thank the many kind donors of the heart sound databases described in Liu *et al* (2016). This work was supported by the Rhodes Trust, Emory University and the International Postdoctoral Exchange Programme of the National Postdoctoral Management Committee of China.

References

American National Standards Institute 2012 Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms: ANSI/AAMI EC57

Ari S, Kumar P and Saha G 2008 A robust heart sound segmentation algorithm for commonly occurring heart valve diseases *J. Med. Eng. Technol.* **32** 456–65

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

Chen T, Kuan K, Celi L and Clifford G D 2009 Intelligent heartsound diagnostics on a cellphone using a hands-free kit *AAAI Spring Symp. on Artificial Intelligence for Development* (Stanford University) pp 26–31

Clifford G D, Liu C Y, Moody B, Springer D B, Silva I, Li Q and Mark R G 2016 Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016 *Computing in Cardiology* **43** 609–12

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

Gamero L G and Watrous R 2003 Detection of the first and second heart sound using probabilistic models *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (Cancun: IEEE) pp 2877–80

Gill D, Gavrieli N and Intrator N 2005 Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model *Computers in Cardiology* (Lyon: IEEE) pp 957–60

Goldberger A L, Amaral L A N, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals *Circulation* **101** e215–20

Gupta C, Palaniappan R, Swaminathan S and Krishnan S 2007 Neural network classification of homomorphic segmented heart sounds *Appl. Soft Comput.* **7** 286–97

Jiang Z and Choi S 2006 A cardiac sound characteristic waveform method for in-home heart disorder monitoring with electric stethoscope *Expert Syst. Appl.* **31** 286–98

Kumar D, Carvalho P, Antunes M and Henriques J 2006 Detection of S1 and S2 heart sounds by high frequency signatures *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (New York: IEEE) pp 1410–6

Laguna P, Jane R and Caminal P 1994 Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database *Comput. Biomed. Res.* **27** 45–60

Leatham A 1975 *Auscultation of the Heart and Phonocardiography* (London: Churchill Livingstone)

Liang H, Lukkarinen S and Hartimo I 1997 Heart sound segmentation algorithm based on heart sound envelolgram *Computing in Cardiology* (Lund: IEEE) pp 105–8

Liu C Y *et al* 2016 An open access database for the evaluation of heart sound algorithms *Physiol. Meas.* **37** 2181–213

Manriquez A I and Zhang Q 2007 An algorithm for QRS onset and offset detection in single lead electrocardiogram records *29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (Lyon: IEEE) pp 541–4

Martinez J P, Almeida R, Olmos S, Rocha A P and Laguna P 2004 A wavelet-based ECG delineator: evaluation on standard databases *IEEE Trans. Biomed. Eng.* **51** 570–81

Moukadem A, Dieterlena A, Hueberb N and Brandtc C 2013 A robust heart sounds segmentation module based on S-transform *Biomed. Signal Process. Control* **8** 273–81

Naseri H and Homaeinezhad M R 2013 Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric *Ann. Biomed. Eng.* **41** 279–92

Nigam V and Priemer R 2005 Accessing heart dynamics to estimate durations of heart sounds *Physiol. Meas.* **26** 1005–18

Oskiper T and Watrous R 2002 Detection of the first heart sound using a time-delay neural network *Computing in Cardiology* (Memphis, TN: IEEE) pp 537–40

Papadaniil C D and Hadjileontiadis L J 2014 Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features *IEEE J. Biomed. Health Inform.* **18** 1138–52

Pedrosa J, Castro A and Vinhoza T T V 2014 Automatic heart sound segmentation and murmur detection in pediatric phonocardiograms *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (Chicago, IL: IEEE) pp 2294–7

PhysioNet/CinC Challenge 2016 2016 Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016 http://physionet.org/challenge/2016/

Rajan S, Budd E, Stevenson M and Doraiswami R 2006 Unsupervised and uncued segmentation of the fundamental heart sounds in phonocardiograms using a time-scale representation *Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (New York: IEEE) pp 3732–5

Ricke A D, Povinelli R J and Johnson M T 2005 Automatic segmentation of heart sound signals using hidden Markov models *Computers in Cardiology* (Lyon: IEEE) pp 953–6

Schmidt S E, Holst-Hansen C, Graff C, Toft E and Struijk J J 2010a Segmentation of heart sound recordings by a duration-dependent hidden Markov model *Physiol. Meas.* **31** 513–29

Schmidt S E, Toft E, Holst-Hansen C and Struijk J J 2010b Noise and the detection of coronary artery disease with an electronic stethoscope *2010 5th Cairo Int. Biomedical Engineering Conf. (CIBEC)* (Cairo: IEEE) pp 53–6

Sedighian P, Subudhi A W, Scalzo F and Asgari S 2014 Pediatric heart sound segmentation using Hidden Markov Model *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (Chicago, IL: IEEE) pp 5490–3

Sepehri A A, Gharehbaghi A, Dutoit T, Kocharian A and Kiani A 2010 A novel method for pediatric heart sound segmentation without using the ECG *Comput. Methods Programs Biomed.* **99** 43–8

Silva I, Behar J, Sameni R, Zhu T, Oster J, Clifford G D and Moody G B 2013 Noninvasive fetal ECG: the PhysioNet/Computing in Cardiology Challenge 2013 *Computing in Cardiology* (Zaragoza: IEEE) pp 149–52

Springer D B 2015a Heart sound segmentation code based on duration-dependant HMM https://github.com/davidspringer/Springer-Segmentation-Code

Springer D B 2015b Mobile phone-based rheumatic heart disease detection *Department of Engineering Science* (Oxford: University of Oxford)

Springer D B, Tarassenko L and Clifford G D 2016 Logistic regression-HSMM-based heart sound segmentation *IEEE Trans. Biomed. Eng.* **63** 822–32

Sun S, Jiang Z, Wang H and Fang Y 2014 Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform *Comput. Methods Programs Biomed.* **114** 219–30

Tang H, Li T, Qiu T S and Park Y 2012 Segmentation of heart sounds based on dynamic clustering *Biomed. Signal Process. Control* **7** 509–16

Tilkian A G and Conover M B 2001 *Understanding Heart Sounds and Murmurs with an Introduction to Lung Sounds* (New York: Elsevier)

Varghees V N and Ramachandran K 2014 A novel heart sound activity detection framework for automated heart sound analysis *Biomed. Signal Process. Control* **13** 174–88

Vazquez-Seisdedos C R, Neto J E, Reyes E J M, Klautau A and de Oliveira R C L 2011 New approach for T-wave end detection on electrocardiogram: performance in noisy conditions *Biomed. Eng. Online* **9** 77

Vepa J, Tolay P and Jain A 2008 Segmentation of heart sounds using simplicity features and timing information *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (Las Vegas, NV: IEEE) pp 469–72

Wang P, Lim C S, Chauhan S, Foo J Y and Anantharaman V 2007 Phonocardiographic signal analysis method using a modified hidden Markov model *Ann. Biomed. Eng.* **35** 367–74

Yan Z, Jiang Z, Miyamoto A and Wei Y 2010 The moment segmentation analysis of heart sound pattern *Comput. Methods Programs Biomed.* **98** 140–50

Zhang Q, Manriquez A I, Medigue C, Papelier Y and Sorine M 2006 An algorithm for robust and efficient location of T-wave ends in electrocardiograms *IEEE Trans. Biomed. Eng.* **53** 2544–52