Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models

Andrew McDonald, Mark Gales, Anurag Agarwal

Department of Engineering, University of Cambridge, Cambridge, UK

Abstract

We describe a recurrent neural network and hidden semi Markov model (HSMM) approach to detect heart murmurs in phonocardiogram recordings. This model forms the 'CUED_Acoustics' entry to the 2022 George B. Moody PhysioNet challenge.

Segmentation of the phonocardiogram is a key preprocessing step for many heart sound algorithms. However, most previous work assumes that heart sound recordings only contain S1 and S2 sounds, leading to poorer segmentations of signals that contain a strong murmur. Our approach applies multiple HSMMs, each making different assumptions about a possible murmur, to produce multiple segmentations of the signal. By comparing the confidence of each HSMM's output, we simultaneously produce a murmur classification and robust segmentation.

On the murmur detection task, our algorithm achieved a training cross-validation score of 0.799 and a validation score of 0.758 (ranked 6th out of 61 teams). On the clinical outcomes task, we achieved a training score of 11040 and a validation score of 9257 (ranked 9th out of 61 teams).

The algorithm is highly sensitive (92.7%) to murmurs and, compared to end-to-end models, provides interpretable results about their location and timing. This makes it a promising tool for symptomatic screening.

1. Introduction

Listening to the chest with a stethoscope (auscultation) is a quick and non-invasive method to screen for cardiac abnormalities. However, auscultation proficiency amongst clinicians varies widely. The sensitivity of a general practitioner in detecting valvular heart disease can be as low as 44% [1]. Automated analysis of stethoscope recordings (phonocardiograms) is a promising solution to improve the consistency and accessibility of auscultation. The George B. Moody PhysioNet Challenge [2] tasked participants with designing algorithms to detect heart murmurs and predict clinical outcomes using paediatric phonocardiograms from a new open-source dataset [3].

2. Methods

A key conclusion of the 2016 PhysioNet challenge on heart sound classification was that feature extraction can be the 'most crucial and important part' of the algorithm [4]. One of the most common feature extraction steps is segmentation, where the start and end of the individual sounds in a phonocardiogram are labelled. This allows the reduction of information from many periodic heartbeats into a single fixed-length feature vector, for input into a subsequent classifier.

Previous state-of-the-art segmentation algorithms such as the work of Springer [5], used to segment recordings in both the 2016 and 2022 challenge datasets, assume a healthy heart sound cycle, which make them susceptible to errors when structural heart disease leads to loud murmurs and weaker S1 or S2 sounds.

The most innovative part of our approach is a segmentation algorithm that localises both healthy S1 and S2 sounds and abnormal murmurs, removing the need for a subsequent classification algorithm [6]. The algorithm uses a recurrent neural network (RNN) to provide observations for multiple hidden semi-Markov models (HSMMs). One of the HSMMs assumes a healthy phonocardiogram whilst the others expect differently-timed systolic murmurs. We then compare the output segmentations to determine a final segmentation and murmur classification.

2.1. Feature Extraction

Each phonocardiogram is first normalised by removing its mean and dividing by the resulting peak amplitude. The log-spectrogram of the signal is then calculated using a Hann window, with length 50 ms and step 20 ms. This gives an effective frequency resolution of 20 Hz and a feature sample rate of 50 Hz, which has been found to be an acceptable trade-off for both segmentation and classification. We crop the spectrogram to the 0-800 Hz range to remove higher frequencies that contain no heart sound information. Each frequency row of the spectrogram is then individually z-score normalised, to further reduce the dynamic range between murmurs and the S1 and S2 sounds.

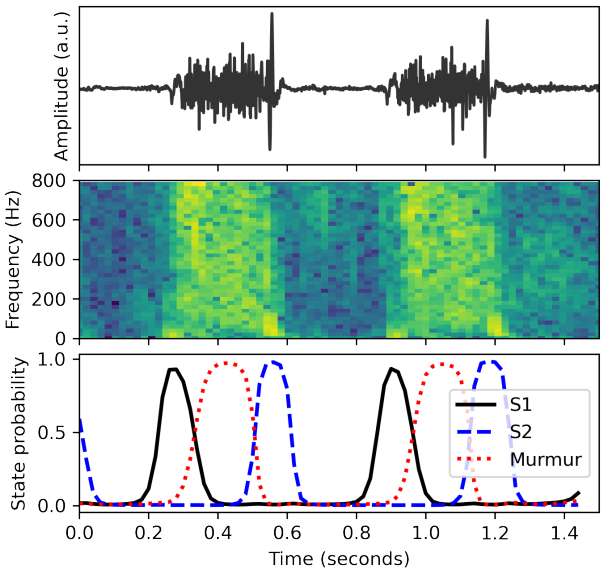

Figure 1. Generation of neural network state predictions. A phonocardiogram (top) is transformed to a normalised spectrogram (middle) that is used as an input to an RNN that predicts heart sound states (bottom).

2.2. Recurrent Neural Network

The distinct heart sound states of the phonocardiogram, $\xi_i \in \{\text{S1, S2, systole, diastole, murmur}\}$, are predicted using a bidirectional RNN. The normalised spectrogram with T windows, $x_{1:T}$, is input to a bidirectional Gated Recurrent Unit (GRU) network with parameters θ_R that predicts the state at time t, q_t, with posterior output $P(q_t = \xi_i | x_{1:T}, \theta_R)$. RNNs have been successfully applied for heart sound segmentation [7], offering improved predictions that model inter-timestep dependencies compared to a simple logistic regression or fully-connected neural network. However, the models deployed in previous approaches predict the four states of a healthy heart sound. Our algorithm is also trained to predict a murmur state (see Figure 1), resulting in a five-state categorical output that better captures the observed features.

A 3-layer bidirectional GRU is used, and the concatenated forward and backward outputs are then fed into a 2-layer fully connected neural network with Tanh activations. This reduces the hidden dimension to the 5-dimension output where a softmax is applied. Dropout is applied between both the GRU and fully-connected layers to reduce overfitting. Table 1 gives the key parameters.

The challenge dataset [3] includes segmentation labels denoting the start and end of the S1, systole, S2 and diastole sounds. To denote the start and end of the murmur sounds, we use the murmur timing information noted by the clinician. For example, if the recording is labelled to contain an early-systolic murmur, we approximate that the first 50% of each systole is the murmur signal. The same logic applies for the mid and late systolic murmurs, whilst for a holosystolic murmur the whole of each systole is labelled as a murmur. A future improvement could be to replicate this labelling for the diastolic murmur signals, although there are very few of these in the dataset.

Parameter	Value
GRU hidden size	60
Number of GRU layers	3
GRU layer dropout	0.1
Fully-connected dropout	0.1
Fully-connected hidden sizes	[60, 40]

Table 1. Hyperparameters chosen for bidirectional GRU segmentation model.

The RNN is trained to predict these modified segmentation labels from the extracted features, using a cross-entropy loss with the Adam optimiser. The loss function is inversely weighted to the frequency of each class label in the dataset, to compensate for the fact that murmurs only appear in some systolic portions of some signals. Stratified 5-fold cross validation (patients stratified according to murmur class) is used to detect overfitting and optimise hyperparameters.

2.3. Parallel Hidden Semi-Markov Models

The RNN predictive output could be immediately used to detect a murmur, by taking a 'greedy' approach and giving a positive result if 'murmur' is ever the most likely posterior state. However, signal noise can impact the predictions of the murmur state and would lead to false positives. Instead, we apply HSMMs to consider the whole signal when computing a murmur prediction.

The RNN predictions are used as observation probabilities for the HSMMs, following a similar structure to the logistic regression and HSMM of Springer [5].

The HSMM is an extension to a standard hidden Markov model that explicitly models the duration of each state. To create these state duration distributions, we follow Springer [5] and first estimate the heart rate of the signal. Springer estimates this by computing the autocorrelation of a smoothed envelope of the heart sound and searching for the highest peak in a specified range. In this work we additionally compute the autocorrelation of the non-diastolic RNN posteriors (the S1, S2, systolic and murmur predictions summed). This leverages the predictive power of the RNN to filter away noise and produce a smoother autocorrelation for improved peak detection. Given the heart rate estimate, the state duration distributions are calculated as in Springer, using normal distributions for the states with means scaled by the heart rate.

Given the RNN observations and the HSMM parameters (state durations and transition matrix), the segmentation of the heart sound signal is calculated using the Springer duration-dependent Viterbi algorithm [5]. Previous work has used a single HSMM that assumes the signal being segmented contains just the major heart sounds. Our approach uses four parallel HSMMs that assume different classifications $(\omega_1, \ldots, \omega_4)$ of the signal:

Normal healthy signal (ω_1) A four state segmentation model with the RNN murmur posterior discarded.
Holosystolic murmur (ω_2) A four state segmentation model, where the murmur posterior replaces the systole posterior.
Early-systolic murmur (ω_3) A five state segmentation model, where the transition matrix requires the S1 state transition to the murmur state and then the systolic state.
Mid-systolic murmur (ω_4) As above, but the model transitions from S1 to systole first.

The predicted classification $\hat{\omega}$ is chosen by calculating a segmentation confidence C_ω for each model by tracing its Viterbi state path, $\hat{q}^{(\omega)}_{1:T}$, through the RNN posteriors:

$$C_\omega = \frac{1}{T}\sum_{t=1}^{T} P\big(q_t = \hat{q}_t^{(\omega)} | x_{1:T}, \theta_R\big) \tag{1}$$

$$\hat{\omega} = \arg\max_{\omega}(C_\omega) \tag{2}$$

The maximal confidence $C_{\hat{\omega}}$ is used as a measure of signal quality. To produce an overall classification for a patient based on multiple individual recordings, we follow a simple criteria that follows what a clinician would do when listening to multiple sites on the chest. If any of the signals are detected as a murmur ($\hat{\omega} \in \{\omega_2, \omega_3, \omega_4\}$), then 'Murmur Present' is predicted. If this is not true and $C_{\hat{w}}$ for any signal falls below a threshold (0.65), 'Unknown' is predicted. Otherwise, 'Murmur Absent' is predicted.

2.4. Prediction of Clinical Outcome

Simply using the murmur prediction to predict abnormal clinical outcome leads to a poor challenge score. Even using the provided ground-truth murmur label to predict clinical outcome leads to a result with poor sensitivity (42%) and a challenge cost of 16083.

Murmurs heard at different locations on the chest have different levels of clinical significance, and the challenge dataset also provides general biometrics such as age, sex and weight. We apply a gradient boosted decision tree to automatically combine this information to predict clinical outcome. For each heart valve recording of a patient, the HSMM confidence difference between the best murmur and normal models is computed, as well as the confidence $C_{\hat{\omega}}$ of the chosen model. Where chest locations have multiple recordings, these values are averaged. We combine this with the patient's age, sex, pregnancy status, and the number of recordings to form the full feature input.

A decision tree is implemented using the CatBoost Python library [8]. The model is trained and optimised using five-fold cross validation, with class weight of 1.8 for the abnormal examples and 1 for the normal examples. The decision tree has a depth of 9, and is trained using a cross-entropy loss. The threshold probability to decide an abnormal result is then chosen to minimise the outcome cost.

Training	Validation	Test	Ranking
0.799	0.758	-	6 / 61

Table 2. Challenge weighted accuracy for the murmur detection task. We used 5-fold cross validation on the public training set and repeated scoring on the validation set.

Class	Cases	Sensitivity (%)	PPV (%)
Present	179	92.7	55.0
Unknown	68	30.9	34.4
Absent	695	77.6	93.1

Table 3. Per-class sensitivity and positive predictive value (PPV, a.k.a. precision) for the murmur detection task, evaluated via 5-fold cross validation on the training data.

3. Results

Table 2 shows the weighted murmur accuracy achieved by the model on the training and validation sets, with Table 3 showing the per-class breakdown of performance on the training set. A total of 13 'Murmur Present' cases were misclassified, all of which are grade 1 systolic murmurs.

Similarly, Table 4 shows the challenge outcome scores achieved by the model on the training and validation sets. On the training set, the model achieves a sensitivity of 84%, a specificity of 31%, and a positive predictive value of 53%.

4. Discussion and Conclusions

Our approaches ranked highly on both tasks, scoring 6th and 9th respectively on the murmur and outcome tasks. The plot of the confidence values in Figure 2 shows a gen-

Training	Validation	Test	Ranking
11040	9257	-	9 / 61

Table 4. Challenge cost metric scores for the clinical outcome identification task, evaluated via 5-fold cross validation (training set) and repeated scoring (validation set).

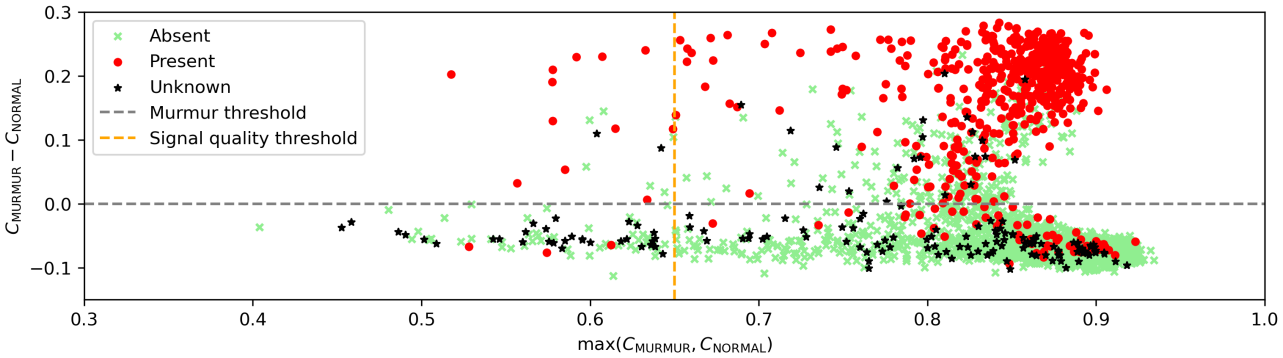

Figure 2. Murmur detection and signal quality check using confidences in murmur ($C_{\text{MURMUR}} = \max(C_2, C_3, C_4)$) and normal HSMM ($C_{\text{NORMAL}} = C_1$) segmentations. If the murmur confidence is greater than the normal confidence (gray line), a murmur is classified. If no murmur detected and the maximum confidence in either HSMM is less than 0.65 (orange line), the signal is deemed poor quality. Otherwise, the signal is classified as normal.

erally strong separation between the murmur and normal classes, and inspection of the classifier false positives suggests some borderline cases. The signal quality threshold is less successful in separating the 'Unknown' class from the rest, but this is not unexpected because the definition of a poor quality recording will be very different from an algorithm and clinician perspective. Recordings with high frequency noise or spikes may be problematic for a human ear but are automatically filtered out by the algorithm.

The outcomes task has been challenging for all participants. The poor sensitivity of even the ground-truth murmur labels at predicting clinical outcome suggests that many of the diseases in the dataset do not carry an audible murmur. The heavy weighting towards sensitivity also forces algorithms to operate with a generally low specificity, which would limit their usefulness in a widespread screening program where false positive referrals will quickly overwhelm secondary care. To improve practical usefulness, it may be pragmatic to focus designs on predicting a certain set of cardiac diseases rather than a general abnormality.

The use of a segmentation-based approach is in contrast to many of the other participants who design end-to-end neural network models. Our approach provides a simultaneous segmentation of the signal which aids interpretability of the diagnoses, and the simple murmur decision criteria should generalise well to future datasets. Future improvements could include modelling of more murmur conditions (late-systolic and diastolic), as well as relaxations of the HSMM durations to better model arrhythmic signals.

Acknowledgments

Andrew McDonald is supported by the UK Medical Research Council (MR/S036644/1).

References

[1] Gardezi SKM, Myerson SG, Chambers J, Coffey S, d'Arcy J, Hobbs FDR, et al. Cardiac auscultation poorly predicts the presence of valvular heart disease in asymptomatic primary care patients. Heart 2018;104(22):1832–1835.

[2] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[3] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[4] Clifford GD, Liu C, Moody B, Millet J, Schmidt S, Li Q, et al. Recent advances in heart sound analysis. Physiological Measurement aug 2017;38(8):E10–E25.

[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] Kay E. Cardiac Acoustics: Understanding and Detecting Heart Murmurs. Ph.D. thesis, University of Cambridge, 2018.

[7] Messner E, Zohrer M, Pernkopf F. Heart Sound Segmentation-An Event Detection Approach Using Deep Recurrent Neural Networks. IEEE Transactions on Biomedical Engineering 2018;65(9):1964–1974.

[8] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: unbiased boosting with categorical features, 2017. URL https://arxiv.org/abs/1706.09516.

Address for correspondence:

Andrew McDonald
Department of Engineering, Trumpington Street, Cambridge, UK
andrewmcdonald@cantab.net