A Markov-Switching Model Approach to Heart Sound Segmentation and Classification

Fuad Noman*, *Student Member, IEEE*, Sh-Hussain Salleh, Chee-Ming Ting, *Member, IEEE*, S. Balqis Samdin, Hernando Ombao and Hadri Hussain

***Abstract*—*Objective:* We consider challenges in accurate segmentation of heart sound signals recorded under noisy clinical environments for subsequent classification of pathological events. Existing state-of-the-art solutions to heart sound segmentation use probabilistic models such as hidden Markov models (HMMs) which, however, are limited by its observation independence assumption and rely on pre-extraction of noise-robust features. *Methods:* We propose a Markov-switching autoregressive (MSAR) process to model the raw heart sound signals directly, which allows efficient segmentation of the cyclical heart sound states according to the distinct dependence structure in each state. To enhance robustness, we extend the MSAR model to a switching linear dynamic system (SLDS) that jointly model both the switching AR dynamics of underlying heart sound signals and the noise effects. We introduce a novel algorithm via fusion of switching Kalman filter and the duration-dependent Viterbi algorithm, which incorporates the duration of heart sound states to improve state decoding. *Results:* Evaluated on Physionet/CinC Challenge 2016 dataset, the proposed MSAR-SLDS approach significantly outperforms the hidden semi-Markov model (HSMM) in heart sound segmentation based on raw signals and comparable to a feature-based HSMM. The segmented labels were then used to train Gaussian-mixture HMM classifier for identification of abnormal beats, achieving high average precision of 86.1% on the same dataset including very noisy recordings. *Conclusion:* The proposed approach shows noticeable performance in heart sound segmentation and classification on a large noisy dataset. *Significance:* It is potentially useful in developing automated heart monitoring systems for pre-screening of heart pathologies.**

***Index Terms*—Dynamic clustering, autoregressive models, regime-switching models, state-space models, Viterbi algorithm.**

I. Introduction

CARDIAC auscultation is a critical stage in the diagnosis and examination of heart functionality. Phonocardiogram (PCG) or heart sound provides a recording of subaudible sounds and murmurs from the heart and allows cardiologists to interpret the closure of the heart valves. Heart sounds can reflect the hemodynamical processes of the heart and provide important screening indications of disease in early evaluation stages. The PCG has been proven as an effective tool to reveal several pathological heart defects such as arrhythmias, valve disease, and heart failure [1]. The goal of this paper is to develop an automatic method for heart sounds analysis, particularly the segmentation and classification of fundamental heart sounds, which is useful to detect heart pathology in clinical applications. Three main problems need to be addressed jointly towards fully automatic heart sound analysis. The first is to detect noise to identify the non-cardiac sounds. Second is to segment heart sounds to localize the four main sound components ($S1$, $systole$, $S2$, and $diastole$). Third is to classify heart sounds into healthy and pathological classes. The performance of the heart sound segmentation is highly dependent on the preprocessing step. This is relatively simple in noise-free recordings. However, in clinical environments, this is difficult due to both endogenous or exogenous in-band noise sources that overlap with the heart sounds frequency range [2]. Accurate localization of the fundamental heart sounds will lead to a more accurate classification of any pathologies in the systolic or diastolic regions [3, 4].

Automatic heart sound segmentation methods proposed in the literature can be categorized into three groups: (1) Envelope-based methods [5–11]; (2) Feature-based methods [12–19]; (3) Machine learning-based methods [20–24]. Refer [1, 3] for further reviews and details of these methods. Machine learning methods based on probabilistic models show an improved accuracy on heart sound segmentation. Gamero and Watrous [25] proposed a hidden Markov model (HMM) approach to detect the $S1$ and $S2$ sounds, using a topology combining two separate HMMs to model the mel-frequency cepstral coefficients (MFCC) of the systolic and diastolic intervals, respectively. In [26], a variable-state embedded HMM was introduced to model the heart sound components using MFCC, Shannon energy, and regression coefficients as features. Gill *et al.* [27] suggested a modified HMM to allow for a smooth transition between states. Sedighian *et al.* [28] also used a homomorphic filtering approach to extract envelograms from the heart sound recordings. Envelope peak detection method was used along with two-states HMM to identify the $S1$ and $S2$ sounds. The method was evaluated on the PASCAL database [29]. Schmidt *et al.* [30] proposed a duration-dependent HMM to model the transition duration of each HMM state, and achieved remarkable performance on a dataset of 113 subjects. This work was extended in [3] using the hidden semi-Markov model (HSMM) with the modified Viterbi algorithm to detect the beginning and end state of the heart sound signal, which was evaluated on a larger dataset of 10,172 seconds of heart sound recordings from healthy and pathological cases including mitral valve prolapse (MVP). While feature-based HMMs remain the current state-of-the-art solution to accurate PCG segmentation, it suffers from several major drawbacks inherent in its modeling mechanism. First, HMMs make an unrealistically strong Markovian assumption that observations are conditionally independent given a state, and thus are unable to account for additional dynamics, e.g.,

F. Noman, S.-H. Salleh and Hadri Hussain are with the School of Biomedical Engineering & Health Sciences, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia (e-mail: mnfuad3@live.utm.my).

C.-M. Ting is with the Medical Devices & Technology Centre, School of Biomedical Engineering & Health Sciences, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, and also the Statistics Program, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia.

H. Ombao and S. B. Samdin are with Statistics Program, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia.

autocorrelation structure typically present in the temporal segment of each heart-sound component. Secondly, conventional HMM does not include a specification for noise, and relies heavily on pre-processing to remove noise and preliminary extraction of robust features to obtain good segmentation performance. This limitation renders it very ineffective when segmenting raw noisy heart sound signals.

Switching linear dynamic system (SLDS) [31, 32] has been introduced as a generalization of HMMs in which each state is associated with a linear dynamical process. It provides a richer framework for modeling complex dynamical phenomena that is both nonlinear and nonstationary. For example, a special case of SLDS - Markov-switching autoregressive (MSAR) model can capture temporal dependencies often found in real-world signals. By repeated returns to a set of simpler (approximately linear) dynamic models, SLDS can effectively model structural changes in signals driven by some underlying time-evolving states that reoccur at certain time intervals [33]. Moreover, the SLDS allows direct modeling of raw signals which has the advantage that the noise can be explicitly modeled. Leveraging on a joint modeling of both raw and noise signals, it offers better noise robustness and was shown to significantly outperform a state-of-the-art feature-based HMM in speech recognition [34]. SLDS has been used in many applications including financial time series [35]; motion tracking [36]; anomaly detection [32]; environmental monitoring [37]. A multivariate MSAR model formulated into a SLDS was employed to track state-related changes in connectivity between brain signals [38]. The switching Kalman filter (SKF) has been applied to electrocardiogram (ECG) signals for ventricular beat detection [39] and apnea bradycardia detection in [40] which showed better performance than HMMs.

In this paper, we develop a unified framework based on MSAR-SLDS models of raw PCG signals with enhanced state inference algorithms to segment the fundamental components of heart sound for subsequent use in classification of heart pathologies. The cyclic repetition of the heart sound components each characterized by different dependence structure can be more concisely described by the MSAR via switching between a set of unique AR models (with different parameters), compared to an HMM assuming conditionally-independent observations. It allows efficient segmentation of raw PCG recordings into quasi-stationary temporal segments according to distinct AR structures (corresponding to frequency contents) in different heart sound states. To enhance noise robustness, we extend the MSAR into a SLDS by incorporating an explicit model to accommodate noise effects in raw signals which are neglected in the HMMs. The raw heart sound signals can be viewed as corrupted version of a latent clean MSAR process that captures the underlying evolving autocorrelations over cardiac cycles. We show that the MSAR-SLDS models significantly outperform HMMs trained on raw signals and are comparable to feature-based HMMs in heart-sound segmentation on noisy recordings.

Under the MSAR-SLDS model, we first employ the SKF with refinement by the switching Kalman smoother (SKS) to infer sequentially the latent heart sound states given the raw signals. However, as in HMM, the MSAR model implicitly assume the duration or sojourn time of each state to be geometrically distributed, i.e., probability of dwelling in a state decreases as the sojourn time increases. This tends to induce rapid regime switching and may not be appropriate for the relatively more enduring heart sound states. Inspired by the idea of HSMM and its superior segmentation performance over HMM [3, 30], we enhance the MSAR-SLDS framework by incorporating a priori information about the expected duration of the heart sound states to improve state decoding. Specifically, we extend the duration-dependent Viterbi algorithm for the MSAR-SLDS model by replacing the independent observation probabilities for HSMM with the filtered probability of the latent MSAR process generated by the SKF to compute the likelihood of the most probable state sequence. The detected heart sound segments by the proposed method were then used in the training of continuous-density HMMs with Gaussian mixtures for heart sound classification. A preliminary version of this work on the segmentation has been reported in [41].

The main contributions of this work are summarized as follows:

1) We, for the first time, exploit an MSAR-SLDS model for accurate heart-sound segmentation based on raw signals without requiring any preliminary feature extraction steps as in conventional HMMs.
2) We propose a novel algorithm combining the duration-dependent Viterbi algorithm with SKF for state inference with an enhanced MSAR-SLDS model that incorporates the state duration of heart sounds, which substantially improves segmentation performance of the standard SKF.
3) We rigorously evaluate the proposed segmentation approach in a unified framework with heart sound classification on one of the largest published datasets, including very noisy recordings (X-Factor). Most previous works on automatic heart sound analysis focused on either segmentation or classification problems using small datasets.

II. Heart Sound Database

Heart sound recordings from a publicly available database released in Physionet/Computing in Cardiology (CinC) Challenge 2016 [1] were used in this study to evaluate the proposed segmentation and classification methods. The Challenge training set includes six databases (*a* through *f*) collected from different research groups in both clinical and nonclinical environments [42]. It consists of 764 subjects with a total of 3153 heart sound recordings, manually labeled by experts into three classes: 2302 normal; 572 abnormal; and 279 *unsure* (poor signal quality or very noisy). The duration of each recording varied from 5.3 s to 122 s with a total length of 19 hours and 73 minutes. The data were recorded using heterogenous equipment at four common locations of aortic, pulmonary, tricuspid and mitral areas, and re-sampled to 2 kHz. Table I summarizes the number of heart-beats in the dataset, where each beat begins at the start of $S1$ sound until the start of the next $S1$ sound, giving a total of 81498 beats (65152 normal and 16346 abnormal). Reference annotations for the four heart sound states were provided only for the clean recordings. Therefore, the unsure recordings were excluded

TABLE I
DISTRIBUTION OF HEART-BEATS IN PHYSIONET DATABASE.

Dataset	Beat count		Total beats	Ignored rec.†
	Normal	Abnormal		
training-*a*	4301	9860	14161	17
training-*b*	2396	589	2985	122
training-*c*	356	1425	1781	4
training-*d*	308	493	801	3
training-*e*	54783	2841	57624	129‡
training-*f*	3008	1138	4146	6⋆
Total	65152	16346	81498	281

Noisy recordings†, including (e00210)‡, (f0043)⋆

from the evaluation for segmentation analysis. For classification, we included the unsure recordings as a separate class, referred as X-Factor as in the ECG classification literature. This is a more challenging task which has not been evaluated for heart sound classification in previous studies.

Since the original Challenge testing set is still not publicly available, we consider the following two schemes for new train-test split out of the Challenge training set.

A. K-Fold Cross-Validation

The entire recordings were split into train and test sets, with and without X-Factor based on a stratified 5-fold cross-validation. Although cross-validation is effective for preventing model over-fitting, the random split may result in the same subjects to appear in both the train and test sets which will falsely inflate classification accuracy. Also note that the 5-fold train-test splits are unbalanced in terms of number of beats.

B. Subject-Oriented Partition

To ensure no overlapping subject affiliation in both the train and test sets, the Challenge dataset in Table I were split such that the test-set contains totally unseen *subject-oriented* recordings from those in the train-set. For training-*b* and training-*e* where the subject ID and information are available (provided in online appendix of the database), we assigned recordings from two different sets of subjects to the train and test sets, respectively. For training-*c* and training-*f* where each recording corresponds to a distinct subject despite that the ID labels are not provided, the recordings were split evenly to the train and test sets. For training-*a* and training-*d* where multiple recordings may belong to the same subjects and subject IDs for these recordings are unknown, the entire training-*a* was included in the train-set and the entire training-*d* in the test-set without splitting. Hence this ensures that subjects in the train-set are exclusively different from that in the test-set. This results in a total of 2072 and 800 normal and abnormal recordings in the train and test sets respectively, as shown in Table II. A total of 2634/147 and 2182/132 X-Factor beats/recordings were also assigned to the train and test sets, respectively. This approach provides a more thorough and realistic performance evaluation of the proposed method in segmentation/classification on unseen heart sound data.

III. METHODS

This section first introduces the MSAR modeling of heart sound and a novel SKF-Viterbi algorithm for heart-sound segmentation. Then, we present HMM-based heart-sound classification based on the derived segmentation boundaries.

TABLE II
DISTRIBUTION OF TRAIN-SET AND TEST-SET (BEATS AND RECORDINGS). NUMBERS IN PARENTHESES INDICATE NOISY (X-FACTOR) CLASS.

Dataset	Challenge set	Normal		Abnormal		**Total**	
		#Beats	#Rec	#Beats	#Rec	#Beats	#Rec
Train	training-*a*	4301(35)	116(1)	9860(438)	276(16)	14161(473)	392(17)
	training-*b*	1217(368)	156(46)	296(119)	37(15)	1513(487)	193(61)
	training-*c*	177(0)	3(0)	710(70)	10(2)	887(70)	13(2)
	training-*e*	41039(1044)	1347(46)	1455(476)	74(19)	42494(1520)	1421(65)
	training-*f*	1502(38)	38(1)	568(46)	15(1)	2070(84)	53(2)
	Total	48236(1485)	1660(94)	12889(1149)	412(53)	**61125**(2634)	**2072**(147)
Test	training-*b*	1179(360)	139(45)	293(126)	36(16)	1472(486)	175(61)
	training-*c*	179(0)	4(0)	715(66)	10(2)	894(66)	14(2)
	training-*d*	308(8)	26(1)	493(33)	26(2)	801(41)	52(3)
	training-*e*	13746(1045)	432(45)	1386(475)	72(16)	15132(1520)	504(63)
	training-*f*	1506(23)	39(1)	570(46)	16(2)	2076(69)	55(3)
	Total	16918(1436)	640(92)	3457(746)	160(40)	**20735**(2182)	**800**(132)

A. Heart Sound Segmentation

Fig. 1 illustrates the proposed framework for heart sound segmentation consisting of three stages: (1.) Pre-processing to assess signal quality and to remove noise and artifacts. (2.) Estimation of MSAR parameters based on dynamic clustering using the reference data labels. (3.) Segmentation using different inference algorithms, i.e., SKF, SKS and fusion of SKF and Viterbi algorithm to estimate the most likely alignment of heart sound states to the observation time points.

1) Pre-processing: Recordings labeled with low signal-quality were discarded [1]. To remove noise sources potentially present in the good quality recordings, signals were filtered using a Butterworth band-pass filter with cut-off frequencies of 25 Hz and 400 Hz. Noise spikes were suppressed using a windowed-outlier filter [30]. Each recording was then normalized by mean subtraction and division by standard deviation.

2) MSAR-SLDS: Modeling the heart sound signal is very challenging because it is nonstationary, nonlinear and periodic time series which consists of repeated heart-beats. Moreover, the clean heart sounds are embedded in various physiological noises and artifacts with a very low signal-to-noise ratio. Let $\mathbf{y} = [\mathrm{y}_1 \ldots, \mathrm{y}_T]'$ be a vector of heart sound time series of length T for the entire recording. We assume an additive noise model for the measured raw heart sound signals as follows

$$\mathrm{y}_t = \mathrm{x}_t + \varepsilon_t \tag{1}$$

where ε_t is an i.i.d. Gaussian observational noise with zero mean and covariance R, $\varepsilon_t \sim N(0, R)$. The underlying switching dynamics of the clean heart sound signals are assumed to follow a MSAR process, a collection of stationary AR processes that alternate among themselves over time according to an indicator variable S_t

$$\mathrm{x}_t = \sum_{p=1}^{P} \varphi_p^{(S_t)} \mathrm{x}_{t-p} + \eta_t \tag{2}$$

where $S_t, t = 1, \ldots, T$ is a sequence of time-varying state variables taking values in a discrete space $j = 1, \ldots, K$; $\{\varphi_p^{(j)}, p = 1, \ldots, P\}$ are the AR coefficients at different lags for state j; and $\eta_t \sim N(0, q)$ is a white Gaussian noise. We assume S_t to follow a hidden Markov chain with transition matrix $Z = [z_{ij}], 1 \leq i, j \leq K$ where

$$z_{ij} = P(S_t = j | S_{t-1} = i)$$

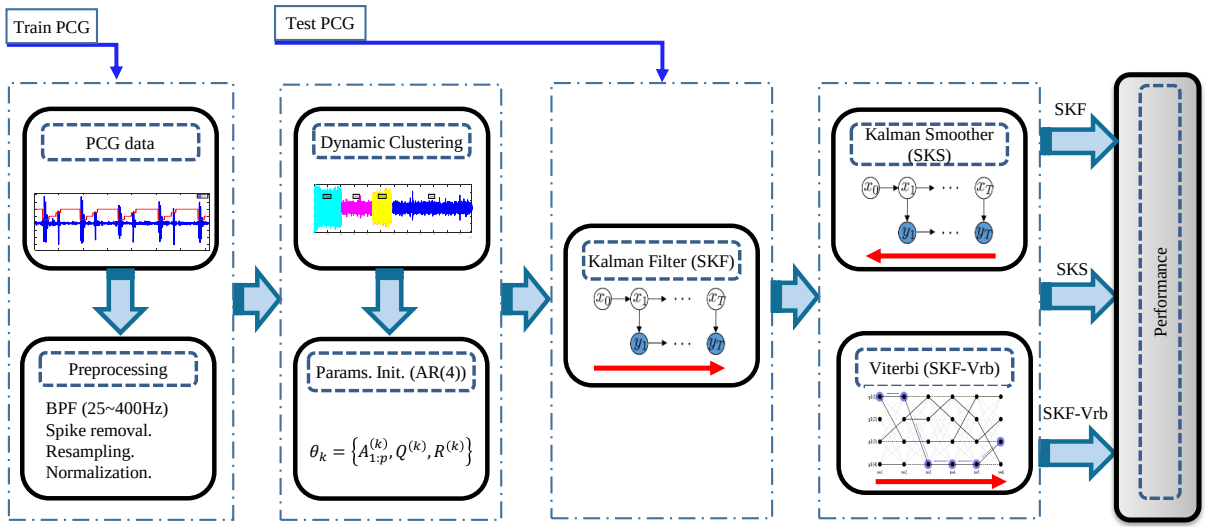

Fig. 1. The proposed MSAR-based framework for heart sound segmentation.

denotes the probability of transition from state i at time $t-1$ to state j at t. Each cardiac cycle of heart sound consists of four fundamental components: $S1$ sound; systolic interval (Sys); $S2$ sound; and diastolic interval (Dia). The heart sound components exhibit distinct dynamic patterns during different time periods, where each can be modeled as a piecewise-stationary AR process of the MSAR model (2). Thus, we use a MSAR with number of states $K = 4$ each corresponding to one of the four components ($j = 1$: $S1$, $j = 2$: Sys, $j = 3$: $S2$ and $j = 4$: Dia). The switching in autocorrelation structure as captured by the state-specific AR coefficients $\varphi_p^{(S_t)}$ between the components is driven by the changes in latent states S_t that indicate which heart-sound component is active at time point t. The segmentation of the heart-sound components can be derived indirectly from the state sequence S_t. We imposed a constraint on the Markovian transition matrix to form a non-ergodic Markov chain with strictly left-to-right topology, allowing only certain pre-specified state transitions according to the temporal order of the heart sound components.

Defining a $P \times 1$ hidden state vector of stacked clean heart sound signals $\mathrm{X}_t = [\mathrm{x}_t, \mathrm{x}_{t-1}, \ldots, \mathrm{x}_{t-P+1}]$, we can formulate the MSAR plus noise model defined in (1)-(2) in a switching linear-Gaussian state-space model, referred as the MSAR-SLDS

$$\mathrm{X}_t = A^{(S_t)}\mathrm{X}_{t-1} + \mathrm{w}_t \quad (3)$$

$$\mathrm{y}_t = C\mathrm{X}_t + \varepsilon_t \quad (4)$$

In the state equation (3), the switching AR(P) process (2) is written as a P-dimensional switching AR(1), where $\mathrm{w}_t = [\eta_t, 0, \ldots, 0]$ is a $P \times 1$ state noise, and $A^{(S_t)}$ is a P matrix of AR coefficients switching according to state variables S_t

$$A^{(S_t)} = \begin{bmatrix} \varphi_1^{(S_t)} & \varphi_2^{(S_t)} & \ldots & \varphi_{P-1}^{(S_t)} & \varphi_P^{(S_t)} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}.$$

In the observation equation (4), the latent MSAR process is observed under noise ε_t as the measured heart sound signals y_t via the $1 \times P$ mapping matrix $C = [1, 0, \ldots, 0]$. We further assume the observation and state noise as white Gaussian processes, i.e. $\varepsilon_t \sim N(0, R^{(S_t)})$ and $\mathrm{w}_t \sim N(0, Q^{(S_t)})$ with

$$Q^{(S_t)} = \begin{bmatrix} q^{(S_t)} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix}.$$

The noise covariance matrices $R^{(S_t)}$ and $Q^{(S_t)}$ are allowed to switch according to S_t. The MSAR model in a state-space form is now fully specified with the model parameters denoted by $\Theta = \{Z, A^{(j)}, Q^{(j)}, R^{(j)}, j = 1, \ldots, K\}$. The estimation algorithms for the unknown state sequence S_t and model parameters Θ are given in the following section.

3) Dynamic Clustering and Model Initialization: To initialize the MSAR model parameters, we first perform the dynamic clustering to group the heart sound time series data that belongs to the same state or component. The clustering was performed by utilizing the heart sound experts annotations attached with the original database. This is followed by fitting a separate stationary AR model to the clustered data of each state to obtain the estimators for the state-specific parameters. Conditioned on the known state sequence derived from the expert's manual annotation labels, we partition temporally the time course of the heart sound recording in the train-set into similar underlying dynamics according to the $K = 4$ heart sound states. Fig. 2 shows an example of clustering a healthy heart sound signal into four dynamic clusters. Note that the time series data of $systoles$ exhibits the similar dynamic structure as that of the $diastole$.

Assuming local stationarity for each of these temporal clusters of heart sound signals, we use this simple procedure to initialize the estimates of the MSAR model parameters. Let $\mathbf{y}^{(j)} = [\mathrm{y}_1^{(j)} \ldots, \mathrm{y}_{T_j}^{(j)}]'$ be $T_j \times 1$ vector of concatenated data being clustered to each heart sound component $j = 1, \ldots, K$, consisting of the y_t with $S_t = j$. We assume the concatenated

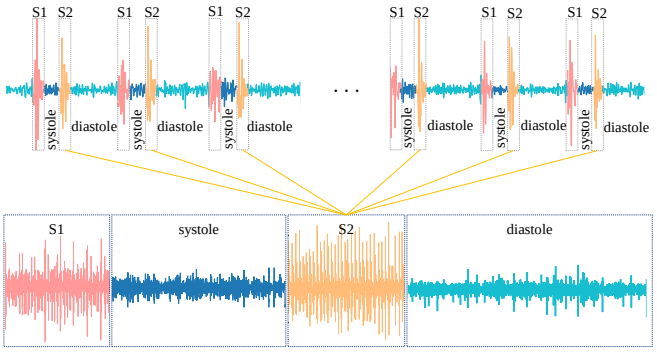

Fig. 2. Dynamic clustering of heart sound into four fundamental components.

time series of each component to follow a distinct stationary AR(P) process

$$\mathrm{y}_t^{(j)} = \sum_{p=1}^{P} \varphi_p^{(j)} \mathrm{y}_{t-p}^{(j)} + \eta_t^{(j)} \tag{5}$$

We compute the initial estimates of the state-specific AR coefficients $\widehat{\varphi}_p^{(j)}$ by a least-square fitting of the AR(P) to $\mathbf{y}^{(j)}$, and the noise variance $\widehat{q}^{(j)}$ based on the estimated residuals $\widehat{\eta}_t^{(j)} = \mathrm{y}_t^{(j)} - \sum_{p=1}^{P} \widehat{\varphi}_p^{(j)} \mathrm{y}_{t-p}^{(j)}$ by $\widehat{q}^{(j)} = 1/T_j \sum_{t=1}^{T_j} \left(\widehat{\eta}_t^{(j)}\right)^2$. Note that the estimators are initialized based on the manual annotations of the heart sound components, which can be subsequently refined based on the switching Kalman filter-derived segmentation. The observation noise variance R is also estimated based on averaged residuals of the fitted AR over sliding-windowed segments of heart sound signal. The state transition probabilities z_{ij} can be initialized by the frequency of transitions from $S_{t-1} = j$ to $S_t = i$.

4) MSAR-based Segmentation Algorithms: Segmenting the heart-sounds can be cast as the problem of estimating the unknown state sequence S_t. Given the sequence of observations $\{\mathrm{y}_t\}_{t=1}^T$, the problem of inference in the switching state-space models is to estimate the posterior probabilities $Pr(S_t = j|\{\mathrm{y}_t\}_{t=1}^T)$ of the hidden state variables S_t. In this paper, we consider three approaches to estimate the state probabilities given the observation sequence. (1) Switching Kalman filter which computes sequentially in a forward recursion the probability densities of the hidden states $P(\mathrm{x}_t|\{\mathrm{y}_t\}_{t=1}^t)$ and $P(S_t|\{\mathrm{y}_t\}_{t=1}^t)$ given observations up to time t; (2) Switching Kalman smoother (or Rauch-Tung-Streibel smoother RTS) computes in a backward recursion refined estimates of densities $P(\mathrm{x}_t|\{\mathrm{y}_t\}_{t=1}^T)$ and $P(S_t|\{\mathrm{y}_t\}_{t=1}^T)$ given the entire observation sequence of length T; (3) Fusion of SKF and extended duration-dependent Viterbi algorithm (SKF-Viterbi) suggested by [3, 30] which decodes the most likely sequence of states given the state probabilities from the one-step ahead Kalman filter predictions $P(S_t = j|M_{t|t}^j)$. The proposed MSAR segmentation methods were initialized by fitting a stationary autoregressive model of order ($P = 4$) on each state observation sequence in a recording-specific manner. The parameters of the MSAR model were computed by averaging parameter estimates overall recordings in the train-set.

a) Switching Kalman Filter (SKF): Algorithm 1 summarizes the procedure of SKF for estimating the hidden

Algorithm 1 : Switching Kalman filter

Inputs: $\mathrm{x}_0^{ij}, P_0^{ij}, M_0^j, \{\mathrm{y}_t\}_{t=1}^T, A, C, R, Q, Z$
Outputs: $M_{t|t}^j$,$\mathrm{x}_{t|t}^j$, $P_{t|t}^j$

for $t = 1, 2, \ldots, T$ **do**
  for $j = 1, \ldots, K$ **do**
    for $i = 1, \ldots, K$ **do**
      $[\mathrm{x}_{t|t}^{ij}, P_{t|t}^{ij}, L_t^{ij}] = \text{Filter}(\mathrm{x}_{t-1|t-1}^i, P_{t-1|t-1}^i, A^j, C, Q^j, R^j)$
    end for
  end for
  for $j = 1, \ldots, K$ **do**
    $[M_{t|t}^j, W_t^{i|j}] = \text{FilterProbs}(L_t^{ij}, Z^{ij}, M_{t-1|t-1}^i)$
    $[\mathrm{x}_{t|t}^j, P_{t|t}^j] = \text{Collapse}(\mathrm{x}_{t|t}^{ij}, P_{t|t}^{ij}, W_t^{i|j})$
  end for
end for

state parameters given the raw heart sound observations $\{\mathrm{y}_t\}_{t=1}^T$ and estimated model parameters for each state $\widehat{\Theta} = \left\{\widehat{Z}, \widehat{A}^{(j)}, \widehat{Q}^{(j)}, \widehat{R}^{(j)}, j = 1, \ldots, K\right\}$. Refer to [43] for further details. Given $\widehat{\Theta}$ and initial state probabilities $M_0^j = [1, 0, \ldots, 0]$, for each time t, a run of K^2 Kalman filters is performed recursively to compute the mean and covariance of the component filtered densities of x_t (denoted as $\mathrm{x}_{t|t}^{ij}$ and $P_{t|t}^{ij}$) for all pairs (i, j) and the corresponding likelihood function L_t^{ij}. The filtered state probability of S_t can be defined by

$$\begin{aligned} M_{t|t}^j &= P(S_t = j|\{\mathrm{y}_t\}_{t=1}^t) \\ &= \sum_i M_{t-1,t|t}^{ij} \end{aligned} \tag{6}$$

where $M_{t-1,t|t}^{i,j} = P(S_{t-1} = i, S_t = j|\{\mathrm{y}_t\}_{t=1}^t)$ is computed from the $M_{t-1|t-1}^i$ at previous time $t-1$ weighted by the likelihood L_t^{ij} and the transition probabilities z_{ij} as follows

$$M_{t-1,t|t}^{ij} = \frac{L_t^{ij} z_{ij} M_{t-1|t-1}^i}{\sum_i \sum_j L_t^{ij} z_{ij} M_{t-1|t-1}^i}$$

After filtering at each time t, the component densities ($\mathrm{x}_{t|t}^{ij}$ and $P_{t|t}^{ij}$) weighted by $W_t^{i|j} = M_{t-1,t|t}^{ij} / M_{t|t}^j$ are collapsed to give the mean and covariance of filtered densities ($\mathrm{x}_{t|t}^j$ and $P_{t|t}^j$).

b) Switching Kalman Smoother (SKS): The SKS provides a refined state estimates from SKF based on both the past and future observations. In a backward recursion, a mixture of K^2 Kalman smoothers is run to compute component smoothed densities of x_t for all pairs (j, k) (with mean $\mathrm{x}_{t|T}^{jk}$ and covariance $P_{t|T}^{jk}$) given the entire observation $\{\mathrm{y}_t\}_{t=1}^T$ based on the filtered densities computed in the SKF. The smoothed state probability of S_t is defined as

$$\begin{aligned} M_{t|T}^j &= P(S_t = j|\{\mathrm{y}_t\}_{t=1}^T) \\ &= \sum_k M_{t,t+1|T}^{jk} \end{aligned} \tag{7}$$

where $M_{t,t+1|T}^{jk} = P(S_t = j, S_{t+1} = k|\{\mathrm{y}_t\}_{t=1}^T)$ can be computed based on the filtered state probabilities $M_{t|t}^j$ and

Algorithm 2 : SKF-Viterbi Algorithm.

Inputs: initials $\pi_0, HR, tSys$

Outputs: q_t.

$[\{M_t^j\}_{t=1}^T] = \text{SKF}(\{\mathrm{y}_t\}_{t=1}^T), A, R, Q, Z, \mathrm{x}_0, P_0, M_0^j)$
Initialization: $[a_{ij}, \delta_1^j, d_{max}]$ =(HR, $tSys$, $\{M_{t|t}^j\}_{t=1}, \pi_0$)
for $t = 2 : T + d_{max} - 1$ **do**
  for $i, j = 1 : K$ **do**
    for $d = 1 : d_{max}$ **do**
      $w_s = t - d, \quad 1 \leq w_s \leq T - 1$
      $w_e = t, \quad 2 \leq w_e \leq T$
      $\delta_t^j = \max_d \Big[\max_{i \neq j}[\delta_{w_s}^i a_{ij}] \; . \; dP_d^j \; . \; \prod_{s=w_s}^{w_e}\{M_{t|t}^j\}_{t=s}\Big]$
      $D_t^j = \arg\max_d \Big[\max_{i \neq j}[\delta_{w_s}^i a_{ij}] \; . \; dP_d^j \; . \; \prod_{s=w_s}^{w_e}\{M_{t|t}^j\}_{t=s}\Big]$
      $\psi_t^j = \arg\max_{1 \leq i \leq K}[\delta_{t-D_t^j}^i a_{ij}]$
    end for
  end for
end for
$T* = \arg\max_t[\{\delta_t^i\}_{t=T}^{T+d_{max}-1}] \qquad 1 \leq i \leq K$
$q_{T*}^* = \arg\max_i[\delta_{T*}^i]$
$t = T*$
while $t > 1$ **do** //Backward Viterbi procedure
  $d^* = D_t^{q_t^*}$
  $\{q\}_{t-d^*}^{t-1} = q_t^*$
  $q_{t-d^*-1}^* = \psi_t^{q_t^*}$
  $t = t - d^*$
end while

the smoothed probabilities $M_{t+1|T}^k$ at $t+1$ as follows

$$M_{t,t+1|T}^{jk} = \frac{M_{t|t}^j z_{jk}}{\sum_j' M_{t|t}^{j'} z_{j'k}} M_{t+1|T}^k$$

Finally, the component densities ($\mathrm{x}_{t|T}^{jk}$ and $P_{t|T}^{jk}$) weighted by $W_t^{k|j} = M_{t,t+1|T}^{jk} / M_{t|T}^j$ are collapsed to give the mean and covariance of the smoothed densities ($\mathrm{x}_{t|T}^j$ and $P_{t|T}^j$).

c) SKF with Viterbi Algorithm (SKF-Viterbi): Under the Markovian assumption in the standard SKF, the sojourn time or dwell time (the number of consecutive time points spent in a specific state before transitioning to other states) is geometrically distributed, i.e., the probability of remaining in a state decreases as the sojourn time increases. This tends to induce unrealistically fast switching states and may not be appropriate for stationary processes such as each heart sound component with possibly long period of time in the same regime. To overcome this limitation, we introduce a two-step procedure by combining the SKF with the duration-dependent Viterbi algorithm introduced by [30] and extended in [3].

The duration-dependent Viterbi algorithm incorporates explicitly the information about each state expected duration (i.e. heart rate —HR, systolic interval —tSys) which are estimated from the heart sound test-set recordings using autocorrelation analysis. This approach allows the self-transitions and ensures state changing at a certain limit of duration corresponding to that of each major component in a heart cycle. The duration probabilities dP are estimated from the data for each of the four heart sound states.

With an initialized δ_1^j, the algorithm computes the state probability in a forward recursion

$$\delta_t^j = \max_d \left[\max_{i \neq j} \quad [\delta_{t-d}^i a_{ij}] \quad dP_d^j \quad \prod_{s=0}^{d-1} \{M_{t|t}^j\}_{t=t-s} \right] \tag{8}$$

for $1 \leq t \leq T$, $1 \leq i, j \leq K$, dP_d^j is the duration probabilities for state j for $1 \leq d \leq d_{max}$ with d_{max} the number of time points for each heart-beat with reference to the estimated heart rate. Note that we incorporate the SKF state probability $M_{t|t}^j = P(S_t = j|\{\mathrm{y}_t\}_{t=1}^t) \propto P(\{\mathrm{y}_t\}_{t=1}^t|S_t = j)P(S_t = j)$ which takes into account the observations up to time t instead of only the current observation $P(\mathrm{y}_t|S_t = j)$ in the original duration-dependent Viterbi algorithm. The state duration argument and the state sequence that maximize (8) are stored in D_t^j and ψ_t^j respectively. The most likely state sequence is obtained and stored in ψ_t^j, $\psi_t^j = \underset{1 \leq i \leq K}{\text{argmax}}[\delta_{t-D_t^j}^i a_{ij}]$.

The psuedocode of the extended Viterbi algorithm is shown in Algorithm (2). Refer to [3] for more details. In Algorithm (2), the δ_t^j is the highest state probability for each state j at time t for all duration probabilities dP_d^j from 1 to d_{max}. The state probabilities are updated only if current δ_t^i is higher than the δ_{t-1}^i in the processing window 1 to d_{max}. The back-tracking procedure is initialized by finding the maximum probability of δ_t^i in the interval $T : T + d_{max} - 1$ after the end of actual signal. The state index that maximizes δ_{T*}^i is stored in $q_{T*}^* = \text{argmax}_i[\delta_t^i]$. The optimal path q_t^* is obtained by back-tracking $\psi_T^{q_t^*}$ and $D_T^{q_t^*}$ such that $q_{t-d^*-1}^* = \psi_{q_t^*}$, where $t = T - 1, \ldots, 1$.

B. Heart Sound Classification

This section presents an automatic classification of healthy and pathological heart sound recordings using HMMs based on the heart-beat segmentation obtained by SKF-Viterbi algorithm. An overview of the proposed classification procedure is shown in Fig. 3.

1) Pre-processing and Feature Extraction: Similar procedure of pre-processing in Section III-A1 was followed, except that the PCG recordings were filtered using Butterworth band-pass filter with cut-off frequencies of 25 Hz and 800 Hz as the murmurs have higher frequencies than the normal heart sounds [44, 45]. The normal and abnormal recordings were then segmented into individual heart-beat cycles (each covers from start of $S1$ sound to the subsequent $S1$ sound) using the SKF-Viterbi algorithm. For the very noisy X-Factor recordings, expert annotations of heart-beat cycles were not provided with the database. We segmented these recordings using non-overlapping windows with duration of 1 s, which approximately corresponds to one complete heart-beat cycle.

For each heart-beat, we computed a sequence of short-time spectral features over sliding windows or frames of 25 ms with 10 ms overlap using a Hamming window. We consider the following two types of spectral features. These features were then used as input to the HMMs.

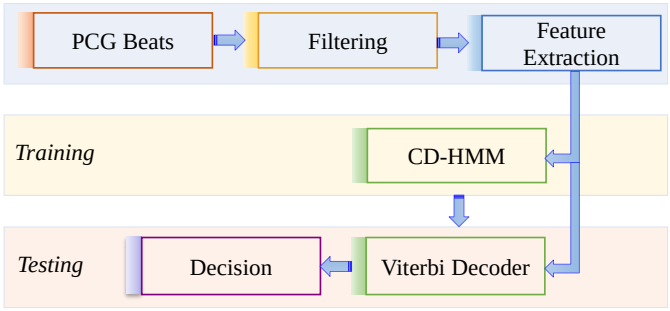

Fig. 3. The heart sound classification system using HMMs.

- *Mel-frequency cepstral coefficients (MFCCs):* MFCCs have been successful for speech analysis and increasingly used for music data. It can provide an effective representation for audio signals including heart sound via frequency warping to a quasi-logarithmic scale that mimics human auditory perception. We extracted 12 static MFCC coefficients, one log-energy value (E), 12 delta coefficients (Δ), and 12 acceleration coefficients (Δ^2), resulting in a vector of 37 features per frame. We used 12 Mel-scale pass-bands between 0–800 Hz with 50% overlapped frequency bands of 105 Hz. The Δ and Δ^2 were calculated with a width of two frames.
- *Time and frequency-domain features:* A set of 16 features were computed from each frame, including 6 features from time and statistical domains (i.e., skewness, kurtosis, average amplitudes, Shannon entropy and all-band power) and 10 features from frequency-domain (i.e., peak-frequency plus 9 features of median spectrum energy extracted from non-overlapping frequency bands in range of 0–800Hz).

2) HMM: The HMM is a probabilistic model that can capture the dynamical changes of the heart sounds by making inferences about the likelihood of being in certain discrete states. We used straightly left-to-right four-state HMMs with Gaussian mixture observation densities. The set of HMM parameters is denoted by $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ where $\boldsymbol{\pi} = [\pi_i]$ with $\pi_i = P[q_1 = S_i], 1 \leq i \leq K$ are the initial state probabilities and $\mathbf{A} = [a_{ij}]$ is $K \times K$ transition matrix with $a_{ij} = P[q_{t+1} = S_i | q_t = S_j], 1 \leq i, j \leq K$. Let $\mathbf{O}_t = [o_{1t}, \ldots, o_{Nt}]'$ be the $N \times 1$ MFCC feature vector at time t. The emission probability $\mathbf{B} = \{b_j(x)\}, 1 \leq j \leq K$ at each state j is defined by a Gaussian mixture model

$$b_j(\mathbf{O}_t) = \sum_{m=1}^{M} c_{jm} N(\mathbf{O}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}), 1 \leq j \leq K \quad (9)$$

where $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$ are respectively the mean vector and covariance matrix of the m-th mixture component with mixture weight c_{jm} at state j. Here, we set the number of mixture components as $M = 16$ per state.

3) Training & Testing: The training and testing of HMMs are illustrated in Fig. 3. HMMs were trained on the sequences of spectral features extracted from segmented heart-beats. The HMM parameters were estimated by maximizing the likelihood function, given the training observation sequences $\mathbf{O}_1, \ldots, \mathbf{O}_T$ of heart-beat segments. The model parameters were initialized using segmental K-means algorithm by first aligning the observations of each heart-beat to the four states via the Viterbi algorithm and then partitioning the observations in each state into 16 mixture components by K-means clustering. This was followed by iterative re-estimation of the parameters via expectation-maximization algorithm (the Baum-Welch algorithm) until convergence. Separate HMMs were trained for the normal and abnormal heart sounds. Given an unknown testing heart sound beat, the Viterbi algorithm was used to compute the approximate likelihood scores for each HMM model based on the most likely state sequence. The testing heart sound signal will be classified to the model with the highest likelihood score.

4) Classification Schemes: We evaluate the performance of HMM in classifying heart sound signals of individual heart-beats (from single cardiac cycle) or long recording (with many cycles) into the normal and abnormal classes correctly. The two partitioning schemes for the train and test sets in the database as detailed in Section II were used for evaluation. We consider experimental evaluation with and without incorporating the noisy (X-Factor) recordings for both the beat-level and recording-level heart sound classification.

In the beat-level classification, each short-segment of single heart-beat was individually assigned to a normal, abnormal, or X-Factor class. In the recording-level classification, decision scores of all heart-beats over a recording were combined by majority voting, where the recording was predicted as abnormal if the proportion of beats assigned to abnormal class is dominant. The beat-level approach substantially expands the number of training instances, which allows the machine learning application to learn more about the heart sound underlying dynamics for each class. The database provides global (recording-level) labels where each record has been assigned to an abnormal or normal class, we assumed all the beats of a given abnormal recording are also abnormal.

IV. Experimental Results

In this section, we present the heart sound segmentation and classification results by the proposed methods on the 2016 Physionet/CinC Challenge data sets described in Section II.

A. Evaluation Metrics

For heart sound segmentation, the performance of the proposed algorithms were measured based on the number of time points over the time course of recordings where the heart sound state labels are correctly predicted relative to ground-truth annotations. We computed evaluation metrics using sensitivity (Se), precision ($Prec$), and F_1 score for individual heart sound state.

$$Se = \frac{TP}{TP + FN} \times 100 \quad (10)$$

$$Prec = \frac{TP}{TP + FP} \times 100 \quad (11)$$

$$F_1 = \frac{2 \times Se \times Prec}{Se + Prec} \times 100 \quad (12)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2925036, IEEE Journal of Biomedical and Health Informatics

To summarize the segmentation performance over all four heart sound states, we compute averages of these measures and the overall accuracy (Acc)

$$Acc = \frac{\sum TP}{N_T} \times 100 \tag{13}$$

where N_T is the total number of time points in recordings of the test-set, $\sum TP$ is the sum of TP values over the four heart sound states for the entire recording. Note that the performance was evaluated in segment-wise manner with zero tolerance between the ground-truth and the estimated labels.

For heart sound classification, we evaluate the performance of the HMM in detecting the abnormal heart sound signals accurately. In addition to Se and $Prec$, we reported specificity (Sp) and Acc defined as follows

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{14}$$

$$Sp = \frac{TN}{TN + FP} \times 100 \tag{15}$$

For classification without involving X-Factor beats or recordings, TP indicates the total beats/recordings correctly identified as abnormal, TN the correct identifications of normal class, FP the instances of normal class identified as abnormal, and FN the number of instances of abnormal class identified as normal. To evaluate performance including the X-Factor class, we used similar metrics, where FN indicates abnormal beats/recordings wrongly identified as normal or X-Factor, FP is the normal and X-Factor beats/recordings wrongly identified as abnormal, and $(TN = N_T - TP - FN - FP)$ where N_T is the total number of beats/recordings in test-set.

B. Results for Heart Sound Segmentation

We compare the performance of the proposed MSAR-SLDS-based segmentation algorithms: SKF, SKS, and SKF-Viterbi in annotating the four heart sound states on recordings in the totally unseen test-set shown in Table II. The results computed based on concatenated recordings in each test subset according to different conditions are given in Table III. We can see that the fusion of SKF and duration-dependent Viterbi algorithm achieved the best segmentation performance on all test sets, achieving a significantly higher average F_1 score of 80.5% compared to 63.3% by SKF and 63.9% by SKS. The refinement of the state estimates by the SKS only gives marginal improvement over the SKF. Both the SKS and SKF-Viterbi algorithms refine the state estimates of SKF, via backward smoothing utilizing the entire observation time course and incorporation of state durations, respectively. However, the margin by which the SKF-Viterbi improves over the SKF is substantially larger than the SKS with less computational effort. Additional experiment shows that the computational time for state decoding of a selected heart-sound recordings with duration 60.16 s for the SKF-Viterbi was 19.52 s, which is lower than SKS with 21.51 s and slightly higher than SKF with 15.06 s (using Matlab software on Windows 10–64bit platform with 3.60 GHz Intel Xeon w-2123 CPU and 64 GB memory). Note that the segmentation accuracy on the normal recordings are higher than that on the abnormal recordings, where heart sounds are corrupted by murmurs arising from different pathological conditions of the cardiovascular system such as coronary artery disease (CAD), mitral regurgitation (MR) and aortic stenosis (AS).

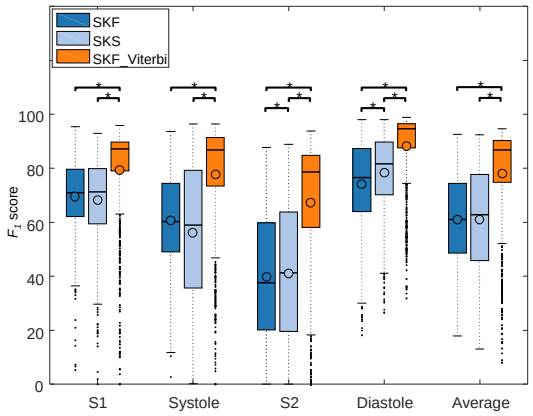

Fig. 4. Box-plots of F_1 scores using different segmentation algorithms for individual heart sound and average over four heart sound components. Each plot is based on all 800 recordings in the test-set. Horizontal braces and asterisks indicate significant difference in performance between pairs of algorithms tested via Wilcoxon signed-rank test at confidence level of 95%.

We further assess the performance in segmenting regimes of each heart sound component in individual recordings. The significant difference in performance were tested by the Wilcoxon signed-rank test. Fig. 4 shows box-plots of F_1 scores obtained by various algorithms for individual heart sounds and average across four heart sounds over the 800 recordings in the entire test-set. We can see the variability in segmentation performance over recordings collected from different subjects with varying lengths and conditions. All algorithms achieved the highest F_1 scores when segmenting the diastolic intervals. The SKF-Viterbi algorithm significantly outperformed others for all heart sounds. However, the SKS estimates performed comparably to the SKF with no significant improvement in performance.

We benchmark the performance of the MSAR-SLDS approach with SKF-Viterbi algorithm by comparison with conventional HMM and the state-of-the-art HSMM [30] using raw signals and pre-extracted features. We trained the HMM and HSMM with Gaussian observation distribution with the four-state Markov chain topology as in the MSAR-SLDS model, using raw heart sound signals and features based on homomorphic eveelogram. Generic and state-dependent Viterbi algorithm were used in state decoding for the HMM and HSMM, respectively. The results are given in Table IV. As expected, both the HMM and HSMM, which do not account for noise effects, performed poorly on raw signals (down-sampled to 50 Hz, worse performance in original sampling frequency). Interestingly, the SKF-Viterbi algorithm even without prior feature extraction can considerably outperform both HMM and HSMM with envelope features. The results imply that the MSAR-SLDS trained merely on raw signals can provide better robustness for heart-sound segmentation under noisy recordings than a feature-based HMM, due to its implicit representation of switching autoregressive structures in heart sounds and explicit incorporation of noise model.

TABLE III
PERFORMANCE OF DIFFERENT SEGMENTATION ALGORITHMS UNDER THE MSAR-SLDS MODEL ON TEST-SET

Challenge set	Disease	#Recordings	#Beats	MSAR-SKS				MSAR-SKF				MSAR-SKF-Viterbi			
				Se	$Prec$	F_1	Acc	Se	$Prec$	F_1	Acc	Se	$Prec$	F_1	Acc
training_*b*	Normal	139	1179	57.4	53.7	53.9	59.4	53.2	50.1	50.9	58.0	68.0	68.6	68.6	73.9
	CAD	36	293	59.7	57.0	57.0	63.0	58.5	56.6	57.1	64.2	66.7	68.9	67.5	73.3
	All	175	1472	57.9	54.3	54.5	60.1	54.2	51.3	52.1	59.3	67.8	69.4	68.4	73.8
training_*c*	Normal	4	179	61.1	61.1	58.2	62.8	56.9	57.9	56.5	63.4	86.3	86.6	86.3	88.3
	MR	7	567	62.1	60.8	60.9	65.4	63.7	63.2	63.3	68.6	83.7	84.9	84.2	87.3
	AS	3	148	60.9	59.6	60.1	66.1	61.9	60.0	60.5	65.2	71.0	75.6	72.8	79.1
	All	14	894	61.8	60.1	60.3	64.9	61.9	61.1	61.3	66.8	82.2	83.7	82.9	86.1
training_*d*	Normal	26	308	64.5	63.3	60.7	67.7	60.1	60.9	59.4	69.7	86.8	87.4	87.0	90.2
	Pathologic	26	493	55.2	51.5	52.0	58.2	50.2	47.3	48.2	55.7	76.3	77.9	77.0	80.8
	All	52	801	58.7	55.2	55.2	61.9	53.9	51.8	52.4	61.2	80.2	81.5	80.9	84.5
training_*e*	Normal	432	13746	69.6	71.4	68.7	72.7	69.1	72.3	70.1	75.3	84.6	86.3	85.3	87.9
	CAD	72	15132	59.0	55.5	56.4	62.2	56.8	54.2	55.1	61.9	72.1	72.7	72.4	77.5
	All	504	28878	68.6	69.4	67.4	71.7	67.9	70.1	68.6	74.0	83.4	85.0	84.0	86.9
training_*f*	Normal	39	1506	50.4	48.9	49.0	53.4	51.4	50.6	50.6	55.6	65.6	66.8	66.1	69.2
	Pathologic	16	2076	42.6	41.4	41.7	47.8	42.0	41.3	41.5	48.3	58.9	60.8	59.7	64.9
	All	55	3582	48.4	46.9	47.1	51.8	48.9	48.1	48.2	53.6	63.8	65.3	64.4	68.0
Average		800	20375	64.7	64.3	**63.3**	67.9	63.8	64.6	**63.9**	69.7	79.8	81.4	**80.5**	83.6

TABLE IV
COMPARISON OF SEGMENTATION PERFORMANCE WITH OTHER METHODS

Method	Features	Se	$Prec$	F_1	Acc
HMM	Envelope	53.5	57.2	51.2	54.5
HMM	Raw	32.9	34.0	27.9	33.3
HSMM[30]	Envelope	77.7	78.2	77.9	81.7
HSMM[30]	Raw	60.7	61.0	60.8	67.5
SKF-Viterbi	Raw	79.8	81.4	**80.5**	83.6

TABLE V
PERFORMANCE COMPARISON USING DIFFERENT INPUT FEATURES TO HMM FOR HEART-SOUND BEAT-LEVEL CLASSIFICATION

Features	TP	FP	TN	FN	Se	Sp	$Prec$	Acc
MFCC+E	14461	2326	14592	2457	85.5	**86.3**	**86.1**	85.9
MFCC+E+Δ+Δ^2	15362	3101	13817	1556	**90.8**	81.7	83.2	**86.2**
Time-Frequency	13394	2966	13952	3524	79.2	82.5	81.9	80.8

C. Results for Heart Sound Classification

In this section, we evaluate the performance of HMM in classifying abnormal heart sound morphology.

1) Comparison of Different Features for HMM: Table V shows the classification results on the unseen test-set for HMM using 12-dimensional MFCCs, MFCCs plus their first- and second-order derivative features, and combination of time- and frequency-domain features. Best results are obtained with MFCCs in terms of Sp and $Prec$ metrics. The inclusion of derivative MFCC features does not lead to improvement for heart sound classification. For this reason, and with the advantage of robust representation of speech and audio signals, we use the MFCC features for the subsequent experiments.

2) Results for Cross-Validation Partitioning: Table VI shows the classification results from 5-fold cross validation without including the X-Factor class. At the beat-level classification (for a total of 81,500 normal and abnormal beats) without X-Factor class, the HMM achieved a reasonably high performance with an average Se of 89.2 ± 2.7, Sp of 84.0 ± 1.0, $Prec$ of 84.8 ± 0.9, and Acc of 86.6 ± 1.5. Note that the performance was still affected by presence of varied levels of noise in the normal/abnormal beats despite the very noisy recordings were excluded from this experiment. Moreover, the beat-level expert annotations of heart diseases are not provided along with the database. This may result in miss-classification of the noisy beat signals as abnormal as noticed from the high FP values. At the recording-level without X-Factor class, the entire heart sound recording was classified either as normal or abnormal based on majority voting of the detected individual beats. The FP rate in detecting the abnormal recordings is high with 94 recordings of the normal class were classified as abnormal. However, the proposed method obtained a Se of 92.5 ± 3.7, Sp of 79.5 ± 2.0, $Prec$ of 81.9 ± 1.4, and Acc of 86.0 ± 1.8. The performance drop slightly for recording-level classification compared to the beat-level classification. This may due to that some 'abnormal' recordings only contained few abnormal beats while most of the beats are with normal characteristics.

Table VII shows results when including X-Factor class for very noisy signals as separate class in addition to the normal and abnormal classes. Different HMM models were trained for normal, abnormal, and X-Factor classes, respectively. The aim of this experiment is to assess the ability of the proposed method to automatically reject the beats/recordings labeled as X-Factor, which is a challenging task. The results show a remarkable classification performance in detecting the X-Factor class accurately.

Only a small percentage of X-Factor beats of 1.0% was misclassified as the abnormal class and 0.2% as the normal. At the beat-level classification with X-Factor, an average Se of 89.5 ± 1.2, Sp of 91.8 ± 1.2, $Prec$ of 84.5 ± 1.8, and Acc of 91.0 ± 0.6 were achieved. Small values of the standard deviations in these measures indicate consistent results across the 5-folds. The performance for the recording-level with X-Factor is better than that without X-Factor, with improvement

TABLE VI
K-FOLD (5-FOLD) CROSS VALIDATION CLASSIFICATION PERFORMANCE ON PHYSIONET CINC DATASET WITHOUT X-FACTOR.

Fold	Beat-level without X-Factor class								Recording-level without X-Factor class							
iterate	TP	FP	TN	FN	Se	Sp	$Prec$	Acc	TP	FP	TN	FN	Se	Sp	$Prec$	Acc
1	12196	2252	10935	991	92.5	82.9	84.4	87.7	440	106	354	20	95.7	77.0	80.6	86.3
2	10454	2057	10325	1928	84.4	83.4	83.6	83.9	395	93	367	65	85.9	79.8	80.9	82.8
3	11804	2094	11056	1346	89.8	84.1	84.9	86.9	440	101	359	20	95.7	78.0	81.3	86.9
4	11839	1860	11227	1248	90.5	85.8	86.4	88.1	432	79	381	28	93.8	82.8	84.5	88.4
5	11874	2133	11215	1474	89.0	84.0	84.8	86.5	420	93	367	40	91.3	79.8	81.9	85.5
Mean	11633	2079	10952	1397	89.2	84.0	84.8	86.6	425	94	366	35	92.5	79.5	81.9	86.0
SD	606	128	331	309	2.7	1.0	0.9	1.5	17	9	9	17	3.7	2.0	1.4	1.8

TABLE VII
K-FOLD (5-FOLD) CROSS VALIDATION CLASSIFICATION PERFORMANCE ON PHYSIONET CINC DATASET WITH X-FACTOR.

Fold	Beat-level with X-Factor class								Recording-level with X-Factor class							
iterate	TP	FP	TN	FN	Se	Sp	$Prec$	Acc	TP	FP	TN	FN	Se	Sp	$Prec$	Acc
1	11984	2292	24504	1414	89.5	91.5	84.0	90.8	436	97	823	24	94.8	89.5	81.8	91.2
2	12041	2140	24316	1187	91.0	91.9	84.9	91.6	444	92	828	16	96.5	90.0	82.8	92.2
3	11568	1817	24459	1570	88.1	93.1	86.4	91.4	420	82	838	40	91.3	91.1	83.7	91.2
4	11341	1878	23776	1486	88.4	92.7	85.8	91.3	416	78	842	44	90.4	91.5	84.2	91.2
5	11389	2599	22527	1174	90.7	89.7	81.4	90.0	440	117	803	20	95.7	87.3	79.0	90.1
Mean	11665	2145	23916	1366	89.5	91.8	84.5	91.0	431	93	827	29	93.7	89.9	82.3	91.2
SD	295	285	742	160	1.2	1.2	1.8	0.6	11	14	14	11	2.4	1.5	1.9	0.7

TABLE VIII
CLASSIFICATION PERFORMANCE FOR SUBJECT-ORIENTED (UNSEEN) TEST-SET (TABLE II).

Classification	Without X-Factor class								With X-Factor class							
scheme	TP	FP	TN	FN	Se	Sp	$Prec$	Acc	TP	FP	TN	FN	Se	Sp	$Prec$	Acc
Beat-level	14461	2326	14592	2457	85.5	86.3	86.1	85.9	14461	2493	31343	2457	85.5	92.6	85.3	90.3
Recording-level	572	164	476	68	89.4	74.4	77.7	81.9	572	163	477	68	89.4	87.3	77.8	88.0

in average Acc from 86.0 ± 1.8 to 91.1 ± 0.7, so do the other metrics. This suggests that the HMM is able to accurately identify the X-Factor recordings, where only one recording of X-Factor was wrongly classified as abnormal.

3) Results for Subject-Oriented Partitioning: Table VIII shows the results for beat-level and recording-level classification on the totally unseen test-set allocated based on subject-oriented partitioning. Interestingly, we can see a substantial improvement in Acc when including the X-Factor class compared to classification without X-Factor class for both the beat-level (from 81.9% to 88.0%) and the recording-level (from 85.9% to 90.3%). This is mainly due to the increases in Sp measures with the inclusion of X-Factor, suggesting that the HMM can accurately identify the X-Factor beats/recordings. For both with and without X-Factor class, there was a considerable drop in performance in terms of $Prec$ in the recording-level classification compared to the beat-level classification. This can be explained by the apparent $Se-Sp$ trade-off where the increments in the Se were at a much lower rate than the drops in Sp. Generally, the results on the subject-oriented test-set are comparable to that on the cross-validation as shown in Table VI and Table VII, with only slight decrease in accuracy for the recording-level classification. This suggests the HMM can perform comparably well on the unseen test-set as on the partially-seen test-set in the cross-validation.

V. CONCLUSION

We developed a new approach for robust heart-sound segmentation using the MSAR model. The MSAR model can rigorously characterizes the switching dependence structure in the cyclical heart sounds. The extension to a SLDS further incorporates an explicit model to account for the noise effects. This overcomes the limitations of the state-of-the-art HMM approach for heart sound modeling, owing to its unrealistic assumption of conditionally-independent observations and lack of a specification for noise. The proposed MSAR-SLDS provides direct modeling of heart sound signals, and thus enable online heart sound segmentation based on noisy raw recordings without preliminary feature extraction as required by HMMs. We introduced a novel fusion of SKF and Viterbi algorithm for the MSAR-SLDS which leverages on the heart-sound state durations to improve segmentation performance. As demonstrated on the large 2016 Physionet/CinC Challenge database, the proposed SKF-Viterbi algorithm is able to accurately segment heart sounds directly based on noisy raw PCG recordings confounded with murmurs and other pathological sounds. It substantially outperforms both the HMM and HSMM when segmenting raw signals and is comparable to the feature-based HSMM. Using the segmented labels to train HMM classifiers for detecting abnormal heart sound, we obtained remarkable accuracy on both the recording and beat-level classification on an unseen test-set including very noisy X-Factor recordings. The use of majority voting scheme in the recording-level classification is appropriate for detecting cardiac pathologies characterized by murmurs which are pervasive across the entire recording. However, it may fail to identify recordings with infrequent non-murmur pathological beats. Nevertheless,

practitioners can utilize complementary information from the beat-level predictions provided by the proposed method in assessing individual beats for clinical diagnosis. Instead of majority voting, future works could consider data-driven selection of the voting threshold and use of soft-thresholding. While this study focuses on binary PCG classification into normal and abnormal based on single signal regardless of its auscultation site, our framework could be extended to take into account information from signals recorded from different auscultation locations which have been shown useful in discriminating different types of heart diseases [46].

REFERENCES

[1] C. Liu, et al., "An open access database for the evaluation of heart sound algorithms," *Physiol Meas*, vol. 37, no. 12, pp. 2181–2213, dec 2016.

[2] D. Kumar, et al., "Noise detection during heart sound recording using periodicity signatures," *Physiol Meas*, vol. 32, no. 5, pp. 599–618, may 2011.

[3] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans Biomed Eng*, vol. 63, no. 4, pp. 822–832, 2016.

[4] D. Springer and L. Tarassenko, "Support vector machine hidden semi-Markov model-based heart sound segmentation," *Comput Cardiol*, 2014.

[5] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *IEEE Computers in Cardiology 1997*, 1997, pp. 105–108.

[6] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proc. 19th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 1997, vol. 4, pp. 1630–1633.

[7] A. Moukadem, et al., "A robust heart sounds segmentation module based on s-transform," *Biomed Signal Process Control*, vol. 8, no. 3, pp. 273–281, 2013.

[8] S. Sun, et al., "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Comput Methods Programs Biomed*, vol. 114, no. 3, pp. 219–230, 2014.

[9] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Syst Appl*, vol. 34, no. 2, pp. 1056–1069, 2008.

[10] Z. Yan, et al., "The moment segmentation analysis of heart sound pattern," *Comput Methods Programs Biomed*, vol. 98, no. 2, pp. 140–150, 2010.

[11] S. Ari, P. Kumar, and G. Saha, "A robust heart sound segmentation algorithm for commonly occurring heart valve diseases," *J Med Eng Technol*, vol. 32, no. 6, pp. 456–465, jan 2008.

[12] H. Naseri and M. R. Homaeinezhad, "Detection and Boundary Identification of Phonocardiogram Sounds Using an Expert Frequency-Energy Based Metric," *Ann Biomed Eng*, vol. 41, no. 2, pp. 279–292, feb 2013.

[13] D. Kumar, et al., "Detection of s1 and s2 heart sounds by high frequency signatures," in *Proc. 28th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2006, pp. 1410–1416.

[14] V. N. Varghees and K. Ramachandran, "A novel heart sound activity detection framework for automated heart sound analysis," *Biomed Signal Process Control*, vol. 13, pp. 174–188, 2014.

[15] J. Pedrosa, A. Castro, and T. T. Vinhoza, "Automatic heart sound segmentation and murmur detection in pediatric phonocardiograms," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2014, pp. 2294–2297.

[16] V. Nigam and R. Priemer, "Accessing heart dynamics to estimate durations of heart sounds," *Physiol Meas*, vol. 26, no. 6, pp. 1005–1018, dec 2005.

[17] J. Vepa, P. Tolay, and A. Jain, "Segmentation of heart sounds using simplicity features and timing information," in *2008 IEEE Intel Conf on Acoustics, Speech and Signal Processing*, mar 2008, pp. 469–472.

[18] C. D. Papadaniil and L. J. Hadjileontiadis, "Efficient Heart Sound Segmentation and Extraction Using Ensemble Empirical Mode Decomposition and Kurtosis Features," *IEEE J Biomed Health Inform*, vol. 18, no. 4, pp. 1138–1152, jul 2014.

[19] A. Gharehbaghi, et al., "An automatic tool for pediatric heart sounds segmentation," in *IEEE Comput Cardiol, 2011*, 2011, pp. 37–40.

[20] T. Oskiper and R. Watrous, "Detection of the first heart sound using a time-delay neural network," in *IEEE Comput Cardiol, 2002*, 2002, pp. 537–540.

[21] A. A. Sepehri, et al., "A novel method for pediatric heart sound segmentation without using the ecg," *Comput Methods Programs Biomed*, vol. 99, no. 1, pp. 43–48, 2010.

[22] C. N. Gupta, et al., "Neural network classification of homomorphic segmented heart sounds," *Applied Soft Computing*, vol. 7, no. 1, pp. 286–297, 2007.

[23] Hong Tang, et al., "Separation of heart sound signal from noise in joint cycle frequencytimefrequency domains based on fuzzy detection," *IEEE Trans Biomed Eng*, vol. 57, no. 10, pp. 2438–2447, oct 2010.

[24] S. Rajan, et al., "Unsupervised and uncued segmentation of the fundamental heart sounds in phonocardiograms using a time-scale representation," in *Proc. 28th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, aug 2006, pp. 3732–3735.

[25] L. Gamero and R. Watrous, "Detection of the first and second heart sound using probabilistic models," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Cancun, Mexico, 2003, pp. 2877–2880.

[26] A. D. Ricke, R. J. Povinelli, and M. T. Johnson, "Automatic segmentation of heart sound signals using hidden markov models," in *IEEE Computers in Cardiology*, 2005, pp. 953–956.

[27] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *IEEE Computers in Cardiology*, 2005, pp. 957–960.

[28] P. Sedighian, et al., "Pediatric heart sound segmentation using Hidden Markov Model," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, aug 2014, pp. 5490–5493.

[29] M. S. Bentley P, Nordehn G, Coimbra M and G. R, "The PASCAL classifying heart sounds challenge 2011 (CHSC2011)," 2011.

[30] S. E. Schmidt, et al., "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiol Meas*, vol. 31, no. 4, pp. 513–529, apr 2010.

[31] R. H. Shumway and D. S. Stoffer, "Dynamic Linear Models with Switching," *J Am Stat Assoc*, vol. 86, no. 415, pp. 763–769, sep 1991.

[32] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Comput*, vol. 12, no. 4, pp. 831–864, apr 2000.

[33] E. Fox, et al., "Nonparametric bayesian learning of switching linear dynamical systems," in *Proc in Nonparametric Bayesian Learning of Switching Linear Dynamical Systems*, 2009, pp. 457–464.

[34] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1850–1858, 2007.

[35] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357, mar 1989.

[36] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in neural information processing systems*, 2001, pp. 981–987.

[37] V. Monbet and P. Ailliot, "Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature," *Comput Stat Data Anal*, vol. 108, pp. 40–51, 2017.

[38] S. B. Samdin, et al., "A unified estimation framework for state-related changes in effective brain connectivity," *IEEE Trans Biomed Eng*, vol. 64, no. 4, pp. 844–858, apr 2017.

[39] J. Oster, et al., "Semisupervised ecg ventricular beat classification with novelty detection based on switching kalman filters," *IEEE Trans Biomed Eng*, vol. 62, no. 9, pp. 2125–2134, 2015.

[40] N. Montazeri, et al., "Switching Kalman filter based methods for apnea bradycardia detection from ECG signals," *Physiol Meas*, vol. 36, no. 9, pp. 1763–1783, sep 2015.

[41] F. Noman, et al., "Heart sound segmentation using switching linear dynamical models," in *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on*, 2017, pp. 1000–1004.

[42] G. D. Clifford, et al., "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *Comput Cardiol, 2016*, Vancouver, BC, Canada, 2016, pp. 609–612.

[43] K. Murphy, "Switching Kalman filters," Tech. Rep., UC Berkeley., 1998.

[44] A. N. Pelech, "The physiology of cardiac auscultation," *Pediatric Clinics*, vol. 51, no. 6, pp. 1515–1535, 2004.

[45] R. F. Schmidt and G. Thews, *Human Physiology*, Springer Berlin Heidelberg, 2 edition, 2012.

[46] D. Bernstein and S. P. Shelov, *Pediatrics for Medical Students*, Wolters Kluwer/Lippincott Williams & Wilkins, 3rd edition, 2012.