Time and Frequency -Based Approach to Heart Sound Segmentation and Classification

Jarno Mäkelä[1,2], Heikki Väänänen[1,2]
[1] RemoteA Ltd, Helsinki, Finland
[2] Aalto University, Espoo, Finland

Abstract

In this study, we propose a decision tree classifier of heart sound signals.

We determined repetitive fundamental heart sound segments based on adaptive similarity value clusterization of the sound signal, and we created a set of filters for decision tree parametrization. Using the filters together with inter-segment timings, we created three sets of markers: a set utilizing both S1 and S2 identification, a set where only one segment was identified, and a set without any identified segment. An individual classification tree was trained for each marker set.

As a result, our classifier attained sensitivity (Se) of 0.66 and specificity (Sp) of 0.92 and overall score of 0.79 for a hidden random (revised) subset.

1. Introduction

Phonocardiograph (PCG) is a recording of the sounds made by the heart. In this study we describe our approach to the Physionet/Computing in Cardiology Challenge 2016 [1]. Our aim was first to reliably identify the most distinct and repeating fundamental heart sound (FHS) segments [2], and then to analyze the energy content of the signal during FHSs and between them on multiple frequency bands.

2. Methods

We divided our approach into a preliminary step and three independently and interdependently iterable steps. The preliminary step was data analysis consisting of electrocardiography (ECG) based model of fundamental heart sound detection from PCG, and introducing PCG-based interdata energy models (convolution kernels or FIR filters) for a proper FHS detection. In the first step a FHS-based highly correlating and consistent energy model detection and decomposition method was introduced. In the second step several PCG markers were defined. In the third step an entropy model based decision tree was formed.

2.1. Electrocardiography-based event detection and domain translation

Utilizing ECG data from set *training-a*, QRS complexes were extracted with strict criteria so that only high-quality events were accepted.

ECG-based events were extracted with a nonlinear windowed and smoothed peak-valley detector and a similar clusterization process we describe in this study as used for FHS clusterization. ECG-based QRS complex detections were translated into PCG domain as potential locations for S1 type FHSs. The actual S1 segment was selected by finding the most common 250 ms window from PCG within 250 ms of the ECG-based QRS complex energy maximum. This was achieved by calculating mutual event similarity values in moving 250 ms window. The similarity value (*sim*) was defined as:

$$sim(A,B) = \begin{cases} \dfrac{cov(A,B)}{cov(A,A)}, cov(A,A) \geq cov(B,B) \\ \dfrac{cov(A,B)}{cov(B,B)}, cov(A,A) < cov(B,B) \end{cases} \quad (1)$$

, where A and B are windowed events, *cov*(x,y) is the covariance of x and y. By aligning the mutual event similarity value maxima on a single event, a per similarity value kernel time domain normalization factor was found. By combining the time domain normalized similarity value maxima, an S1 event was defined. The combining criteria included a stop criterion that only correlating enough kernels were used. The S1 events were ranked by multiplying the count of used kernels with the average of local similarity value maximal values.

The same approach was repeated for S2 with the assumption that S2 is in [QRS+0.4*RR/2, QRS+0.5*RR/2], where RR is the local QRS interval.

2.2. Time-frequency surfacing using limited time discrete Fourier transform

Utilizing the FHS event input from the previous step, and later from other iterative sources and the reference

ISSN: 2325-887X DOI:10.22489/CinC.2016.166-530

event set, time-frequency surfaces were formed. For each sample in data in [S1-200 ms, S1+800 ms] and [S2-200 ms to S2+800 ms], a discrete Fourier transform in a 250 ms wide SRS flattop window was calculated:

$$w(n) = 1.0 - 1.93 * \cos(\frac{2\pi n}{N-1}) + 1.29 * \cos(\frac{4\pi n}{N-1}) - 0.388 * \cos(\frac{6\pi n}{N-1}) + 0.028 * \cos(\frac{8\pi n}{N-1}) \quad (2)$$

, where *N* is the window width in samples.

Events within 500 ms to data boundaries were discarded. From each of these per data transforms both an average transform surface (*avg*) and standard deviation (*stdev*) were stored. The stored transforms were combined as representative surfaces in three patient group contexts: all data and data classified either as normal or abnormal. The aim of the representative surface was to denote the interdata concurrent frequency components near FHS segments. This was achieved by normalization:

$$avg_{com}(t,f) = \frac{\sum_{n=1}^{N} avg_n(t,f) * \frac{1}{stdev_n(t,f)}}{\sum_{i=1}^{N} \frac{1}{stdev_n(t,f)}} \quad (3)$$

, where N is the total number of data. The resulting surfaces (avg_{com}) were interpreted per training set to verify the differences between normalities and abnormalities. The deviation value was used as an inverse quality marker.

A set of convolution kernels was constructed by subtracting the average baseline frequency components from the frequency components at both FHSs (S1, S2) and by committing inverse Fourier transform for each subtracted set. These convolution kernels represented concurrent average energy signatures of S1 and S2 in each of the three patient group contexts.

2.3. Phonocardiograph event detection and decomposition

The actual PCG event detection (trigger) consisted of two sequential steps: an energy norm based trigger and a concurrent segment extractor. Our energy norm was defined by first filtering data with predefined convolution kernel. The resulting translated signal (X_f) was then nonlinearly squared utilizing the Blackman window function $w_{blackman}$ [3]:

$$X_e(x) = \sum_{i=0}^{N-1} (X_f(x - N/2 + i) * w_{\text{blackman}}(i))^2 \quad (4)$$

, where *x* is a data sample, *N* is the desired window length and X_e is the resulting energy signal.

For detecting preliminary events, a per-data threshold value was defined by arranging local energy minimums and maximums and by selecting the nth most representing value as limit. After this, local maximums above the threshold value were chosen as preliminary events.

To extract FHS segments, similarity values (see Equation 1) for each preliminary event were calculated in a 250 ms wide window in the same way as described in the ECG-based event detection. A preliminary set of clusters was formed by calculating the maximum similarity value near each preliminary event. If similarity value was above a static cut-off limit, the event was assigned to a preliminary cluster. The preliminary clusters were ordered by the average similarity value of their assigned events. The less similar half of preliminary clusters was discarded. The resulting set of preliminary clusters was trialed in order to form final clusters. Adaptive criteria was introduced: first, a cluster must consist of at least three events; second, the time-bias-corrected similarity value of the events must remain above 0.8; third, an event can belong only to one cluster. The preliminary clusters were tried until either all preliminary clusters were used or both FHS segment clusters S1 and S2 were found. If all preliminary clusters were used without a clear distinction between S1 and S2, the data was labeled to contain only a single heart sound segment cluster (*ss*). The certainty was defined by dynamical intracluster minimal event similarity value limit which is always above 0.8 and above 0.95 in most cases. If no cluster reached the certainty limit, the data was labeled untriggered (*remainder*). If two clearly distinguishable clusters were found, they were labeled as S1 and S2 (*s1s2*). This is if the shortest distance (*d*) between an event in one cluster and an event in the other cluster must be in a reasonable window:

$$0.2 * SS < d < 0.8 * SS, d < 500 \quad (5)$$

, where *SS* is the shortest distance in milliseconds within an event cluster. In addition it was verified that the clusters didn't contain any events within event distance less than 1.25 times *d*. If no such event was found, the preceding cluster was labeled as a S1 segment and the subsequent cluster as a S2 one.

2.4. PCG markers

Using the set of filters and inter-segment timings, we created three sets of markers: a set utilizing both S1 and S2 identification (*s1s2*), a set where only a single heart sound (S, assumed to be either S1 or S2) is identified (*ss*), and a set without any detected repetitive heart sound (*remainder*). Set sizes are shown in Table 1.

The markers in the *s1s2* set utilized five data segments per cardiac cycle: s1: [S1-100 ms, S1+100 ms], s2: [S2-100 ms, S2+100 ms], s1_s2: [S1-150 ms, S1+SS-150 ms], s2_s1 [S2+150 ms, S2+SS-d-150 ms] and base [S1-125 ms, S1-75 ms], where S1 and S2 are the times of the identified S1 and S2 respectively, and the *SS* and *d* the minimal *SS* and *s1s2* distances as already described. For the *ss* set three segments per FHS were defined: s: [S-100 ms, S+100 ms], s_s [S+150 ms, S+SS-150 ms] and base [S1-125 ms, S1-75 ms]. In *remainder* set, the signal was divided into sequential 3 second segments (seg).

After defining the segments, the minimum and maximum standard deviation values in 100 ms windows (STD) were calculated for each segment. The median values over all the similar segments (over all the beats in *s1s2* and *ss* sets, and over all 3 second windows in *remainder* set) defined the final min and max estimates. The procedure of these min and max estimates was then repeated after filtering the data with a four pole Chebyshev bandpass filters in varying bandwidths.

All the min, max and max-min estimates were used as markers (ABS), and the differences of all but the base-markers and the base-markers were calculated to produce the baseline corrected markers (CORR). In addition we created the normalized markers (NORM), by dividing all the frequency limited (filtered) markers by the otherwise similar, but not filtered markers and the relative markers (REL), where other markers were divided with the s1, s2 and s markers. Finally we created a set of time based width markers (WIDTH) for each of the FHSs - S1, S2, and S (for the *ss* set) - by defining the samples where the STD dropped below 30, 40, 50, 60 or 70 % the maximum value before and after each FHS, and by taking again the median value over all the similar FHSs.

This resulted in a total of 3818 markers for the *s1s2* set, 1811 markers for the *ss* set and 429 markers for the *remainder* set.

2.5. Entropy-based decision tree classifier

We used binary tree classifier where each node uses one PCG marker and limit value. The marker and limit value combinations were selected by minimizing the split entropy values [4], which were normalized by adding the weight on abnormal recordings so that the sum weights over the normal and abnormal sets were equal. The classifier was trained initially for discriminating first set *training-f*, and then set *training-b* from any other set (*trunk*). At this phase all the data was used, and only the marker set *remainder* was utilized. After that the training was continued for separating the abnormal from normal cases in three separate trees with full available marker sets: one tree for *s1s2*, *ss*, and *remainder* sets each. The training was continued until predefined stop criterion was reached: either there were 16 times more weighted hits from one set than from another, or the total number of weighted hits was less than 1 % of the total weighted data size. The stop criteria was defined by finding the best k-fold cross-validation values in the training set. The search for the minimal split entropy value for each node was done simply by testing all the marker - value combinations.

The final classifiers had all two nodes for set separation (*trunk*: nodes 0 and L in Table 2) and 14 nodes in *s1s2* tree, 26 nodes in *ss* tree and 36 nodes in *remainder* tree for classifying the abnormal from normal. 57 % of the leafs were classified as normal, 41 % as abnormal, and 2 % as unknown - a leaf was defined as unknown, if there were less than 3 times more weighted hits in either of the sets (normal or abnormal). See the most important nodes in Table 2.

3. Results

After several iterations we achieved a score of 0.79 (*Se* = 0.69 and *Sp* = 0.92) on a hidden test subset.

The separation capability of our FHS segment extractor was found sufficient as seen in Table 3, where accuracy of the event extractor is compared to the Challenge provided hand corrected reference annotations. If we discard events detected within the ranges that were defined as noise in the reference, we achieved a total FHS detection accuracy of 99.8 %, if *s1s2* and *ss* sets were combined so that an extracted event is aligned with either S1 or S2 reference. By observing the consistency of segmentation within data in the *ss* set, we achieved an inconsistency rate of 2.46 % meaning that in that fraction of data the selected marker did not separate S1 from S2 consistently but selected both in an inconsistent fashion.

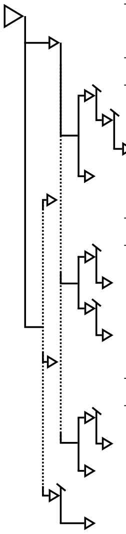
Table 1. Distribution of Challenge Phase II training set data in heart sound decomposition sets by patient groups.

Event set	Normals	Abnormals	Total
s1s2	1402	203	1605
ss	761	292	1053
remainder	325	170	495

Table 3. Event detection precision (Prec) compared to reference events. Prec is defined as TP / (S1+S2+FP), where TP is S1 for s1, S2 for s2 and S1 or S2 for ss. Noise is the number events detected inside reference noise annotations. False positive (FP) reflects the detected events that don't match any reference event.

Event	S1	S2	Noise	FP	Prec
s1	21793	413	417	14	0.981
s2	611	17369	657	76	0.962
ss	8976	516	517	19	0.998

Table 2. The most important nodes in the three classifier trees. WHAT, WHERE, TO and HOW are type descriptors for node parameters. $\mathrm{Freq}_{\mathrm{low}}$ and $\mathrm{freq}_{\mathrm{high}}$ are the filter bandpass frequencies. Limit is the binary limit of the left and right separator. N and Abn are abbreviations for normal and abnormal patient groups. See the text for more accurate description.



	WHAT	WHERE	TO	HOW	$\mathrm{freq}_{\mathrm{low}}$	$\mathrm{freq}_{\mathrm{high}}$	limit	N left	Abn left	N right	Abn right
trunk											
-	NORM	seq	-	max	300	500	0.154				
L	NORM	seq	-	min	0	25	0.596				
s1s2											
LL	NORM	s2s1	-	min	100	200	0.157	141	153	1179	16
LLL	REL	s2s1	s1	minmax	400	600	0.502	62	3	79	150
LLLR	ABS	s1s2	-	min	750	850	0.586	63	53	16	97
LR	NORM	seg	-	min	25	50	0.275	48	0	15	31
R	ABS	s2s1	-	all	-	-	133.5	19	0	0	3
ss											
LL	NORM	ss	-	min	75	100	1.33	69	191	484	15
LLL	REL	ss	s	minmax	500	700	0.486	24	2	45	189
LR	NORM	seg	-	min	25	50	0.256	119	13	50	60
LRR	WIDTH	s	-	50%	125	150	0.052	47	22	3	38
R	ABS	seq	-	min	125	150	0.172	29	1	10	12
remainder											
LL	NORM	seg	-	min	87.5	112.5	0.145	31	68	27	1
LLL	ABS	seg	-	minmax	0	25	442	12	63	19	5
LR	ABS	seg	-	min	350	400	0.601	220	63	0	8
R	NORM	seg	-	min	12.5	37.5	0.218	21	0	26	30
RR	NORM	seg	-	min	225	250	0.020	25	15	1	15

4. Discussion

Our approach was based on methods we had implemented for other signal domains, and we were pleased to confirm that our event detection algorithm was accurate also in PCG context. The main implementation effort was porting the actual software code from our heterogenous computing environment (OpenCL) to the PhysioNet Challenge Entry setup. We think it is fair to state that many of our operations in preliminary studies would have been very frustrating to run without the capability to immerse parallel computing.

We had only limited success in balancing the data sets, and we were surprised by the notable differences between the sets. Our markers seemed more capable in separating the data sets from each other, than in separating the normal patient group from the abnormal one: Single node separated the set *training-f* with 99.2 % accuracy and the set *training-b* with the 96.3 % accuracy from the other sets and we ended up using these nodes also as the first nodes in the final classification tree. We tested also several balancing setups, where we weighted each data set based on their number of patients per number of recordings ratio, but those didn't seem to improve our classifier accuracy.

To conclude, we feel that our weakest point was too simple classifier approach. The binary tree used only one marker at a time, and it would have required more advanced markers that were less influenced by measurement setup. Reliable patient identification with PCG data from multiple sources is likely to require other than only time-frequency based approach, or at least substantial adjustments to balance the measurement position and device-specific differences.

References

[1] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AE, Syed Z, Schmidt SE, Papadaniil CD, Hadjileontiadis L, Naseri H, Moukadem A, Dieterlen A, Brandt C, Tang H, Samieinasab M, Samieinasab MR, Sameni R, Mark RG, Clifford GD. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(9).

[2] Leatham A. Auscultation of the Heart and Phonocardiography. Second edition. Churchill Livingstone, 1975.

[3] Blackman R. B. TJW. The measurement of power spectra. Dover Publications, 1958.

[4] Pavlopoulos SA Stasis AC LE. A decision treebased method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. Biomedical engineering online 2004;1(3).

Address for correspondence:

Jarno Mäkelä
RemoteA Ltd
Lars Sonckin kaari 10-16
02600 Espoo, Finland
jarno.makela@remotea.com