Time-Frequency Analysis of Phonocardiogram for Classifying Heart Disease

Rohan Banerjee[1], Swagata Biswas[1], Snehasis Banerjee[1], Anirban Dutta Choudhury[1], Tanushyam Chattopadhyay[1], Arpan Pal[1], Parijat Deshpande[1] and Kayapanda M Mandana[2]

[1] Research and Innovations, Tata Consultancy Services Ltd.
[2] Fortis Hospital, Kolkata, India

Abstract

Analysis of heart sounds is a popular research area for non invasive identification of several heart diseases. This paper proposes a set of 88 time-frequency features along with five different methodologies for classifying normal and abnormal heart sounds. State of the art approach was applied for segregating the fundamental heart sounds. Apart from a baseline two class classifier, separate classifiers for long and short heart sounds were also explored in order to get rid of the dependency of features on the duration of the recordings. Finally, a three class classifier was explored to deal with the noisy data present in the dataset. Both balanced and unbalanced sets were considered for crating of the training models. A comparative analysis showed that, out of all the methodologies, the three class classifier based approach produces the most optimum performance by simultaneously yielding high values of both sensitivity and specificity.

1. Introduction

Automatic classification of normal and abnormal heart sounds has been extensively studied over the last few decades. Heart sound signals, commonly known as phonocardiogram (PCG) is typically captured using a digital stethoscope and is known to carry useful information regarding many cardiac abnormalities.The state of the art techniques employ a number of steps for classifying normal and abnormal heart sounds, including pre-processing of noisy data, identification of the fundamental heart sounds, followed by feature extraction and classification. Wavelet based features [1] and spectral features, obtained from FFT [2] was widely used in literature to identify cardiac abnormalities. More complex features like Mel frequency Cepstral Coefficients (MFCCs) were also investigated in [3] using Hidden Markov Model. However, due to the vulnerability of PCG towards ambient noise in audible range, variation in sensor quality and the location of data acquisition, automatic classification of PCG is a challenging task till date.

An extensive corpus of PCG was provided in the Physionet 2016 challenge for classifying normal and abnormal heart sounds. The dataset is detailed in [4]. In summary, a total of 3153 heart sounds, including 2488 normal and 665 abnormal recordings are available in the corpus, partitioned in six subsets. In this paper, we have proposed five methodologies for robust classification of normal/abnormal heart sounds using machine learning approach. Our contributions are 1) inspecting a wide list of time-frequency features for heart sound classification, 2) identifying separate feature sets to deal with the variation in data length, 3) creating a three class classifier to identify the noisy data. Rest of the paper is organized as follows, Section 2 describes different features used in this paper for classification. The five proposed methodologies are detailed in Section 3, followed by experimental results and conclusions in Section 4 and 5 respectively.

2. Segmentation and Feature Extraction

Each complete cycle inside a PCG signal typically contains two prominent heart sounds, namely S1 and S2. S1 precedes the systole whereas S2 precedes the diastole region. Accurate segregation of the fundamental Heart sounds (FHSs) is considered as the major prerequisite of any kind of analysis job dealing with PCG. There is a vast literature available ([5], [6]) for automatic segregation of hearts sounds. All recordings in the dataset is sampled at 2000 Hz. Raw PCG is further down sampled at 1000 Hz, in order to segregate four cardiac states (S1, systole, S2 and diastole) using the logistic regression based HSMM approach, developed by Springer et al [7]. A wide list PCG features were extracted subsequently.

A total of 88 features were explored in this paper. First 20 time domain features are related to the arithmetic mean and standard deviation of the intermediate distance between different cardiac states and are detailed in [4]. These features contain information regarding individual heart beat as well as heart rate variability (HRV). Feature 21 measures the standard deviation of the successive differences between adjacent NN intervals. Feature indices 22 to

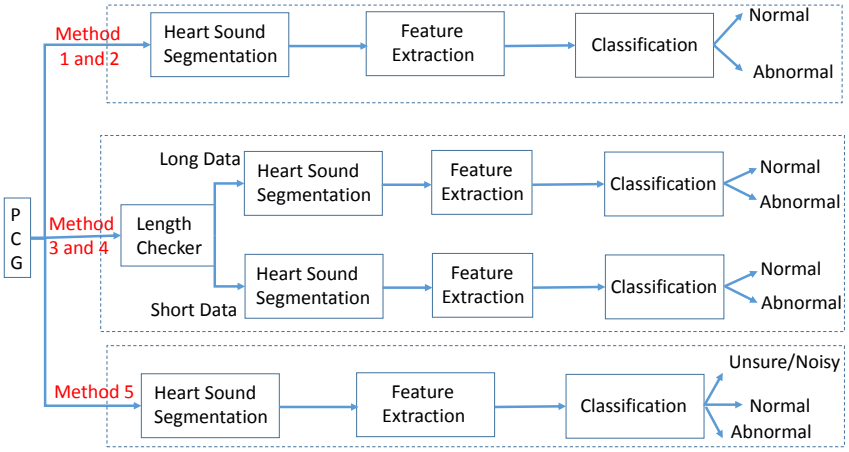

Figure 1. Block Diagram of Proposed Methodology.

37 measures the normalized spectral power within the frequency range of 0-20 Hz, 20-40 Hz, 40-60 Hz and 60-80 Hz respectively for S1, systole, S2 and diastole regions. Features 38 to 45 are the magnitude and phase angles of the first four poles of the diastolic regions, modeled using autoregressive (AR) model. The diastolic portion is sub segmented into non-overlapping windows of 50 ms and each window is modeled with a 10th order autoregressive (AR) model. Finally the median across all the sub-segments are considered as representative feature values. Features 46 to 53 represent the mean and standard deviation of the spectral centroid across all S1, systole, S2 and diastole segments in a recording. Features 54 to 60 represent the spectral power between 0-100 Hz in five equal frequency bands as well as the mean and standard deviation of the spectral centroid for the entire spectrum for all complete cardiac cycles. The next 26 features are the mean and standard deviation of 13 dimensional MFCC coefficients. To extract these features, the entire signal is broken into 250 ms windows with 100 ms overlapping using hamming window. The signals are analysed upto 300 Hz for extracting the coefficients. The final two features are wavelet related features, extracted from the diastolic portion. The diastolic portion is decomposed upto third level using 'Reverse biorthogonal 3.9' (rbio 3.9) mother wavelet. The median values of the mean and the standard deviation of the third level detailed coefficients across all the segments are included in the feature list.

3. Methodologies

A total of five different methodologies have been explored in this paper. The optimum feature list for each of them is selected from the exhaustive lists of 88 features by ranking them based on Maximal Information Coefficients (MIC) [8] scores. Our different methods (as shown in Figure 1) are detailed subsequently.

3.1. Method 1: Baseline Two Class Classifier

In the baseline approach, a simple two class classifier is designed for classifying normal and abnormal heart sounds. It was observed that each subset(a to f) of the entire dataset provided in[4] is highly unbalanced. Hence, in order to ensure a balanced training, all instances of the minor class, along with equal representation of the other member class is drawn at random from each of the subset. This random under-sampling, resulted in a total of 944 recordings from the entire set. Top 31 most significant features were selected for performance evaluation.

3.2. Method 2: Baseline Method on Unbalanced Recordings using all Features

Due to highly unbalanced ratio of normal and abnormal classes in each of the subsets, random under-sampling of majority class leads to removal of a significant number of observations. Hence, in method 2, we applied the baseline method on the entire corpus of unbalanced recordings. Further, the analysis is done using all 88 features to mitigate the effect of possible information loss occurred due to exclusion of certain feature in method 1.

3.3. Method 3: Separate Models for Long and Short Recordings

It was observed in method 1 and 2 that, the overall classification accuracy obtained on the long recordings (minimum duration of longer than 10 seconds) is significantly higher compared to the short recordings (duration less than 10 seconds). It was also observed that, small recordings are mostly present in one of the partitions (set b) of the entire dataset. The signal quality of the small recordings are also very poor compared to others. We defined the following

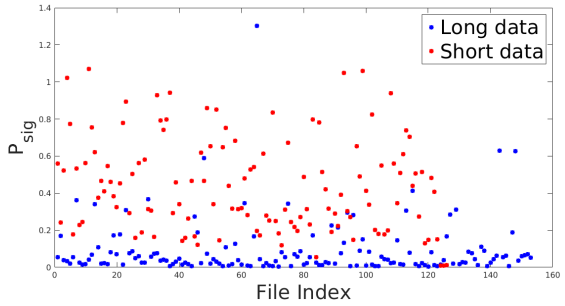

Figure 2. Signal Quality Comparison between Long and Short Data.

metric (P_{sig}) to measure the signal quality of the PCGs.

$$P_{sig} = median(P_{dia}/max(P_{s1}, P_{S2})) \quad (1)$$

For a noisy recording, the S1 and S2 peaks are often suppressed by the noise, induced in systole and/or diastole. Due to its prolonger duration, the diastolic portion is generally more vulnerable. Thus the ratio between diastolic power (P_{dia}) and power of S1 or S2 becomes comparable. Resulting in a higher value of P_{sig}. The value of P_{sig} should ideally be lesser for a recording of good signal quality. Figure 2, shows the values of P_{sig} of more than 120 long and short recordings, drawn at random from the entire dataset. It is evident that the small recordings are generally nosier than the longer recordings owing to the lesser values of P_sig. It is also evident that certain features, mostly related to HRV are often not captured properly in the recordings of very shorter duration. Thus the feature lists for long and short recordings for classifying abnormal heart sounds are expected to be different. So, instead of a single two class classifier, we decided to create two separate classifiers for long and short data respectively. A total of 684 long and 260 short recordings are available in the balanced subset of 944 recordings used in method 1. We selected equal representations of normal and abnormal cases from each of them in order to create two balanced subset. Finally, separate sets of features are selected for each cases, for designing of the classifiers.

3.4. Method 4: Separate Models But Unbalanced Long Recordings

This is a logical extension of method 3. Similar to method 2, here we explore all possible PCG features on the entire unbalanced dataset. However, we found that, this only improves the performance of the long recording. The performance on short recordings actually gets degraded. Thus in this method, we modify the analysis on long recordings only, by utilizing the entire feature list. The model for the short recordings remains the same as used in the previous method.

Table 1. Performance Comparison between Proposed Methodologies

Method	Se		Sp		$MAcc$	
	$mean$	std	$mean$	std	$mean$	std
1	0.78	0.036	0.79	0.058	0.78	0.034
2	0.60	0.039	0.97	0.006	0.78	0.020
3	0.79	0.044	0.78	0.045	0.78	0.036
4	0.95	0.092	0.72	0.034	0.84	0.148
5	0.80	0.042	0.90	0.015	0.85	0.025

3.5. Method 5: Three class classifier for Noisy Data

A significant portion of the entire dataset is largely corrupted due to human speech, background noise and the frictional noise generated due to the motion of the stethoscope on human body. In all our previous methods, all recordings were categorized into one of normal and abnormal classes regardless their signal quality. Ground truth signal quality for each signal was also provided along the training set in a binary form. This shows that 279 out of a total 3153 recordings are very poor for analysis even by expert annotators. Thus we created a three class classifier which provides a scope of identifying the noisy data and mark them as unsure along with classifying the rest as normal or abnormal. Six more features were derived for the noisy signals and are combined with the previous 88 features. the new features are - standard deviation of ratio between 1) diastolic and S1 power, 2) diastolic and S2 power, 3) mean of ratio between S1 and S2 power, median of ratio between 4) diastolic and S1 power, 5) diastolic and S2 power and 6) Kurtosis of the envelop of the autocorrelated PCG signal. All 94 features are applied on the entire dataset of 3153 recordings for performance evaluation.

4. Experimental Results

The popular ensemble learning method Random Forest (RF) is used for creating the learning models and classification. The number of decision trees in the forest is optimized during training. All our results in this paper are reported using 5-fold cross validation technique. The performance is evaluated in terms of three metrics detailed in [4], 1) $Sensitivity(Se)$ 2) $Specificity(Sp)$ and $MAcc = (Se + Sp)/2$. All unsure predictions, obtained in method 5 are marked as correct in the scoring system, if the ground truth signal quality is poor and incorrect otherwise.

Table 1 shows a comparative analysis among all the five methodologies explored in this paper. For all the cases, the performance metrics are reported in terms of $mean \pm std$ values obtained across all folds of the 5-fold cross vali-

dation technique. It can be concluded that overall performance ($MAcc$) of the first three methodologies is quite similar. However, due to training on a balanced dataset, method 1 and 3 generate more unbiased classifiers, resulting in sensitivity and specificity scores close to each other. Method 4 shows a significant improvement in mean $MAcc$ over the first three methods owing to very high sensitivity. However, the overall classification score is still fairly unstable as evident in high standard deviation values across all matrices.

A high value of sensitivity and specificity can simultaneously be achieved in method 5. In spite of being trained on an unbalanced dataset, addition of new features for identifying the noisy recordings is found to improve the accuracy significantly over the other methods. A possible reason may be, treating the noisy recordings as a separate class, reduces the anomaly in both normal and abnormal classes, thereby improving the overall training.

In our application, sensitivity measures the fraction of abnormal heart sounds out of all the test cases, getting correctly detected by the classifier. Specificity on the other hand, measures the fraction of normal heart sounds that are being correctly identified. Since, we are dealing with a screening system, a high value of sensitivity is always required to ensure that most of the abnormal heart sound gets detected by the system. Thus, in spite of a lesser accuracy compared to method 5, method 4 is the expected to come out to be a suitable method for developing a screening system due to yielding a mean sensitivity score of 0.9. However, if both sensitivity ans specificity are equally important, method 5 comes out to be the most optimum approach.

5. Conclusion

This paper deals with classification of normal and abnormal heart sounds using machine learning approach. Several time and frequency domain PCG features have been explored in this context. Five different methods have been investigated for performance comparison. The dataset provided in Physionet Challenge 2016 has been used for performance evaluation via 5-fold cross validation approach. Results show that separate training models, for long and short recordings can improve the sensitivity of the system. Results also show that, the overall accuracy can also be improved by incorporation a three class classifier to identify the noisy data. In this paper, we have used state of the art techniques for noise cleaning and segregation of fundamental heart sounds from raw PCG. Our future work includes, improving those state of the art techniques in order to perform better even on noisy signals. We are also planning to enhance the existing feature list and comparing the outcome of the proposed classifiers used in this paper with other popular learning techniques for further improvement.

References

[1] Akay YM, Akay M, Welkowitz W, Kostis J. Noninvasive detection of coronary artery disease. IEEE Engineering in Medicine and Biology Magazine Nov 1994;13(5):761–764.

[2] Bhatikar SR, DeGroff C, Mahajan RL. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. Artificial intelligence in medicine 2005; 33(3):251–260.

[3] Wang P, Lim CS, Chauhan S, Foo JYA, Anantharaman V. Phonocardiographic signal analysis method using a modified hidden markov model. Annals of Biomedical Engineering 2007;35(3):367–374.

[4] 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).

[5] Huiying L, Sakari L, Iiro H. A heart sound segmentation algorithm using wavelet decomposition and reconstruction. In Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE, volume 4. IEEE, 1997; 1630–1633.

[6] Moukadem A, Dieterlen A, Hueber N, Brandt C. A robust heart sounds segmentation module based on s-transform. Biomedical Signal Processing and Control 2013;8(3):273–281.

[7] Springer DB, Tarassenko L, Clifford GD. Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2016;63(4):822–832.

[8] Banerjee R, Sinha A, Choudhury AD, Visvanathan A. Photoecg: Photoplethysmographyto estimate ecg parameters. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014; 4404–4408.

Address for correspondence:

Anirban Dutta Choudhury
Tata Consultancy Services
Building 1B, Ecospace
Plot - IIF/12, New Town, Rajarhat
Kolkata - 700160,West Bengal
India
anirban.duttachoudhury@tcs.com