Clustering SCG Events using Unsupervised Machine Learning

Peshala T. Gamage[1], Md Khurshidul. Azad[1], Amirtaha Taebi[2], Richard H. Sandler[1], Hansen A. Mansy[1]

1. Biomedical Acoustic Research Lab, University of Central Florida, Orlando, Florida, USA
2. Department of Biomedical Engineering, University of California Davis, USA
{peshala@knights., khurshid@knights., hansen.mansy}@ucf.edu, ataebi@ucdavis.edu

Abstract— Seismocardiographic (SCG) signal morphology is known to be affected by cardio-pulmonary interactions, which introduce variability in the SCG signal. Hence, grouping of SCG signals according to their respiratory phase can reduce their morphological dissimilarity. In addition, correlating SCG with pulmonary phases may provide more insights into the nature of cardio-pulmonary interactions. This study uses unsupervised machine learning to cluster SCG events based on their morphology. Here, K-means clustering was employed using the time domain amplitude as the feature vector. The method is applied on measured SCG data from 5 male subjects (Age: 30±5.8 years). The mean Silhouette values for different number of clusters suggested that optimal clustering was reached when SCG waveforms were divided into two groups. Using respiratory flow information, SCG waves were labeled as inspiratory vs. expiratory or high vs. low lung volume. The SCG clusters were then compared with these labels and purity values were calculated. The distributions of clustered SCG events in relation to respiratory flowrate and lung volume phases showed consistent trends in all subjects. Results suggested that grouping SCG based on lung volume phases would yield more homogeneous groups and, hence, would keep SCG variability (within each group) to a minimum. The demonstrated utility of the proposed machine learning approach in identifying respiratory phases from SCG waveforms may obviate the need for simultaneous respiratory measurements.

Keywords- Seismocardiography (SCG), unsupervised machine learning, K-means clustering, cardiorespiratory, respiratory flow, lung volume

I. INTRODUCTION

Seismocardiography (SCG) is the measurement of cardiac related vibrations on the chest surface that are produced by mechanical activities of the heart, primarily caused by valve closure and opening, blood momentum changes and myocardial movements [1, 2]. In addition to being a potential low cost non-invasive measurement of heart function, SCG can provide important information about the interactions between cardiovascular and pulmonary systems [3, 4]. In this context, many studies have focused on characterization of SCG signal features in comparison with to other cardiac diagnostic methods [4] and in relation to different cardiac pathologies [5]. Moreover, some studies have focused on the study of time and frequency features of SCG to gain better understanding of heart function and used classification of SCG events for diagnosis of heart pathologies [6, 7].

As SCG signals are associated with the mechanical movement (rather than electrical activity) measured over chest surface, SCG signal morphology is affected by different factors such as respiration (e.g., changes in lung volume), heart rate and cardiac contractility [8, 9]. These factors may cause signal variabilities that mask subtle SCG changes that may be of diagnostic value. To reduce these variabilities, SCG waveforms can be grouped into different groups (with each group having similar waveform morphology). This can help provide more accurate signal features, which may increase the diagnostic value of SCG.

A few studies reported the effects of respiration on SCG morphology. One study suggested that SCG events can be categorized as inspiratory or expiratory [10]. By calculating morphological dissimilarities between SCG events, a recent study [4] showed that SCG morphology is more dependent on lung volume (which may correlate with intra-thoracic pressure) than respiratory flow direction (inspiration vs. expiration). Effects of respiration on SCG include two interrelated mechanisms: 1) changes in the heart position (due to the movement of the heart, diaphragm and lungs) with respect to the SCG sensor and 2) intra-thoracic pressure changes, which lead to cardiac filling alterations. For example, inspiration draws more blood into the right heart (due to the induced negative intrathoracic pressure) and increased right heart output into the more compliant lungs. Conversely, expiration exerts positive pressure on the lungs and thereby inhibits right heart filing. These and other mechanisms can cause complex changes to the SCG morphology.

Machine learning (ML) is a convenient tool to classify SCG events based on their morphological features without a need to have full understanding of the underlying mechanisms. Other classification methods may provide more insights into the effect of cardio-pulmonary interaction on SCG morphology. A few studies have employed ML to classify SCG waveforms (into inspiration/expiration or high lung volume/low lung volume) using supervised machine learning classifiers such as support vector machine (SVM) and random forest (RF) methods [3, 11, 12]. In these supervised methods, SCG grouping is decided *a priori* and the algorithm is trained to provide optimum classification accuracy. However, the training and test data may contain mislabeled waveforms. Hence, the accuracy of these classifiers is not necessarily indicative of the grouping purity.

This study focuses on implementing the K-means algorithm for clustering SCG events using unsupervised ML. In contrast to supervised ML, unsupervised ML is capable of classifying input data into optimally separated clusters with no training. Here, separation of each cluster is such that the cluster is optimally internally similar while differences between clusters are maximized. After SCG events are clustered into different groups, the phase of respiration (i.e., lung volume, inspiration, or expiration) of each event is examined to give insights into the effect of respiratory phases on SCG morphology thereby improving the utility of SCG monitoring for cardiac conditions (such as heart failure deterioration).

II. SCG MEASUREMENTS

SCG signals were acquired using a tri-axial accelerometer (Model: 356A32, PCB Piezotronics, Depew, NY) placed on the chest surface at the 4th intercostal space near the left lower sternal border. The accelerometer was affixed using double sided medical-grade tape such that the measured z-component of the acceleration was normal to the chest surface (i.e., dorso-ventral component). A spirometer (Model: A-FH-300, iWorx Systems, Inc., Dover, NH) was used to simultaneously measure respiratory flow rate via a mouthpiece. The time integral of respiratory flowrate was used to determine the corresponding lung volume changes. A sampling rate of 10 kHz was used for data acquisition.

The acquired SCG signals were filtered (band pass 0.5-50 Hz) to remove respiratory sounds and low frequency noise. SCG events (SCG signals during each heart cycle) were found using matched filtering with a template of manually identified SCG event from the same recording. All identified SCG events had the same length (~ 700 milliseconds). These events were time-aligned with respect to the first SCG peak (i.e., SCG1, where SCG1 aligns with the traditional phonocardiographic or stethoscope appreciated first heart sound, S1) (Fig. 1). More information about SCG pre-processing can be found in previous studies [3, 4, 6].

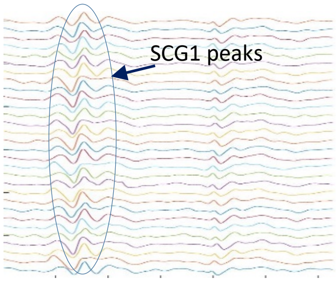

Fig. 1. Filtered and aligned SCG events that are used for clustering analysis. (SCG events are the segmented signals from the measured chest wall acceleration normal to chest surface)

Table 1. Subjects information

Age (years)	30 ± 5.8
Height (cm)	173.35 ± 11.25
Weight (kg)	75.58 ± 17.57
BMI	26.52 ± 3.4

After Institutional Review Board (IRB) approval, measurements were taken from 5 healthy male subjects. Subject summary is shown in Table 1.

III. K-MEANS CLUSTERING

In the current study K-means is used to cluster SCG waveforms using their time amplitudes as features. K-means clustering is an unsupervised machine learning algorithm which can be used to cluster input data into different groups based on the dissimilarity in input features. K-means clustering is beneficial over other clustering methods such as hierarchical clustering due to its simplicity and compatibility with large number of feature variables [13]. This algorithm partitions input observations (events) into K number of mutually exclusive clusters. The user should input a feature vector of size (1×n) for each event and an integer value for the number of clusters “k”. The algorithm starts by randomly selecting k number of centroids $\{C_1, \ldots, C_j, \ldots C_k\}$ unless they are specified by the user. Here, each centroid C_j is an (1×n) array. At the initial step, each observation is assigned to the nearest centroid (cluster) after calculating the distance from each centroid (each observation is assigned to the nearest centroid). Then the centroids are updated to the ensemble average calculated within each cluster. This process is repeated until the cluster assignment is converged. The convergence is usually monitored by “total sum of distances” (TOD), which is the total summation of distances between each observation from its cluster centroid. In some cases, clustering can converge to a local minimum since K-means algorithm heavily depends on the initial conditions. However, this issue can be eliminated by monitoring the TOD value for several different starting conditions (initial centroid locations) where some starting conditions will show a higher TOD value after convergence while some will converge to a low TOD value. Also, by monitoring the number of iterations taken for convergence, a good starting condition can be chosen for a faster execution.

In the current study, only the z-component of the SCG signal was considered for clustering analysis. Segmented SCG events were down sampled from 10 kHz to 320 Hz and amplitudes of these SCG events were used as feature vectors. Each down sampled SCG event contained 201 points. Hence, the feature vector was an array of size (1×201).

To determine the optimum number of clusters, the average Silhouette value of the clustering was analyzed. The Silhouette value (*Si*) of the point i in the clustering can be expressed as,

$$Si = (b_i - a_i)/\max(a_i, b_i) \tag{1}$$

where a_i denotes the mean distance measured from the i th point to the other points in the same cluster (that point i belongs to) and b_i is the average of the minimum distances measured from i^{th} point to all other points in each of the other clusters. Si values range from -1 to 1 where a positive Si value closer to 1 indicates a point is well inside in its own cluster (away from other clusters) and a negative value indicates that the point is away from its own cluster [14]. Hence, mean Si value of all points can be used as a measure to find the optimum number of clusters. Fig. 2 shows the mean Si values calculated for different number of clusters for all 5 subjects. The results showed that two clusters had the highest Si value, and hence, the optimum clustering of this SCG data.

IV. Results

The purity of the clustering was analyzed by comparing clustered results with two types of labeling: Respiratory flow direction (inspiration (INS) vs. expiration (EXP)); and lung volumes phase (high lung volume (HLV) vs. low lung volume (LLV)). Labeling started by time locating the peak of each SCG on the respiratory cycle. Then the event was labeled according to when it occurred during the cycle. For example, inspiration vs. expiration phases were determined based on the breathing flow directions while low and high lung volumes were based on the current lung volume compared to the mean lung volume over the entire recording time. These labeling criteria were selected as suggested in previous studies [4, 6]. Purity values were calculated (below) to quantify how well each labeling agrees with the clustering results. Equation 2 was used to find the clustering purity.

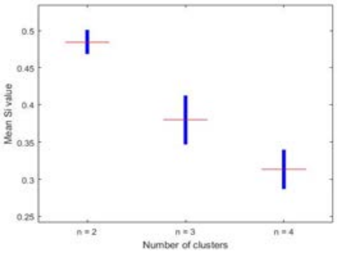

Fig. 2. Mean silhouette values for different number of clusters.

$$Purity = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Where, *TP* is the number of true positives and *TN* is the number of true negatives. (e.g. *TP* and *TN* are the number of events that are correctly labeled as HLV and LLV, respectively). Similarly, *FP* and *FN* indicate the number of false positives and false negatives, respectively.(e.g. *FP* and *FN* are the number of events that are incorrectly labeled as HLV and LLV, respectively).

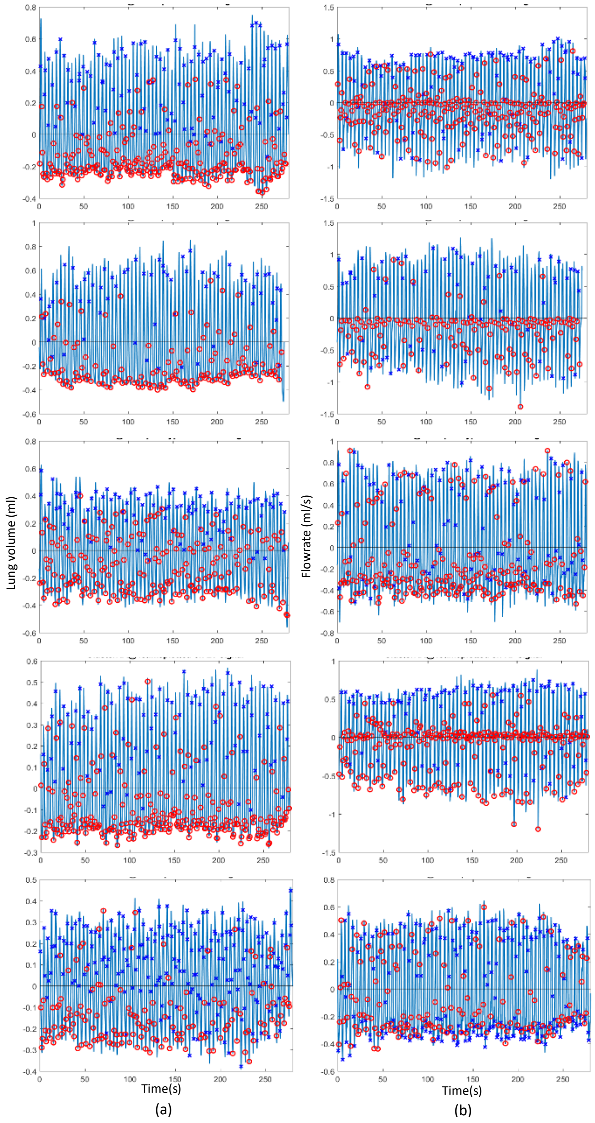

Fig. 3. Locations of clustered SCG events plotted on a) lung volume (left column) and b) respiratory flow signals (right column) for the 5 study subjects.

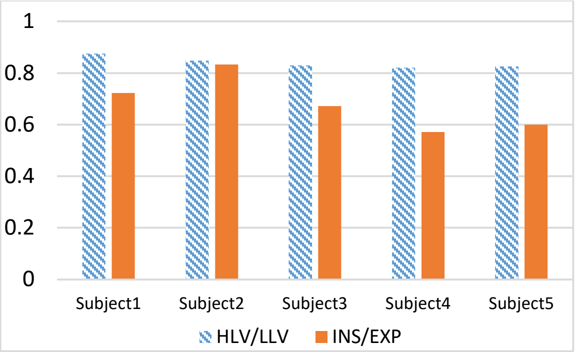

Fig. 4. Purity values for each subject for HLV/LLV and INS/EXP labelling.

Fig. 3 shows the location of clustered SCG peak of each event on respiratory flowrate and lung volume waveform. Here, the waveform of the respiratory flowrate and lung volume were considered such that positive values indicate INS and HLV while negative values indicate EXP and LLV phases on respective waveforms. While cluster distributions on respiratory flow rate were irregular on positive (INS) and negative (EXP) regions, a more organized distribution pattern can be found on lung volume waveforms. Across all the subjects, the majority of cluster 1 events (indicated in blue ×) occupied HLV region and the majority of cluster 2 events (indicated in red o) occupied LLV region. Less events of each cluster were placed in the opposite phase. The latter are identified as "mismatched events" for the purpose of discussion in this paper. The cluster distribution results presented in Fig. 3 are further clarified in Fig. 4 that shows the purity results for labels HLV/LLV and INS/EXP. It can be seen in the figure that for all subjects, the labelling criteria HLV/LLV is associated with a higher purity than INS/EXP.

To further analyze the possible patterns in clustering SCG, the event distribution in each cluster was separately studied. Here, SCG events were divided into four respiratory phases, namely: HLV-INS, HLV-EXP, LLV-EXP and LLV-INSP. These four phases can be seen on the lung volume signal shown in Fig. 5 (a). The percentage of each cluster that occur at each of the 4 phases are shown in Fig. 5 (b) for all 5 subjects. The results shown in Fig. 5 (b) re-organizes the cluster distribution results shown in Fig. 3 to provide more information about the distribution of correctly and incorrectly labeled events. It can be seen that a higher percentage of incorrectly labeled events in the HLV cluster (shown as a blue striped cluster) are in LLV-INS region than LLV-EXP. Similarly, a higher percentage of mislabeled events in the LLV cluster (solid orange cluster) are in HLV-EXP phase than the HLV-INSP.

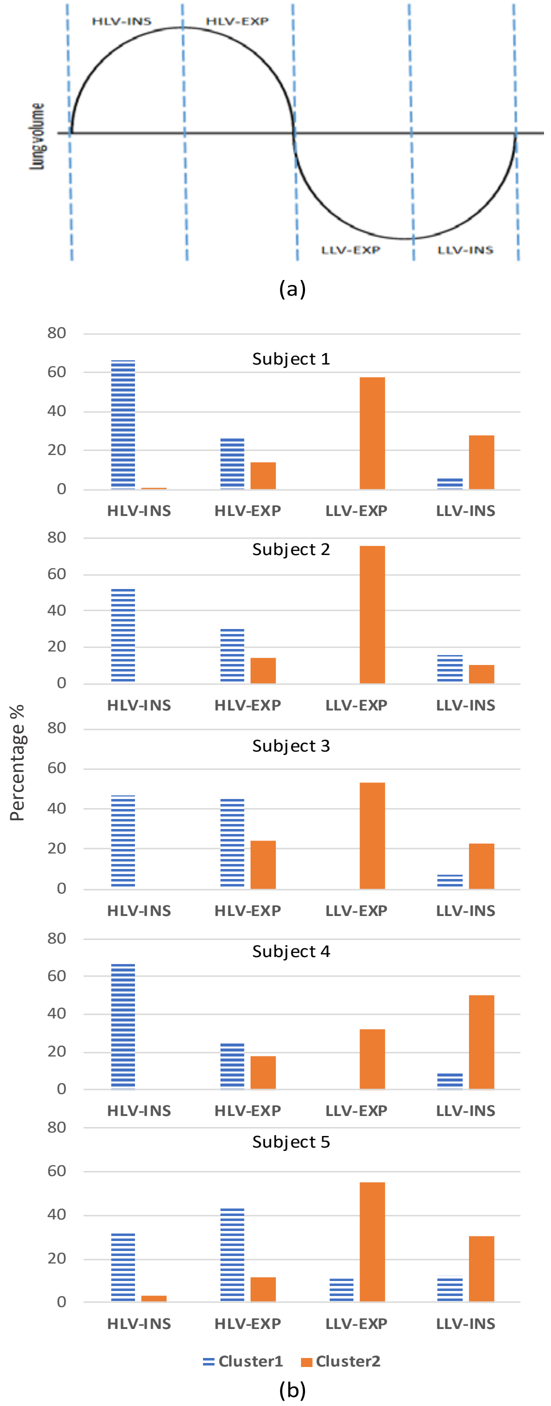

Fig. 5. (a) Four respiratory phases identified on the lung volume signal (b) corresponding percentages of SCG event distribution over four categories in each cluster.

V. Conclusion

In this study, K-means clustering was used to group SCG signal morphologies into different clusters by inputting the time domain feature vector, which contains the amplitude values of each SCG event. Comparing the mean Silhouette value for different numbers of clusters suggested that SCG morphology is optimally divided into 2 clusters. The Purity of the clustering for two labelling criteria based on respiration was calculated. The results showed that separating SCG signal morphology based on lung volume criteria is more accurate than separation based on respiratory phase (inspiration vs. expiration). These findings are in agreement with previous studies that compared SCG waveform dissimilarity based on lung volume vs. on respiratory phase [4, 6]. Further, we have demonstrated, for the first time, the likely utility of machine learning approaches to accomplish this separation thereby obviating the need for actual simultaneous respiratory measurement.

Future studies will involve defining appropriate cut-off regions on respiratory flowrate and lung volume waveforms to optimally separate SCG events based on their morphology. These studies will help accurate diagnosis of heart diseases using SCG as well as to enhance the understanding of SCG genesis.

Acknowledgments

This study was supported by NIH R44HL099053.

Hansen A Mansy and Richard H Sandler are part owners of Biomedical Acoustics Research Company, which is the primary recipient of the above grant, as such they may benefit financially as a result of the outcomes of the research work reported in this publication.

References

[1] V. Gurev, K. Tavakolian, J. Constantino, B. Kaminska, A. P. Blaber, and N. A. Trayanova, "Mechanisms underlying isovolumic contraction and ejection peaks in seismocardiogram morphology," *Journal of Medical and Biological Engineering,* vol. 32, p. 103, 2012.

[2] I. Korzeniowska-Kubacka, B. Kuśmierczyk-Droszcz, M. Bilińska, B. Dobraszkiewicz-Wasilewska, K. Mazurek, and R. Piotrowicz, "Seismocardiography-a non-invasive method of assessing systolic and diastolic left ventricular function in ischaemic heart disease," *Cardiology Journal,* vol. 13, pp. 319-325, 2006.

[3] A. Taebi, B. E. Solar, and H. A. Mansy, "An Adaptive Feature Extraction Algorithm for Classification of Seismocardiographic Signals," *arXiv preprint arXiv:1803.10343,* 2018.

[4] A. Taebi and H. A. Mansy, "Grouping similar seismocardiographic signals using respiratory information," in *Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE*, 2017, pp. 1-6.

[5] C. A. Wick, J.-J. Su, J. H. McClellan, O. Brand, P. T. Bhatti, A. L. Buice, *et al.*, "A system for seismocardiography-based identification of quiescent heart phases: implications for cardiac imaging," *IEEE Transactions on Information Technology in Biomedicine,* vol. 16, pp. 869-877, 2012.

[6] A. Taebi, "Characterization, Classification, and Genesis of Seismocardiographic Signals," 2018.

[7] A. Taebi and H. A. Mansy, "Time-frequency distribution of seismocardiographic signals: A comparative study," *Bioengineering,* vol. 4, p. 32, 2017.

[8] O. T. Inan, P.-F. Migeotte, K.-S. Park, M. Etemadi, K. Tavakolian, R. Casanella, *et al.*, "Ballistocardiography and seismocardiography: a review of recent advances," *IEEE J. Biomedical and Health Informatics,* vol. 19, pp. 1414-1427, 2015.

[9] K. Tavakolian, G. Portacio, N. R. Tamddondoust, G. Jahns, B. Ngai, G. A. Dumont, *et al.*, "Myocardial contractility: A seismocardiography approach," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 3801-3804.

[10] K. Tavakolian, A. Vaseghi, and B. Kaminska, "Improvement of ballistocardiogram processing by inclusion of respiration information," *Physiological Measurement,* vol. 29, p. 771, 2008.

[11] B. E. Solar, A. Taebi, and H. A. Mansy, "Classification of seismocardiographic cycles into lung volume phases," in *Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE*, 2017, pp. 1-2.

[12] V. Zakeri, A. Akhbardeh, N. Alamdari, R. Fazel-Rezai, M. Paukkunen, and K. Tavakolian, "Analyzing seismocardiogram cycles to identify the respiratory phases," *IEEE Transactions on Biomedical Engineering,* vol. 64, pp. 1786-1792, 2017.

[13] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *ICML*, 2001, pp. 577-584.

[14] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis* vol. 344: John Wiley & Sons, 2009.