Joint Training of Hidden Markov Model and Neural Network for Heart Sound Segmentation

Francesco Renna[1], Miguel L. Martins[2], Miguel Coimbra[2]

[1] Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto, Portugal
[2] INESC TEC, Faculdade de Ciências da Universidade do Porto, Portugal

Abstract

In this work, we propose a novel algorithm for heart sound segmentation. The proposed approach is based on the combination of two families of state-of-the-art solutions for such problem, hidden Markov models and deep neural networks, in a single training framework.

The proposed approach is tested with heart sounds from the PhysioNet dataset and it is shown to achieve an average sensitivity of 93.9% and an average positive predictive value of 94.2% in detecting the boundaries of fundamental heart sounds.

1. Introduction

Cardiac auscultation is a first line of assessment of the cardiac activity that is particularly attractive for the application in underprivileged scenarios, due to its low cost, simplicity, and ability to detect several heart conditions [1]. Each recorded heartbeat is composed by two fundamental sounds: the first sound (S1), that is generated by vibrations of the mitral and tricuspid valves at the beginning of the systole, and the second sound (S2), that is generated by the closure of the aortic and pulmonary valve at the beginning of the diastole.

Heart sound segmentation consists in detecting the location and boundaries of fundamental heart sounds and, consequently, systolic and diastolic intervals, in each heartbeat. The identification of these four fundamental segments in each heart cycle plays a key role in the analysis of the phonocardiogram (PCG), i.e., the heart sound signal, as it allows the detection and localization of extra sound components (e.g., the third and fourth heart sounds, murmurs, ejection clicks, etc.) and it allows the extraction of useful information from the analysis of the morphology of the waveforms associated to the S1 and S2 sounds.

Current state-of-the art solutions for heart sound segmentation can be roughly divided into two classes. The first class leverages statistical models to include prior information about the sequential nature of PCG signals, mainly hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs). For both models, each data point composing a heart sound recording is mapped into a hidden state. In particular, [2] has introduced the use of HSMMs for PCG segmentation, which allow explicit modeling of the statistics of the time spent by the system in each state, i.e., the sojourn time. Then, other works have improved further the performance of HSMM-based segmentation algorithms, by considering a modified Viterbi algorithm that addresses boundary conditions [3], or by proposing methods to adapt sojourn time and emission distributions to the specificities of each PCG signal [4, 5].

A second class of segmentation algorithms is based on the used of deep learning architectures. Deep convolutional neural networks (CNNs) have been applied to envelopes extracted from heart sound recordings for heart sound segmentation in [6]. Also, deep learning sequential models, namely recurrent neural networks (RNNs), have been leveraged to segment PCG signals, by keeping track of the temporal dependencies embedded in a PCG signal [7]. More recent deep learning approaches for heart sound segmentation have focused on enhancing the capacity of deep learning models in modeling the sequential behavior of heart sounds. In particular, [8] has considered the use of a bidirectional long short-term memory (LSTM) in conjunction with an attention mechanism, which enables to identify the most salient aspect of the signal, thus providing enhanced robustness against noisy and irregular recordings. Finally, [9] has proposed the use of a temporal-framing adaptive network which is trained with a specific transition loss and is able to perform dynamic inference, thus adapting to irregular heart sound behaviors.

In general, recent results in heart sound segmentation have shown that more robust results can be obtained when highly discriminant deep learning models are accompanied by mechanisms able to jointly take into account the semi-periodic sequential nature of PCG signals.

In this work, we propose a novel heart sound segmentation framework which combines the benefits of both HMM-based approaches and more recent deep learning methods. Inspired by results in automatic speech recognition [10], a hybrid model-based/data-driven approach is proposed, which incorporates the strong ability of HMMs in modeling explicitly the semi-periodic nature of heart sounds and the high discriminative power of deep neural networks (DNNs). The overall hybrid model is trained end-to-end using a gradient-based approach.

In particular, this work contains the following contributions:

1. A novel heart sound segmentation framework, based on the joint training of a HMM and a DNN;
2. The test of the proposed method and comparison with state-of-the-art approaches based on HSMMs [3] and deep CNNs [6] over the PhysioNet dataset.

2. Methods

In this section, we describe the proposed approach for heart sound segmentation, providing details about input pre-processing, the definition of the hybrid HMM-DNN framework, training method, and post-processing.

2.1. Pre-processing

Heart sound recordings are first filtered with a Butterworth filter of order two with pass-band $[25, 400]$ Hz and then processed with the spike removal algorithm presented in [2]. Then, the four evnvelopes considered in [3, 6] are extracted from the filtered signals (homomoprhic envelope, Hilbert envelope, power spectral density envelope, and wavelet envelope), and normalized to have zero mean and unit variance.

We denote with $\mathbf{x}_t \in \mathbb{R}^4$, for $t = 1, \ldots, T$ the 4-dimensional signal obtained consdiering the four envelopes extracted from a given PCG with length T and with s_t, for $t = 1, \ldots, T$, the corresponding state label, where $s_t \in \{0, \ldots, L-1\}$ and L represents the total number of possible signal states. In this work, we consider $L = 4$, where the possible PCG states are S1, systole, S2, and diastole.

2.2. Hybrid model

Our objective is to model the PCG signal via an underlying HMM whose emission probabilities are modeled via a DNN. In contrast with previous work appeared in the literature about combining the outputs of DNNs with a HMM [6], we define a learning strategy to *jointly* train the HMM and the DNN at the same time, using a set of annotated PCG recordings, thus allowing the sequential information contained in the HMM to be used in the traininig phase of the DNN.

We first define input feature maps for the DNN which are obtained by collecting the signal vectors contained in an observation window as follows:

$$\mathbf{o}_t = [\mathbf{x}_{t-F}, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_{t+F-1}] \in \mathbb{R}^{4\times 2F}, \qquad (1)$$

for some integer F.

On denoting by $s_1^T = [s_1, \ldots, s_T]$ and by $\mathbf{o}_1^T = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$ the state sequence and the observed feature map sequence associated to a PCG of length T, respectively, we can characterize the hybrid model used for segmentation in terms of their joint probability as follows:

$$P(\mathbf{o}_1^T, s_1^T) = p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} P(\mathbf{o}_t | s_t), \qquad (2)$$

where p_{s_1} is the probability of being in state s_1 at the first time step, p_{s_{t-1},s_t} is the transition probability from the state visited at time $t-1$ to the state visited at time t, and $P(\mathbf{o}_t|s_t)$ is the emission probability of the feature map $\mathbf{o}_t$ given that the signal is in state s_t at time t.

In order to combine a deep learning model within the HMM framework defined by (2), we assume that a DNN provided with an observation feature map $\mathbf{o}_t$ generates outputs $y_{t,s_t,\mathbf{\Theta}}(\mathbf{o}_t)$, which are estimates of the probabilities $P(s_t|\mathbf{o}_t)$, where $\mathbf{\Theta}$ is a vector containing the trainable parameters of the DNN. In the following, we will drop the explicit dependence of $y_{t,s_t,\mathbf{\Theta}}(\mathbf{o}_t)$ from $\mathbf{\Theta}$ and $\mathbf{o}_t$, to simplify the notation.

Then, we can express the joint probability in (2) as

$$P(\mathbf{o}_1^T, s_1^T) = p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} \frac{P(\mathbf{o}_t)}{P(s_t)} \prod_{t=1}^{T} y_{t,s_t}. \qquad (3)$$

Moreover, the marginal probability of the observation sequence $\mathbf{o}_1^T$ is obtained by summing the joint probability in (2) over all possible state sequences of length T, $s_1^T \in \mathcal{S}$,

$$P(\mathbf{o}_1^T) = \sum_{s_1^T \in \mathcal{S}} p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} \frac{P(\mathbf{o}_t)}{P(s_t)} \prod_{t=1}^{T} y_{t,s_t}. \qquad (4)$$

Note that the marginal probability in (4) can be efficiently computed using the forward-backward algorithm [11].

2.3. Training

We assume that we have access to a training set containing N PCG recordings with the corresponding state sequences, i.e., we have access to the set of pairs $\{(\mathbf{o}_1^T)_n, (s_1^T)_n\}_{n=1}^N$. Then, the objective of training is to find the set of parameters of the model that are optimal according to a given criterion. We denote by $\mathbf{\Psi} = [\boldsymbol{\pi}, \mathbf{\Gamma}, \mathbf{\Theta}]$ the set all parameters of the model, where the vector $\boldsymbol{\pi}^{\mathrm{T}} = [p_0, \ldots, p_{L-1}]$ collects the initial state probabilities of the underlying Markov model, the elements of the matrix $\mathbf{\Gamma} \in \mathbb{R}^{L\times L}$ are $\mathbf{\Gamma}_{\ell,\ell'} = p_{\ell,\ell'}$, i.e., the state transition probabilities of the underlying HMM, and $\mathbf{\Theta}$ are the parameters of the DNN. We adopt the maximum mutual information (MMI) criterion for training [10], i.e., we search for the model parameters that maximize the following objective function:

$$\mathcal{L}(\mathbf{\Psi}) = \sum_{n=1}^{N} \log P\left((\mathbf{o}_1^T)_n, (s_1^T)_n\right) - \log P\left((\mathbf{o}_1^T)_n\right), \quad (5)$$

where $P\left((\mathbf{o}_1^T)_n, (s_1^T)_n\right)$ and $P\left((\mathbf{o}_1^T)_n\right)$ are computed using (3) and (4), respectively.

The optimal model parameters are searched using a gradient-based optimization approach.

2.4. Inference and post-processing

At inference time, PCG test signals are pre-processed according to the steps described in Section 2.1. Then, the trained model is applied to the features maps obtained from the pre-processed test data and the corresponding outputs y_{t,s_t} are used to approximate the emission probabilities $P(\mathbf{o}_t|s_t)$. Finally, the output sequence of states $\hat{s}_1^T$ is computed using the Viterbi algorithm [11].

3. Experiments

The proposed hybrid HMM-DNN segmentation algorithm is compared with the HSMM-based method from [3] and the CNN-based approach of [6]. The performance from the considered methods is tested via 10-fold cross-validation over the available heart sound dataset, by assuring that, at each iteration, sounds from patients contained in the test set are not contained in the training set.

The DNN considered for the experiments is a simple CNN starting with three blocks of one dimensional convolutions with rectified linear unit (ReLU) activation functions followed by max-pooling layers. A kernel size of 3 with stride of 1 is used throughout all convolutional layers and they stack 8, 16, and 32 filters, respectively. The max-pooling layers have a kernel size and stride of 2. The bottleneck features are passed through a 25% dropout layer and fed to a single hidden dense layer of size 64 using a ReLU activation function. The output layer uses a softmax activation function returning the estimated conditional state probability densities $P(s_t|\mathbf{o}_t) \in \mathbb{R}^L$.

The dimension of the DNN input feature maps is $F = 32$. The underlying Markov model parameters $\boldsymbol{\pi}$ and $\mathbf{\Gamma}$ were first estimated with a maximum likelihood approach and then kept fixed while estimating the DNN parameters $\mathbf{\Theta}$. These are obtained by maximizing the objective function $\mathcal{L}(\mathbf{\Psi})$ using the Adam algorithm [12] , with learning rate 10^{-4}. The maximum number of epochs for training is fixed to 50, and early stopping is implemented by extracting 10% of the training data for validation and retaining the network weights corresponding to the model that guarantees the highest validation accuracy among all epochs.

The performance of the considered segmentation algorithms in determining the position of the fundamental heart sounds S1 and S2 is evaluated in terms of their sensitivity (S) and positive predictive value (P_+). Such metrics are computed according to the description in [2], where true positives are counted when the mismatch between the center of a sound in the estimated sequence and in the ground truth sequence is lower than 60 ms. All performance metrics are computed for each recording in the test set and then averaged over the test set. Finally, the values corresponding to the different 10 test subsets are reported.

The heart sounds used for the experiments were taken from the dataset made publicly available for the PhysioNet/CinC challenge 2016. In particular, we considered 792 heart sounds recorded from 135 patients in different clinical and non-clinical environments.[1] From those, 181 sounds are collected from patients with pathological heart lesions (most commonly mitral valve prolapse), as assessed by echocardiography. The remaining 246 sounds are collected from healthy patients. Sound recordings are sampled at 1 kHz. The annotations provided with the dataset are obtained from the analysis of synchronous ECG recordings.

In Fig. 1 are reported the values of sensitivity (S) and positive predictive value (P_+) for the proposed method and the algorithms in [3] and [6]. It is possible to note that the proposed methods slightly outperforms the segmentation algorithms considered for comparison both in terms of sensitivity and positive predictive value. These results can be explained by the enhanced capability of the proposed hybrid model in embedding explicitly domain knowledge regarding the sequential nature of the PCG signal when compared to the application of a CNN classifier followed by Viterbi decoding.

Such behavior can be observed in the segmentation example reported in Fig. 2, where it is possible to observe that early embedding of the HMM in the CNN training allows to avoid segmentation errors leading to output sequences with reduced physiological significance, characterized by exaggeratedly long or short diastolic periods.

4. Conclusion

In this paper, a hybrid model-based/data-driven approach for heart sound segmentation was presented. The proposed framework consists in the joint training of a HMM, able to explicitly embed sequential information about heart sound states, and a highly discriminative DNN, via a mutual information maximization criterion.

[1] The sounds are available online at https://PhysioNet.org/physiotools/hss/.

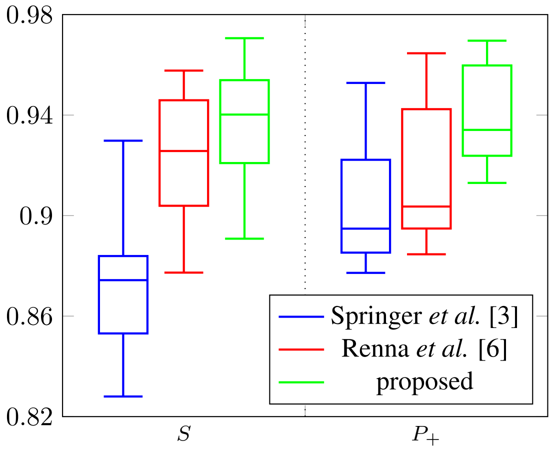

Figure 1. Sensitivity (S) and positive predictive value (P_+) of the HSMM-based method in [3] (blue, left), CNN-based method in [6] (red, center), and proposed method (green, right).

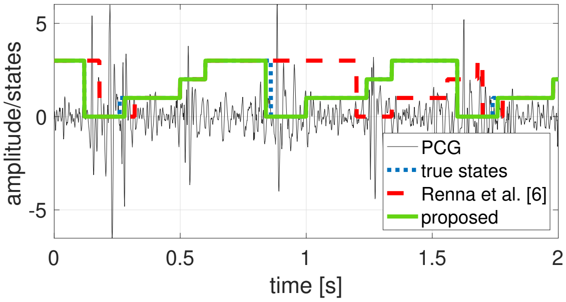

Figure 2. Segmentation example reporting the values of the ground truth state sequence (blue, dotted line) and the output sequences of the CNN-based approach in [6] (red, dashed line) and the proposed approach (green, solid line).

The proposed approach is shown to provide segmentation performance superior to current, more complex CNN architectures and HSMM approaches. This is motivated by the high flexibility of the hybrid model in striking a balance between explicit sequential modeling and discrminiative power.

Acknowledgements

This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020. This work is also a result and funded by the projects: DigiScope2 (POCI-01-0145-FEDER-029200 - PTDC/CCI-COM/29200/2017), funded by Fundo Europeu de Desenvolvimento Regional (FEDER), through Programa Operacional Competitividade e Internacionalização (POCI), and UID/EEA/50008/2019. The work of F. Renna is also funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Individual Call - CEECIND/01970/2017.

References

[1] Mendis S, Puska P, Norrving B, Organization WH, et al. Global atlas on cardiovascular disease prevention and control. World Health Organization, 2011.

[2] Schmidt S, Holst-Hansen C, Graff C, Toft E, Struijk JJ. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. Physiological measurement 2010;31(4):513–529.

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

[4] Oliveira J, Renna F, Coimbra MT. Adaptive sojourn time HSMM for heart sound segmentation. IEEE Journal of Biomedical and Health Informatics 2019;23(2):642–649.

[5] Oliveira J, Renna F, Coimbra M. A subject-driven unsupervised hidden semi-Markov model and Gaussian mixture model for heart sound segmentation. IEEE Journal of Selected Topics in Signal Processing 2019;13(2):323–331.

[6] Renna F, Oliveira JH, Coimbra MT. Deep convolutional neural networks for heart sound segmentation. IEEE Journal of Biomedical and Health Informatics 2019; 23(6):2435–2445.

[7] Messner E, Zöhrer 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] Fernando T, Ghaemmaghami H, Denman S, Sridharan S, Hussain N, Fookes C. Heart sound segmentation using bidirectional lstms with attention. IEEE Journal of Biomedical and Health Informatics 2020;24(6):1601–1609.

[9] Wang X, Liu C, Li Y, Cheng X, Li J, Clifford GD. Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration. IEEE Transactions on Biomedical Engineering 2021;68(2):650–663.

[10] Fritz L, Burshtein D. Simplified end-to-end MMI training and voting for ASR. arXiv170310356 2017;.

[11] Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[12] Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv14126980 2014;.

Address for correspondence:

Name: Francesco Renna
Full postal address: Rua do Campo Alegre 1021/1055,
4169-007 Porto, Portugal
E-mail address: frarenna@dcc.fc.up.pt