Scaled Unsupervised Arctic Bioacoustics: Joint Narwhal Call & Click Train Indexing

Hervé Glotin[1,2,3,6], Marius Bartcus[1,2,6], Vincent Roger[1,2,6], Faicel Chamroukhi[1,2,4,6], Julie Patris[1,2,6], Kim Juniper[5,6], Pascale Giraudet[2,6], Jennie Knopp[7], Kristin Westdal[7], Louie Porta[7].

1) Aix Marseille Univ., CNRS, ENSAM, LSIS, 13397 Marseille, France
2) Université de Toulon, 83957 La Garde, France
3) Institut Universitaire de France, 75005 Paris, France
4) Lab. P. Painlevé, Univ. Lille, CNRS, 59655 Villeneuve d'Ascq, France
5) Ocean Networks Canada, Victoria University, Canada
6) SABIOD, MI MASTODONS, CNRS, http://sabiod.org, France
7) Oceans North Canada ONG, Canada

SABIOD LSIS Research Report 12.2015

Contents

1 Introduction

One of the Oceans North Canada project aims to survey the Arctic harbours that is "one of the world's least disturbed marine ecosystems and plays a crucial role in moderating the planet's climate". The warming ocean is altering this marine ecosystem in ways never before. Moreover the survey of arctic anthropophony becomes an urgent matter, because the disappearance of ice is now providing unprecedented access for industrial development. In this purpose, Oceans North Canada is recording nearly continuously since 2013, the sonic and ultrasonic soundscapes of the Baffin Arctic harbours with several hydrophones. They also record metadata with camera and thermometers, yielding to several TeraBytes per year of data. The analyses are done in collaboration with Pr Juniper's team, involving new generations of underwater sensors at UVic's leading underwater observatories VENUS and NEPTUNE, and in collaboration with SABIOD's partners based at Toulon university: the DYNI research group. This last one has developed algorithms for passive acoustic monitoring over the past ten years, including recently unsupervised infinite HMM bioacoustic analysis.

We think that such advanced signal clustering methods allows scaled monitoring populations of various species, hence helps in the monitoring of anthropophony.

We present in this report our new method for jointly analyse of the calls and the clicks trains of the Narwhal (Monodon monoceros). This approach will allow to discriminate the present species, like Beluga whale (Delphinapterus leucas) or Killer whale (Orcinus orca). These target species are known to produce both frequency modulated manations (e.g. whistles) and short-duration, broadband clicks and longer-duration.

These analyses of the Oceans North Canada non labeled long term sequences (from 2013 to 2015 on 4 stations) aim at discovering different behaviours of the species in quiet conditions and during anthropic activities (boat traffic, offshore installations...). This can be seen as a problem of unsupervised sound unit classification as in [7]. More recently we developed an unsupervised automatic clustering approach [1] to decompose and classify the sounds of cetaceans. Our proposal for Narwhal indexing is based on Hierarchical Dirichlet Process for Hidden Markov Model (HDP-HMM) [12, 5] that is an instance of infinite Hidden Markov Model (iHMM) [2].

In this first report, we demonstrate preliminary analyses on few days or continuous recordings, yielding to a segmentation designed to allow a fine grain study of the long term recordings of cetaceans, that will be correlated to the metadata (temperature, anthropic noise...) and used to recognize and cluster similar voiced signals, thus to characterize the behaviour of the target species.

This report is organized as follows: first we present the material, then we analyse the spectral content, bit rate and frequency sampling of the recordings. In section 4 we introduce two audio features, that we use to train the iHMM model. Then in section 6 we show and comment our preliminary results.

2 Material

The material is property of Oceans north Canada. The hydrophones and camera that have been installed at various locations, within Darnley Bay Fish Camp near ($69^\circ 22' 22.25''$N, $123^\circ 32' 25.92''$W), Tippi (69°30′31.53”N, 124°22′38.14”W), and Pond Inlet near Baffin (See the two next figures).

Initially these hydrophones were intended to provide stand-alone recordings from two disparate locations, given that the hydrophones are separated by up to 100 km between the study sites. However, the UTC time of deployment for each hydrophone was documented, allowing for the synchronization between the obtained audio and visual data.

Figure 1: The general situation of the recording stations in Arctic north Canada, from West at Darney Bay (Paulatuk) to East (Pond Inlet).

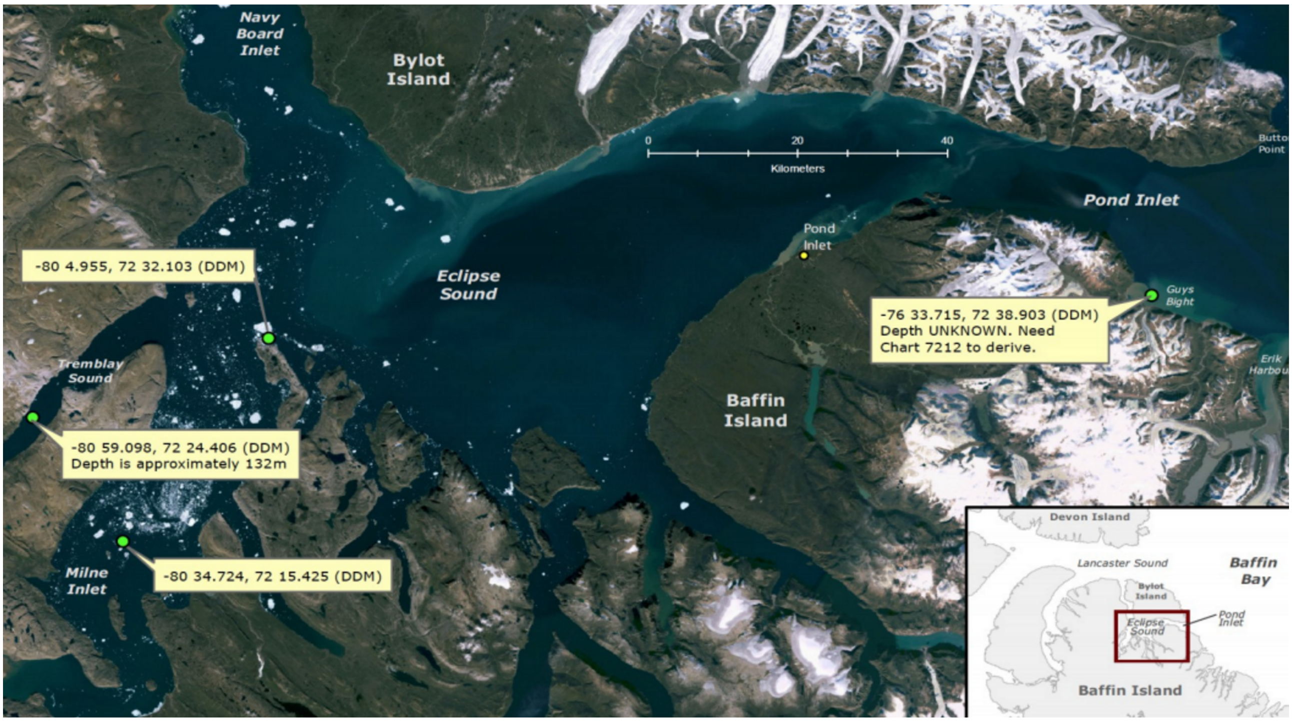

Figure 2: Zoom of the situation of some hydrophone stations near Pond Inlet (after ONorth-Canada).

In this preliminary report, we worked on the data received from Oceans North Canada in early December 2015. They are from Tremblay1, Tremblay 2, Low Isl. 1 and Low Isl 2, and

Fish Camp. They are recording at high sampling rate : 192 kHz for most of them. The Total corpora is more than 5 Tera Octet.
The computations of this study have been conducted on the HPC of SABIOD Dyni LSIS at Toulon (256 Go RAM, 20 cpu), allowing to deliver the results nearly 24 times faster than real-time.

3 Preliminary temporal and spectral analyses

3.1 Data quality

We analysed the spectrum quality of several samples of the given data set. We inspected many hours of recordings and checked if an electronic noise (spectral ray) would be present. We also checked if the bins sequence is correct (continuous and derivable wave form), because it is usual that medium quality digital recorder (DAQ) are loosing or swapping bins.
The conclusions are that the recordings of Oceans North Canada are of high quality. There is no swapped nor missing bins. However there is often clipped (saturated) recordings, which produces alterations into the spectrum. Nevertheless, we are able to post-process the data in order to remove most of these artifacts (by wave form normal thresholding).
The recordings are sometimes completely silent and one can assume the DAQ is not correctly working. But Arctic is one of the last silent area in the world, and we confirm that DAQ is correctly operating because sometimes a boat noise is suddenly arising with a typical wave form (see next figure). This kind of event is demonstrating that the DAQ is working fine, with very low self-noise.

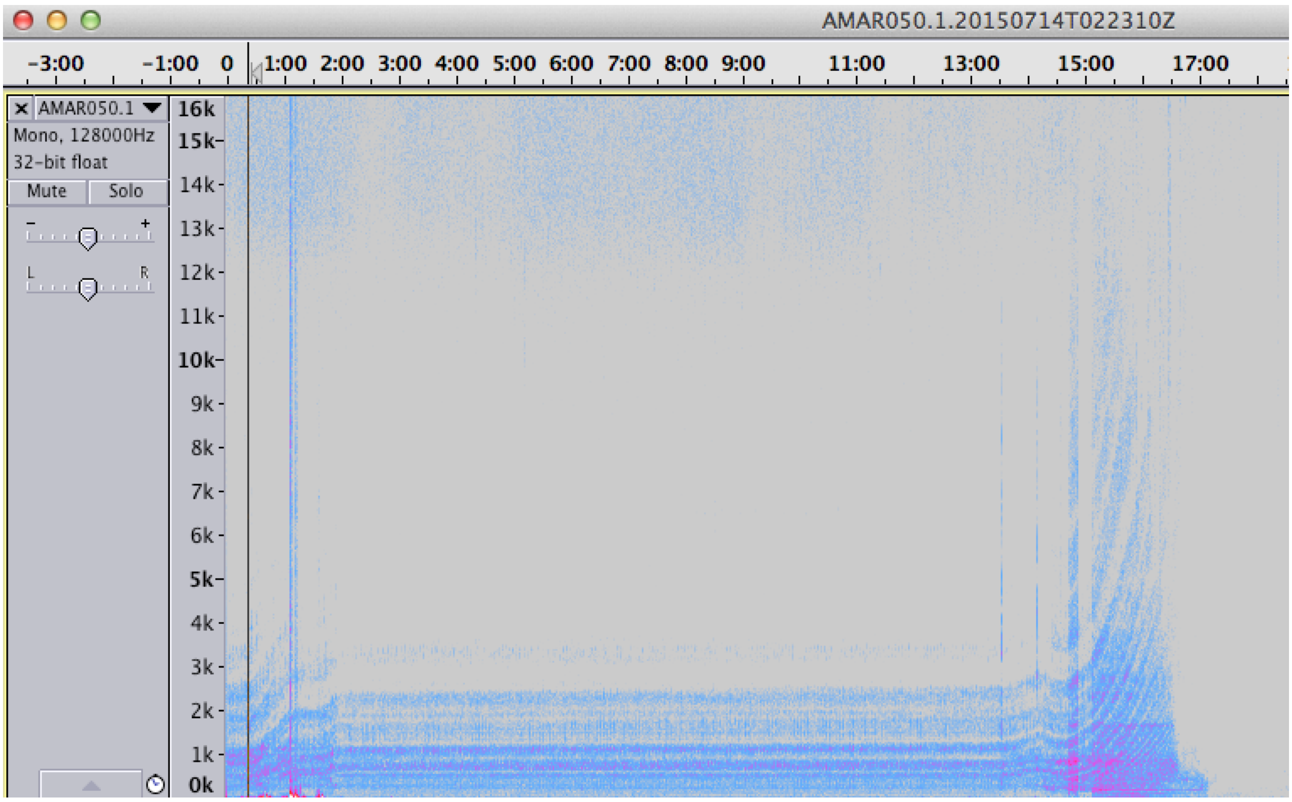

Figure 3: Evidences of the very low self-noise of the recorder: the boat noise is very high, then after the boat engine stops, only a very low SNR self noise of the recorder is present.

3.2 Narwhal identification

We present in the next Table the characteristics of the present cetaceans in the area.

Species	Peak frequency (kHz)	Band width (kHz)	Duration (microseconds)	Peak to peak source level (dB)	Sample location	References
Delphinapterus leucas (beluga)	100-115	30-60	50-80	225	Bay	Au, 1985; 1987
Monodon monceros (narwhal)	40	27	250	218	Sea	Mohl, 1990
Orcinus orca (killer whale)	14-20	4	210	178	Pool	Evans, 1973

Figure 4: Characteristics of the impulses of the main target species, after (Au & Hastings, 2008).

We summarize in the next figures the spectrum and time-frequency content found in most of the Oceans North Canada recordings. We see that the time-frequency, the spectral and the temporal characteristics are fitting with the Narwhals click trains definition given in this report, including the voicing of individuals. Thus we confirm that the recordings are of correct quality for Narwhals population study.

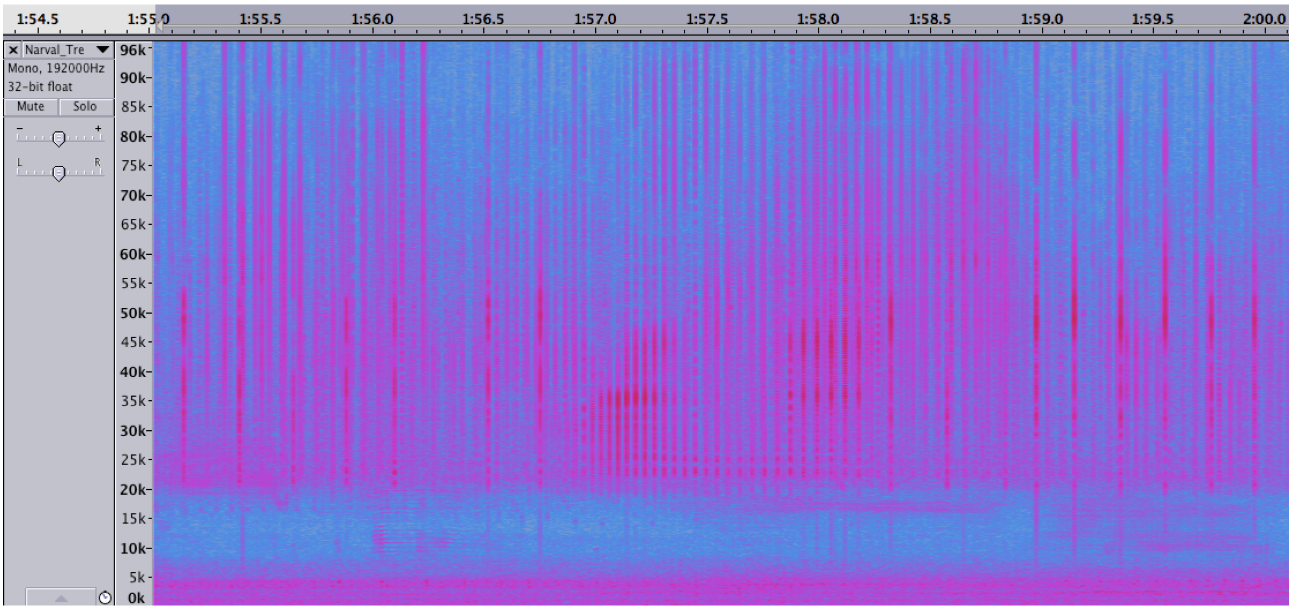

Figure 5: Time-frequency analysis of the Tremblay 1 recording of the 2015 august 4 at 5 pm. It shows series of click trains, and some voiced calls near 10 kHz at date 1:56, which are fitting with Narwhal characteristics.

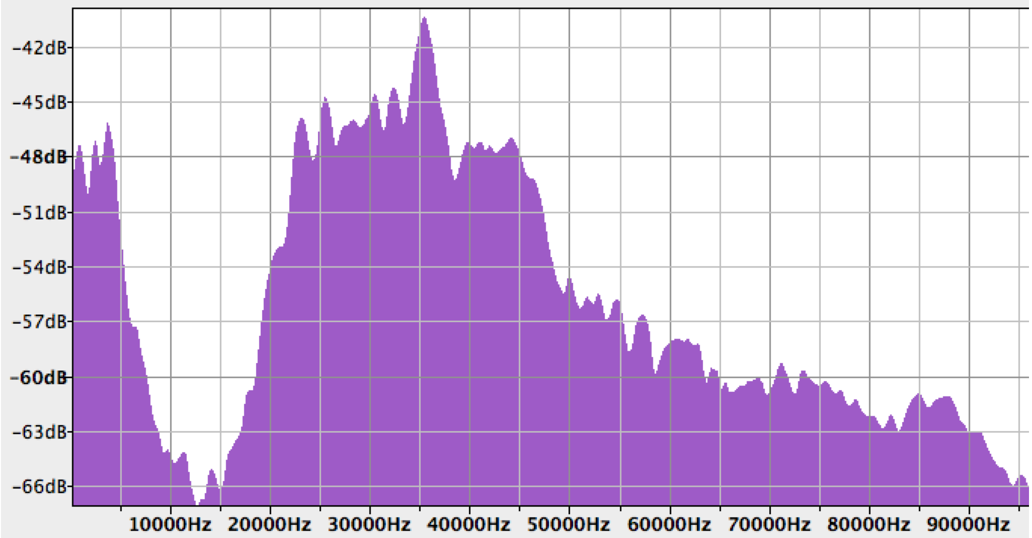

Figure 6: Average spectrum of the recording of 2015 august 4, 5 pm, showing a frequency peak near 36 kHz, and 25 kHz large (-6 dB), compatible with the Narwhal characteristics.

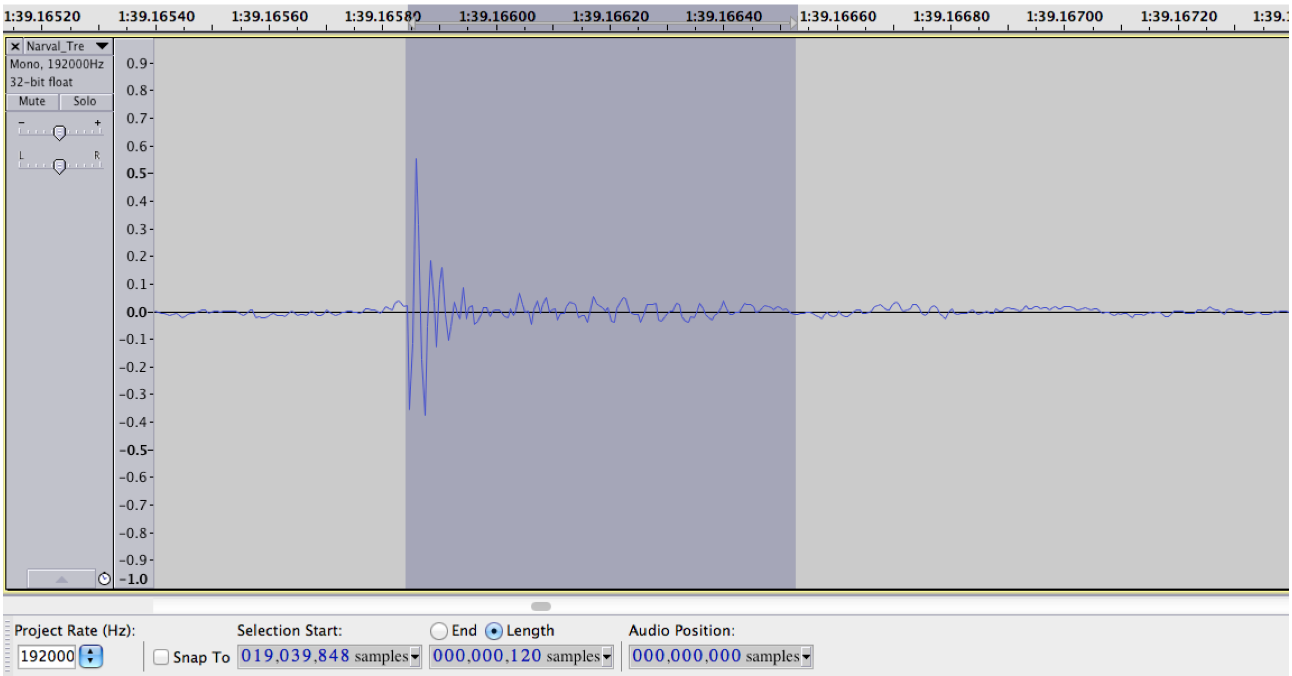

Figure 7: Zoom on a single click of the same recording, its duration is nearly 300 microseconds as the Narwhal.

4 Efficient acoustic features definition for scaled cetacean monitoring

We are looking for low CPU complexity acoustic features for this scaled monitoring study. First, a high pass filter is processed to reduce the noise (set at 1.030 kHz as suggested by [6]). First we extract the Mel-frequency Cepstral Coefficients (MFCC) features that compress the voicing part of the signal. Secondly, because Narwhals and other species emit also broadband clicks, containing temporal coda, we extract the Inter-Click-Interval with our dedicated fast implementation (Glotin et al. 2015).

4.1 Characterisation of voiced cetacean emissions by MFCC

The MFCC are features that represent and compress short-term power spectrum of a sound. It follows the Mel scale, which simulate the mammal auditory system. To extract the MFCC, we: 1) preemphasis trebles, 2) hamming windowing, 3) Fast Fourier Transform, 4) use triangular overlapping windows (following the Mel scale), 5) log each Mel frequencies, and 6) compress by discrete cosine transform. The MFCC implementation can be processed 10 times faster than real time (SigPro, LabRosa, and more recently DYNI).

In our approach we use windows of 1 second and no overlap (to accelerate the process) for the hamming windowing. We focus on frequencies below 3.110 kHz according to [6], because narwhals does less sing on higher frequencies.

4.2 Fast hierarchical ICI estimation of click trains (FICI)

In order to represent impulse signal (e.g. click trains), we propose a fast computation of the Interval Inter Click (FICI). We compute a local auto-correlation function accelerated by hierarchical decimation. It has been defined for continuous real time offshore monitoring of cetaceans with low signal to noise ratio emissions, to discriminate between anthropic noises during a road construction (Route du Littoral offshore anthropophony and cetacean survey, La Réunion 2014-2017, Glotin et al. 2015).

The main idea is that a regular ICI defines a particular behaviour. The behavior of Narwhal may be represented by different FICI clusters.

4.3 Discussion

We developed two kinds of features allowing to represent Narwhal behavior when they sing or emit click trains. We will then train an infinite HMM (iHMM) that allow unsupervised learning of the different kind of sounds. The iHMM model are without any a priori parameters, except the given features. The next section presents the iHMM, that is computed afterwards.

5 Unsupervised analysis by iHMM

5.1 Definition of iHMM

This section resumes the theory of iHMM. The Hidden Markov Model (HMM) are one of the most successful statistical models widely used in many applications involving complex sequential data [3, 9]. Defining HMM, the states $\mathbf{z} = (z_1, \ldots, z_n)$ are linked throw a state transition matrix and each observation $\mathbf{x}_t$, at time t is drawn independently of other observations conditional to z_t. This model has a finite number of states, thus $z_t \in \{1 \ldots K\}$, where K is the number of states in the model.

Thus, HMM framework generally requires the number of hidden states that are mostly unknown. This induces to a difficult model selection problem. To overcome this problem we propose to use a Bayesian non-parametric (BNP) version of the Hidden Markov Model, that is able to infer the number of hidden states from the data. To derive such so called infinite HMM (iHMM), a Dirichlet Process [4] is used. The generative process for the HMM is given as follows: the current state z_t indexes the specific row of the transition matrix. The probabilities of that row are mixing proportions to the choice to the next state z_{t+1}. The observation $\mathbf{x}_{t+1}$ is drawn from the mixing component of the state z_{t+1}. Thus one can say that each value of the current state z_t, in HMM, produces a mixture model and HMM can be seen as a set of mixture models. To derive the HMM for the BNP version, a set of DP for each state value is considered.

These sets of DP priors over the transition matrix is called the hierarchical Dirichlet process (HDP) [12, 5]. A HDP prior distribution provides a potential countability infinite number of hidden states and tackles the challenging problem of model selection for the HMM, inferring the number of states automatically from the proposed data. This model takes the name of Hierarchical Dirichlet Process for Hidden Markov Model (HDP-HMM), or can be also called as the infinite Hidden Markov Model [2]. We represent the graphical representation of the HDP-HMM in Fig. 8.

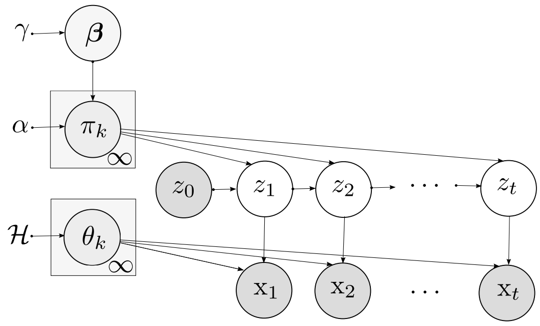

Figure 8: Graphical representation of the infinite Hidden Markov Model (iHMM).

5.2 Advanced signal decomposition by iHMM

Assuming an distinguished initial state z_0 the generative process for HDP-HMM (Eq. 1) defines distributions of the parameters from Fig. 8, $\forall k = 1, 2, \ldots$ and for each time steps t.

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

where the parameter $\boldsymbol{\beta}$ is distributed according to the stick-breaking construction formalism [10], noted by GEM(.)[8, 11], with it's concentration parameter γ. The transition matrix is distributed according to the Dirichlet Process with the concentration parameter α and the base measure $\boldsymbol{\beta}$ that is itself a Dirichlet Process, thus we obtain a hierarchical Dirichlet process distribution on the transition matrix. The states $\mathbf{z}$ are distributed according to a multinomial distribution Mult(.); the parameters of the model are drawn independently, according to the conjugate prior distribution $\mathcal{H}$; the observed data likelihood density is $F(\boldsymbol{\theta}_{z_t})$, where we assume that the unique parameters space of θ_{z_t} being equal to $\boldsymbol{\theta}_k$. We suppose that the data likelihood is a Gaussian density $\mathcal{N}(\mathbf{x}_t|\boldsymbol{\theta}_k)$, with the emission parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ respectively the mean vector and the covariance matrix drawn from a conjugate Normal inverse-Wishart distribution ($\mathcal{H} = \mathcal{NIW}(.)$).

From the first point of view, the HDP-HMM model resolves the problem of advanced signal decomposition providing precise information regarding frequency energies with respect to time, identifies species, behaviour and enhancing populations studies. From the other point of view modeling data with the HDP-HMM offers a great alternative for standard HMM to tackle the challenging problem of selecting the number of states, identifying the the unknown number of hidden units from the Mel Frequency Cepstral Coefficients (MFCC). The experimental results demonstrate the good performance of the proposed Bayesian nonparametric approach.

6 First results of the iHMM

In this section we show the pre-results of the iHMM or more precisely HDP-HMM trained on three different kind of features computed on **Tremblay Data 1** and **Tremblay Data 2**. We did compute also the model on Low Isl1, 2, but the lake of clean bioacoustic events on this last one are not yet generating simple illustrations.

6.1 iHMM(MFCC): iHMM segmentation based on MFCC

Fig. 9 shows the partition obtained after learning of the iHMM(MFCC) on Tremblay 1 dataset. The HDP-HMM converged to a set of 26 sound states (or sound units, one color each).

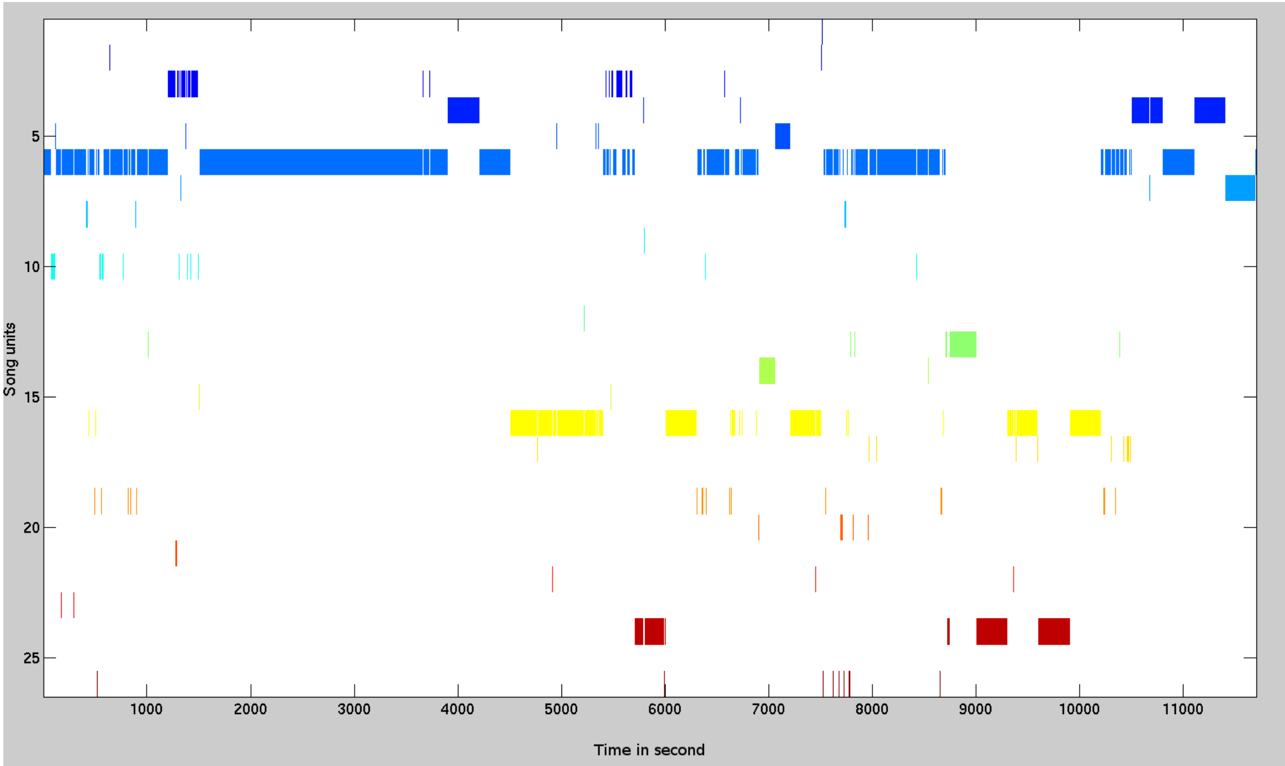

Figure 9: Obtained song units by applying the HDP-HMM on 190 minutes (3 hours) of **Tremblay Data 1** (MFCC features). It shows clear structures in yellow, red, orange or light blue that are not permanent. They are detailed in next figure at the scale of 5 minutes.

Fig. 10 illustrates the five minutes recorded the 20150804 170000 from Tremblay 1 that can be listened at url `sabiod.org/tremblyoneONC20150804at5pm.wav`.

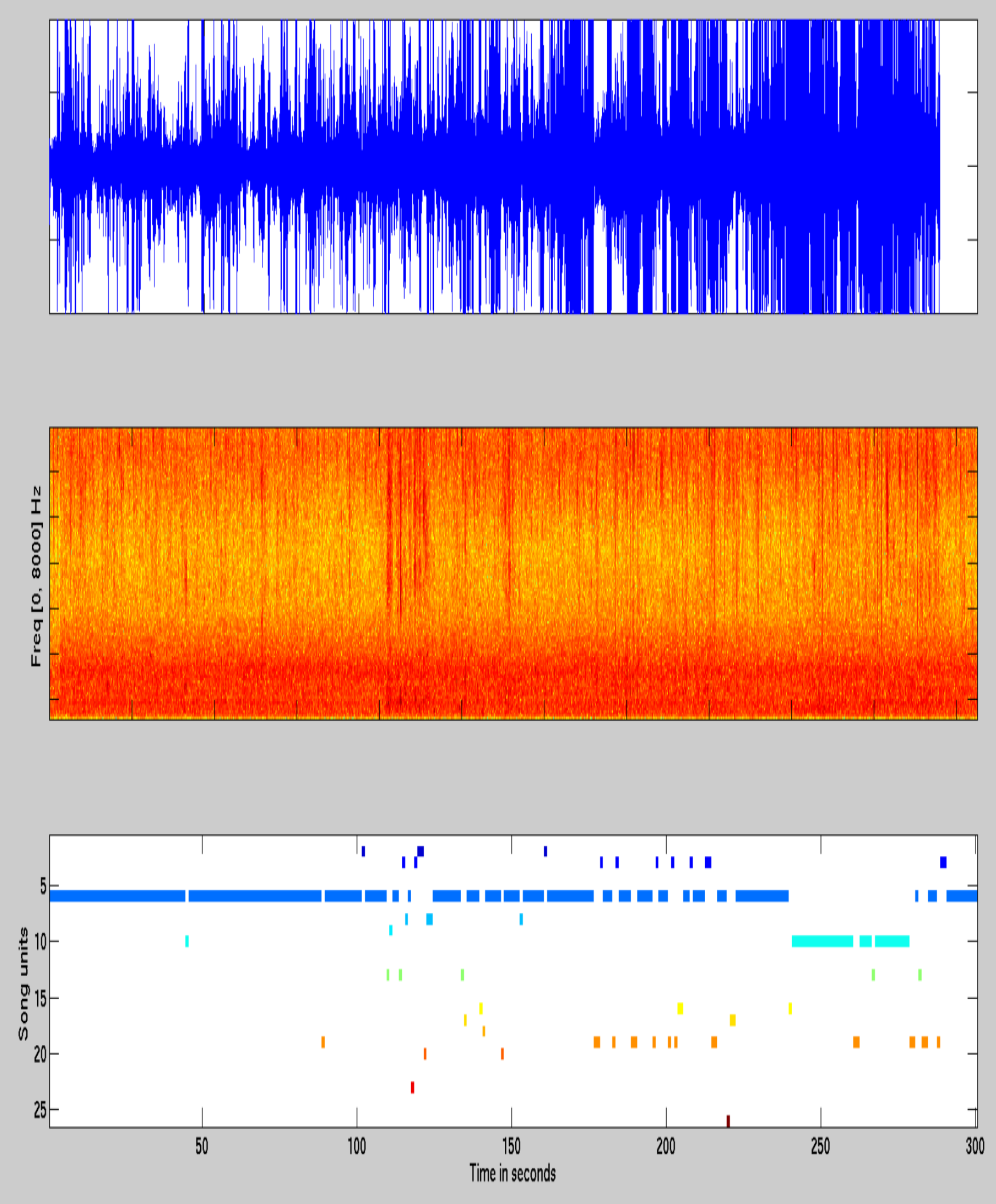

Figure 10: The wave form (top), the spectrogram (middle), and the respective obtained song units (bottom) for the 5 extracted minutes of the MFCC features of **Tremblay Data 1**. We see that the different states are not overlapping.

We now learn the iHMM on the merged Tremblay data sets: **Tremblay Data 1** union **Tremblay Data 2**, for a total of more than **3 DAYS of continuous data.**
It shall be more accurate than previous ones, because trained on more data. It took few days to be trained on our HPC, and the obtained model is very interesting. The resulted sound units or partition is given in Fig. 11, and shows interesting

variations of six main units (numbered 3, 8, 9, 20, 21, 28), similarly on Tremblay 1 and Tremblay 2. The sequences of the different kind of sounds are changing in nights and days, and more analyse shall also demonstrate changes with moon phases.

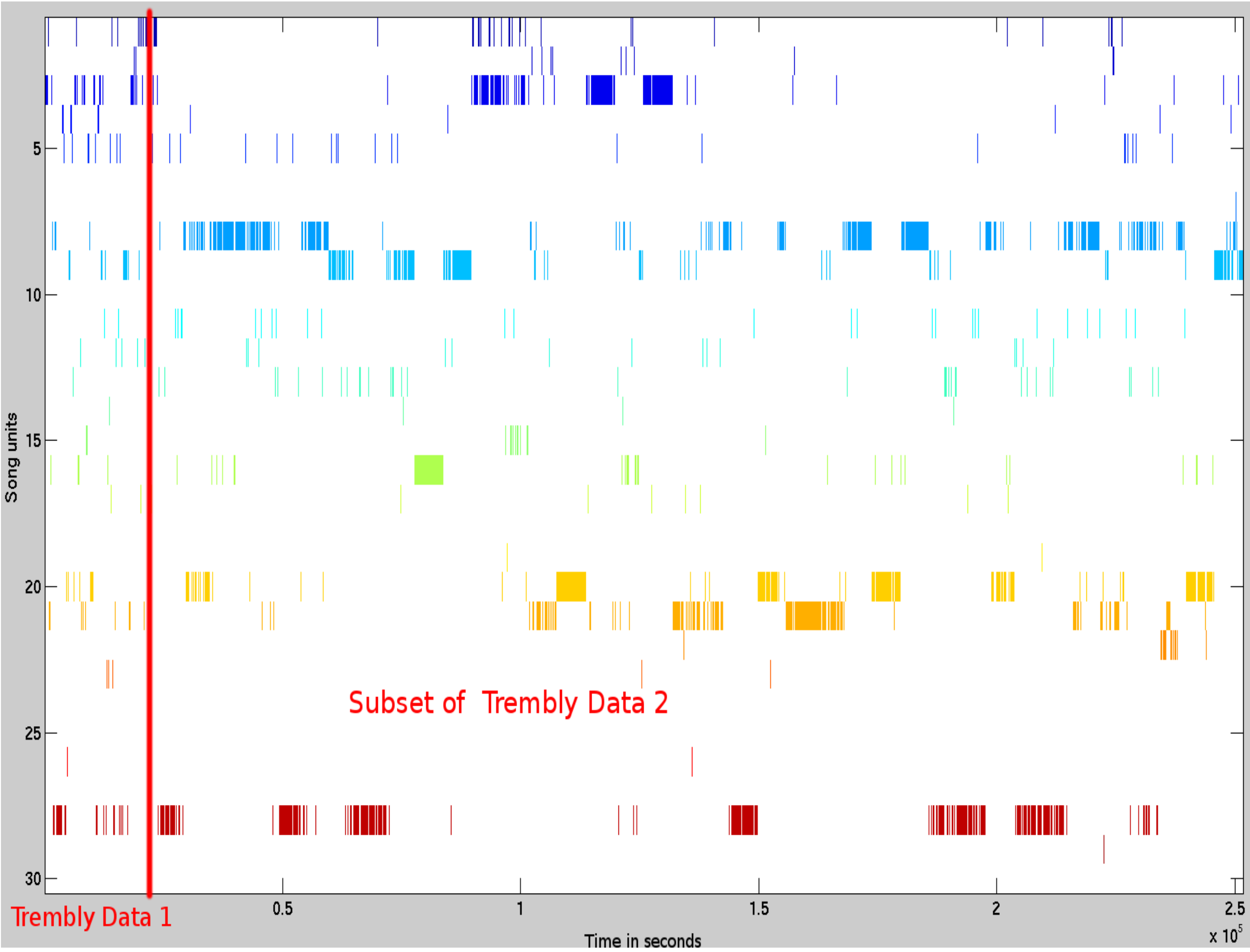

Figure 11: Long term (3 days) decompositions of iHMM(MFCC) trained on merged **Tremblay 1** and **Tremblay 2** data sets.

6.2 iHMM(FICI): iHMM segmentation based on FICI

Narwhals produce clicks to communicate, as other cetaceans, they may produce coda or variation of click trains correlated with their behaviour. Therefore we computed iHMM(FICI) on the Tremblay data set. A first result is given in Fig. 12, showing that iHMM converged to only 15 sound types. and shows that our method determines 15 different units, that are less numerous than in the case of the MFCC features. It can be assumed that the spectral content contains more information (more sound types) than click trains (we can assume that social communications are in majority emitted as voicing patterns than pulsed patterns).

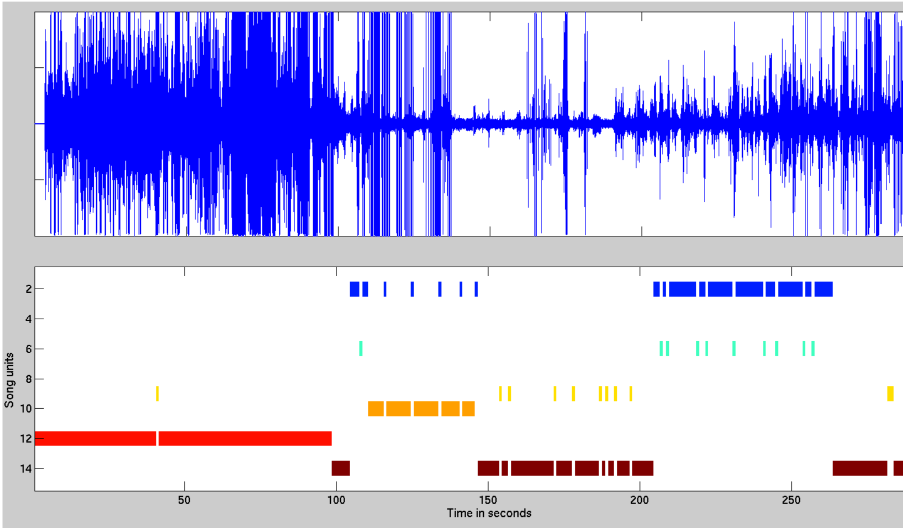

Figure 12: The wave form (top) and the respective obtained song units (bottom) from iHMM(FICI) trained on Tremblay. Here we only plot 5 minutes from **Trembly Data 1**. It reveals 15 sound types, and only 4 are active in this 5 minutes (one color per type). The different types are consistent with the global wave form, but are more detailed and stable that just an eye survey would tell.

6.3 iHMM(MFCCFICI): iHMM on concatenated MFCC and FICI

The results of iHMM(MFCC) and iHMM(FICI) show different segmentations because MFCC and FICI are complementary features coding respectfully stationary versus temporal structures. In order to have a generic representation, we propose to concatenate these two features together into one feature that we call MFCCFICI (simple vector concatenation of MFCC with FICI). Then we train a new iHMM(MFCCFICI). The results is given in the Fig. 13. As it was expected it represents a mix between the obtained partition of the MFCC and FICI features.

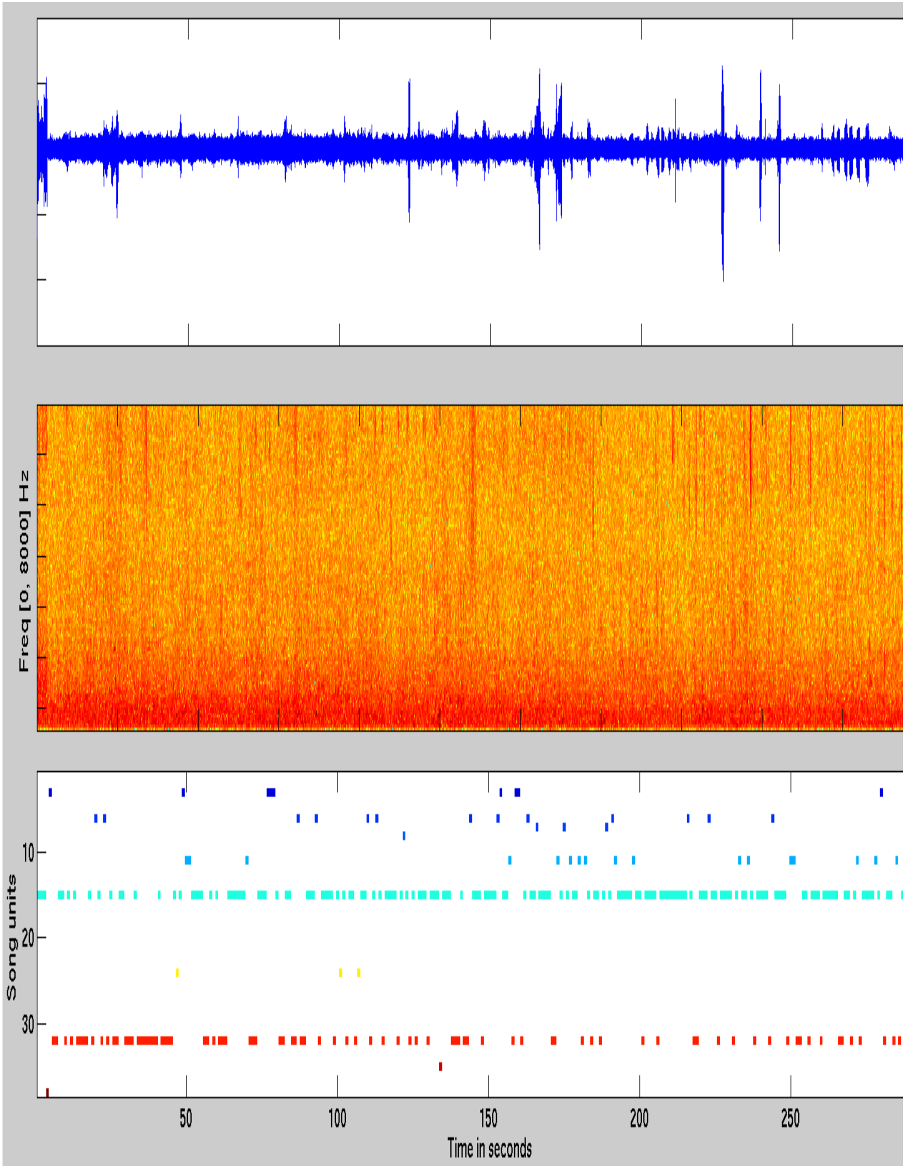

Figure 13: The wave form (top), the spectrogram (middle) and (bottom) the respective obtained song units from iHMM(MFCCFICI) trained on **Trembly Data 1**, illustrated on only 5 extracted minutes from **Trembly Data 1**.

Interestingly, the total number of units generated by iHMM(MFCCFICI) are larger than the number of units from iHMM(MFCC) and iHMM(FICI). In the illustrated example of 5 minutes we show that only two units are activated (units 15 and 32). This may tell that there is two different kinds of emissions, and possibly two individuals. More analyses are required to better characterize iHMM(MFCCFICI).

6.4 Quality of the models versus quality of the features

The quality of the Model / Features can be estimated by their likelihood, as for all kind of generative model. The highest the likelihood is, the best are the Model and Features, and the more the representation of the data is accurate. Thus, we give in the table 1 the log(likelihood) of each model.

Let us note L(F(D)) the likelihood of an iHMM model trained on feature F on a data set D. First this table shows a similar likelihood for the models iHMM(FICI)

trained on Tremblay 1 versus Tremblay 2 (we note L(FICI(Tremblay 1))) is similar to L(FICI(Tremblay 2)). Then we can assume that these two sets are of equal quality respectfully to the FICI features.

Second, we see that the L(FICI(Tremblay1)) $>$ L(MFCC(Tremblay1)).
This could be due to the fact that the MFCC are computed only on the frequencies under 3110Hz which is the maximum frequency of narwhals calls, thus MFCC do not code a lot of the information conveyed into the click trains. On the contrary the FICI feature is computed on the raw data.
Also we demonstrated in previous sections that the click train is generating less sound units, then we can assume that iHMM requires less data in quantity and quality to converge to an accurate representation. We can also suppose that our parametrisation of MFCC with a frame window of 1 second is not optimal. Thus the final parametrisation will the optimized to maximize the likelihood of each model. It would take few days of optimization and then the segmentation will be more accurate, and the behaviour of the Narwhal shall be easier to analyse.

Dataset	Features	Log(likelihood(iHMM))
Tremblay 1	iHMM(FICI)	-7 465
Tremblay 2	idem	-9 356
Tremblay 1	iHMM(MFCC)	-144 370
Tremblay 1	iHMM(MFCCFICI)	-157 650

Table 1: The log-likelihood values of the models shown in previous sections, on the two acoustic features and their concatenation.

7 Conclusion and Perspectives

According to this preliminary study, the quality of the recordings (sampling rate, bit rate, Signal to Noise Ratio, frequency bandwidth...) is compatible for at least species identification at Tremblay recordings and automatic joint calls and click trains segmentation. The Tremblay recordings have a lot of Narwhal calls and clicks, allowing interesting bioacoustic pattern and behaviour studies. We could recommend for future recordings to decrease the gain of the DAQ at Tremblay 1 and 2. But the gain at Low Isl 1 and 2 shall not be decreased: these recordings seem not to contain many bioacoustic contact. We'll look in details the bathymetry in this area.

We would recommend Oceans North Canada to record in stereo (2 x 192 kHz) to allow sources separation and animal counting, and also azimuth estimation of each source even with a small 2 meters aperture, as we did in Toulon Bombyx to compute 3D localisation of Physeter with our stereo JASON system and using surface reflexion virtual hydrophones[1].
[2]

We notice that Low Is. Data 1 and Low Is. Data 2 do not contain so much interesting sounds than the other data set. This can be due to the fact that the hydrophones are not enough powered, or at a too low gain, or masked, or simply

[1] A demonstration of 3D localisation is available at http://sabiod.org/vamos

[2] Toulon JASON DAQ system (¡ 1 000 $) requires only 1 W for stereo (up to 8 channels) up to 1 MHz sampling Rate.

the animals are not visiting this point.
We demonstrate that our developed FICI acoustic feature is generating consistent sound classes with iHMM(FICI). This feature is complementary to MFCC feature, and their concatenation gives a third kind of summary of the recordings. The ensemble of these decompositions can yield to classification between different acoustic activities, and thus behaviour analysis, estimation of the complexity of the interaction between individuals, correlation to external events (biologic, geologic and anthropic).

Thus, future works will be:
- Enhance our model with features selection by likelihood maximisation (3 months),
- Forward the model on the whole data set 2013+2014+2015 for a total of 4To (4 months),
- Statistics on cetacean presence (1 month),
- Conspecifics and heterospecifics bioacoustics correlated with anthropic noise, time, moon phases, weather, temperature (1 month).

Acknowledgments

This preliminary study has been conducted thanks to the support of SABIOD MI CNRS, IUF Glotin's grant, and PNPC Pelagos VAMOS grant. The recordings have been distributed with restriction from Oceans North Canada.

References

[1] Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015.

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

[3] L. Derrode, S. Benyoussef and W. Pieczynski. Contextual estimation of hidden markov chains with application to image segmentation. In *ICASSP*, pages 15–19, Toulouse, May 2006.

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

[5] 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*, pages 312–319, New York, NY, USA, 2008. ACM.

[6] Marianne Marcoux. *Narwhal communication and grouping behaviour: a case study in social cetacean research and monitoring.* PhD thesis, Department of Natural Resource Sciences McGill University, Monreal, February 2011.

[7] 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.

[8] J. Pitman. Combinatorial Stochastic Processes. Technical Report 621, Dept. of Statistics. UC, Berkeley, 2002.

[9] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

[11] Yee Whye Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. 2010.

[12] 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.