Using Hidden Markov Models For Sleep Disorderd Breathing Identification

Tarik Al-ani a,*, Yskandar Hamam b, Redouane Fodil b, Frédéric Lofaso b, Daniel Isabey b

a Groupe E.S.I.E.E., Cité Descartes, BP 99 - 93162 Noisy-le-Grand Cedex, France
b Unité de Physiopathologie et Thérapeutique Respiratoires
INSERM U492, Faculté de Médecine, 8, rue du Général Sarrail, 94010, Créteil Cedex, France

Abstract

In this work, an automatic diagnosis system based on Hidden Markov Models (HMMs) is proposed to help clinicians in the diagnosis of sleep apnea syndrome. Our system offers the advantage of being based on solid probabilistic principles rather than a predefined set of rules. Conventional and new simulated annealing based methods for the training of HMMs are incorporated. The inference method of this system translates parameter values into interpretations of physiological and pathophysiological states. The interpretation is extended to sequences of states in time to obtain a state-space trajectory. Some of the measurements of the respiratory activity issued by the technique of polysomnography (brain activity, respiratory activity, oxygen levels, and cardiac activity) are considered for off-line and on-line detection of the different sleep apnea syndromes: obstructive, central and hypopnea. Experimental results using respiratory clinical data and some future perspectives of our work are presented.

Keywords: Diagnosis; Sleep Apnea; Stochastic modeling; Training

1. Introduction

Our aim is to build an automatic on-line/off-line sleep apnea diagnosis system. Sleep apnea refers to episodes of non-breathing events lasting more than 10 seconds [2]. This happens during sleep and may sometimes occur more than 300 times a night. Based on a study of healthy subjects, it was suggested that more than 30 apneas per night should be considered abnormal [8]. These episodes may in a transient manner, awaken the patient resulting into sleep fragmentation. As a result, the level of oxygen may drop to dangerously low levels which may then result in cardiac arrhythmias (irregular heart beats) which may be fatal. Additionally, individuals with sleep apnea are much more prone to heart attacks and strikes. Mainly, there are three syndromes of sleep apneas: obstructive [8], central [19] and hypopnea [9]. The consequences of sleep apnea include depression, irritability, sexual dysfunction, learning and memory difficulties and falling asleep while driving or at work.

The most important diagnostic tool in any medical condition is for the physician to take the time to obtain a good history and physical examination. A chest x-ray along with laboratory tests are usually performed to evaluate other possible contributing factors, such as diabetes or hypothyroidism. The definitive diagnostic exam is a polysomnogram where the patient stays in a sleep laboratory or even at home overnight while measurements of his brain activity, respiratory activity, oxygen levels, and cardiac activity are performed. A sleep apnea recording is a very time consuming task consisting of expert visual evaluation all 10 minutes pieces of approximately 8 hour recording with a setting of many channels: airway flow, oesophageal pressure, gastric pressure, EEG, EMG, EOG and other signals, Figure 1.

E-mail addresses: t.alani@esiee.fr (T. Al-ani), y.hamam@esiee.fr (Y. Hamam), fodil@im3.inserm.fr (R. Fodil), f.lofaso@rpc.ap-hop-paris.fr (F. Lofaso) isabey@im3.inserm.fr (D. Isabey).
Effective treatment eliminates snoring and apnea. The success of the treatment is measured by the reduction of respiratory disturbance to normal levels, by the elimination of symptoms such as fatigue and depression, and the patient's subjective feeling of well-being. The main standard treatment accepted by physicians and others trained in sleep disorders medicine is the continuous application of positive air pressure through a mask during sleep (Continuous Positive Airway Pressure (CPAP)). The polysomnogram based measurements task of an approximately 8 hours is very time consuming, expensive and tedious.

![Figure 1. A polysomnography recording example.](image)

The increasing sophistication of medical devices is generally advantageous in breathing-care environments. Increasing both safety and efficiency of the treatment is quite important. It is especially so in the case of difficult, unexpected and time-critical situations [12], [20], [17]. Thus, off-line or on-line automation of monitoring and diagnosis of breathing-care unit patients becomes an important objective. Many approaches have been developed in the diagnosis field. These approaches are based on expert or rule based systems (including fuzzy logic systems) [16 et al., 1993] and model based systems [10], [11]. Expert systems have had a large impact on automated diagnosis. Expert systems, however, have several weaknesses, mainly due to the difficulty of building and maintaining large rule bases and due to the fact that inference algorithms are not well-suited for real-time action or fast response. A current trend is to use models instead of rules. The advantage of this approach is that a true predictive inference may be obtained. In general, most of these approaches have traditionally been based on linear modeling techniques, which restricts the type of practical situations that may be modeled. However, obtaining linear or non linear analytical models of many patient monitoring domains is a difficult problem. A good approach is to construct empirical models using machine learning based modeling system for forming empirical models at different levels of abstraction according to the modeling objectives specified by human experts. The modeling process may be considered as a learning activity using machine learning techniques. These techniques use a clinical data base to construct generalized or individual models for each patient. Based on this approach, the objective of our work is to construct an automated off-line or on-line diagnosis system of the different apnea syndromes. This system interprets sequences of data such as clinical measurements issued by the technique of polysomnography.
In recent years, new inference methods have been introduced to deal with sequences of data which traditionally cannot be dealt with efficiently due to problems such as missing, incomplete, and uncertain information. Our diagnosis approach is based on a stochastic training methodology for constructing automated diagnosis architecture. This approach uses Hidden Markov Models (HMMs) (see section 2). These models are based on an extension of the concept of Markov chains to include cases where the observation is a probabilistic function of the state.

The inference method of this system translates parameter values into interpretations of physiological and pathophysiological states. If the interpretation is extended to sequences of states in time, a state-space trajectory may be obtained. The identification of state-space trajectories is a useful concept in diagnosis because some disorders may only be distinguished from each other by time sequences of pathophysiological states. The probabilistic concept of HMMs captures the uncertainty inherent in state interpretation.

A major interest of using HMMs is their capabilities for predictive inference. This inference makes our diagnosis system useful in the evaluation of treatment plans, in the optimization of treatment plans, in the predictive alarming, for example in anticipation of non desirable parameter values and pathophysiological states and management of computational resources based on demand estimates.

The algorithms used in our diagnosis system are all developed in Scilab environment [14]. However, these algorithms may be very easily translated to Matlab or C language for off-line or on-line applications.

In Section 2 a brief introduction to hidden Markov models is given. In section 3, the basic methodology and structure of the diagnostic system are introduced. Section 4 gives some results of our approach using a real clinical data. The conclusions of this paper are presented in section 5.

2. Basic Hidden Markov Models (HMMs)

Hidden Markov models (HMM) have been widely applied in automatic speech recognition. In this field, signals are encoded as temporal variation of short time power spectrum [13]. HMM applications are now being extended to many fields such as pattern recognition, signal processing, modeling and control of dynamic systems. In general, the HMMs approach seems to be promising in other fields such as in medical diagnosis and in other types of system monitoring [4], [15]. They are well suited for the classification of one or two dimensional signals. A HMM is a double stochastic process with one underlying process that is not observable but may be estimated through a set of processes that produce a sequence of observations. They may be used for the treatment of problems where information is uncertain and incomplete. Their use necessitates two stages [13]: a training stage where the stochastic process is estimated through extensive observation sequences and a detection stage where the model may be used in real-time to obtain sequences of maximum probability. HMM models owe their success to the existence of many algorithms which are efficient and reliable [3], [6], [13], [14]. The use of the trained HMM in real-time necessitates the use of an efficient algorithm which gives the state sequence of maximum probability. The Viterbi algorithm [18] or the forward algorithm [13] fulfills this need.

A HMM is a finite model that describes a probability distribution over an infinite number of possible sequences. The HMM is composed of some number of hidden states, which might correspond to pathophysiological states like breathing events. Each state "emits" observations according to observation emission probabilities (e.g. clinical data using the technique of polysomnography), and the states are interconnected by state transition probabilities. Starting from some initial state, a sequence of states is generated by moving from one state to another according to the transition probabilities until an end state is reached. Each state then emits observations according to that states' emission probability distribution, creating an observable sequence of observations. HMMs provide an effective
approach for incorporating the temporal context. They may be employed directly using the measured clinical signals of the process.

By applying off-line training procedures, different models may be constructed using only clinical measurements. Based on these models, off-line or on-line diagnosis and prediction may be achieved. In this work, only the measurements of the respiratory activity issued by the technique of polysomnography (upper airway flow, oesophageal pressure and gastric pressure) are considered.

A Hidden Markov Model is defined by the triplet $\lambda = (\Pi, A, B)$, where $\Pi$, $A$ and $B$ are the initial state distribution vector, matrix of state transition probabilities and matrix of measurement probability distribution, respectively.

$$
\Pi = [\pi_1, \pi_2, \ldots, \pi_N], \pi_i = P(q_t = i), \quad i, j \in \{1, 2, \ldots, N\}, t \in \{1, 2, \ldots, T\}.
$$

$$
A = [a_{ij}], \quad a_{ij} = P(q_{t+1} = j|q_t = i), \quad B = [b_j(O_t)], \quad b_j(O_t) = P(O_t|q_t = j),
$$

The observation may be either discrete or continuous. In this paper, it is considered as continuous Gaussian observation. In this case, $b_j(O_t) = N(M_j, \Sigma_j)$, where $M_j$ is the observation mean vector in the state $j$ and $\Sigma_j$ is the observation covariance matrix in state $j$. Thus $\lambda = \{\Pi, A, M, \Sigma\}$.

In general, at each instant of time the model is in one of the states $q_t = i$. It outputs $O_t$ with probability $b_i(O_t)$ and then jumps to state $q_{t+1} = j$ with probability $a_{ij}$. The model may be obtained off-line by training. The state transition matrix represents the structure of the HMM.

### 2.1 Structure of HMMs

The structure of the models defines the constraints on the elements in the state transition matrix $A$. The left-to-right models [13] are used to represent temporal sequences like in our case. An example of three state HMM is shown in Figure 2. They have the following properties: 1. The first observation is produced when the Markov chain is in the first state. 2. The last observation is generated when the Markov chain is in the final state. 3. There is not possibility to come back to a precedent state.

![Figure 2: A left-to-right HMM example. Three hidden states $S_1$, $S_2$, and $S_3$ with their corresponding emitted observations $O_1$, $O_2$, and $O_3$, respectively.](image)

### 2.2 Training and recognition problems

In Patient monitoring, we may represent the evolution of the process by some hidden states corresponding to event 1, event 2, ..., event $n$. For this process, the observations may be continuous online or off-line clinical data. To use HMMs, two phases should be considered: training phase and recognition phase.

**Training phase**: Build one HMM (estimate the parameters $\Pi$, $A$ and $B$) for each of the respiratory cases $w$ ($w = 1, \ldots, W$) to be modeled. For example, in patient monitoring, a multiple observation sequences corresponding to event $w$ is used to estimate the optimum parameters of its model. Event $k$ ($k=1, \ldots, W$) may be a normal breathing, obstructive apnea, hypopnea, etc.
Recognition or detection phase: For each hidden sequence in the test set characterized by an observation sequence \( O = O_1, O_2, ..., O_T \), the occurrence probability

\[
P_w = P(O | \lambda^w)
\]

is calculated for each model. The unknown observation sequence \( O \) is then classified as the process

\[
w^* = \arg \max_w P_w.
\]

For the recognition phase, the forward probability function that evaluates probability of a partial observation sequence at node i may be used [13].

3. Basic methodologie

The diagnosis system, Figure 3, is organized in such a manner to be interactive with the user. It is organized as three phases: training phase, state interpretation phase and detection phase.

A. Training phase:

Using a given training data set, a corresponding set of models is constructed using one of the following two main training approaches: Iterative Baum-Welch algorithm [13], or Simulated annealing algorithm [6], [7].

Baum-Welch algorithm use the most popular Expectation-Maximization (EM) algorithm [5] to perform Maximum-Likelihood of the HMMs parameters. When the number of observations is very large, the EM algorithm yields very satisfactory results in most applications. However, this algorithm converge to local maxima and its limiting position is independent on initialization. When small sample sizes are available, the objective function of the maximization becomes highly irregular and the EM algorithm can perform poorly. To overcome this shortcoming, our training algorithm uses simulated annealing method [1]. Based on the optimal trajectory of the state, it implements a finite coding of the solution space. It is applied to both discrete and continuous observations. It needs no specific initialization of the HMM parameters or the algorithm parameters (initial and final temperatures) by the user, the cooling schedule being general and applicable to any specific model. The parameters of the algorithm are derived from theoretical considerations. The algorithm is implemented within Scilab environments [14]. The cost evaluations of the algorithm are made independent of the problem size in order to minimize the computation time.

Our simulated annealing algorithm may be briefly presented as follows:

1. Given the observation sequence \( O \), a state sequence is generated at random;
With the observations that are then assigned to each state, the mean and the variance of the sample are calculated;

2. The probability of generating O with the guessed state sequence is then calculated. This is called the initial cost;

3. Afterwards one state in the sequence is changed at random and the probability that this new state sequence has generated O is calculated, after changing the mean and the variance of the changed states;

4. If the new probability (Pnew) is greater than the old one (Pold), or the probability of acceptance \( P_{\text{accept}} = \exp((P_{\text{new}} - P_{\text{old}})/\text{Temp}) \) is high enough, the change is saved. Temp is the temperature;

5. The probability of acceptance decreases with decreasing temperature so that for small values of Temp only better solutions are accepted.

Whatever the chosen training algorithm, in our system, we use the given polysomnographic data \( O = O_1, O_2, \ldots, O_T \) to train one HMM for each breathing event \( k \) (e.g. normal breathing, obstructive apnea, snoring, etc.).

B. Segmentation of the training data sequence into different pathophysiological state sequence - State interpretation phase:

Given training clinical observation sequences and their Hidden Markov Models, this phase allows to segment using Viterbi algorithm [6] and interpret, off-line, by an expert the pathophysiological state sequence (e.g. snoring, inspiration, expiration, obstructive apnea, etc.), Figure 4. The interpretation of each state will be useful for on-line detection and interpretation of the on-line state sequence. The identification of state-space trajectories is useful concept in diagnosis since some disorders can only be distinguished from each other by the sequences of pathophysiological states that they follow in time.

![Figure 4. An example of data segmentation of the air flow into different states and their interpretations. Event model contains three states: state 1: week airflow, state 2: snoring during expiration and state 3: snoring during inspiration.](image)
C. Off-line or on-line Detection phase

Once the HMMs are available, the detection of hidden pathophysiological states in a given observation sequence is possible. This may be done by two consecutive steps. At step 1 select the model that gives the maximum likelihood and at step 2: detect the pathophysiological state either using Viterbi dynamic programming algorithm for off-line detection or by calculating the forward probability for on-line detection. In the second case a sliding window is used on the observation sequence.

4. Case study

As an example, the respiratory activity diagnosis is considered to test the on-line/off-line robustness and reliability of our approach. Our experiment concern Sleep Apnea class modeling and detection. The main types of apneas are obstructive, central and mixed. Real data was used from different patients to study our approach. For apnea detection (pathophysiological state), only the measurements of the respiratory activities, upper airway flow, oesophageal pressure and gastric pressure, are considered. These measurements are stochastic in nature and patient dependent. We considered that the number of events to be modeled and their characteristics (labeling) are known a priori. In order to take into account the temporal nature and the non stationarity of the training signals, a multiple observation sequences corresponding to each event are used to estimate its optimum model parameters. Some training sequences for three events are given in Figure 5.

Figure 5. Some training sequences for three events. From top to down: normal breathing, obstructive hypopnea with snoring and one of many events (episodes) corresponding to obstructive apnea (non breathing events lasting more than 10 seconds). FLOW: airway flow (L/s), PESO: oesophageal pressure (CmH2o), PGAST: gastric pressure (CmH2o).
Figure 6 gives a Scilab simulation example for obstructive hypopnea event using 2 states: inspiration-expiration with snoring. Of course, the number of states in each event is chosen by an expert.

Figure 7 gives a simulation example of the on-line detection case.

The detection and the recognition of pathophysiological states using four HMMs for the events (normal breathing, snoring, obstructive apnea, centrale apnea) were satisfactory with 90% to 100% recognition rates. The given recognition rates are obtained by using five observation sequences for each event to train the different HMMs.

Figure 6. Obstructive hypopnea event using 2 states. From top to down: airway flow, oesophageal pressure, gastric pressure and the resulting of two states trajectory corresponding to the inspiration-expiration with snoring (state 1) and the end of expiration with snoring (state 2).
5. Conclusions and perspectives

In this paper, a brief description of the elements of a HMM based diagnosis system in Scilab environment was presented. Clearly, a training based diagnosis system, such as our system, is an interdisciplinary field which lies at the intersection of clinical medicine, computer science, AI and biomedical engineering. As a result, the use of this system necessitates a balanced distribution of clinicians and computer scientists.

The HMMs may be used to provide a predictive estimate of pathophysiological states. To use them, some considerations must be satisfied:

- The process is described by a first order Markov chain.

- A prior knowledge of the pathophysiological states (labeling) and their recordings data are available for model training purpose.

Our approach is now being tested on many different breathing data for robustness and reliability.

As stated in [21], the current techniques of investigating patients with suspected sleep disordered breathing are inadequate. The Obstructive apnea episodes are not usually difficult to detect even when only a basic measure of respiratory effort such as thoracic and abdominal movement is used. On the other hand, correctly identifying obstructive hypopneas and episodes of upper airway resistance needs a sensitive measure of airflow and inspiratory effort. The measurement of swings in pleural pressure by oesophageal manometry is the current gold standard techniques for detecting changes in respiratory effort. However, the placement of an oesophageal catheter is often uncomfortable and unacceptable, it may modify the upper airway dynamics, and some believe that it contributes to the sleep disturbance during the sleep study. Furthermore, this technique is available in only a proportion of sleep laboratories and, if performed, adds significantly to the cost of the sleep study. For all these reasons, other new techniques for detecting and classifying sleep apneas and other breathing disorders are developed using mainly the Electroencephalogram (EEG) [22] or Pulse Transit Time (PPT) [21]. We are presently investigating the use of the airway flow with these noninvasive signals, to construct and detect automatically the different types of apnea in real time. All the algorithms used in our diagnosis
system are developed in Scilab environment [14]. However, for the reasons of portability, these algorithms may be very easily translated to Matlab or C language for off-line or real time applications. Both Scilab and Matlab include real time tools in order to implement the developed procedures on hardware.

References