Heart Sound Classification Based on Temporal Alignment Techniques

José Javier González Ortiz, Cheng Perng Phoo, Jenna Wiens
Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

Abstract

The ability to accurately stratify patients at risk of adverse cardiovascular outcomes using heart sound recordings could result in earlier treatment and improved patient outcomes. However, there remain several challenges associated with risk stratifying patients based on the phonocardiogram (PCG) alone. First, inter-patient differences can make it challenging to learn a model that generalizes well across patients. Second, heterogeneity introduced by the collection environment of the recordings can render a classifier trained on one population useless when applied to another. To address these challenges we explore the use of temporal alignment techniques, in particular dynamic time warping (DTW). Using DTW we compare heart sounds within and across subjects/recordings. These DTW based features, coupled with widely used spectral MFCC coefficients, serve as input to a linear SVM. Applied to the held-out test set our classifier obtained a test score of 82.4%, suggesting that temporal alignment techniques can effectively reduce the effects of inter-patient variability and mitigate the differences introduced by heterogeneous data collection environments.

1. Introduction

In cardiac auscultation an examiner uses a stethoscope to listen for unique and distinct sounds, that provide important data regarding the condition of the heart. Modern recording equipment captures these heart sounds as a phonocardiogram (PCG). In principle, these recordings could be used to automatically monitor patients and diagnose cardiac abnormalities. Yet, while auscultation is a common practice in patient exams, PCGs are not widely used clinically, where echocardiograms and electrocardiograms are more prevalent. This is due, in part, to the lack of robust algorithms for automatically classifying PCGs. To address this issue, the 2016 PhysioNet/CinC Challenge focused on the development of algorithms to classify PCGs collected from both clinical and nonclinical environments [1].

Robust PCG classification algorithms must accurately identify cardiac abnormalities across patients and across diverse recording environments. To address challenges associated with inter-patient variability we borrow techniques that have been successfully applied in speech processing and ECG analysis, where similar issues arise [2–4]. In particular, we explore the use of dynamic time warping (DTW) in measuring similarity between heartbeats from the same subject and across subjects. Our experiments show that such DTW-based features can mitigate the differences introduced by heterogeneous data collection environments and improve classification performance, especially when training and test populations differ.

2. Methods

In this section we present our supervised learning system for classifying PCGs as either normal of abnormal. We begin by describing the signal segmentation, then move on to feature extraction and lastly explain the learning algorithm.

2.1. Segmentation

As a first step, we segment the PCG recording into the fundamental heart sounds: S1 and S2 in addition to the systolic and diastolic intervals. These four intervals make up the heart cycle states. Segmentation is an essential step in the automatic analysis of PCGs, allowing one to uncover the underlying physiological structure of the signal and recognize abnormalities within physiologically meaningful regions. Here, we use the state-of-the-art segmentation algorithm introduced by Springer *et al.* [5].

2.2. Feature Engineering

Next, we apply several transformations to the segmented heart cycle states of each record, obtaining features pertaining to time intervals, spectral analysis and morphology.

2.2.1. Time Interval Features

From the segmented recordings, we first extract features pertaining to heart sound intervals. We compute statistics for the length of each heart cycle state, as in [1]. These were the baseline features provided in the Challenge, but such timing data can contain important information regarding cardiac arrhythmias.

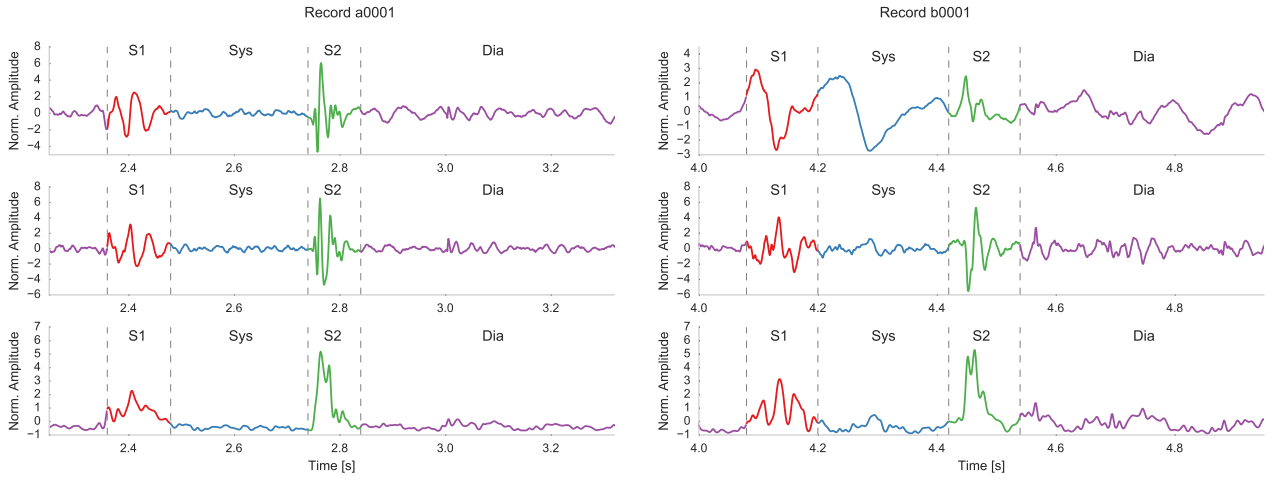

Figure 1: Comparison of DTW preprocessing techniques in sample intervals from different training populations. *Top*: Untouched recording, *Middle*: High Pass Filter (HPF), *Bottom*: Homomorphic envelogram

2.2.2. Spectral Analysis: MFCC Features

To analyze the spectral content of the signals we compute the discrete wavelet transform (DWT) for each RR interval. From this, we extract the mean and standard deviation of each approximation and detail coefficient. This captures the frequency content of the signal, which is known to be useful in PCG classification [6]. Here, we chose a wavelet transform over a discrete Fourier transform, since PCGs are highly non-stationary.

In addition, we compute Mel-Frequency Cepstral Coefficients (MFCCs) for each of the heart cycle states. We calculate mean and standard deviation of these coefficients for each record to capture variability within each filterbank interval. Our choice of MFCCs was inspired by their frequent use in the speech recognition domain. MFCCs offer several benefits over wavelets such as decorrelated coefficients, which often perform better in linear models. Moreover, researchers have had success using MFCCs with HMMs for automatic auscultation classification [7].

2.2.3. Morphology Analysis: DTW Features

The features described above capture frequency content and timing of a signal, but may fail to capture variability in morphology that may be symptomatic of cardiac abnormalities. To capture this variability we consider the temporal representation of the PCG. Since classic pairwise measures such as Euclidean distance are susceptible to temporal distortions [8], we use dynamic time warping (DTW). DTW finds an optimal alignment between two time-dependent sequences [4] by warping the sequences in a nonlinear fashion. This allows similar time series that are locally out of phase to be optimally aligned and compared in a meaningful way. DTW has been widely used for tasks such as speech recognition and ECG analysis, resulting in good empirical performance [3, 9, 10]. Here, we use DTW to compare the morphology of heart sounds within a subject and across subjects.

Preprocessing: Before computing DTW distances, we preprocess the PCGs to reduce noise and the effect of characteristics specific to the recording environment. Empirical analysis revealed that applying a high-pass Butterworth ($f_c = 25$ Hz, $N = 3$) filter effectively reduced noise without affecting the core morphology of clean signals. To further reduce noise we perform an envelope computation over the PCG [5, 11]. We perform a final z-standardization to make the signals equally distributed, with zero mean and unit variance. When computing DTW distances, we experimentally selected a 10% Sakoe-Chiba constraint, and for the DTW normalization strategy we resampled the records to a common median length [9].

Intra-DTW: Several cardiac conditions are manifested by higher than usual variability in the shape and frequency of the heartbeats in the ECG. To capture this intra-patient variability using PCGs, we compute intra-PCG heartbeat DTW distances. First, we construct a representative (i.e., medoid) heartbeat for a given record. This is the heartbeat whose average DTW distance to all the other in-record heartbeats is minimal. Then we compute pairwise distances between this medoid beat and all other beats in the record. We transform these distances into a set of five features by taking the mean, standard deviation, and first, second and third quartiles.

However, this does not capture the evolution of the signal across consecutive heartbeats. Therefore, we also extract the same features described above for consecutive DTW distances.

Inter-DTW: Intra-DTW features will fail to capture abnormalities that manifest consistently. To cope with this

limitation we also compute *inter*-patient DTW distances. These features aim to capture canonical patterns based on a beat's similarity to a set of template heartbeats. We construct templates by first clustering the medoid beats for each population and class label (normal and abnormal) and then extracting the centroid of each cluster. For this, we use spectral clustering with DTW for the affinity measure. For each record and template, we compute DTW distances to all the in-record heartbeats. Mean and standard deviation of these distances are used as features.

2.3. Classifier

Given the features described in the previous section, we learn a linear classifier to separate normal from abnormal recordings. We use SVMs since others have shown them to be effective when applied to this kind of task [12]. Before learning the classifier, we apply zero-one min-max scaling to the features in order to reduce bias towards any one dimension. In addition, we employ asymmetric cost parameters to handle the class imbalance present in the Challenge data [13]. While we explored several different kernels, here we focus on results pertaining to a linear model. We optimized the cost parameters using crossvalidation on the training data. The precise training/testing setup is described in more detail below.

3. Experiments & Results

To measure the effectiveness of the proposed DTW-based features in classifying PCGs, we ran a number of experiments using the Challenge data [1], which consists of PCG recordings from 6 populations (a through f). In this section we present our results and discuss how we addressed one of the main technical challenges: the extent to which the data varied across populations.

3.1. Robustness of DTW Features

To assess how robust the proposed DTW-based features are to inter-population differences, we computed histograms for the DTW features from each population and constructed kernel density estimates. Figure 2 compares these estimates for an intra-DTW based feature (lower plot) to those for an MFCC-based feature (upper plot). While both plots exhibit variability across populations, the lower plot illustrates how DTW-based features can reduce inter-population variability. The trend shown in Figure 2 was consistent across the other DTW and MFCC features.

3.2. Test Setup + Classification Results

Next, we compared how different feature set combinations performed when applied to the given classification task. We evaluated performance of each classifier using the challenge score described in [1]. The Challenge consisted of both given labeled data and hidden test data. Due

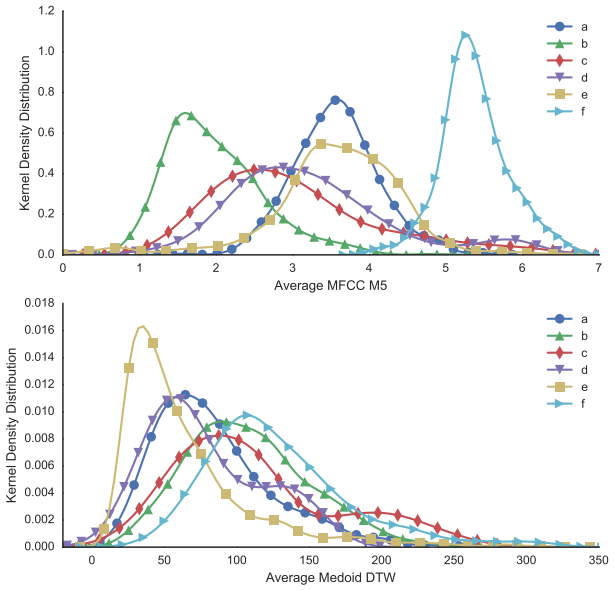

Figure 2: Comparison of Population Kernel Density Estimates for an average MFCC coefficient and the average intra-DTW distance.

to the limited number of test submissions, we conducted several experiments on the given data alone. For this, we considered separate experimental setups that differ in the way data are split into training and validation sets. Below we summarize the four main setups we considered. It's important to note that in selecting model hyperparameters, we used analogous splits during crossvalidation.

Balanced – Include 70% of each population in the training set; validate on the remaining 30%.

Balanced except *a*, *f* – Include 70% of each population, except a and f, in the training set; validate on the remaining 30% of b-e, 40% of a and 10% of f.

Balanced except *f* – Train on all of a, and 70% of b-e; validate on 100% of f.

LOPO(*p*) – Train on all populations except population p, validate on p. For performance evaluation, we define two aggregate metrics: $\overline{L}$, the average LOPO score and $\overline{wL}$ weighted average LOPO score.

$$\overline{L} = 1/N \sum_p \text{LOPO}(p) \qquad \overline{wL} = \sum_p 1/|p| \cdot \text{LOPO}(p)$$

Results for each validation scheme and the challenge hidden test set are presented in Table 1. The first row corresponds to a classifier trained on the baseline interval features combined with commonly used wavelet features [1, 5, 6]. Comparing the performance of this row to the second row, it is clear that MFCC features are better at capturing important differences. In experiments not shown here we found the gain in performance from the interval features to be limited, thus we omit these features

Table 1: Results using various input features in different validation sets and the Challenge test set. Balanced splits are averaged across 20 iterations and include standard deviations. Min and max LOPO metrics are included for $\overline{L}$ and $\overline{wL}$.

Features	Balanced	Bal. except a, f	Bal. except f	$\overline{L}$	$\overline{wL}$	$[L_{\min}, L_{\max}]$	Challenge
Interval, Wavelet	74.22 ± 0.63	73.19 ± 0.86	76.41 ± 0.49	58.27	52.35	[48.20, 76.50]	78.1
Interval, MFCC	77.68 ± 0.48	74.77 ± 0.72	79.66 ± 0.41	60.90	54.91	[51.70, 73.80]	
MFCC, $\text{DTW}_{\text{inter}}$	$\mathbf{85.73 \pm 0.48}$	78.05 ± 1.04	79.72 ± 0.42	66.03	64.64	[58.50, 75.70]	79.5
MFCC*, $\text{DTW}_{\text{intra}}$	85.18 ± 0.74	78.97 ± 0.95	$\mathbf{84.89 \pm 0.43}$	**68.37**	**68.81**	**[61.10, 77.40]**	**82.4**
MFCC, $\text{DTW}_{\text{intra}}$, $\text{DTW}_{\text{inter}}$	85.63 ± 0.42	$\mathbf{79.98 \pm 0.87}$	84.42 ± 0.49	66.95	67.78	[60.60, 75.30]	78.9

$*$ This set of features includes also Systole and Diastole in addition to S1 and S2

from the remainder of our experiments. The last three rows of Table 1 describe different combinations of MFCC features (applied to S1 and S2 intervals) and inter-DTW, intra-DTW. While inter-DTW based features perform better in a balanced setting (where populations appear in both training and test sets), intra-DTW based features are more robust to inter-population differences and obtain higher scores in LOPO based metrics. Moreover, the best performance in the hidden challenge test set is achieved when combining MFCC features with the intra-DTW features. We hypothesize that using intra-DTW features leads to better generalization since they capture the variability within a record while eliminating differences caused by heterogeneous recording environments.

4. Conclusion

In this paper we proposed two novel approaches to using time alignment techniques in PCG classification. When combined with spectral MFCC features, both approaches consistently improved the performance of the classifier, even in the presence of significant population differences between training and validation sets. In particular, we found that intra-patient DTW measurements produced quasi-homogeneously distributed features across populations and successfully captured intra-PCG variability. While we considered only a simple linear classifier, we suspect that the proposed features could prove useful in more complex approaches to time series classification, e.g. HMM, RNN and LSTM.

Acknowledgments

We thank Dr. Hitinder Gurm for his assistance in understanding common heart sound abnormalities and how auscultation based diagnosis is performed. This research program is supported by the National Science Foundation (NSF award numbers CNS-1330142 and IIS-1553146). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF.

References

[1] Liu C, Springer D, Li Q, Moody B, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(9).

[2] Wiens J, Guttag JV. Active learning applied to patient-adaptive heartbeat classification; NIPS, 2010.

[3] Tuzcu V, Nas S. Dynamic time warping as a novel tool in pattern recognition of ECG changes in heart rhythm disturbances. In 2005 IEEE International Conference on Systems, Man and Cybernetics, volume 1. IEEE, 2005; 182–186.

[4] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics speech and signal processing 1978;26(1):43–49.

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

[6] Lee JJ, Lee SM, Kim IY, et al. Comparison between short time fourier and wavelet transform for feature extraction of heart sound. In TENCON 99. Proceedings of the IEEE Region 10 Conference, volume 2. IEEE, 1999; 1547–1550.

[7] Chauhan S, Wang P, Lim CS, et al. A computer-aided mfcc-based hmm system for automatic auscultation. Computers in Biology and Medicine 2008;38(2):221–233.

[8] Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In KDD workshop; 1994.

[9] Ratanamahatana CA, Keogh E. Everything you know about dynamic time warping is wrong. In Third Workshop on Mining Temporal and Sequential Data. 2004; .

[10] Syed Z, Guttag JV. Identifying patients at risk of major adverse cardiovascular events using symbolic mismatch; NIPS, 2010.

[11] Gupta CN, Palaniappan R, Swaminathan S, et al. Neural network classification of homomorphic segmented heart sounds. Applied Soft Computing 2007;7(1):286–297.

[12] Maglogiannis I, Loukis E, Zafiropoulos E, et al. Support vectors machine-based identification of heart valve diseases using heart sounds. Computer methods and programs in biomedicine 2009;95(1):47–61.

[13] Morik K, Brockhausen P, Joachims T, et al. Combining statistical learning with a knowledge-based approach-a case study in intensive care monitoring; ICML, 1999.