Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis

Marius Bartcus, Faicel Chamroukhi, Hervé Glotin

***Abstract*—Hidden Markov Models (HMMs) are one of the most popular and successful models in statistics and machine learning for modeling sequential data. However, one main issue in HMMs is the one of choosing the number of hidden states. The Hierarchical Dirichlet Process (HDP)-HMM is a Bayesian non-parametric alternative for standard HMMs that offers a principled way to tackle this challenging problem by relying on a Hierarchical Dirichlet Process (HDP) prior. We investigate the HDP-HMM in a challenging problem of unsupervised learning from bioacoustic data by using Markov-Chain Monte Carlo (MCMC) sampling techniques, namely the Gibbs sampler. We consider a real problem of fully unsupervised humpback whale song decomposition. It consists in simultaneously finding the structure of hidden whale song units, and automatically inferring the unknown number of the hidden units from the Mel Frequency Cepstral Coefficients (MFCC) of bioacoustic signals. The experimental results show the very good performance of the proposed Bayesian non-parametric approach and open new insights for unsupervised analysis of such bioacoustic signals.**

I. INTRODUCTION

Hidden Markov Models (HMM) [1] are one of the most successful models in statistics and machine learning for sequential data including acoustic recognition [1]. The usually used algorithm to learn the model is the Expectation-Maximization (EM) algorithm [2], also known as Baum-Welch in HMMs [3]. One main issue in HMMs is the one of selecting the number of hidden states, required by EM. This model selection problem can be addressed by cross validation techniques or information selection criteria such as the Bayesian Information Criterion (BIC) [4], the Akaike Information Criterion (AIC) [5], the Integrated Classification Likelihood criterion (ICL)[6], etc. which select an HMM with a number of states from a pre-estimated HMMs with varying number of states. The Bayesian Non-Parametric (BNP) approach for HMMs [7] gives a well-principled alternative to standard HMMs. This alternative is known as the infinite HMM (IHMM)[8]. It provides a principled way to infer the number of states from the data in an automatic way as the learning proceeds. The BNP approach for HMMs relies on Hierarchical Dirichlet Process (HDP) to define a prior over the states [7]. It is known as the Hierarchical Dirichlet Process for the Hidden Markov Models (HDP-HMM) [7]. HDP-HMM infers it from the posterior distribution considered the number of states as a hidden parameter. The HDP-HMM parameters can be estimated by MCMC sampling techniques such as Gibbs sampling [7]. Note that the standard HDP-HMM Gibbs sampling has the limitation of an inadequate modeling of the temporal persistence of states [9]. This problem has been addressed in [9] by relying on a sticky extension which allows a more robust learning. Other solutions for the inference of the hidden Markov model in this infinite state space models are using the Beam sampling [10] rather than Gibbs sampling.

We investigate the BNP formulation for the HMM, that is the HDP-HMM into a challenging problem of unsupervised learning from bioacoustic data. The problem consists of extracting and classifying, in a fully unsupervised way, unknown number of whale song units. We use the Gibbs sampler to infer the HDP-HMM from the bioacoustic data.

The paper is organized as follows. In Section II, we give a brief description of the finite Hidden Markov Model. Then, Section III describes the infinite Hidden Markov Model and Subsection III-A introduces the Hierarchical Dirichlet Process. A representation of the Hierarchical Dirichlet Process in terms of the Chinese Restaurant Franchise (CRF), is given in Subsection III-B. Section III-C describes the generative process, and Section III-D describes the Gibbs sampling algorithm used for the inference. Finally, Section IV presents the experimental results for the humpback whale song signals.

II. THE HIDDEN MARKOV MODEL

The finite Hidden Markov Model (HMM) [1] is very popular due to its rich mathematical structure and it stability to model sequential data, namely acoustic data. It assumes that the observed sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathbb{R}^d$ is the multidimensional observation at time t, is governed by a hidden state sequence $\mathbf{z} = (z_1, \ldots, z_T)$, where z_t represents the hidden state of $\mathbf{x}_t$ and takes its values in a finite set $\{1, \ldots, K\}$, K being the possibly unknown number of states. The generative process of the HMM can be described in general by the following steps. Starting with the first time step state, z_1 is distributed according to the initial transition distribution. Then, the current state z_t is distributed according to the transition distribution given the previous state (z_{t-1}) . Finally, given the state z_t, the observation x_t is generated from the emission distribution $F(\boldsymbol{\theta}_{z_t})$ of that state. This generative process for the HMM can be summarized as in Equation (1).

$$\begin{aligned} z_1 &\sim \pi_1 \\ z_t|z_{t-1} &\sim \pi_{z_{t-1}}, \ \forall t > 1 \\ \mathbf{x}_t|z_t &\sim F(\boldsymbol{\theta}_{z_t}) \end{aligned} \quad (1)$$

For example, $F(\boldsymbol{\theta}_{z_t})$ can take a Gaussian (normal) distribution, denoted as $\mathcal{N}(\mathbf{x}; \boldsymbol{\theta}_{z_t})$, where the emission parameters are the mean vector and the covariance matrix

Marius Bartcus and Faicel Chamroukhi are with Aix Marseille University (AMU), CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France and University of Toulon (UTLN), CNRS, LSIS, UMR 7296, 83957 La Garde, France. Hervé Glotin is with AMU, UTLN and Institut Universitaire de France (IUF), 75005 Paris, France. Contacts: {bartcus,chamroukhi,glotin}@univ-tln.fr

978-1-4799-1959-8/15/$31.00

$\boldsymbol{\theta}_{z_t} = \{\boldsymbol{\mu}_{z_t}, \boldsymbol{\Sigma}_{z_t}\}$. The joint distribution for the hidden states $\mathbf{z}$ and the observations $\mathbf{X}$ can be given by Equation (2).

$$p(\mathbf{z}, \mathbf{X}|\pi_1, \boldsymbol{\pi}, \boldsymbol{\theta}) = p(z_1)p(\mathbf{x}_1|z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})p(\mathbf{x}_t|z_t) \quad (2)$$

where T is the number of observations. The estimation of the Hidden Markov Model parameters $(\pi_1, \boldsymbol{\pi}, \boldsymbol{\theta})$, that are, the initial state transition, the transition matrix, and respectively the emission parameters, is in general, estimated in a maximum likelihood estimation (MLE) framework by using the Expectation Maximization (EM) algorithm, also known as the Bauch-Welch algorithm [1]. However, for the finite HMM, the number of unique states K is required to be known a priori for the EM algorithm. This model selection issue can be addressed in a two-stage scheme by using model selection criteria such as the Bayesian Information Criterion (BIC) [4], the Akaike Information Criterion (AIC) [5], the Integrated Classification Likelihood criterion (ICL)[6], etc to select a model from a pre-estimated HMMs with varying number of states.

The HDP-HMM [7][11][9] is a Bayesian non-parametric alternative for HMMs that offers a good alternative to select the number of clusters from the data as the learning proceeds, rather than a two-stage strategy.

III. The Infinite Hidden Markov Model

The Bayesian Non-Parametric (BNP) [12][13] alternative is a principled way to tackle the challenging problem of model selection in HMMs. Due the fact that, the transitions of states takes independent priors, there is no coupling across transitions between different states [8], therefore DP [14] is not sufficient to extend HMM to an infinite model. The Hierarchical Dirichlet Process (HDP) prior [7] over the transition matrix[8] tackle this issue and extends the HMM to the infinite state space model.

A. The Hierarchical Dirichlet Process

A Dirichlet Process (DP) [14] is a prior distribution over distributions. It can be denoted as $\mathrm{DP}(\alpha, G_0)$ and has two parameters, the scaling parameter α and the base measure G_0. However DP is not sufficient to extend HMM to an infinite state space model. When the data has a related but different generative process, the Hierarchical Dirichlet Process (HDP) prior is used to extend the HMM to an infinite state space HDP-HMM [7]. A HDP assumes that the base measure for a set of DP $G_k \sim \mathrm{DP}(\alpha, G_0), \forall k = 1, \ldots K$ is itself sampled from a DP with parameters (γ, H), that is $G_0 \sim \mathrm{DP}(\gamma, H)$.

The Chinese Restaurant Process plays a great role in the representation of the Dirichlet Process and HDP-HMM, by giving a metophor to the existence of a restaurant with possible infinite tables (clusters) that customers (the observations) are siting in that restaurant. An alternative of such a representation for the Hierarchical Dirichlet Process can be described by the Chinese Restaurant Franchise process.

B. The Chinese Restaurant Franchise (CRF)

The Chinese Restaurant Franchise (CRF) gives a representation for the Hierarchical Dirichlet Process (HDP) by extending the Chinese Restaurant Process (CRP) [15][16][17] to a set of (J) restaurants rather than a single restaurant. Suppose a patron of Chinese Restaurant creates many restaurants, strongly linked to each other, by a franchise wide menu, having dishes common to all restaurants. As a result, J restaurants are created (groups) with a possibility to extend each restaurant to an infinite number of tables (states) at witch the customers (observations) sit. Each customer goes to his specified restaurant j, where each table of this restaurant has a dish that shares between the customers that sit at that specific table. However, multiple tables of different existing restaurants can serve the same dish. Figure 1 represents one such Chinese Restaurant Franchise Process for 2 restaurants. One can see the customers $\mathbf{x}_{ji}$ entered the restaurant j and takes the place of one table t_{ji}. Each table has a specific dish k_{jt} that can be also common for different restaurants (this case we have 2 restaurant representation).

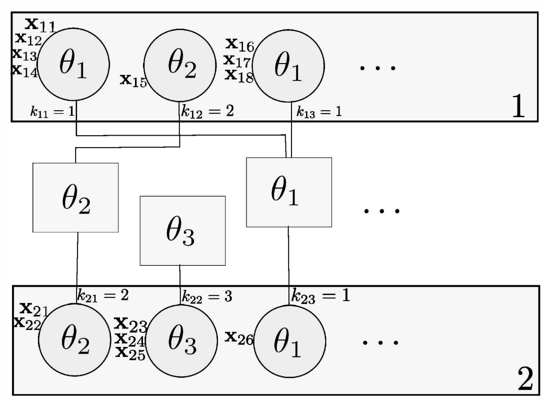

Fig. 1. Representation of a Chinese Restaurant Franchise with 2 restaurants. The clients $\mathbf{x}_{ji}$ are entering the jth restaurant ($j = \{1, 2\}$), sit at table t_{ji} and chose the dish k_{jt}.

The generative process of the Chinese Restaurant Franchise can be formulated as follows. For each table a dish is assigned with $k_{jt}|\beta \sim \beta$, where β is the rating of the dish served at the specific restaurant j. The table assignment of the jth restaurant for the ith customer is then drawn. Finally the observations, $\mathbf{x}_{ji}$, or the customers i that enters the restaurant j are generated by a distribution $F(\boldsymbol{\theta}_{k_{jt_{ji}}})$. The generative process for CRF is given in Equation (3).

$$\begin{aligned} k_{jt}|\beta &\sim \beta \\ t_{ji}|\tilde{\boldsymbol{\pi}}_j &\sim \tilde{\boldsymbol{\pi}}_j \\ \mathbf{x}_{ji}|\{\boldsymbol{\theta}_k\}_{k=1}^{\infty}, \{k_{jt}\}_{t=1}^{\infty}, t_{ji} &\sim F(\boldsymbol{\theta}_{k_{jt_{ji}}}) \end{aligned} \quad (3)$$

A graphical model of such a process can be seen in the following Figure (2).

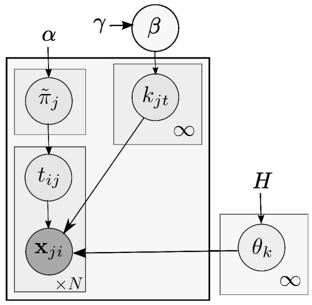

Fig. 2. Graphical representation of the Chinese Restaurant Franchise (CRF).

More details for derivation and inference of the Chinese Restaurant Franchise (CRF) and the use of it in the Hierarchical Dirichlet Process could be found in [7][18] and [9].

C. Model Definition

Hierarchical Dirichlet Process gives the possibility to have distributions over hyper-parameters by making the models more flexible. The coupling between transition matrix allows a higher level to DP prior over the parameters.

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{Dir}(\gamma/K, \ldots, \gamma/K) \\ \boldsymbol{\pi}_k &\sim \text{Dir}(\alpha\boldsymbol{\beta}) \end{aligned} \tag{4}$$

$\boldsymbol{\pi}_k$ being the transition matrix for the specific group k and $\boldsymbol{\beta}$ the prior hyper-parameter.

Let G_k describes both, the transition matrix $\boldsymbol{\pi}_k$ and the emission parameters $\boldsymbol{\theta}_k$, the infinite HMM can be described by the following generative process:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \text{GEM}(\gamma) \\ \boldsymbol{\pi}_k|\alpha, \boldsymbol{\beta} &\sim \text{DP}(\alpha, \boldsymbol{\beta}) \\ z_t|\boldsymbol{\pi}_k &\sim \text{Mult}(\boldsymbol{\pi}_k) \\ \boldsymbol{\theta}_k|H &\sim H \\ \mathbf{x}_t|z_t, \{\boldsymbol{\theta}_k\}_{k=1}^{\infty} &\sim F(\boldsymbol{\theta}_{z_t}) \end{aligned} \tag{5}$$

where $\boldsymbol{\beta}$ is a hyperparameter for the DP [19] that is distributed according to the stick-breaking construction noted GEM$(.)$; z_t is the indicator variable of the HDP-HMM that are sampled according to a multinomial distribution Mult$(.)$; the parameters of the model are drawn independently, according to a conjugate prior distribution H [1]; $F(\boldsymbol{\theta}_{z_t})$ is a data likelihood density, where we assume the unique parameter space of $\boldsymbol{\theta}_{z_t}$ being equal to $\boldsymbol{\theta}_k$. Suppose the observed data likelihood is a Gaussian density $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\theta}_k)$ where the emission parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ are respectively the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. According to [20][21], the prior over the mean vector and the covariance matrix is a conjugate Normal-Inverse-Wishart distribution, denoted as $\mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$, with the hyper-parameters describing the shapes and the position for each mixture densities: μ_0 is the mean of the mixtures should be, κ_0 the number of pseudo-observations supposed to be attributed, and ν_0, Λ_0 being similarly for the covariance matrix. In the generative process given in Equation (5), $\boldsymbol{\pi}$ is interpreted as a double-infinite transition matrix with each row taking a Chinese Restaurant Process (CRP), thus, in the HDP formulation "the group-specific" distribution, $\boldsymbol{\pi}_k$ corresponds to "the state-specific" transition where the Chinese Restaurant Franchise(CRF) defines distributions over the next state. As a consequence it was defined the infinite state space for the Hidden Markov Model. The graphical model for the infinite Hidden Markov Model is representated in figure 3.

[1]A conjugate prior distribution over $\boldsymbol{\theta}$ is a prior distribution for which the posterior distribution over $\boldsymbol{\theta}$ remains in the same family of the prior distribution.

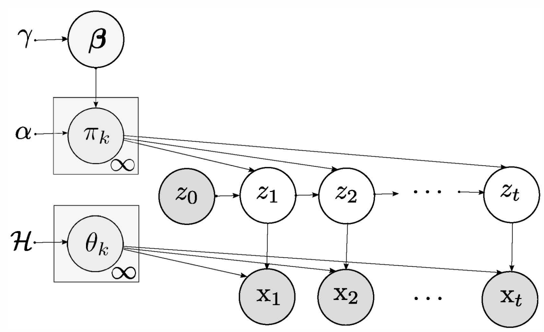

Fig. 3. Graphical representation of the infinite Hidden Markov Model (IHMM).

D. Inference of the infinite Hidden Markov Model

In this paper we investigate the inference of the Hidden Markov Model in the infinite state space (the HDP-HMM) with the Gibbs sampling algorithm. The base idea of the Gibbs sampling is to estimate the posterior distributions over all the parameters from the generative process of HDP-HMM given in (5).

The Gibbs sampling algorithm is briefly summarized in the pseudo-code (1) that computes $\mathcal{O}(K)$ probabilities for each of t states, therefore it has a $\mathcal{O}(TK)$ computational complexity. The main idea to inference the HDP-HMM is to estimate the hidden states of the observed data $\mathbf{z} = (z_1, \ldots z_T)$. This step needs computing two factors: the first is the conditional likelihood $p(\mathbf{x}_t|\mathbf{x}_{\backslash t}, z_t = k, \mathbf{z}_{\backslash t}, H)$ and the second factor $p(z_t|\mathbf{z}_{\backslash t}, \boldsymbol{\beta}, \alpha)$ computed as in Equation (10).

$$p(z_t = k|\mathbf{z}_{\backslash t}, \boldsymbol{\beta}, \alpha) \propto \begin{cases} (n_{z_{t-1},k} + \alpha\beta_k)\frac{n_{k,z_{t+1}} + \alpha\beta_{z_{t+1}}}{n_{k.} + \alpha} & \text{if } k \leq K,\ k \neq z_{t-1} \\ (n_{z_{t-1},k} + \alpha\beta_k)\frac{n_{k,z_{t+1}} + 1 + \alpha\beta_{z_{t+1}}}{n_{k.} + 1 + \alpha} & \text{if } k = z_{t-1} = z_{t+1} \\ (n_{z_{t-1},k} + \alpha\beta_k)\frac{n_{k,z_{t+1}} + \alpha\beta_{z_{t+1}}}{n_{k.} + 1 + \alpha} & \text{if } k = z_{t-1} \neq z_{t+1} \\ \alpha\beta_k\beta_{z_{t+1}} & \text{if } k = K + 1 \end{cases} \tag{10}$$

where n_{ij} is the number of transitions from state i to the state j, excluding the time steps t and $t-1$; $n_{.i}$ and $n_{i.}$ is the number of transition in and respectively out of state i and K is the number of distinct states in $\mathbf{z}_{\backslash t}$.

Second, sampling of the global transition distribution $\boldsymbol{\beta}$ is given by a Dirichlet distribution where $m_{.k}$ represents the number of clusters k, respectively one can say $m_{.k} = \sum_{j=1}^{K} m_{jk}$ [7][22]. Afterwards, the transition distribution $\boldsymbol{\pi}_k$, is sampled according to the Dirichlet distribution that is followed by the sampling of the emission parameters $\boldsymbol{\theta}_k$.

Algorithm 1 Gibbs sampling for the HDP-HMM

Inputs: The observations $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and the # of Gibbs samplings n_s

1: Initialize a random hidden state sequence $\mathbf{z}_0 = (z_1, \ldots, z_T)$.
2: **for** $q = 1$ **to** n_s **do**
3: **for** $t = 1$ **to** T **do**
4: 1. Sample the state z_t from

$$p(z_t = k|\mathbf{X}, \mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha, H) \propto p(\mathbf{x}_t|\mathbf{x}_{\setminus t}, z_t = k, \mathbf{z}_{\setminus t}, H) \; p(z_t = k|\mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha) \quad (6)$$

5: 2. Sample the global transition distribution

$$\boldsymbol{\beta} \propto \text{Dir}(m_{.1}, \ldots, m_{.K}, \gamma) \quad (7)$$

6: 3. Sample a new transition distribution

$$\boldsymbol{\pi}_k \propto \text{Dir}(n_{k1} + \alpha\beta_1, \ldots n_{kK} + \alpha\beta_K, \alpha \sum_{i=K+1}^{\infty} \beta_i) \quad (8)$$

7: 4. Sample the emission parameters $\boldsymbol{\theta}_k$.

$$\boldsymbol{\theta}_k \propto p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{z}, H, \boldsymbol{\theta}_{\setminus t}) \quad (9)$$

8: **end for**
9: 4. Eventually update the hyper-parameters α, γ.
10: **end for**

Outputs: The states assignments $\hat{\mathbf{z}}$ and the emission parameter vector $\hat{\boldsymbol{\theta}}_k$.

Assuming that the observed data takes a Gaussian distribution, the emission parameters to be estimated are the mean vector and the covariance matrix, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. These model parameters conditional on the data $\mathbf{X}$, states $\mathbf{z}$ and the prior distribution $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$ are sampled according to their posterior distributions.

Finally, the hyper-parameters α and γ, because of their lack of the strong beliefs, are sampled according to a Gamma distribution [7], [8], [10].

IV. Experiments on humpback whale song data

In this experiment, we consider the HDP-HMM in a challenging problem of humpback whale song decomposition.

The analysis of such complex signals that aims at discovering the call units (which can be considered as a kind of whale alphabet), can be seen as a problem of unsupervised call units classification as in [23]. We therefore reformulate the problem of whale song decomposition as a unsupervised sequential data class problem. Contrary to the approach used in [23], in which the number of states (call units in this case) has been fixed manually, here, we apply the Hierarchical Dirichlet Process for Hidden Markov Model (HDP-HMM) method to learn the complex bioacoustic data, to find the sates of the whale song, and automatically infer the number of states from the data. The idea is to highlight the effectiveness of using such a model over the difficult problem of discovering whale song units.

The used data are available in the framework of our SABIOD project publicly. They consist of MFCC parameters of 8.6 minutes of a Humpback whale song recordings produced at few meters distance from the whale in La Reunion - Indian Ocean. The data comprises 51336 observations with 39 features. A dimension reduction pretreatment with a PCA technique was made. We choose to retain 13 features of the data, since it was sufficient to capture more then 95% of the cumulative percentage of the variance. We used the Gibbs inference algorithm for Hierarchical Dirichlet Process for Hidden Markov Model which runs for 30000 samplings.

Figure (4) shows the state sequences partition, for all 8.6 minutes of humpback whale song data, obtained by the Gibbs sampling, with the maximum number of states proposed, a priori, to be equal to $L = 30$. One can see that the number of states estimated by the HDP-HMM Gibbs sampling is 6.

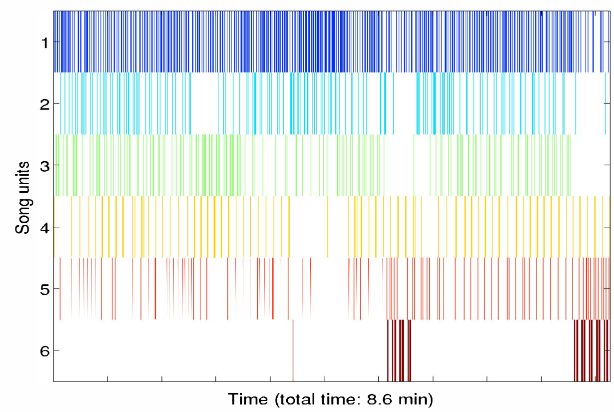

Fig. 4. The state sequences obtained by the Gibbs sampling inference approach for HDP-HMM.

For a more detailed information, the whole signal of the humpback whale song was separated by several parts of 15 seconds each. All the spectrograms of the humpback whale song and their corresponding obtained state sequence partitions, as well as the associated song are made available in the demo: http://sabiod.univ-tln.fr/workspace/IHMM_Whale_demo/. This demo highlights the interest of using the Bayesian non-parametric HMM for unsupervised structuring whale signals. Three examples of the humpback whale song, with 15 seconds duration each, are presented and discussed in this paper (see Figures (5), (6), and (7)).

Figure (5) represents the spectrogram and the corresponding state sequence partition obtained by the HDP-HMM Gibbs inference algorithm, where the selected starting time point, in the whole signal, is 60 seconds. One can see that the state 1 corresponds to the sea noise. Another thing to say is that the state 6 is not present in this time range. This can be also seen in the full length state sequence partition shown in figure (4), where the 6-th unit appears firstly at 234 seconds.

Figure (6) represents the spectrogram and the respective state sequence partition obtained by the HDP-HMM Gibbs inference algorithm, for the signal part starting at 255 seconds, is temporal location close to the middle of the humpback sound recording. The sea noise, which we can see in unit 1, is predominant noise in this time step. The song unit 2, 3 and 4 song unit can be also seen in this song time range.

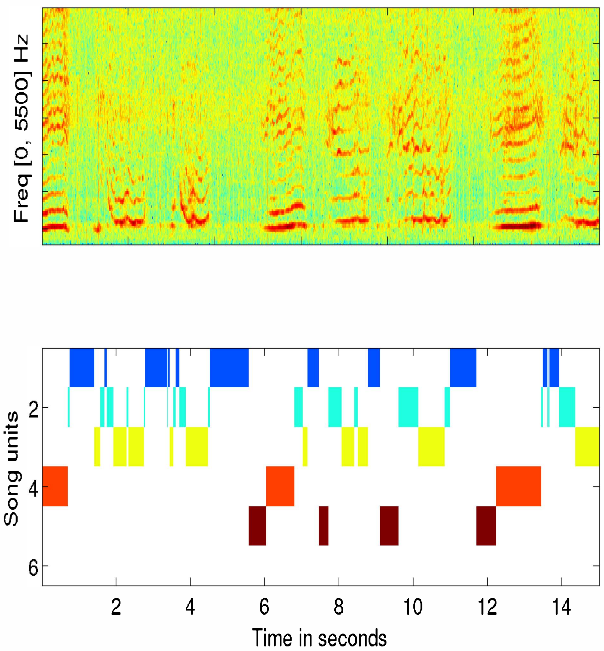

Fig. 5. The spectrogram of the whale song (top), starting with 60 seconds and the obtained state sequences (bottom) by the Gibbs sampling inference approach for the HDP-HMM.

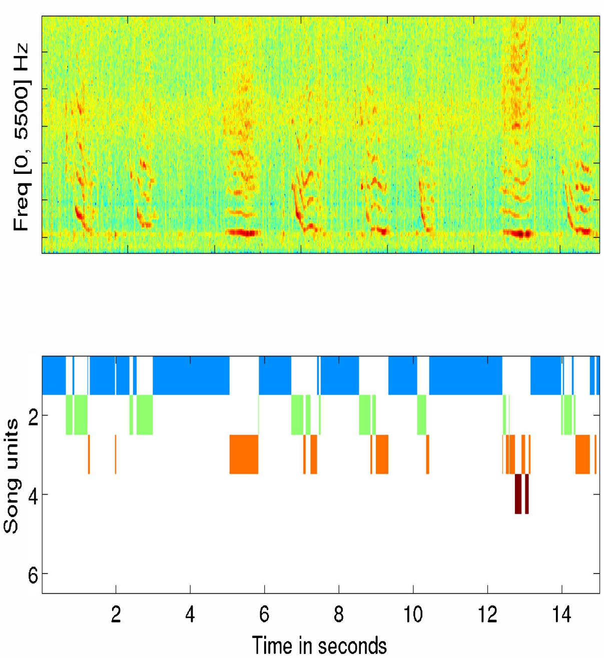

Fig. 6. The spectrogram of the whale song (top), starting with 255 seconds and the obtained state sequences (bottom) by the Gibbs sampling inference approach for the HDP-HMM.

Figure (7) represents the spectrogram and the respective state sequences obtained by the HDP-HMM Gibbs inference algorithm, for a starting point at 495 seconds, which is close to the end of the humpback sound recording. In this time range the 6-th sound unit is the predominant one. Moreover, the sound unit 1 remains the sea noise.

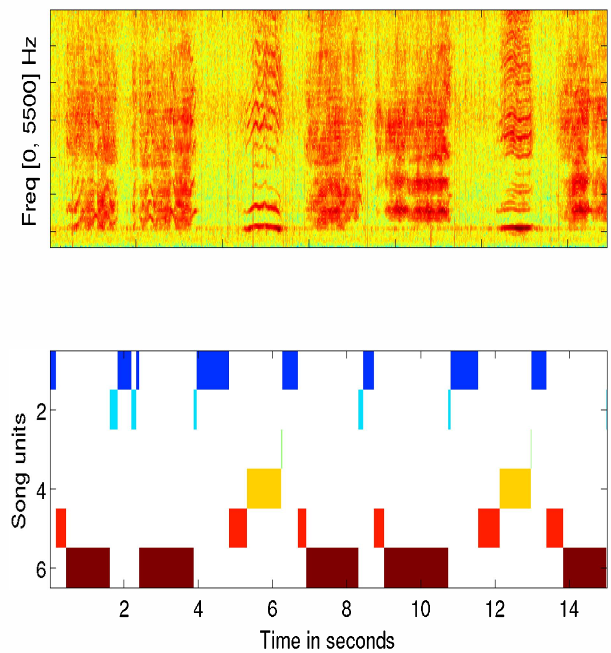

Fig. 7. The spectrogram of the whale song (top), starting with 495 seconds and the obtained state sequences (bottom) by the Gibbs sampling inference approach for the HDP-HMM.

All the obtained state sequences partitions fit very well the spectral patterns. We note that the estimated state 1 is the silence. The state 2 fits the up and down sweeps. State 3 fits low and high fundamental harmonics sound units, the fourth state fits for numerous harmonics sound. The fifth state is the silence, generally continued by some another sound unit, this can be due to the fact that there where not a sufficient number of Gibbs samplings. For a longer learning the fifth state should be merged with the first state. Finally, the state 6 is a very well separated song unit that is a very noisy and broad sound. The analysis is discriminative on the structure.

With this method we can evaluate the representation of units or sequence of units, for example, we compute the optimal local alignment using the Smith-Waterman algorithm. Thus for each frame of 15 sec and each position in the complete signal we get the alignment score (with the following costs : alignment=4, mutation=-3, insertion=-1, deletion=-1). The results demonstrate that several sequences of few seconds are regularly repeated in the complete sequence, as depicted in the literature [23]. We see in the figure 8 that the temporal pattern of the third frame of 15 seconds (available on line [2] and the corresponding wav file [3]) is regularly repeated in the complete sequence at times index 50, 150, 290, 410 and 460 (i.e. 0.8, 2.5, 4.8, 6.8, 7.6 minutes). This

[2]http://sabiod.univ-tln.fr/workspace/IHMM_Whale_demo/GIBBS_seg3IJCNNdemo.png

[3]http://sabiod.univ-tln.fr/workspace/IHMM_Whale_demo/GIBBS_seg3IJCNNdemo.wav

result is consistent with the alignment analysis processed on humpback songs recorded in Pacific Ocean [24]. The interest of our approach is to propose here a complete automatic analysis, whereas usual sequence decomposition is using a priori information. Precise and objective sequence comparisons are mandatory because they reveal information on the whale origin [24].

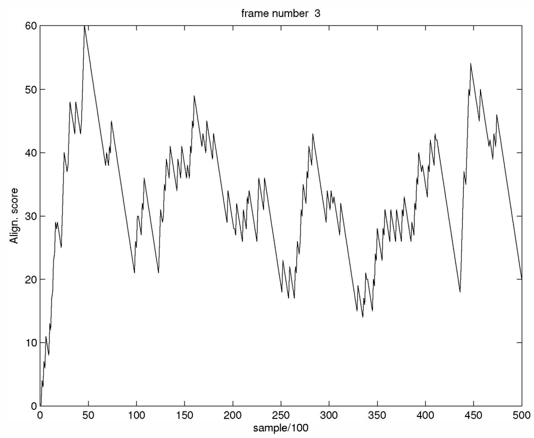

Fig. 8. The optimal local alignment score of the third frame of 15 seconds on the complete song of 8 minutes (abscissa is in feature samples / 100). We see clearly a modulation that demonstrates a repetition of the frame pattern.

The HDP-HMM method was compared to the Sticky HDP-HMM that include a parameter for self transition bias, and places a prior over this parameter, thus the expected probability of self transition is increased by an amount proportional to a value $\kappa > 0$. The Sticky HDP-HMM (with a small $\kappa = 0.1$) was investigated over the same data set (with the same features). The Gibbs also runs for 30000 samplings. The resulting state sequences partition, for all 8.6 minutes of humpback whale song data, obtained by the Stick HDP-HMM is illustrated in figure (9). One can see that the number of states estimated by the Stick HDP-HMM Gibbs sampling is 8, however the states 1,2,3 and 6 have a small number of observations assigned (respectively 33, 72, 50 and 57), thus this states could be merged with other states for a longer Gibbs sampler.

An example of the obtained result is illustrated on figure 10 that is the sound spectrogram equivalent to illustrated figure 6. We see that state 5 represents the sea noise, the other states are the whale song units.

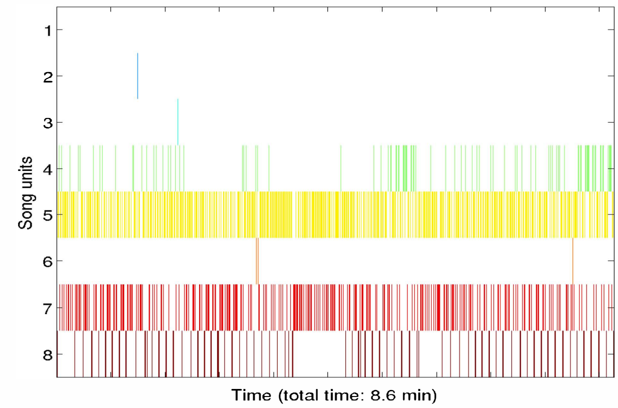

Fig. 9. The state sequences obtained by the Gibbs sampling inference approach for Stick HDP-HMM.

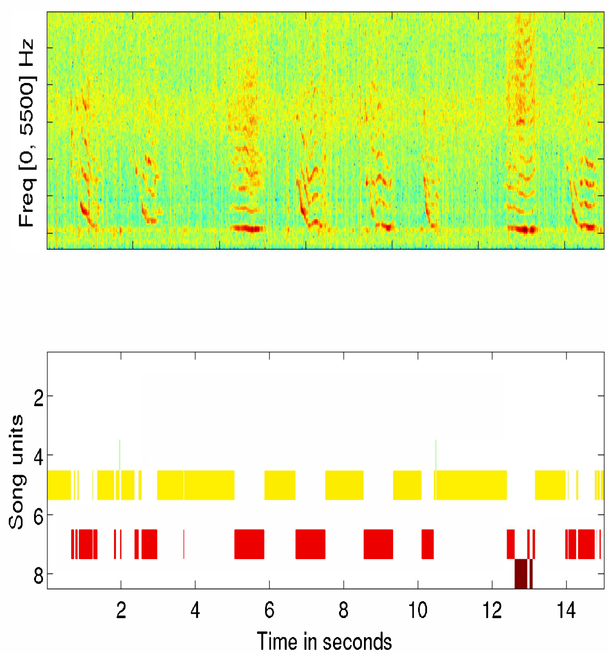

Fig. 10. The spectrogram of the whale song (top), starting with 255 seconds and the obtained state sequences (bottom) by the Gibbs sampling inference approach for the Stick-HDP-HMM.

V. Conclusion and Future works

We investigated the Hierarchical Dirichlet Process for Hidden Markov Model in a challenging problem of unsupervised learning from complex bioacoustic data. The Gibbs sampling algorithm of the HDP-HMM was applied to this real world data. The possible hidden whale song units of the humpback whale signals were accurately recovered in fully automatic way. This result would be very interesting for biopopulation studies such as in [24]. The results highlight the interest of using the Bayesian non-parametric approach for the Hidden Markov Model. In a future work, we will propose the possibility to make the eigenvalue decomposition for the covariance matrix for the emission density of the HMM, more flexible models could appear in term of different volumes, orientations and shapes for each states.

Acknowledgments

This work was supported by the Scaled Acoustic Biodiversity project (SABIOD), MASTODONS MI CNRS http://sabiod.org. We thank Dr. J. Razik for providing the MFCC and Pr. Gianni PAVAN from CIBRA/University of Pavia (It) for his comments on the results regarding the real whale song data.

REFERENCES

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of The Royal Statistical Society, B*, vol. 39(1), pp. 1–38, 1977.

[3] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[4] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[6] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[8] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden markov model," in *Machine Learning*. MIT Press, 2002, pp. 29–245.

[9] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *ICML 2008: Proceedings of the 25th international conference on Machine learning*. New York, NY, USA: ACM, 2008, pp. 312–319.

[10] J. Van Gael, Y. Saatci, Y. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," in *Proceedings of the 25th international conference on Machine learning*. ACM New York, NY, USA, 2008, pp. 1088–1095.

[11] E. Fox, "Bayesian nonparametric learning of complex dynamical phenomena," PhD. Thesis, MIT, Cambridge, MA, 2009.

[12] C. P. Robert, *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[13] N. Hjort, H. C., P. Muller, and S. G. Waller, *Bayesian Non Parametrics*. Cambrige University Press, 2010.

[14] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[15] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probab. Theory Related Fields*, vol. 102, no. 2, pp. 145–158, 1995.

[16] J. G. Samuel and D. M. Blei, "A tutorial on bayesian non-parametric model," *Journal of Mathematical Psychology*, vol. 56, pp. 1–12, 2012.

[17] F. Wood, T. L. Griffiths, and Z. Ghahramani, "A non-parametric bayesian method for inferring hidden causes." in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

[18] Y. W. Teh and M. Jordan, *Hierarchical Bayesian Nonparametric Models with Applications*. Cambridge, UK: Cambridge University Press, 2010.

[19] J. Sethurman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4(2), pp. 639–650, 1994.

[20] H. S. S. Andrew Gelman, John B. Carlin and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

[21] F. Wood and M. J. Black, "A nonparametric bayesian alternative to spike sorting," *Journal of neuroscience methods*, vol. 173, no. 1, pp. 1–12, 2008.

[22] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[23] F. Pace, F. Benard, H. Glotin, O. Adam, and P. White, "Subunit definition and analysis for humpback whale call classification," *Applied Acoustics*, vol. 71, no. 11, pp. 1107 – 1112, 2010.

[24] E. C. Garland, A. W. Goldizen, M. L. Rekdahl, R. Constantine, C. Garrigue, N. D. Hauser, M. M. Poole, J. Robbins, and M. J. Noad, "Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale," *Current Biology*, vol. 21, no. 8, pp. 687 – 691, 2011.