BIOACOUSTIC SEGMENTATION BY HIERARCHICAL DIRICHLET PROCESS HIDDEN MARKOV MODEL

Vincent Roger
DYNI, LSIS UMR CNRS, Machine Learning
AMU, University of Toulon, ENSAM
La Garde, France
vincent-roger@etud.univ-tln.fr

Marius Bartcus
DYNI, LSIS UMR CNRS, Machine Learning
AMU, University of Toulon, ENSAM
La Garde, France
marius.bartcus@gmail.com

Faicel Chamroukhi
LMNO UMR CNRS, Statistics and Data Science
University of Caen
Caen, France
faicel.chamroukhi@unicaen.fr

Hervé Glotin
DYNI, LSIS UMR CNRS, Machine Learning
AMU, University of Toulon, ENSAM
La Garde, France
glotin@univ-tln.fr

ABSTRACT

Understanding the communication between different animals by analysing their acoustic signals is an important topic in bioacoustics. It can be a powerful tool for the preservation of ecological diversity. We investigate probabilistic models to analyse signals issued from real-world bioacoustic sound scenes. We study a Bayesian non-parametric sequential models based on Hierarchical Dirichlet Process Hidden Markov Models (HDP-HMM). The model is able to infer hidden states, that are referred here as song units. However, using such a model raise one main issue: defining the number of hidden states the model has to learn. In bioacoustic problems we often do not know the number of song units (unlike in human speech recognition). Hence, we work with the Hierarchical Dirichlet Process (HDP)-HMM, which is a Bayesian non-parametric (BNP) model that offers a way to tackle this challenging problem. We focus our work on unsupervised learning from bioacoustic data. It consists in simultaneously finding the structure of hidden song units and automatically infer the unknown number of the hidden states to represent the data. Two real bioacoustic sound scene applications are investigated in this work: on whale and multi-species birds segmentation. The learning of these models is proceeded by using Markov-Chain Monte Carlo (MCMC) sampling techniques on Mel Frequency Cepstral Coefficients (MFCC) of audio signals. The results show an interesting song unit segmentation of the bioacoustic signals and open new insights for unsupervised analysis of such signals. This paper illustrates the potential of chunking non-human animal signals into structured parts. This can yield to a new species representation and help experts to better understand the behaviour of such species as Kershenbaum et al. (2014) wanted.

1 INTRODUCTION

Acoustic communication is common in the animal world where individuals communicate with sequences of some different acoustic elements (Kershenbaum et al., 2014). An accurate analysis is important in order to give a better identification of some animal species and interpret the identified song units in the course of time. In this paper, we automatically model the sequence of a non-human signal and determine their acoustic song units. As highlighted in Kershenbaum et al. (2014), the way according to which non-human acoustic sequences can be interpreted can be summarized as shown in Fig 4. We distinguish four common properties that are used to define potential criteria for segmenting such signals into song units. The first way, shown in Fig 4(A), consists in separating the signals using silent gaps. The second way, shown in Fig 4(B), consists in separating the signals according to the changes in the acoustic properties in the signal. The third way, shown in Fig 4(C) consists in grouping similar sounds separated with silent gaps as a single unit. The last common

way, shown in Fig 4(D) consists in separating signal in organized sound structure, considered as fundamental units.

Acoustic units can be determined either manually (e.g. from spectrogram representation), or automatically (e.g. based on a model). Manual segmentation is time consuming and not possible for a large acoustic dataset. That is why automatic approaches are needed. Furthermore, in bioacoustic signals, the problem of segmenting signals of many species, is still an issue, including for bioacoustic. Hence, a well-principled learning system based on unsupervised approach can help them to have a better understanding of bioacoustics species. In this context, we investigate statistical latent data models to automatically identify song units.

First, we study Hidden Markov Models (HMMs) (Rabiner & Juang, 1986). Which are the gold standard for sequential data, and thus could be relevant for acoustic data modeling and segmentation. The typically used algorithm to learn the model is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), also known as Baum-Welch in HMMs (Baum et al., 1970). The main issue with HMMs is the one of selecting the number of hidden states. Because of the lack of knowledge on non-human species, it is hard to have this number. This rises a model selection problem, which can be addressed by information selection criteria such as BIC, AIC (Schwarz, 1978; Akaike, 1974), which select an HMM with a number of states from pre-estimated HMMs with varying number of states.

Such approaches are limited because they require learning multiple HMMs. On the other hand, non-parametric derivations of HMMs constitute a well-principled alternative to address this issue. This approach is more flexible than using a Bayesian non-parametric (BNP) formulation for HMMs (Teh et al., 2006), also called the infinite HMM (iHMM) (Beal et al., 2002). It allows to infer the number of states (segments, units) from the data. The BNP approach for HMMs relies on Hierarchical Dirichlet Process (HDP) to define a prior over the states (Teh et al., 2006). It is known as the Hierarchical Dirichlet Process for the Hidden Markov Models (HDP-HMM) (Teh et al., 2006). The HDP-HMM parameters can be estimated by MCMC sampling techniques such as Gibbs sampling. The standard HDP-HMM Gibbs sampling has the limitation of an inadequate modeling of the temporal persistence of states (Fox et al., 2008). This problem has been addressed by Fox et al. (2008) by relying on a sticky extension which allows a more robust learning. Hence, we have a model to separate non-human signals into states that represent different activities (song units) and exploring the inference of complex data such as bioacoustic data in surroundings cases is not yet resolved.

In this paper, we investigate the BNP formulation of HMM, that is the HDP-HMM, into two challenges involving real bioacoustic data. First, a challenging problem of humpback whale song decomposition is investigated. The objective is the unsupervised structuration of whale bioacoustic data. Humpback whale songs are long cyclical sequences produced by males during the reproduction season which follows their migration from high-latitude to low-latitude waters. Singers from the same geographical region share parts of the same song. This leads to the idea of dialect (Helweg et al., 1998). Different hypotheses of these songs were emitted (Medrano et al., 1994; Frankel et al., 1995; Baker & Herman, 1984; Garland et al., 2011). Next, we investigate a challenging problem of bird song unit structuration. Catchpole & Slater (1995); Kroodsma & Miller (1996) show how birds sing and why birds have such elaborate songs. However, analysing bird song units is difficult due to the transientness of typical bird chirps, the large behavioural intra-class variability, the small amount of examples per class, the presence of wildlife noise, and so forth. As shown later in the obtained segmentation results, such automatic approaches allow large-scale analysis of environmental bioacoustics recordings

1.1 Related work

Discovering the call units (which can be considered as a kind of non-human alphabet) of such complex signals can be seen as a problem of unsupervised call units classification as Pace et al. (2010).

Picot et al. (2008) also tried to analyse bioacoustic songs using a clustering approach. They implemented a segmentation algorithm based on Payne's principle to extract sound units from a bioacoustic song. In this paper we reformulate the problem of song decomposition as an unsupervised data classification problem. Contrary to the approach used by Pace et al. (2010), in which the number of states (call units in this case) has been fixed manually, or the one used by Picot et al. (2008), where

a K-means algorithm is used for automatic classification and then automatically define the optimal number of classes by maximizing the Davies Bouldin criterion.

Our approach is based on a probabilist approach on the MFCC; it is a non-parametric formulation, that is well-suited to the problem of automatically inferring the number of the states in the data. In the next section we describe the real-world bioacoustic challenges we used and explain our approach.

2 DATA AND METHODS

The data used represent the difficulties of bioacoustic problems, especially when the only information linked to the signal is the species name. Thus, we have to determine a sequence without ground truth.

2.1 HUMPBACK WHALE DATA

Humpback whale song data consist of a recording (about 8.6 minutes) produced at few meters from the whale in La Reunion - Indian Ocean (NIP, 2013), at a frequency sample of 44.1kHz, 32 bits, one channel.

We extract MFCC[1] features from the signal, with pre-emphasis: 0.95, hamming window, FFT on 1024 points (nearly 23ms), frameshift 10 ms, 24 Mel channels, 12 MFCC coefficients plus energy and their delta and acceleration, for a total of 39 dimensions as detailed in the NIPS 2013 challenge (NIP, 2013) where the signal and the features are available. The retained data for our experiment are the 51336 first observations.

2.2 MULTI-SPECIES BIRD DATA

Bird species song data from Fernand Deroussen Jerome Sueur of Musee National d'Histoire Naturelle (F. Deroussen, 2006), consists of a training and a testing set (not used here). Theses sets were designed for the ICML4B challenge[2]. The recordings have a frequency sample of 44.1kHz, 16 bits, one channel. The training set is composed of 35 recordings, 30 seconds each taken from 1 microphone. Each record contains 1 bird species in the foreground for a total of 35 different birds species.

The feature extraction for this application is applied as follows. First, a high pass filter is processed to reduce the noise (set at 1.000 kHz to avoid noises). Then, we extract the MFCC features with windows of 0.06 seconds and shift of 0.03 seconds, we keep 13 coefficients, with energy as first parameter, to be compact and sufficient accurate, considering only the vocal track information and removing the source information. Also, we focus on frequencies below 8.000 kHz, because of the alterations into the spectrum. We obtain 34965 observations with 13 dimensions each for train set, that is used to learn our model.

2.3 METHOD: BIOACOUSTIC UNSUPERVISED LEARNING FOR SIGNAL REPRESENTATION

To solve bioacoustic problems and finding the number of call units we propose to use the HDP-HMM model to model the complex bioacoustic data. Our approach automatically discovers and infers the number of states from the non-human song data.

In this paper we present two applications on bioacoustic data. We study the song unit structuration, for the humpback whale and for the multi-species birds signal.

In the next section we give a brief description of the Hidden Markov Model and it's Bayesian non-parametric alternative used in our bioacoustic signal representation applications.

[1]The MFCC are features that represent and compress short-term power spectrum of a sound. It follows the Mel scale.

[2]http://sabiod.univ-tln.fr/nips4b/challenge2.html

3 BAYESIAN NON-PARAMETRIC ALTERNATIVE FOR HIDDEN MARKOV MODEL

The finite Hidden Markov Model (HMM) is very popular due to its rich mathematical structure and its stability to model sequential data (e.g. acoustic data). It assumes that the observed sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ is governed by a hidden state sequence $\mathbf{z} = (z_1, \ldots, z_T)$, where $\mathbf{x}_t \in \mathbb{R}^d$ is the multidimensional observation at time t and z_t represents the hidden state of $\mathbf{x}_t$ taking values in a finite set $\{1, \ldots, K\}$, K being the possible number of states, that is unknown. The generative process of the HMM can be described in general by the following steps. First, z_1 follows the initial distribution π_1. Then, given the previous state (z_{t-1}), the current state z_t follows the transition distribution. Finally, given the state z_t, the observation $\mathbf{x}_t$ follows the emission distribution $F(\boldsymbol{\theta}_{z_t})$ of that state. The HMM parameters, that are the initial state transition (π_1), the transition matrix ($\boldsymbol{\pi}$), and the emission parameters ($\boldsymbol{\theta}$) are in general estimated in a maximum likelihood estimation (MLE) framework by using the Expectation-Maximization (EM) algorithm, also known as the Bauch-Welch algorithm (Baum et al., 1970) in the context of HMMs.

Therefore, for the finite HMM, the number of states K is required to be known a priori. This model selection issue can be addressed in a two-stage scheme by using model selection criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Akaike Information Criterion (AIC) (Akaike, 1974), the Integrated Classification Likelihood criterion (ICL) (Biernacki et al., 2000), etc to select a model from pre-estimated HMMs with varying number of states. Such approaches is limited, it requires learning N HMMs, N being sufficiently high to have an equivalent of a non parametric approach. Regardless this, a non parametric approach is more efficient because it theoretically tends to an infinite number of states. Thus, we use a Bayesian non-parametric (BNP) version of the HMM, that is able to infer the number of hidden states from the data. It is more flexible than learning multiple HMM, because in bio-acoustic problems the model have to characterize multiple species/individuals, thus it possibly tends to a large number of hidden states. Thence, exploring the inference of complex data such as bioacoustic data in surroundings cases is new.

The BNP approach for the HMM, that is the infinite HMM (iHMM), is based on a Dirichlet Process (DP) (Ferguson, 1973) needs to be used. However, due to the transitions of states take independent priors, there is no coupling across transitions between different states Beal et al. (2002), therefore the DP is not sufficient to extend the HMM to an infinite model. The Hierarchical Dirichlet Process (HDP) prior distribution on the transition matrices over countability infinite state space, derived by Teh et al. (2006), extends the HMM to the infinite state space model and is briefly described in the next subsection.

3.1 HIERARCHICAL DIRICHLET PROCESS (HDP)

Suppose the data subdivided into J groups, each produced by a related, yet distinct process. The HDP extends the DP by an hierarchical Bayesian approach such that a global Dirichlet Process prior $\text{DP}(\alpha_0, G_0)$ is drawn from a global prior G_j, where G_0 is itself a Dirichlet Process distribution with two parameters, a base distribution H and a concentration parameter γ. The generative process of the data with the HDP can be summarized as follows. Suppose data $\mathbf{X}$, with $i = 1, \ldots, T$ observations that is grouped into $j = 1, \ldots, J$ groups. Note that the observations of the group j are given by $\mathbf{X}_j = (\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots)$, all observations of group j being exchangeable. Assume each observation is drawn from a mixture model, thus each observations $\mathbf{x}_{ji}$ is associated with a mixture component, with parameter θ_{ij} . Note that from the DP property, we observe equal values in the components θ_{ij}. Now, giving the model parameter θ_{ji}, the data $\mathbf{x}_{ji}$ is drawn from the distribution $F(\theta_{ji})$. Assuming a prior distribution G_j over the model parameters associated for group j, $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots)$, we can define the generative process in Eq. (1).

$$
\begin{aligned}
G_0|\gamma, H &\sim \text{DP}(\gamma, H), \\
G_j|\alpha_0, G_0 &\sim \text{DP}(\alpha_0, G_0),\ \forall j \in 1, \ldots, J, \\
\theta_{ji}|G_j &\sim G_j,\ \forall j \in 1, \ldots, J \text{ and } \forall i \in 1, \ldots, T, \\
\mathbf{x}_i|\theta_{ji} &\sim F(\mathbf{x}_i|\theta_{ji}), \forall j \in 1, \ldots, J \text{ and } \forall i \in 1, \ldots, T.
\end{aligned}
\tag{1}
$$

The Chinese Restaurant Process (CRP) (Pitman, 1995) is a representation of the Dirichlet Process that results from a metaphor related to the existence of a restaurant with possible infinite tables

(clusters) where customers (the observations) are sitting in it. An alternative of such a representation for the Hierarchical Dirichlet Process can be described by the Chinese Restaurant Franchise (CRF) process by extending the CRP to multiple restaurants that share a set of dishes.

The idea of CRF is that it gives a representation for the HDP by extending the Chinese Restaurant Process 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 with 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.

3.2 The hierarchical Dirichlet process for the hidden Markov model (HDP-HMM)

The HDP-HMM uses a HDP prior distribution providing a potential countability infinite number of hidden states and tackles the challenging problem of model selection for the HMM. This model is a Bayesian non-parametric extension for the HMM also presented as the infinite Hidden Markov Model (Beal et al., 2002). To derive the HDP-HMM model we suppose a doubly-infinite transition matrix, where each row corresponds to a CRP. Thus, in a HDP formalism, the groups correspond to states, with CRP distribution on next states. CRF links these states distributions.

We assume for simplicity a distinguished initial state z_0. Let G_j 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:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \text{GEM}(\gamma), \\ \boldsymbol{\pi}_k|\alpha, \boldsymbol{\beta} &\sim \text{DP}(\alpha, \boldsymbol{\beta}), \\ z_t|z_{t-1} &\sim \text{Mult}(\boldsymbol{\pi}_{z_{t-1}}), \\ \boldsymbol{\theta}_k|H &\sim H, \\ \mathbf{x}_t|z_t, \{\boldsymbol{\theta}_k\}_{k=1}^{\infty} &\sim F(\boldsymbol{\theta}_{z_t}). \end{aligned} \tag{2}$$

where,
$\boldsymbol{\beta}$ is a hyperparameter for the DP (Sethuraman, 1994) distributed according to the stick-breaking construction noted GEM$(.)$;
z_t is the indicator variable of the HDP-HMM that follows a multinomial distribution Mult$(.)$;
the emission parameters $\boldsymbol{\theta}_k$, are drawn independently, according to a conjugate prior distribution H;
$F(\boldsymbol{\theta}_{z_t})$ is a data likelihood density with the unique parameter space of $\boldsymbol{\theta}_{z_t}$ 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 Gelman et al. (2003), 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 components: μ_0 is the mean of Gaussian 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 Eq. (2), $\boldsymbol{\pi}$ is interpreted as a double-infinite transition matrix with each row taking a CRP. Thus, in the HDP formulation "the group-specific" distribution, $\boldsymbol{\pi}_k$ corresponds to "the state-specific" transition where the CRF defines distributions over the next state. In turn, Fox et al. (2008) showed that HDP-HMM inadequately models the temporal persistence of states, creating redundant and rapidly switching states and proposed an additional hyperparameter κ that increase the self-transition probabilities. This is named as sticky HDP-HMM. The distribution on the transition matrix of Eq. (2) for the sticky HDP-HMM is given as follows:

$$\boldsymbol{\pi}_k|\alpha, \boldsymbol{\beta} \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_k}{\alpha + \kappa}\right), \tag{3}$$

where a small positive $\kappa > 0$ is added to the k^{th} component of $\alpha\boldsymbol{\beta}$, thus of self-transition probability is increased by κ. Note that setting κ to 0, the original HDP-HMM is recovered. Under such assumption for the transition matrix, Fox et al. (2008) proposes an extension of the CRF to the

Chinese Restaurant Franchise with Loyal Customers. A graphical representation of (sticky) HDP-HMM is given in Fig 5.

The inference of the infinite HMM (the (sticky) HDP-HMM) with the Block Gibbs sampler algorithm is given in Algorithm 3 of Supplementary Material in Fox et al. (2008) paper. The base idea of this sampler is to estimate the posterior distributions over all the parameters from the generative process of (sticky) HDP-HMM given in Eq. (2). Here, the CRF with Loyal Customers, hyperparameter κ of the transition matrix can be sampled in order to increase the self-transition probability.

Hence, the HDP-HMM model resolves the problem of advanced signal decomposition using acoustic features with respect to time. It allows identifying song units (states), behaviour and enhancing populations studies. From the other point, modelling data with the HDP-HMM offers a great alternative of the standard HMM to tackle the challenging problem of selecting the number of states, identifying the unknown number of hidden units from the used features (here: MFCC). The experimental results show the interest of a such approach.

4 EXPERIMENTS

In this section we present two applications on bioacoustic data. We study the song unit structuration, for the humpback whale signal and for multi-species birds signals.

4.1 HUMPBACK WHALE SOUND SEGMENTATION

The learning of the humpback whale song, applied via the HDP-HMM, is done with the Blocked Gibbs sampling. A number of iterations was fixed to $N_s = 30000$ and a truncation level, that corresponds to the maximum number of possible states in the model (being sufficient big to approximate it to an infinite model), is fixed to $L_k = 30$. The number of states estimated by the HDP-HMM Gibbs sampling is 6.

The Fig 6 shows the state sequences partition, for all 8.6 minutes of humpback whale song data, obtained by the HDP-HMM Gibbs sampling. For more detailed information, the result of the whole humpback whale signal segmentation is separated by several parts of 15 seconds. All the spectrograms of the humpback whale song and the obtained segmentation are made available in the demo: `http://sabiod.univ-tln.fr/workspace/ICLR2017/whale/`. This demo highlights the interest of using a BNP formulation of HMMs for unsupervised segmentation of whale signals. Three examples of the humpback whale song, with 15 seconds duration each, are presented and discussed in this paper (see Fig 1).

Figure 1 represents the spectrogram and the corresponding state sequence partition obtained by the HDP-HMM Gibbs inference algorithm. They respectively represent examples of the beginning, the middle and the end of the whole signal. All the obtained state sequence partitions fit the spectral patterns. We note that the estimated state 1 fits the sea noise, state 5 also fits sea noise, but it is right before units associated to whale songs. The presence of this unit can be due to an insufficient number of Gibbs samples. For a longer learning the fifth state could be merged with the first state. State 2 fits the up and down sweeps. State 3 fits low and high fundamental harmonic sounds, state 4 fits for numerous harmonics sound and state 6 fits very noisy and broad sounds. Fig 7 shows two spectrograms extracted from the 6^{th} song unit (left) and from the 2^{nd} song unit (right) of the whole humpback whale signal. We can see that the units fit specific patterns on the whole signal.

Pr. Gianni Pavan (Pavia University, Italy), an undersea NATO bioacoustic expert analysed the results on the humpback whale song segmentation, during his stay at DYNI in 2015. He validated the proposed representation. This highlight the interest of learning BNP model on a single species. Next, we will see how such model reacts with multiple species.

4.2 BIRDS SOUND SEGMENTATION

In this section we describe the obtained bird song unit segmentation. We segment the bird signals into song units by learning the HDP-HMM model on the training set (containing 35 different species). The main goal is to see if a such approach can model multiple species. Note that in this set, we assume there is no multiple species singing at the same time.

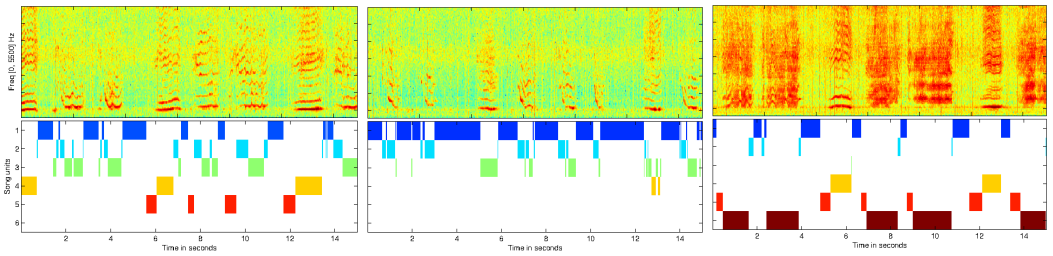

Figure 1: Obtained song units starting at 60 seconds (left), 255 seconds (middle) and 495 seconds (right). The spectrogram of the whale song (top), and the obtained state sequence (bottom) by the Blocked Gibbs sampler inference approach for the HDP-HMM. The silence (unit 1 and 5) looks well separated from the whale signal. Whale up and down sweeps (unit 2), harmonics (unit 3 and 4) and broad sounds (unit 6) are also present. See fig 13 for bigger figures.

For this application, we considered 145000 Gibbs iterations and a truncation level of 200 for the maximum number of states. We suppose them to be sufficiently big for this data problem. Moreover, we use one mixture component per state, that appeared to give satisfactory results and we use a sticky HDP-HMM with the hyper-parameter κ set to 0.1.

We discovered 76 song units with this method. For more detailed information over the signal, we separated the whole train set into parts of 15 seconds each. All the spectrograms and the associated segmentation obtained are made available in the demo: `http://sabiod.univ-tln.fr/workspace/ICLR2017/bird/`. Fig 2 contains three examples of bird bioacoustic segmentation. The species are: *Carduelis chloris* (left), *Luscinia megarhynchos* (middle) and *Parus caeruleus* (right).

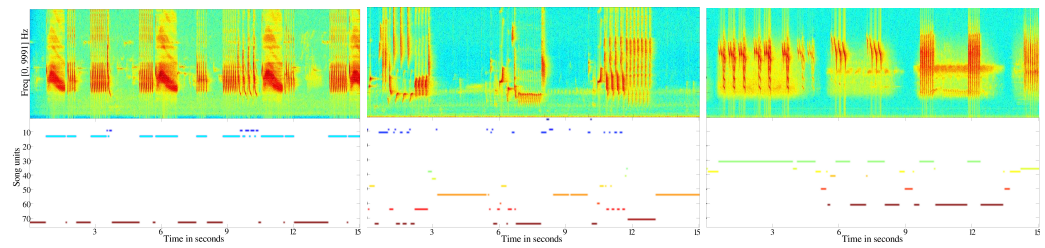

Figure 2: Obtained song units on 15 seconds of *Carduelis chloris* bird song. The spectrogram of the bird song (top), and the obtained state sequence (bottom) by the Blocked Gibbs sampler inference approach for the HDP-HMM. Three song units persist in this sound. The silence looks well separated from the bird signal. Furthermore, we can denote that song units fit well the song and similarity in spectrum persists for each song unit. See fig 14 for bigger figures.

4.2.1 Evaluation of the bird result

To evaluate the bird results, we used a ground truth produced by Simone Clemente (bird expert). We ask him to segment each recording of the dataset according to the different patterns on the signal. Then we compare this ground truth with the segments produce by the model using NMI (Strehl & Ghosh, 2002) which calculates shared information between two clustering sets.

First, we compute the NMI score between these two segmentations and we obtain a score of 0.490. Thus, the global segmentation from the model isn't near from a segmentation done by an expert. Second, we compute the NMI score for each species to see the possible mistake done by the model. Tab 1 shows the different scores obtained with a resulting mean score of 0.367. The highest score is 0.680 (corvus corone) and the lowest score is 0.003 (garrulus glandarius). Thus, for some species, the model has difficulties to segment the data. Sometimes, it uses less states than the expert to segment the data: for the oriolus oriolus (golden oriole), the model identifies 12 song units versus 50 identified by the expert. Nevertheless, the model also uses more states than the expert to segment the data: for the fringilla coelebs (chaffinch), the model identifies 15 song units versus 3 identified

by the expert. In other cases, the model can't differentiate 2 distinct vocalizes if they have close frequencies (phylloscopus collybita and columba palumbus), background and foreground species (streptopelia decaocto). This can be due to the feature used or an insufficient number of iterations of the Gibbs sampling. For most of species, the model and the ground truth have similar patterns observable on Fig 8.

To improve the model, we can investigate better feature representation for species with different acoustic characteristics. We can also improve noise reduction which could be useful for background activities. Nevertheless, the application highlights the interest of using BNP formulation of HMMs for unsupervised segmentation of bird signals.

5 CONCLUSIONS

In this work, we relied on real world bioacoustic applications and propose to use a BNP formulation of HMMs to realize a representation of bio-acoustic signals. It is a response for (Kershenbaum et al., 2014). We investigated this approach on real-world bioacoustic signals from two challenges. The demo for the two applications are available online.

The obtained signal segmentation on the bioacoustic data is recovered in a fully automatic way and is proceeded by a Hierarchical Dirichlet Process for Hidden Markov Model on MFCC. The BNP formulation give an estimate number of cluster needed to segment the signal and our experiments highlight the interest of such formulation on bioacoustic problems. Furthermore we compare the segmentation obtained for birds with the segmentation from an expert and the model using NMI, the results are promising. We describe a full bioacoustic perspective in the annexes.

However, the model used is computational expensive and not suitable larger dataset. Future work will consist in studying methods that should accelerate the MCMC sampling and dealing with larger data problems, like variational inference (Jordan et al., 1999) or stochastic variational inference used for HMMs (Foti et al., 2014).

Future work will consist in considering our segmentation results for classification task, where the goal is to identify the species.

REFERENCES

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

C. Scott Baker and Louis M. Herman. Aggressive behavior between humpback whales (Megaptera novaeangliae) wintering in Hawaiian waters. *Canadian Journal of Zoology*, 62(10):1922–1937, 1984.

L.E. 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*, 41:164–171, 1970.

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pp. 29–245. MIT Press, 2002.

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*, 22(7):719–725, 2000.

C K Catchpole and P J B Slater. *Bird Song - Biological Themes and Variations*. Cambridge University Press, 1995.

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*, 39(1):1–38, 1977.

F. Jiguet F. Deroussen. La sonotheque du Museum: Oiseaux de France. 2006.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364.

Nicholas Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden Markov models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3599–3607. Curran Associates, Inc., 2014.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. An HDP-HMM for systems with state persistence. In *ICML 2008: Proceedings of the 25th international conference on Machine learning*, pp. 312–319, New York, NY, USA, 2008. ACM.

A. S. Frankel, C. W. Clark, L. M. Herman, and C. M. Gabriele. Spatial distribution, habitat utilization, and social interactions of humpback whales, Megaptera novaeangliae, off Hawai'i, determined using acoustic and visual techniques. *Canadian Journal of Zoology*, 73(6):1134–1146, 1995.

Ellen C. Garland, Anne W Goldizen, Melinda L. Rekdahl, Rochelle Constantine, Claire Garrigue, Nan Daeschler Hauser, M. Michael Poole, Jooke Robbins, and Michael J. Noad. Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current Biology*, 21(8):687–691, 2011.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

David A. Helweg, Douglas H. Cato, Peter F. Jenkins, Claire Garrigue, and Robert D. McCauley. Geographic Variation in South Pacific Humpback Whale Songs. *Behaviour*, 135(1):pp. 1–27, 1998.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, November 1999.

Arik Kershenbaum, Daniel T Blumstein, Marie A Roch, Çağlar Akçay, Gregory Backus, Mark A Bee, Kirsten Bohn, Yan Cao, Gerald Carter, Cristiane Cäsar, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 2014.

D.E. Kroodsma and E.H. Miller. *Ecology and evolution of acoustic communication in birds*. Comstock Bks. Comstock Pub., 1996. ISBN 9780801482212.

L. Medrano, M. Salinas, I. Salas, P. Ladrón de Guevara, A. Aguayo, J. Jacobsen, and C. S. Baker. Sex identification of humpback whales, Megaptera novaeangliae, on the wintering grounds of the Mexican Pacific Ocean. *Canadian Journal of Zoology*, 72(10):1771–1774, 1994.

Proc. Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data, USA, 2013. NIPS Int. Conf.

Federica Pace, Frederic Benard, Herve Glotin, Olivier Adam, and Paul White. Subunit definition and analysis for humpback whale call classification. *Applied Acoustics*, 71(11):1107 – 1112, 2010.

G. Picot, O. Adam, M. Bergounioux, H. Glotin, and F.-X. Mayer. Automatic prosodic clustering of humpback whales song. In *New Trends for Environmental Monitoring Using Passive Systems, 2008*, pp. 1–6, Oct 2008.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995. ISSN 0178-8051.

Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

ANNEXES

ACKNOWLEDGEMENTS

We would like to thanks Pr. Gianni Pavan (Pavia University, Italy) and Simone Clemente for their bioacoustic points of view. We also want to thanks Virgil Tassan for his re-reading.

BIOACOUSTICIAN DISCUSSION

One of the main topics in ecological acoustics is the development of unsupervised methods for automatic detection of vocalized species, which would help specialists in ecological works during their monitoring activities. Although some works already have reach good classification percentage [3] [4], there's a lack of methodologies available for works focused on real world data, and with further applications in ecology and wildlife management. One of the major bottlenecks for the application of these methodologies is their inability to work under heavy complex acoustic environment, where different taxa may sing together or conversely, their extreme sensitivity which may result in an "over classification" due to the high degree of variability insight many repertoire of the vocal species.

Our unsupervised method for automatic annotation of bioacoustics sequences seems to overtake these obstacles, by the identification of specie-specific pattern, and seems to be not influenced by the inter-individual variation whiting the song's structure. Furthermore, during the study, we have worked whit recording that may contain more than a species vocalizing, or partial overlapping from different species and specimen, and even in these circumstance our model has shown a great ability of categorisations and generalisations, identifying the main pattern and almost all the other sequences recorded (including silence).

Good examples are provided, from the birds' dataset, by recordings including the common wood pigeon (Columba palumbus; Figure 3 (top)), or from the files containing golden oriole (Oriolus oriolus) as a main species. Their vocalisations are partially overlapped by vocalisation of different species, such as Eurasian blue tit (Cyanistes caeruleus, figure 3 (bottom)) or cricket, and is clear in the analysis, the distinction between the bird vocalising and other animals in the background. Finally, a last example could be taken from the carrion crow (Corvus corone, figure 3 (middle) recordings. Here, multiple specimens are vocalising at the same time, but it does not influence the efficiency of our method.

By the other hand, in presence of species with complex acoustics behaviours, such as the common nightingale (Luscinia megarhynchos) the method could lead to identify as different classes each one of the different phrasing elements in the song. This last approach could be useful in order to perform behavioural analysis focused on the identification of hidden significance in the songs, but may loose his benefit when the main achievement is an ecological topic. We finally consider this method as a promising tool, as we move forward, but further analysis will be needed to build, an efficient tool whit relevant application in the acoustic monitoring and conservation of biodiversity.

[3] Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning; Dan Stowell and Mark D Plumbley; 2014; PeerJ

[4] Automatic classification of a taxon-rich community recorded in the wild; Potamitis and Ilyas; 2014; PloS on

Species	NMI Score
sturnus_vulgaris	0.467
turdus_philomelos	0.398
emberiza_citrinella	0.534
certhia_brachydactyla	0.417
columba_palumbus	0.352
picus_viridis	0.602
anthus_trivialis	0.332
phasianus_colchicus	0.272
cuculus_canorus	0.205
sylvia_atricapilla	0.405
corvus_corone	0.68
phylloscopus_collybita	0.267
streptopelia_decaocto	0.306
turdus_viscivorus	0.417
dendrocopos_major	0.481
erithacus_rubecula	0.394
pavo_cristatus	0.437
fringilla_coelebs	0.565
aegithalos_caudatus	0.202
turdus_merula	0.395
branta_canadensis	0.339
parus_palustris	0.521
sitta_europaea	0.332
alauda_arvensis	0.169
prunella_modularis	0.476
oriolus_oriolus	0.316
carduelis_chloris	0.385
phoenicurus_phoenicurus	0.291
strix_aluco	0.2
parus_caeruleus	0.413
parus_major	0.27
motacilla_alba	0.105
luscinia_megarhynchos	0.497
troglodytes_troglodytes	0.407
garrulus_glandarius	0.003
mean	0.367

Table 1: NMI score for the obtained segmentation using HDP-HMM.

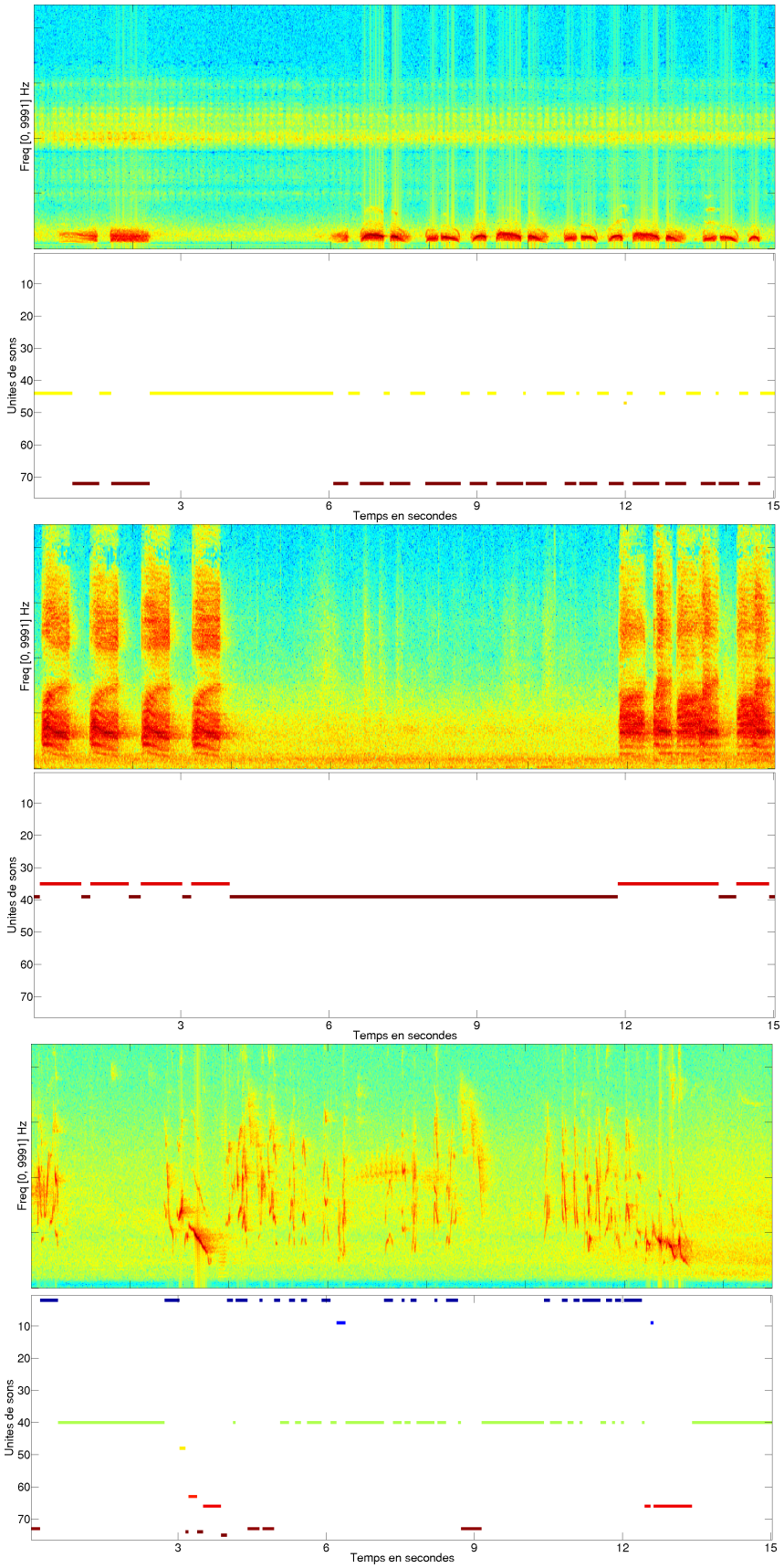

Figure 3: The spectrogram of the bird song and the obtained state sequence by the Blocked Gibbs sampler inference approach for the HDP-HMM. Columba palumbus with cricket noises activities (top); different specimens of Corvus corone (middle); Oriolus oriolus partially overlapped whit Cyanistes caeruleus on the backgound (bottom).

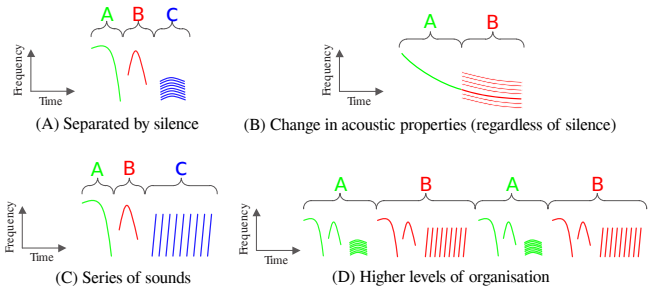

Figure 4: Acoustic common way used to divide spectrogram into units.

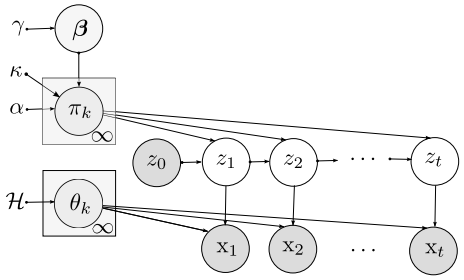

Figure 5: Graphical representation of sticky Hierarchical Dirichlet Process for Hidden Markov Model (HDP-HMM).

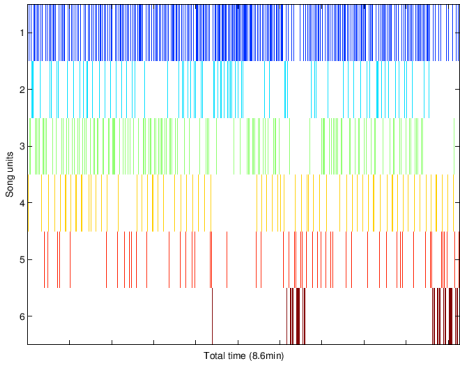

Figure 6: State sequence for 8.6 min of humpback whale song obtained by the Blocked Gibbs sampling inference approach for HDP-HMM.

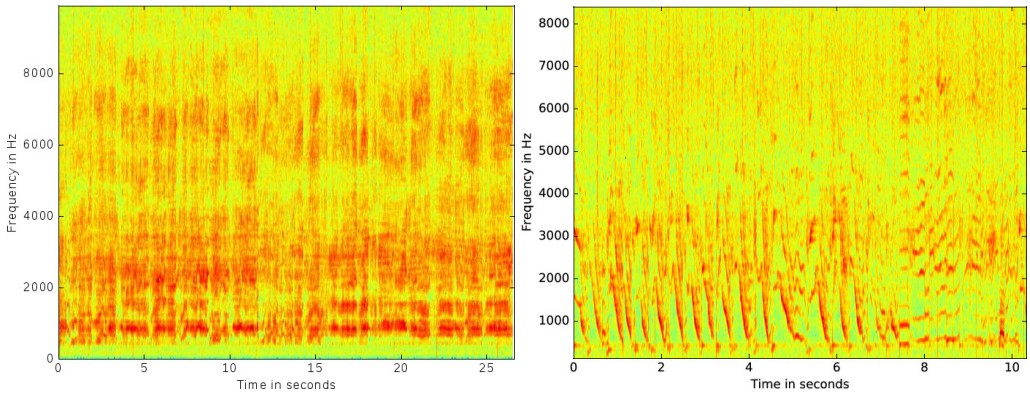

Figure 7: Spectrograms of the 6^{th} whale song unit (left) and 2^{nd} song unit (right).

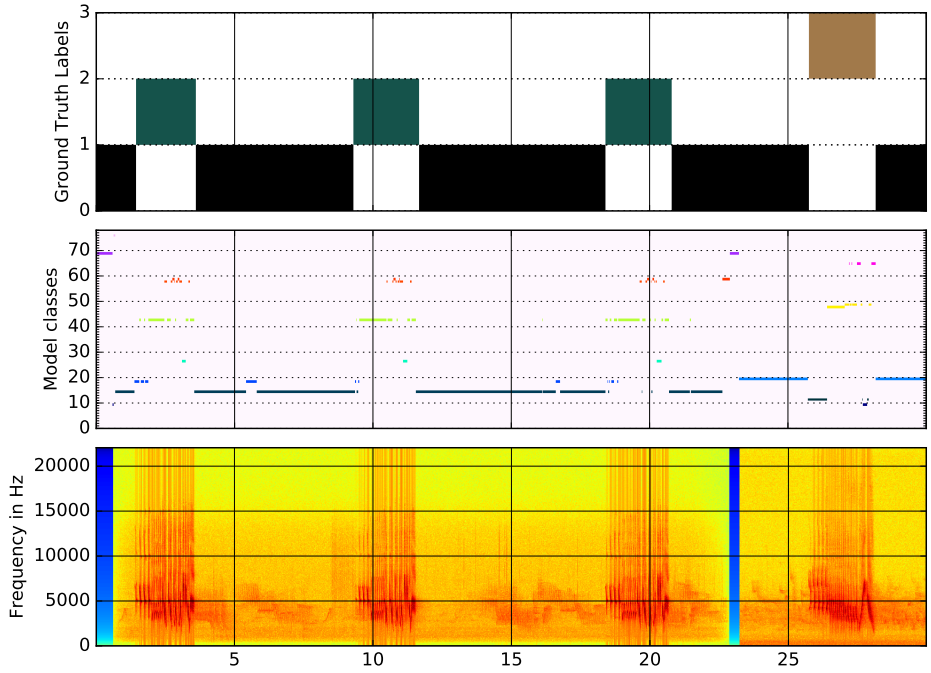

Figure 8: Fringilla coelebs results. The first graph represents the labelled ground truth over 30s where label 0 is always the ”none” label and the other labels are not the same from one graph to an other. The second graph represents our model with the 76 classes. The last one is the spectrogram.

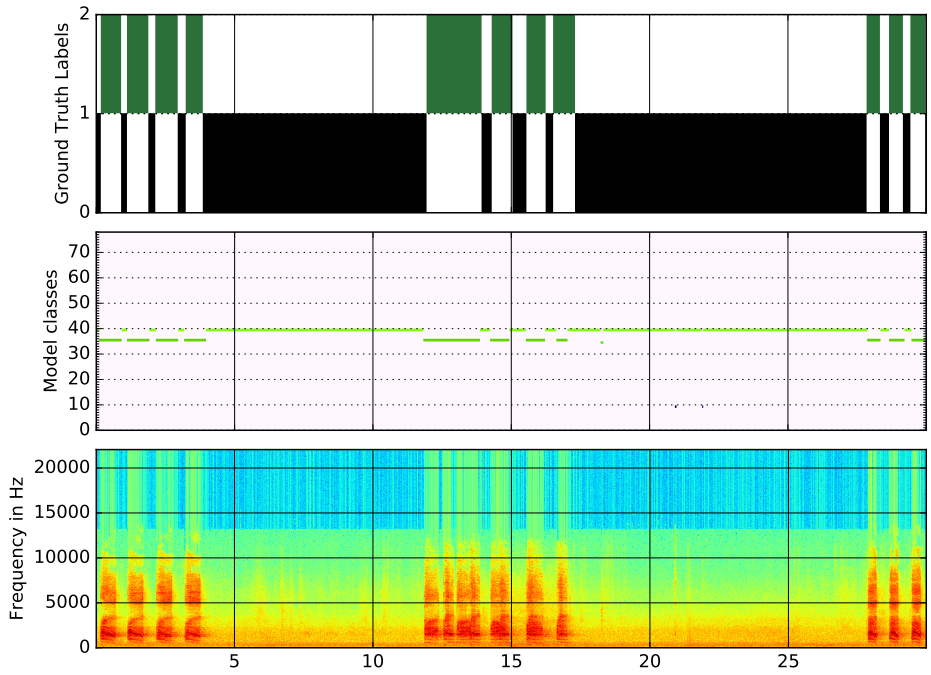

Figure 9: Corvus corone results. See caption of 8 for details.

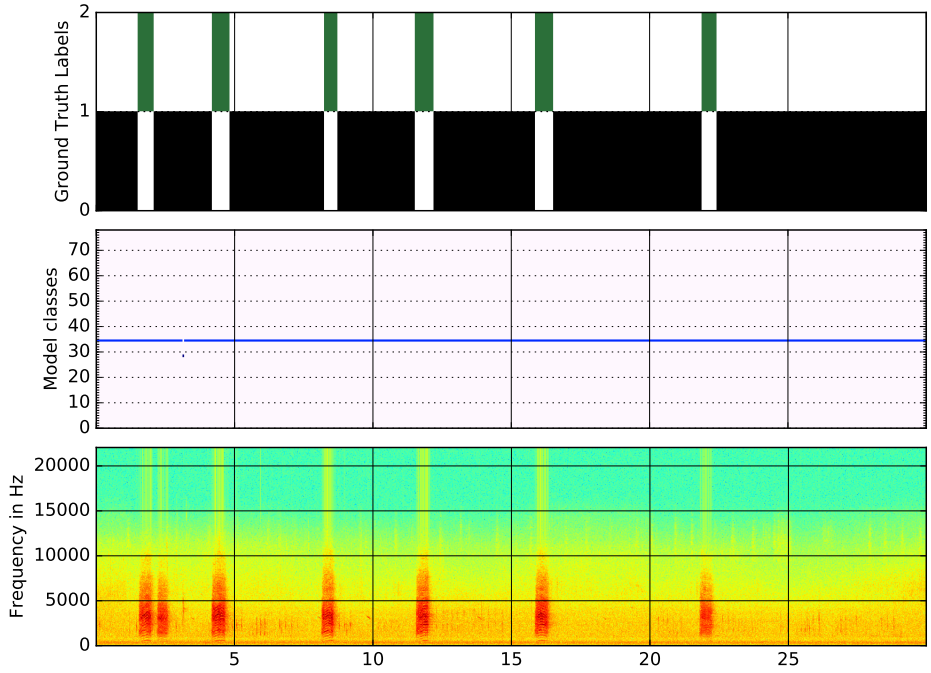

Figure 10: Garrulus glandarius results. See caption of 8 for details.

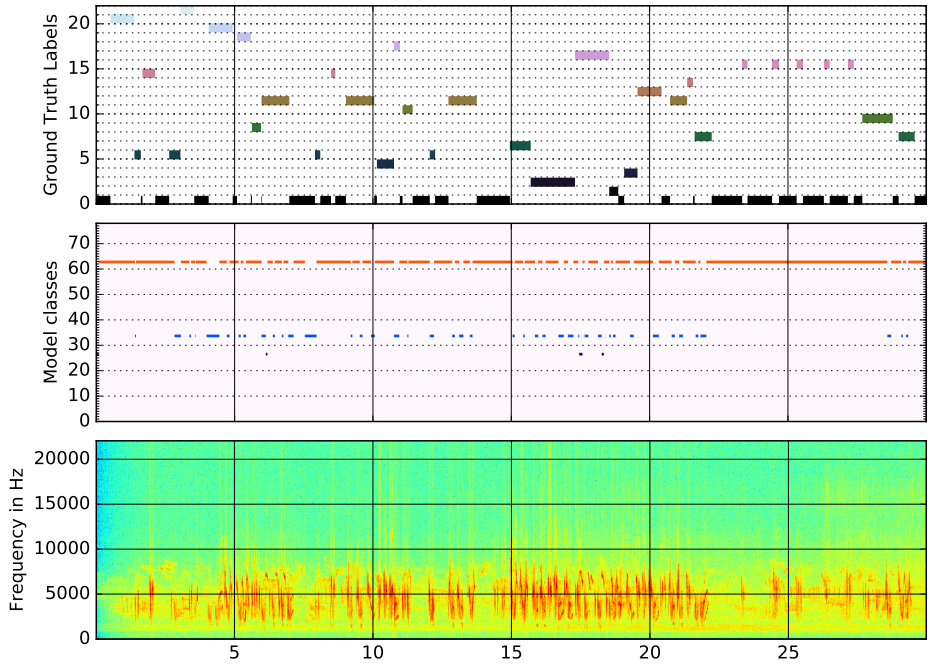

Figure 11: Motacilla alba results. See caption of 8 for details.

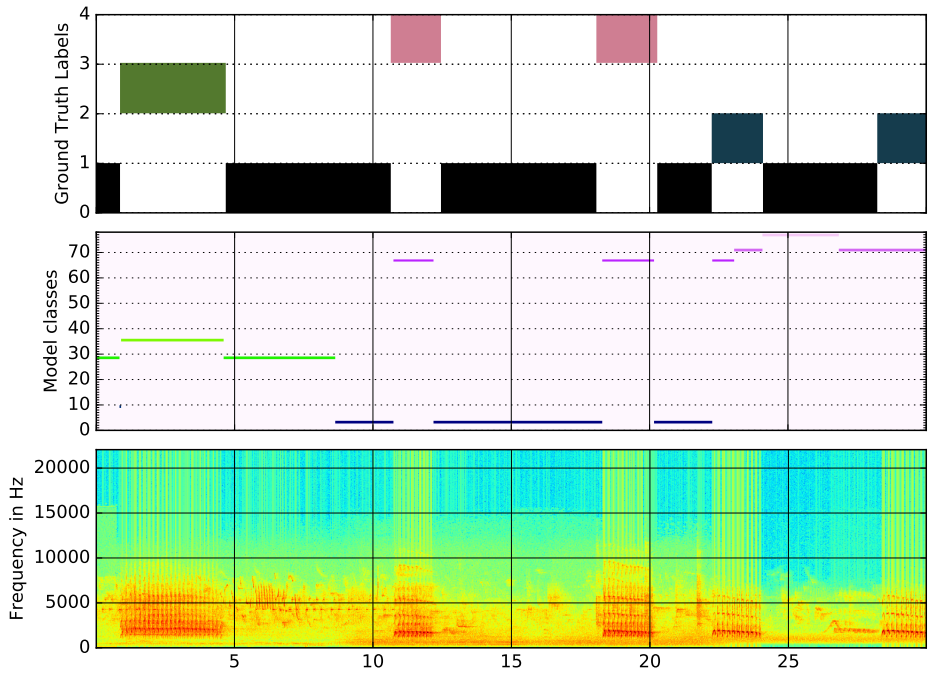

Figure 12: Picus viridis results. See caption of 8 for details.

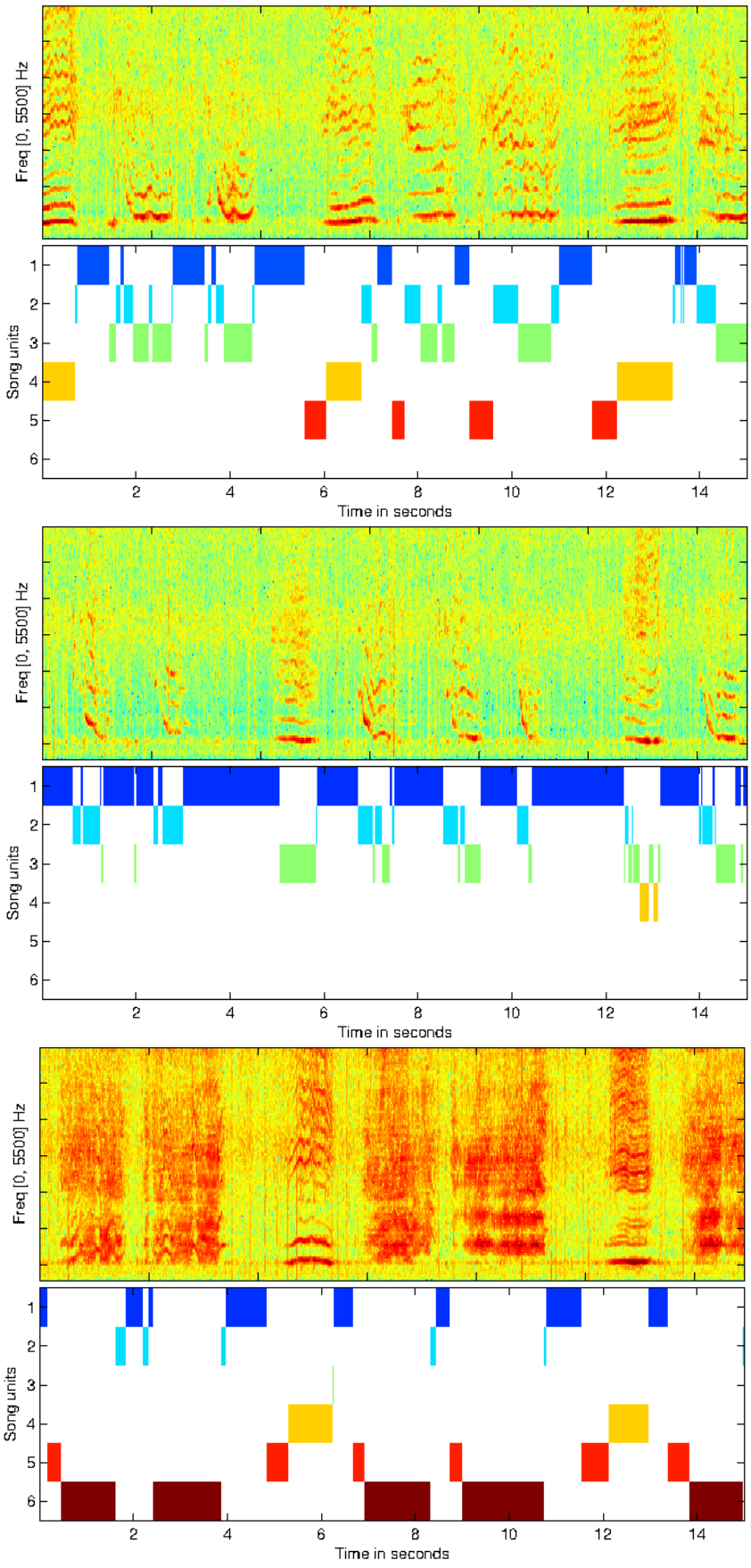

Figure 13: See caption of 1 for details.

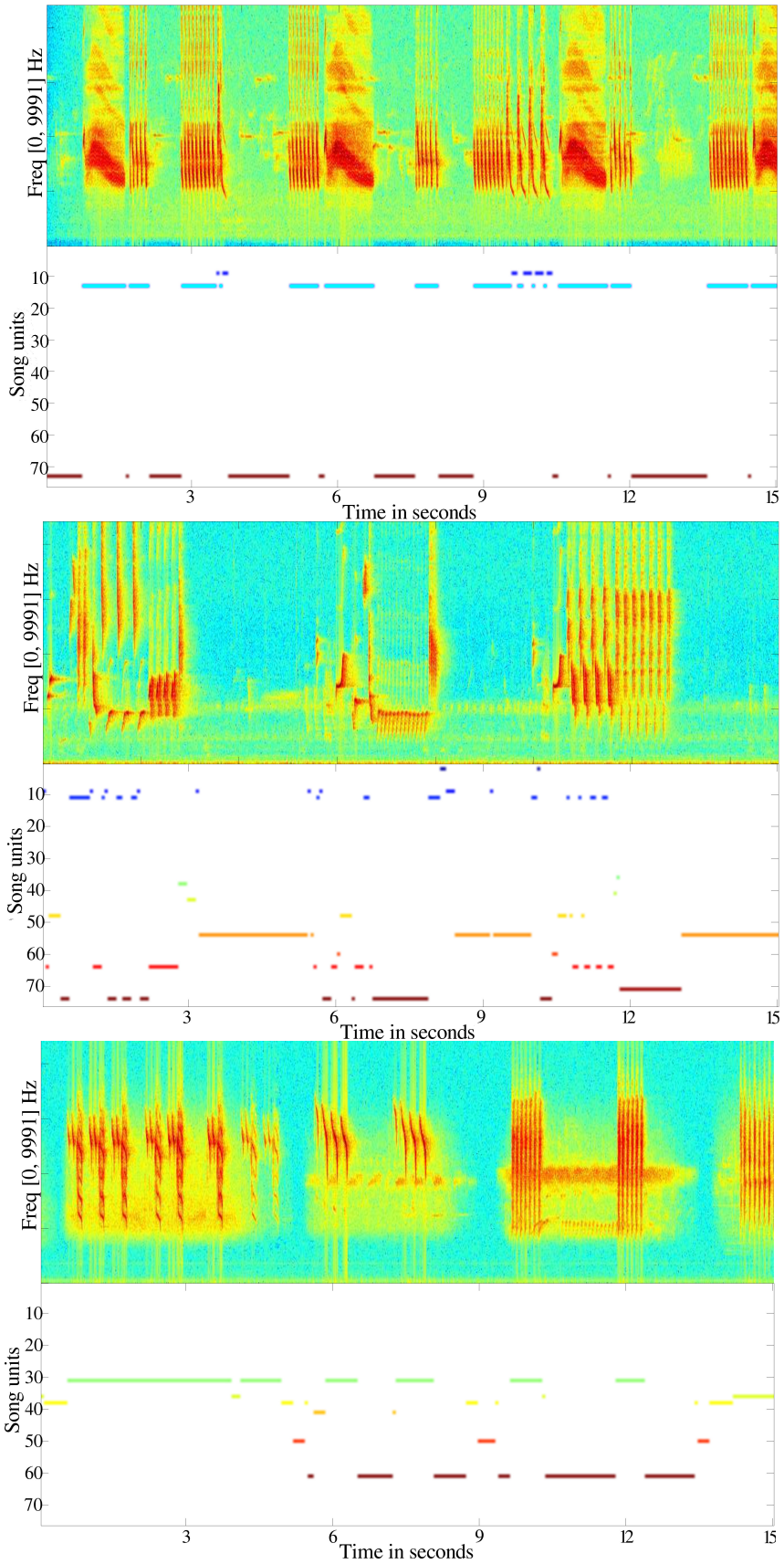

Figure 14: See caption of 2 for details.