

Coupled Hidden Markov Model-Based Method for Apnea Bradycardia Detection

N. Montazeri Ghahjaverestan, *Student Member, IEEE*, S. Masoudi, *Student Member, IEEE*,
M. B. Shamsollahi, *Senior Member, IEEE*, A. Beuchée, P. Pladys, D. Ge, *Member, IEEE*,
and A. I. Hernández, *Member, IEEE*

Abstract—In this paper, we present a novel framework for the coupled hidden Markov model (CHMM), based on the forward and backward recursions and conditional probabilities, given a multidimensional observation. In the proposed framework, the interdependencies of states networks are modeled with Markovian-like transition laws that influence the evolution of hidden states in all channels. Moreover, an offline inference approach by maximum likelihood estimation is proposed for the learning procedure of model parameters. To evaluate its performance, we first apply the CHMM model to classify and detect disturbances using synthetic data generated by the FitzHugh–Nagumo model. The average sensitivity and specificity of the classification are above 93.98% and 95.38% and those of the detection reach 94.49% and 99.34%, respectively. The method is also evaluated using a clinical database composed of annotated physiological signal recordings of neonates suffering from apnea-bradycardia. Different combinations of beat-to-beat features extracted from electrocardiographic signals constitute the multidimensional observations for which the proposed CHMM model is applied, to detect each apnea bradycardia episode. The proposed approach is finally compared to other previously proposed HMM-based detection methods. Our CHMM provides the best performance on this clinical database, presenting an average sensitivity of 95.74% and specificity of 91.88% while it reduces the detection delay by -0.59 s.

Index Terms—Apnea-bradycardia (AB), coupled hidden Markov model (CHMM), electrocardiography (ECG), forward-backward (FB) algorithm, hidden Markov model (HMM).

I. INTRODUCTION

EPISODES of apnea-bradycardia (AB), consisting of a respiratory pause (apnea) accompanied by oxygen desatu-

ration and a significant drop in heart rate (bradycardia), are commonly observed during the first weeks of life of preterm newborns. These episodes compromise tissue perfusion, affecting the normal development of the neurological and cardiovascular systems of the preterm infant and may even be lethal [1], [2]. The detection and characterization of AB episodes during continuous cardio-respiratory monitoring in neonatal intensive care units is thus of primary importance.

Although many methods have been presented in the literature for the detection of adult apnea events [3]–[5], the detection of AB episodes on preterm infants requires specific processing, integrating an analysis of the dynamics of heart rate through the event. The most common approaches for heart rate characterization in this context are simply based on the detection of bradycardia, by applying a fixed or an adaptive threshold on heart rate time-series [6], [7]. In our previous works, we have proposed different methods to improve the characterization of heart rate dynamics of AB episodes using abrupt-change detection methods [8], and different kinds of unidimensional hidden Markov models (HMM) [9]–[11]. In this paper, we present a significant improvement of our previous methods, by proposing a new methodological framework for the characterization of multivariate time-series dynamics, based on a particular kind of Bayesian network (BN), called the coupled hidden Markov model (CHMM). We also present a novel AB detection method for preterm infants, integrating a phase of multivariate feature extraction from the ECG, and a phase of time-series characterization through the proposed CHMM.

HMM is a particular type of BN, in which it is assumed that a sequence of discrete hidden states generates a sequence of observations of the same length. The topology of the state network and their probabilistic pattern rule the transition from one state to another. The state transition and initial state distribution govern the evolution of the unobserved hidden states, whereas the observation probability controls the observation data, for a given state sequence [12]. Although HMMs are, by definition, single-process models, different approaches have been proposed to use them for the analysis of multivariate data. One common approach is to consider the observation data as a multidimensional random variable. Although these approaches provide interesting results in some applications [13], [14] and [10], the single-process nature of HMM can not be adapted easily to other cases, where observations are indeed generated by distinct underlying processes, such as in vision, speech recognition [15] or forensic analysis [16].

Manuscript submitted in September 2014 and published in March 2016.
Dr. Alfredo Hernandez : alfredo.hernandez@inserm.fr

N. Montazeri Ghahjaverestan is with the Biomedical Signal and Image Processing Laboratory, School of Electrical Engineering, Sharif University of Technology, Tehran 11356-9363, Iran and also with LTSI, University of Rennes 1 and also with INSERM, Rennes F-35000, France (e-mail: montazeri.nasim@gmail.com).

S. Masoudi and M. B. Shamsollahi are with the Biomedical Signal and Image Processing Laboratory, School of Electrical Engineering, Sharif University of Technology, Tehran 11356-9363, Iran (e-mail: masoudi@ee.sharif.edu; mbshams@sharif.edu).

A. Beuchée and P. Pladys are with CHU Rennes, Rennes F-35000, France, and also with the INSERM, Rennes F-35000, France, and also with the LTSI, University of Rennes 1, Rennes F-35000, France (e-mail: alain.beuchee@chu-rennes.fr; patrick.pladys@universal.fr).

D. Ge and A. I. Hernández are with the INSERM, Rennes F-35000, France and also with the LTSI, University of Rennes 1, Rennes F-35000, France (e-mail: di.ge@univ-rennes1.fr; alfredo.hernandez@univ-rennes1.fr).

In order to overcome these limitations, CHMM has been proposed by Brand in [15], [17] as a generalization of HMM. In a CHMM, each channel (Markov chain) is associated with an univariate observation to represent its underlying generation process, and transition probabilities depend on the current state of all channels. However, this structure implies that the state space grows exponentially with respect to the number of channels [17]. In order to cope with this complexity, Brand proposed a simplification, considering a factorization of the transition matrix

$$P(S_t^c | S_{t-1}^1, S_{t-1}^2, \dots, S_{t-1}^C) = \prod_{c=1}^C P(S_t^c | S_{t-1}^c) \quad (1)$$

where S_t^c denotes the state of channel c at time t . According to Brand's assumption, the state conditional probability in the left side of (1) is substituted by the product of all marginal conditional probabilities. This model has been successfully applied in the field of sensor fusion, such as in forensic electronics [16], genetics [18], audio-visual speech recognition systems [19] and target tracking [20].

In another study, Rezek *et al.* computed the forward-backward (FB) parameters without considering Brand's assumption in (1) and derived the maximum likelihood (ML) estimators for the CHMM parameters using the expectation maximization (EM) algorithm [21]. They considered the CHMM as a one-channel HMM with an ordered C-fold state formed by (S^1, S^2, \dots, S^C) , with a state space of $N = \prod_{c=1}^C n_c$ and a transition matrix of dimension $N \times N$. The algorithm's complexity reaches $O(TM^{2C})$ with $M = n_c$. Zhong *et al.* proposed another approach in [22], [23], in which the transition probability is expressed under the form of a weighted sum of marginal conditional probabilities, associated with a rather complex estimation method for the model parameters considering the normalization of weights.

All the previous methods are based on the recursive relations of FB variables, for which the learning procedure can be complicated and time consuming, such as in [22]. In this paper, we propose an original framework for CHMM with a considerable reduction in the complexity of the FB algorithm. Section II presents the methodological contributions regarding the proposed CHMM framework, addressing the problems of likelihood evaluation, state sequence optimization and model parameter learning. Section III presents the strategy of applying the proposed framework for detection of AB. In Section IV, the proposed method is quantitatively compared with Rezek's approach [21], as well as with classical HMM and its generalization, hidden semi Markov model (HSMM) [24], using simulated and a real database. Finally, Section V presents the conclusions of this study.

II. PROPOSED CHMM FRAMEWORK

Let's note $\{S_1^c, S_2^c, \dots, S_{M(c)}^c\}$ to be the state space of channel c in a CHMM and let q_t^c and o_t^c , $t = 1, 2, \dots, T$ be the state and the observation of channel c at time t , respectively. C is the total number of channels in the CHMM model. Also, let $a_{nm}^{c,c} = P(q_t^c = S_m^c | q_{t-1}^c = S_n^c)$ denotes the probability of transition to state m in channel c at time t , subjected to being in state

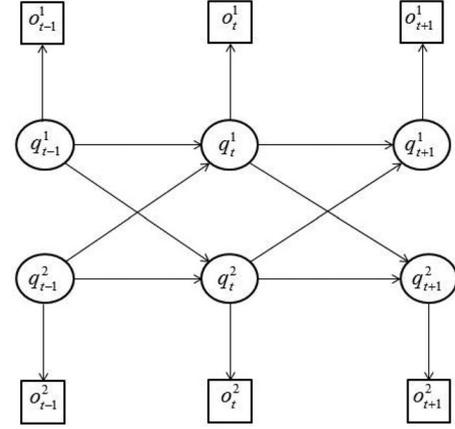


Fig. 1. BN representation for the proposed CHMM showing the probabilistic relations among the states and observations for the particular case of a two-channel CHMM.

n in channel c' at time $t-1$. The probability of the observation is written as $b_m(o_t^c) = P(o_t^c | q_t^c = S_m^c)$, where o_t^c may be either discrete or continuous. We also define $o_t = \{o_t^1, o_t^2, \dots, o_t^C\}$. In this paper, $b_m(o_t^c)$ is assumed to be normally distributed, and characterized by its mean (μ_m^c) and standard deviation (σ_m^c). For simplicity, we also note $v_t^c(m) \equiv \{q_t^c = S_m^c\}$, so that $a_{nm}^{c,c} = P(v_t^c(m) | v_{t-1}^c(n))$. The structure of the interchannel coupling is depicted in Fig. 1 for the two-channel case.

In order to define the FB parameters, first consider the probability of observation of each channel at time t given all the previous observations

$$\begin{aligned} P(o_t^c | o_{1:t-1}) &= \sum_{m=1}^{M(c)} P(v_t^c(m), o_t^c | o_{1:t-1}) \\ &= \sum_{m=1}^{M(c)} b_m(o_t^c) P(v_t^c(m) | o_{1:t-1}). \end{aligned} \quad (2)$$

Following [24], we define the forward parameter as

$$\alpha_{t|x}^c(m) = P(v_t^c(m) | o_{1:x}) \quad (3)$$

where for $x = t-1, t, T$, the above quantity is termed as predicted, filtered and smoothed probability, respectively. The forward recursion based on the predicted probability is determined by

$$\begin{aligned} \alpha_{t|t-1}^c(m) &= P(v_t^c(m) | o_{1:t-1}) \\ &= \sum_{n_1=1}^{M(1)} \sum_{n_2=1}^{M(2)} \dots \sum_{n_C=1}^{M(C)} P(v_t^c(m), v_{t-1}^1(n_1), v_{t-1}^2(n_2), \\ &\quad \dots, v_{t-1}^C(n_C) | o_{1:t-1}) \\ &= \sum_{n_1=1}^{M(1)} \sum_{n_2=1}^{M(2)} \dots \sum_{n_C=1}^{M(C)} P(v_t^c(m) | v_{t-1}^1(n_1), v_{t-1}^2(n_2), \\ &\quad \dots, v_{t-1}^C(n_C), o_{1:t-1}) \\ &\quad \times P(v_{t-1}^1(n_1), \dots, v_{t-1}^C(n_C) | o_{1:t-1}) \end{aligned} \quad (4)$$

where the first term can be simply calculated by Brand's assumption in (1). $o_{1:t-1}$ can be omitted since knowing all the previous states, it does not add any information to estimate $v_t^\zeta(m)$

$$P(v_t^\zeta(m)|v_{t-1}^1(n_1), \dots, v_{t-1}^C(n_C), o_{1:t-1}) \\ = \prod_{c=1}^C P(v_t^\zeta(m)|v_{t-1}^c(n_c)). \quad (5)$$

For the second term we show in Appendix B that the states of channels given the observations $o_{1:t-1}$ are independent

$$P(v_{t-1}^1(n_1), \dots, v_{t-1}^C(n_C)|o_{1:t-1}) = \prod_{c=1}^C P(v_{t-1}^c(n_c)|o_{1:t-1}). \quad (6)$$

Substituting (1), (5) and (6) in (4), the following recursion can be obtained:

$$\alpha_{t|t-1}^\zeta(m) = \sum_{n_1=1}^{M(1)} \sum_{n_2=1}^{M(2)} \dots \sum_{n_C=1}^{M(C)} \prod_{c=1}^C P(v_t^\zeta(m)|v_{t-1}^c(n_c)) \\ \times P(v_{t-1}^c(n_c)|o_{1:t-1}). \quad (7)$$

Note that $P(v_t^\zeta(m)|v_{t-1}^c(n_c)) \neq P(v_t^\zeta(m)|v_{t-1}^c(n_c), o_{1:t-1})$. Hence, the summations can be exchanged with the product.

$$\alpha_{t|t-1}^\zeta(m) = \prod_{c=1}^C \sum_{n_c=1}^{M(c)} P(v_t^\zeta(m)|v_{t-1}^c(n_c)) P(v_{t-1}^c(n_c)|o_{1:t-1}) \\ = \prod_{c=1}^C \sum_{n_c=1}^{M(c)} a_{n_c m}^{c\zeta} \alpha_{t-1|t-1}^c(n_c) \\ = \prod_{c=1}^C \sum_{n_c=1}^{M(c)} a_{n_c m}^{c\zeta} \alpha_{t-1|t-2}^c(n_c) \tilde{b}_{n_c}^c(o_{t-1}) \quad (8)$$

where $\tilde{b}_m^\zeta(o_t)$ is defined as the ratio of the filtered probability, $\alpha_{t|t}^\zeta(m)$, over the predicted probability $\alpha_{t|t-1}^\zeta(m)$.

$$\tilde{b}_m^\zeta(o_t) \triangleq \frac{\alpha_{t|t}^\zeta(m)}{\alpha_{t|t-1}^\zeta(m)} = \frac{P(o_t|v_t^\zeta(m), o_{1:t-1})}{P(o_t|o_{1:t-1})} \\ = \frac{P(o_t|v_t^\zeta(m), o_{1:t-1})}{\prod_{c=1}^C P(o_t^c|o_{1:t-1})}. \quad (9)$$

The last equality is based on the observation decomposition (cf., Appendix A)

$$P(o_t|o_{1:t-1}) = \prod_{c=1}^C P(o_t^c|o_{1:t-1}). \quad (10)$$

Assuming the effects of other channels, the nominator of (9) is summarized as the product of distinct channel effects as follows:

$$P(o_t|v_t^\zeta(m), o_{1:t-1}) = \\ \underbrace{\sum_{n_1=1}^{M(1)} \dots \sum_{n_C=1}^{M(C)} P(o_t, v_t^1(n_1), \dots, v_t^C(n_C)|v_t^\zeta(m), o_{1:t-1})}_{C-1 \text{ (except channel } \zeta)} \quad (11)$$

$$= \underbrace{\sum_{n_1=1}^{M(1)} \dots \sum_{n_C=1}^{M(C)} P(o_t^1|v_t^1(n_1)) P(o_t^2|v_t^2(n_2)) \dots P(o_t^C|v_t^C(n_C))}_{C-1 \text{ (except channel } \zeta)} \\ \times P(v_t^1(n_1), \dots, v_t^C(n_C)|v_t^\zeta(m), o_{1:t-1}). \quad (12)$$

Since the probability calculation of o_t^c only needs the corresponding state, i.e., $v_t^c(n_c)$, we can establish the following equation:

$$P(o_t|v_t^\zeta(m), o_{1:t-1}) \\ = b_m(o_t^\zeta) \underbrace{\sum_{n_1=1}^{M(1)} \dots \sum_{n_C=1}^{M(C)} b_{n_1}(o_t^1) \dots b_{n_C}(o_t^C)}_{C-1 \text{ (except channel } \zeta)} \\ \times P(v_t^1(n_1), \dots, v_t^C(n_C)|v_t^\zeta(m), o_{1:t-1}). \quad (13)$$

For achieving a recursion, we utilize the following simplification [also necessary to prove (6)]:

$$P(v_t^1(n_1), \dots, v_t^C(n_C)|o_{1:t-1}) = \prod_{c=1}^C P(v_t^c(n_c)|o_{1:t-1}). \quad (14)$$

Then we have

$$P(v_t^1(n_1), \dots, v_t^C(n_C)|v_t^\zeta(m), o_{1:t-1}) = \\ \frac{\prod_{c=1}^C P(v_t^c(n_c)|o_{1:t-1})}{P(v_t^\zeta(m)|o_{1:t-1})} = \prod_{c=1, c \neq \zeta}^C \alpha_{t|t-1}^c(n_c). \quad (15)$$

Hence, using (2), we have

$$P(o_t|v_t^\zeta(m), o_{1:t-1}) = b_m(o_t^\zeta) \prod_{c=1, c \neq \zeta}^C \sum_{n_c=1}^{M(c)} b_{n_c}(o_t^c) \alpha_{t|t-1}^c(n_c) \\ = b_m(o_t^\zeta) \prod_{c=1, c \neq \zeta}^C P(o_t^c|o_{1:t-1}). \quad (16)$$

Accordingly, $\tilde{b}_m^\zeta(o_t)$ can be calculated by

$$\tilde{b}_m^\zeta(o_t) = \frac{b_m(o_t^\zeta)}{P(o_t^\zeta|o_{1:t-1})}. \quad (17)$$

Backward parameter is defined in a way that the probability of being in state m of channel c given all the samples of observations, i.e., smoothed probability, could be easily calculated by the product of forward and backward parameters.

$$\alpha_{t|T}^\zeta(m) = \alpha_{t|t-1}^\zeta(m) \times \beta_t^\zeta(m) \quad (18)$$

which leads to

$$\beta_t^\zeta(m) \triangleq \frac{P(v_t^\zeta(m)|o_{1:T})}{P(v_t^\zeta(m)|o_{1:t-1})} \\ = \tilde{b}_m^\zeta(o_t) \frac{P(o_{t+1:T}|v_t^\zeta(m), o_{1:t})}{P(o_{t+1:T}|o_{1:t})}. \quad (19)$$

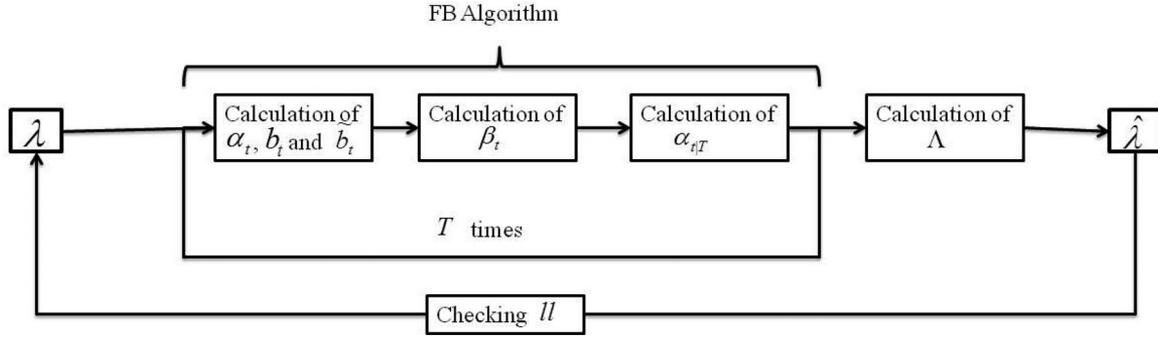


Fig. 2. Block diagram of training procedure. E step: assuming the set of model parameter λ , $\Lambda_t^{c'c}(m', m)$ and $\alpha_{t|T(m)}$ are computed for $m, m' = 1, 2, \dots, M$ and $t = 1, 2, \dots, T$ using FB algorithm. M step: calculation of the model parameters using (35) and (37)–(39).

The backward recursion is derived based on the next transition in channel ς as follows:

$$\begin{aligned} \beta_t^\varsigma(m) &= \tilde{b}_m^\varsigma(o_t) \sum_{n=1}^{M(\varsigma)} \frac{P(o_{t+1:T}, v_{t+1}^\varsigma(n) | v_t^\varsigma(m), o_{1:t})}{P(o_{t+1:T} | o_{1:t})} \\ &= \tilde{b}_m^\varsigma(o_t) \sum_{n=1}^{M(\varsigma)} \frac{P(o_{t+1:T} | v_{t+1}^\varsigma(n), v_t^\varsigma(m), o_{1:t})}{P(o_{t+1:T} | o_{1:t})} \\ &\quad \times P(v_{t+1}^\varsigma(n) | v_t^\varsigma(m), o_{1:t}). \end{aligned} \quad (20)$$

Considering (19) and $P(o_{t+1:T} | v_{t+1}^\varsigma(n), v_t^\varsigma(m), o_{1:t}) = P(o_{t+1:T} | v_{t+1}^\varsigma(n), o_{1:t})$, the first term in the summation can be identified as $\beta_{t+1}^\varsigma(n)$. The second term can be written as

$$\begin{aligned} &P(v_{t+1}^\varsigma(n) | v_t^\varsigma(m), o_{1:t}) \\ &= \underbrace{\sum_{n_1=1}^{M(1)} \dots \sum_{n_C=1}^{M(C)} P(v_{t+1}^\varsigma(n), v_t^1(n_1), \dots, v_t^C(n_C) | v_t^\varsigma(m), o_{1:t})}_{\text{except } \varsigma} \\ &= \underbrace{\sum_{n_1=1}^{M(1)} \dots \sum_{n_C=1}^{M(C)} P(v_{t+1}^\varsigma(n) | v_t^1(n_1), \dots, v_t^C(n_C), v_t^\varsigma(m), o_{1:t})}_{\text{except } \varsigma} \\ &\quad \times P(v_t^c(n_c), \dots, v_t^C(n_C) | v_t^\varsigma(m), o_{1:t}) \\ &= a_{mn}^{\varsigma\varsigma} \prod_{c \neq \varsigma} \sum_{n_c=1}^{M(c)} a_{n_c n}^{\varsigma c} \alpha_{t|t-1}^c(n_c) \tilde{b}_{n_c}^c(o_t). \end{aligned} \quad (21)$$

Thus, the backward recursion is given by

$$\begin{aligned} \beta_t^\varsigma(m) &= \tilde{b}_m^\varsigma(o_t) \sum_{n=1}^{M(\varsigma)} \{ \beta_{t+1}^\varsigma(n) a_{mn}^{\varsigma\varsigma} \\ &\quad \times \prod_{c=1, c \neq \varsigma}^C \sum_{n_c=1}^{M(c)} a_{n_c n}^{\varsigma c} \alpha_{t|t-1}^c(n_c) \tilde{b}_{n_c}^c(o_t) \}. \end{aligned} \quad (22)$$

The initial condition of parameters are $\beta_T^c(m) = \tilde{b}_m^c(o_T)$ and $\alpha_{1|0}^c = \pi_m^c b_m^c(o_1)$.

We resume hereby the three main problems treated in the context of HMM.

Problem 1: Evaluation of the likelihood of an observation sequence, given the model parameters

$$P(o_{1:T}) = P(o_1) \prod_{t=2}^T P(o_t | o_{1:t-1}) \quad (23)$$

where

$$P(o_t | o_{1:t-1}) = \prod_{c=1}^C P(o_t^c | o_{1:t-1}) \quad (24)$$

the latter can be calculated by (2).

Problem 2: Finding the optimal state m_c^* that generates the observations. We can use the maximum *a posteriori* estimation

$$m_c^*(t) = \arg \max_m P(v_t^c(m) | o_{1:T}) = \arg \max_m \alpha_{t|T}^c \quad (25)$$

to construct the optimal sequence of channel c : $Q_c^* = \{m_c^*(1), m_c^*(2), \dots, m_c^*(T)\}$.

Problem 3: Learning of model parameters $\lambda = \{a_{m'm}^{c'c}, \mu_m^c, \sigma_m^c, \pi_m^c\}$ that best fit the observation $\{o_{1:T}\}$. The Baum–Welch algorithm [25] can be applied, using the maximum likelihood (ML) criterion. As a special case of the estimation-maximization (EM) algorithm, the likelihood is estimated using the FB parameters in the E-step while the M-step uses expected counts of transitions and observations to reestimate the parameters λ (see Fig. 2). This procedure is performed iteratively until convergence.

Liporace generalizes the Baum–Welch algorithm to a larger class of distributions of observation probability than just Gaussian distribution [26]. In the following, we present a generalization of the reestimation method inspired from the Liporace paper. The likelihood probability of the observations $\{o_{1:T}\}$ given the model parameters λ can be written as

$$P_\lambda(o_{1:T}, \mathcal{S}) = \sum_{\mathcal{S}} P_\lambda(o_{1:T}, \mathcal{S}) \quad (26)$$

while $P_\lambda(o_{1:T}, \mathcal{S})$ can be written as

$$P_\lambda(o_{1:T}, \mathcal{S}) = \prod_{c=1}^C \{ \pi_1^c b_1(o_1^c) \times \prod_{t=2}^T \prod_{c'=1}^C a_{m_{t-1} m_t}^{c'c} b_{m_t}(o_t^c) \} \quad (27)$$

where \mathcal{S} is the set of the state sequences of all channels. The objective is to maximize $P_\lambda(o_{1:T})$ over all parameters λ . The

algorithm starts with an initial guess of λ_0 , and then updates it to ensure that $P_{\lambda_{k+1}}(o_{1:T}) \geq P_{\lambda_k}(o_{1:T})$ in each iteration. As in the EM algorithm structure, an *auxiliary function* $Q(\lambda, \bar{\lambda})$ is used, defined by (see [26])

$$Q(\lambda, \bar{\lambda}) = \sum_{\mathcal{S}} P_{\lambda}(o_{1:T}, \mathcal{S}) \log(P_{\bar{\lambda}}(o_{1:T}, \mathcal{S})). \quad (28)$$

Extra normalization conditions include

$$\sum_{m=1}^{M(c)} \bar{a}_{m'm}^{c'} = 1, \text{ and } \sum_{m=1}^{M(c)} \bar{\pi}_m^c = 1 \quad (29)$$

for all channels (c) and states of channel (m'). These constraints are integrated using the Lagrange multiplier method to yield the maximization problem

$$\begin{aligned} \lambda_{\text{new}} = \underset{\lambda}{\text{arg max}} & Q(\lambda, \bar{\lambda}) + \sum_{c=1}^C \theta_c \left(\sum_{m=1}^{M(c)} \bar{a}_{m'm}^{c'} - 1 \right) \\ & + \sum_{c=1}^C \varepsilon_c \left(\sum_{m=1}^{M(c)} \bar{\pi}_m^c - 1 \right) \end{aligned} \quad (30)$$

where θ_c, ε_c are the Lagrange parameters.

A description of the reestimation procedure is detailed as follows [26].

A. Reestimation of Transition Matrices

Differentiate (30) with respect to each $\bar{a}_{m'm}^{c'}$ to obtain

$$\sum_{\mathcal{S}} P_{\lambda}(o_{1:T}, \mathcal{S}) \sum_{t \in \mathfrak{T}_{m'm}^{c'}} \frac{1}{\bar{a}_{m'm}^{c'}} - \theta_c = 0 \quad (31)$$

where $\mathfrak{T}_{m'm}^{c'} = \{t : q_{t-1}^c = S_{m'}, q_t^c = S_m\}$. Interchange the order of summations as in [26] to get

$$\begin{aligned} \bar{a}_{m'm}^{c'} \theta_c &= \sum_{t=1}^T \sum_{\mathcal{S} \in \mathfrak{S}_{m'm}^{c'}(t)} P(o_{1:T}, \mathcal{S}) \\ &= \sum_{t=1}^T P(o_{1:T}, q_{t-1}^c = S_{m'}, q_t^c = S_m) \end{aligned} \quad (32)$$

where $\mathfrak{S}_{m'm}^{c'}(t) = \{\mathcal{S} : q_{t-1}^c = S_{m'}, q_t^c = S_m\}$ denotes the set of state sequence having state m' at time $t-1$ and state m at time t . Summing over all states of channel c , we obtain θ_c

$$\theta_c = \sum_{k=1}^T \sum_{m=1}^{M(c)} P(o_{1:T}, q_{k-1}^c = S_{m'}, q_k^c = S_m). \quad (33)$$

Furthermore, we define $\Lambda_t^{c'}(m', m) = P(q_{t-1}^c = S_{m'}, q_t^c = S_m | o_{1:T})$ as the conditional smoothed transition probability

$$\begin{aligned} \Lambda_t^{c'}(m', m) &= P(v_{t-1}^{c'}(m'), v_t^c(m) | o_{1:T}) \\ &= \frac{P(v_{t-1}^{c'}(m'), v_t^c(m), o_{t:T} | o_{1:t-1})}{P(o_{t:T} | o_{1:t-1})} \\ &= \beta_t^c(m) \bar{a}_{m'm}^{c'} \alpha_{t-1|t-2}^{c'}(m') \bar{b}_{m'}(o_{t-1}^c). \end{aligned} \quad (34)$$

It is then straightforward to achieve the reestimation for the transition probability

$$\bar{a}_{m'm}^{c'} = \frac{\sum_{t=1}^T \Lambda_t^{c'}(m', m)}{\sum_{t=1}^T \sum_{m''=1}^{M(c)} \Lambda_t^{c'}(m', m'')}. \quad (35)$$

B. Reestimation of $\bar{\pi}_m^c$

Adding the appropriate constraint for initial probability of being in state m to the auxiliary function, we have

$$\frac{\partial}{\partial \bar{\pi}_m^c} \{Q(\lambda, \bar{\lambda}) - \varepsilon_c (\sum_{m=1}^{M(c)} \bar{\pi}_m^c - 1)\} = 0 \quad (36)$$

$$\begin{aligned} \sum_{\mathcal{S}} P_{\lambda}(o_{1:T}, \mathcal{S}) \frac{1}{\bar{\pi}_m^c} - \varepsilon_c &= 0 \\ \bar{\pi}_m^c \varepsilon_c &= P(o_{1:T}, q_1^c = S_m) \\ \varepsilon_c &= \sum_{m=1}^{M(c)} P(o_{1:T}, q_1^c = S_m). \end{aligned}$$

Thus, using (18), the reestimation of initial probability is straightforward

$$\bar{\pi}_m^c = \frac{\alpha_{1|T}^c(m)}{\sum_{m'=1}^{M(c)} \alpha_{1|T}^c(m')}. \quad (37)$$

C. Reestimation of $\bar{\mu}_m^c$ and $\bar{\sigma}_m^c$

As previously mentioned, the density functions of observations $b_m(o_t^c)$ are assumed to be Gaussian functions. Applying the derivation on auxiliary function with respect to the components of $\bar{\mu}_m^c$, we can obtain

$$\bar{\mu}_m^c = \frac{\sum_{t=1}^T \alpha_{t|T}^c(m) o_t^c}{\sum_{t=1}^T \sum_{m'=1}^{M(c)} \alpha_{t|T}^c(m')} \quad (38)$$

$$\bar{\sigma}_m^c = \frac{\sum_{t=1}^T \alpha_{t|T}^c(m) (o_t^c - \bar{\mu}_m^c)^2}{\sum_{t=1}^T \sum_{m'=1}^{M(c)} \alpha_{t|T}^c(m')}. \quad (39)$$

III. EVALUATION METHOD

In this section, we present the method applied in this study to evaluate the performance of the proposed CHMM framework, in the contexts of classification and online detection, using both simulated and real datasets. Quantitative performance evaluation is based on common performance metrics and accomplished in two steps: 1) training the CHMM models using a train dataset and 2) performing classification or online detection on the test dataset. Each channel represents a 1-D observation. To initialize states parameters, the k-means clustering method is first applied to the range of observations amplitudes in training data. Each cluster is represented by a Gaussian distribution to characterize the observation probability $b_m(o_t^c)$ while the number of clusters is equal to the number of states.

In the following, we describe the proposed approach for applying the CHMM for the classification of simulated time-series. Then, we explain how the proposed framework can be applied

for online detection. Note that the optimum number of states in Markovian models are found by Bayesian information criterion method.

A. Classification

In previous works [13], the problem of classifying a time-series into one of K classes has been addressed by defining a set of K competing models (M_1, \dots, M_K), for which a learning dataset corresponding to each class (L_1, \dots, L_K) is used to estimate the model parameters of each model ($\lambda_1, \dots, \lambda_K$). Then, in the test phase, each time-series of the test dataset (O) is analyzed by calculating the log-likelihood using each model: $\log P(O | Q^*, \lambda_k), j \in \{1, 2, \dots, K\}$, where Q^* represents the optimal state sequence. The classification result is obtained by choosing the class corresponding to the model presenting the maximum log-likelihood.

A similar approach is applied in this paper for the proposed CHMM. However, in order to cope with the multichannel nature of the CHMM, the overall log-likelihood for a CHMM corresponding to class k will be obtained by summing the log-likelihoods in all channels with their optimal state sequence Q_c^*

$$ll_k = \sum_{c=1}^C \log\{P(o^c | Q_c^*, \lambda_k)\}. \quad (40)$$

B. Online Detection

The classification application can also be extended to online detection, as in [13]. The data can be divided into overlapping moving windows, and the classification procedure is applied to each window. One of the K classes (i.e., class k) is defined to represent the event of interest. The difference of log-likelihood of class k from other classes can be obtained

$$ll_{\text{total}}^{kj}(t) = ll_k(t) - ll_j(t) \quad (41)$$

where $j \in \{1, 2, \dots, K\} - \{k\}$. In case of multiple channels as in our proposed CHMM, the above equation can be rewritten as

$$ll_{\text{total}}^{kj}(t) = \sum_{c=1}^C (ll_k^c(t) - ll_j^c(t)). \quad (42)$$

An event corresponding to class k takes place in a window containing time t if the following condition is satisfied:

$$ll_{\text{total}}^{kj}(t) > \delta_{kj} \quad (43)$$

where δ_{kj} is a constant threshold that should be optimized.

C. Performance Evaluation

We evaluate the performance of the classifiers and the detectors using the most common metrics found in the literature: the receiver operating characteristic (ROC) curve and its area under curve (AUC) defined by sensitivity, and specificity of classification. To express how successfully a detector recognizes events without missing them, sensitivity [SEN = TP/(TP + FN)] is used. Likewise, specificity [SPC = TN/(TN + FP)] measures how exclusively it does not detect a wrong event. TP, FP, TN and FN denote the number of true positives, false positives, true negatives and false negatives, respectively. False positive

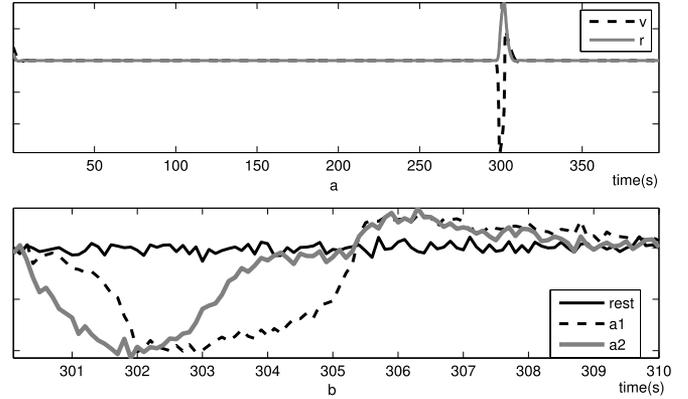


Fig. 3. (a) Simulated signals obtained from the FitzHugh–Nagumo model (state variables v and r) with a disturbance applied in 300–305 (s) without noise. (b) Focusing on v dimension with 5-dB additive noise, in the rest condition and during activations with different dynamics (a_1 and a_2). Note that different dynamics are obtained when using parameters a_1 and a_2 , while the amplitude of the responses are similar.

rate error can also be defined as $1 - \text{SPC}$. In a classification task, $AC = (\text{TP} + \text{TN}) / (N + P)$, where $N = \text{TN} + \text{FN}$ and $P = \text{TP} + \text{FP}$, is also calculated to evaluate the accuracy of the method. Moreover, time delay is another crucial metric for the overall system performance evaluation in the case of early detection of a desired event, and is defined as the difference between the detected onset and the annotated onset of an event.

An ROC is traced using different threshold values in (43). The point with optimum detection performance and related threshold are chosen using a criterion called perfect detection (PD) defined as the maximum product of sensitivity and specificity.

All the methods are executed five times with different records in the training and test sets, in order to cross-validate the results. Therefore, the mean and variance over all executions are reported.

D. Evaluation Datasets

1) *Simulated Data*: The simulated data is generated with the FitzHugh–Nagumo model, defined by the following differential equations:

$$\begin{aligned} \frac{dv}{dt} &= 3(v - \frac{1}{3}v^3 + r + I) \\ \frac{dr}{dt} &= -\frac{1}{3}(v - a + 0.8r) \end{aligned} \quad (44)$$

where the variables r and v are set to their “rest” values (fixed point) and disturbances are injected to the system by changing the value of I from 0 to 1. The dynamics of the system depend on the value of parameter a , which is assumed in this test to be a random variable with a uniform probability density function $a_1 \sim \mathcal{U}(0.58, 0.62)$ for the first class and $a_2 \sim \mathcal{U}(0.78, 0.82)$ for the second class. The impact of changing parameter a in the simulated data is illustrated in Fig. 3. The resemblance between the time series in class 1 and class 2 makes the classification a difficult issue. An appropriate classifier for such problem needs to correctly differentiate the dynamics of these time series and not only their instantaneous amplitudes. In order to evaluate

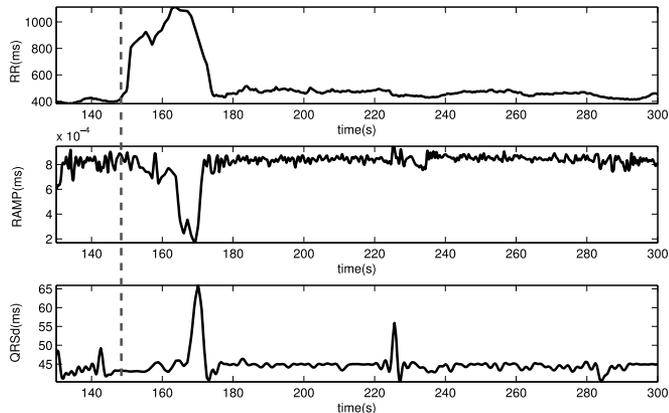


Fig. 4. Example of time series extracted from test real data. The annotated onset of the AB event is shown by dashed grey line.

the proposed approach, 200 sequences of 400 s duration at a sampling frequency of 10 Hz are generated with a disturbance introduced during 300–305 s [see Fig. 3(a)]. 40 segments of 300–310 s are used for training [cf., Fig. 3(b)] and the rest of them for evaluation.

Concerning online detection, train data is similar to that used for the classification case. However, the test data contain the whole 400-s-duration signals and are processed with 10-s sample by sample sliding window. Finally, a white Gaussian noise was added to achieve SNR value of 5 dB to all the database. Moreover, since the range of signal amplitude variations in the rest condition is lower than in dynamic conditions, the number of states of the rest class is less.

2) *Real Data*: For real data analysis, we use a database acquired during our previous works and consisting of 236, one-lead ECG segments from 32 preterm infants, hospitalized in the NICU at the University Hospital of Rennes, France [14]. ECG signals were acquired at 400 Hz. This observational study was approved by the ethical committee of the University Hospital of Rennes and a written consent was obtained from the parents of each infant. All data were anonymized at acquisition.

Each acquired ECG segment was denoised by a combination of low-pass and notch filters for discarding the baseline and the noise of 50 Hz, respectively. Individual beats for each segment were detected using the Pan and Tompkins algorithm [27]. Finally, three ECG features were extracted in a beat-to-beat basis, by applying a wavelet-based beat delineator described in detail in [13]. The first obvious feature is the time interval between two consecutive beats (RR interval), representing the cardiac period, or the inverse of the heart rate, since a bradycardia event is usually observed as a marked increase of cardiac period. The time series representing the evolution of these cardiac periods are called the RR series. Furthermore, the amplitude of R-wave on the ECG (RAMP) has been considered since it is modulated by the respiratory activity [28]. Some works have also reported an increase in the duration of the QRS complex (QRSd) preceding episodes of AB, so this feature has also been extracted [29]. The total set of features that will be used as observations to evaluate the CHMM framework are thus: the RR interval, R-wave amplitude (RAMP) and QRS duration (QRSd). The

TABLE I
COMPARISON OF THE CROSS-VALIDATION RESULTS CORRESPONDING TO MARKOVIAN METHODS FOR OPTIMAL STATE NUMBER FOR DYNAMICS CLASSIFICATION

Method	Class	# States	SEN(%)	SPC(%)	AC(%)
HMM	a_1	7	55.63 ± 37.30	91.56 ± 18.69	79.58 ± 10.13
	a_2	7	83.13 ± 37.39	77.50 ± 18.63	79.38 ± 9.96
	(v) <i>rest</i>	3	99.38 ± 0.77	100 ± 0	99.79 ± 0.26
HMM	a_1	3	86.25 ± 5.23	90.63 ± 2.30	89.17 ± 1.59
	a_2	3	81.25 ± 4.59	92.94 ± 2.52	89.04 ± 1.55
	(v/r) <i>rest</i>	2	99.63 ± 0.34	100 ± 0	99.88 ± 0.11
HSMM	a_1	7	67.00 ± 12.67	84.50 ± 3.81	78.67 ± 4.74
	a_2	7	69.00 ± 7.62	80.50 ± 6.82	76.67 ± 4.64
	(v) <i>rest</i>	3	94.00 ± 3.79	100 ± 0	98.00 ± 1.26
HSMM	a_1	3	100 ± 0	74.44 ± 1.26	82.96 ± 0.84
	a_2	3	48.88 ± 2.52	100 ± 0	82.96 ± 0.84
	(v/r) <i>rest</i>	2	100 ± 0	100 ± 0	100 ± 0
CHMM	a_1	5	89.56 ± 31.49	96.75 ± 8.65	94.35 ± 11.26
	a_2	5	99.69 ± 0.53	93.25 ± 17.80	95.40 ± 11.85
	(v/r) <i>rest</i>	3	90.75 ± 16.99	100 ± 0	96.92 ± 5.66
our	a_1	5	96.04 ± 9.69	97.19 ± 4.04	96.81 ± 3.43
CHMM	a_2	5	93.98 ± 6.26	95.38 ± 4.49	95.58 ± 4.00
(v/r) <i>rest</i>	3	95.13 ± 2.98	100 ± 0	98.38 ± 0.99	

sampling frequency of these extracted signals is transformed to 10 Hz, using well-known interpolation techniques. Fig. 4 depicts an example of these features before, during and after an AB episode.

The objective here is to detect, with the maximum detection performance and the lowest detection delay, all events of AB of the database. Two datasets are constructed for the learning phase, consisting of segments of 7 s duration: LS1: composed of 30 segments taken randomly and beginning with a bradycardia event. LS2: consisting of 30 segments taken randomly from the normal parts of the series (without any AB event). The length of these segments (7 s) corresponds to the average time measured from the beginning of the bradycardia to the peak RR value within the bradycardia episodes [9]. Two models are trained, corresponding respectively to bradycardia and normal segments. Then, the evaluation of the trained models for detecting the onset of bradycardia is applied with a sliding window of size $T = 7$ s. The test dataset includes 40 sets of RR-RAMP-QRSd time-series including mostly normal activity and one or more AB events.

IV. RESULTS

In this section, we report both the classification results on simulated data and the results of detection of disturbance in simulated data and bradycardia arrhythmia in preterm infants. Comparisons are made with respect to existing Markovian-based models such as HMM [12], HSMM [24] and the CHMM proposed by Rezek *et al.* [21].

A. Classification of Dynamics in Simulated Data

A classification is performed on simulated data, by defining three classes: a_1 , a_2 and a rest condition. We further assume that the number of states of the two competing models is equal and

