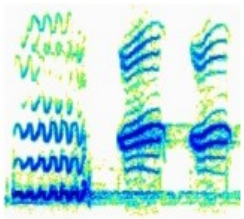
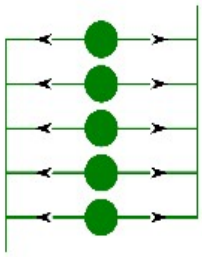
Neural Information Processing Systems Foundation

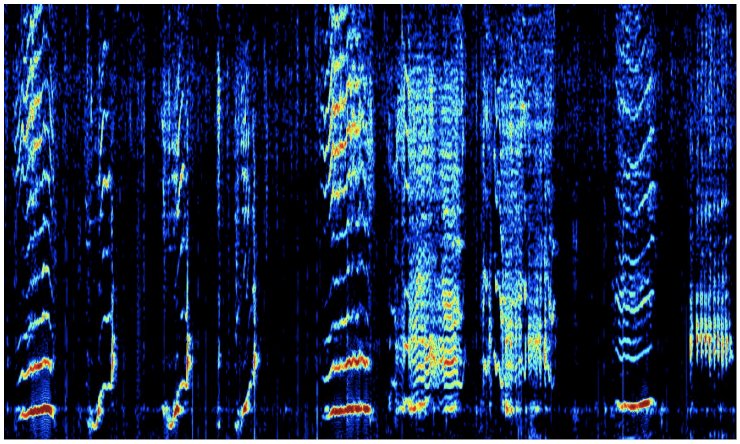
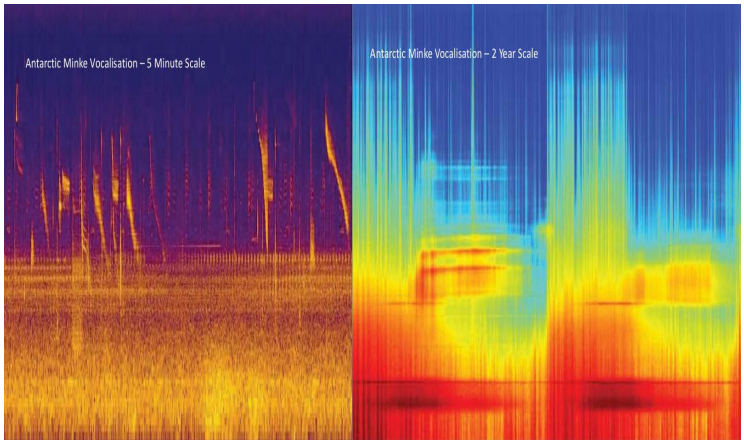
SABIOD.ORG Scaled Acoustic Biodiversity Mastodons Big Data

Neural Information Processing Scaled for Bioacoustics -from Neurons to Big Data-

Proceedings of NIPS4B, international workshop joint to NIPS, USA, 2013

Glotin H., LeCun Y., Artières T., Mallat S., Tchernichovski O., Halkias X.
Toulon, New-York, Paris

http://sabiod.univ-tln.fr

ISBN: 979-10-90821-04-0

We acknowledge CNRS MASTODONS MI for their support to SABIOD.ORG project which co-organized this workshop.
Proceedings compiled with the help of R. BALESTRIERO, Toulon university.
We thank O. Dufour for flyers edition

December 2013 - La Garde

Legend of the cover :

Left image : spectrogram of the challenge 2 humpback whale song, produced

by Vincent Lostanlen, ENS.

Right picture : Minke whale Fourier time-frequency representation, on 5

minutes scale (left) versus on two years scale (right), showing season effect

and global frequency shift [from L. Kindermann 2013, in this book]. It can be

found in the section 9.2

Glotin H. & Halkias X : CNRS, LSIS, USTV, Toulon University, FR
LeCun Y. : New York University, USA
Artières T. : Sorbonne University, Paris 6, FR
Mallat S. : Ecole polytechnique, FR
Tchernichovski O. : Hunter College, USA

These book is available on line at http//sabiob.org/NIPS4B2013_book.pdf

It shall be cited as :

Proc. of Neural Information Processing Scaled for Bioacoustics: from neurons to Big Data, 2013, Glotin H., LeCun Y., Artieres T., Mallat S., Tchernichovski O., Halkias X., joint to NIPS Conf., http://sabiod.org/NIPS4B, ISSN 979-10-90821-04-0

Bibtex =
@proceedings{procNIPS4B2013
title={Proc. Neural Information Processing Scaled for Bioacoustics, from neurons to Big Data},
year={2013},
author={Glotin H., LeCun Y., Arti\`eres T., Mallat S., Tchernichovski O., Halkias X.},
organization={NIPS Int. Conf.},
address={USA},
note={http://sabiod.org/NIPS4B},
key={ISSN 979-10-90821-04-0},
}

Contents

Video supports: Most of the talks are available in video at http://sabiod.univ-tln.fr/nips4b/

Authors list (not including all participants)

Names	Emails	Affiliation
Adam Olivier	*olivier..adam@u-psud.fr*	Univ Orsay and Sorbonne Paris 6, CNRS LAM CNPS, FR
Andrews Russ	*russ_andrews@alaskasealife.org*	Univ. of Alaska Fairbanks, USA
Artieres Thierry	*thierry.artieres@lip6.fr*	Univ. Sorbonne Paris 6, LIP6, FR
Balestriero Randall	*randallbalestriero@gmail.com*	Univ. Toulon, FR
Bartcus Marius	*marius.bartcus@gmail.com*	Univ. Toulon, CNRS LSIS, FR
Bas Y.	*ybas@biotope.fr*	LIP6, Univ. Paris 6, FR
Berthommier Frederic	*frederic.berthommier@gipsa-lab.fr*	Grenoble-Alpes Univ., CNRS, FR
Boe Louis-Jean	*louis-jean.boe@gipsa-lab.fr*	Grenoble-Alpes Univ., CNRS, FR
Cazau Dorian	*cazaudorian@aol.com*	Univ. Sorbonne Paris 6, CNRS LAM, FR
Chamroukhi Faicel	*chamroukhi@univ-tln.fr*	CNRS, ENSAM, LSIS, FR
Chen Wei	*chenwei@i2r.a-star.edu.sg*	Institute for Infocomm Research, Singapore
Doh Yann	*yanndoh.m2@gmail.com*	Univ. Toulon and Sorbonne Paris 6, CNRS LSIS LAM, FR
Dufour Olivier	*olivierlouis.dufour@gmail.com*	Aix-Marseille Univ., CNRS, ENSAM, LSIS, FR
Elie Julie E.	*julie.elie@berkeley.edu*	Berkeley Univ. USA
Gervaise Cedric	*cedric.gervaise@gipsa-lab.grenoble-inp.fr*	Grenoble-INP, FR
Giraudet Pascale	*giraudet@univ-tln.fr*	Toulon Univ., Dept Biology, FR
Girin Laurent	*laurent.girin@gipsa-lab.fr*	Grenoble-Alpes Univ., CNRS, FR
Girolami Mark A.	*m.girolami@ucl.ac.uk*	Univ. College London, UK
Glotin Hervé	*glotin@univ-tln.fr*	CNRS LSIS, USTV, Toulon, FR
Halkias Xanadu	*xanadu.halkias@univ-tln.fr*	CNRS LSIS, USTV, Toulon, FR
Hoeberechts Maia	*maiah@uvic.ca*	Univ. of Victoria, Canada
Horaud Radu	*radu.horaud@inria.fr*	INRIA Grenoble Rhone-Alpes, FR
Janvier Maxime	*maxime.janvier@inria.fr*	INRIA Grenoble Rhone-Alpes, FR
Jones Kate	*kate.e.jones@ucl.ac.uk*	Univ. College London, UK
Kemp Caralyn	*caralyn@kemputer.com.au*	Aix-Marseille Univ., CNRS, FR
Keogh Eamonn	*eamonn@ucr.edu*	UC Riverside, USA
Kindermann Lars	*lars.kindermann@awi.de*	Bremerhaven, Germany
Lasseck Mario	*Mario.Lasseck@mfn-berlin.de*	Museum für Naturkunde Berlin, Germany
Le Bot Olivier	*olivier.le-bot@gipsa-lab.grenoble-inp.fr*	Grenoble-INP, FR
LeCun Yann	*hongtam@cs.nyu.edu*	New York Univ., USA
Lee Dong-Hyun	*sayit78@gmail.com*	Germany
Legou Thierry	*thierry.legou@lpl-aix.fr*	Aix-Marseille Univ., CNRS, FR
Lewandowski Achim	*achim@oefai.at*	Vienna, Austria
Li Xiaohui	*lixh@i2r.a-star.edu.sg*	Technology and Research (A*STAR), Singapore
Lostanlen Vincent	*vincent.lostanlen@ens.fr*	ENS, FR
Loza Mencia Eneldo	*eneldo@ke.tu-darmstadt.de*	Technische Universitat Darmstadt, Germany
Mallat Stephane	*stephane.mallat@ens.fr*	Ecole polytechnique, FR
Mars Jerome	*jerome.mars@gipsa-lab.grenoble-inp.fr*	Grenoble-INP, FR
Massaron Luca	*lucamassaron@gmail.com*	Independent marketing research director, data scientist Verona,IT
Mathias Delphine	*delphine.mathias@gmail.com*	Univ. J. Fourier,UMR CNRS GIPSA, FR
Mercado III Eduardo	*emiii@buffalo.edu*	Univ. at Buffalo, SUNY Buffalo, NY 14031, USA
Mirebrahim Hamid	*smire002@ucr.edu*	UC Riverside, USA
Mishchenko Ales	*ales.mishchenko@univ-tln.fr*	LSIS-DYNI, FR
Nam Jinseok	*nam@kdsl.informatik.tu-darmstadt.de*	Technische Universitat Darmstadt, Germany
Ntalampiras Stavros	*stavros.ntalampiras@jrc.ec.europa.eu*	Joint Research Center, EC
Paris Sebastien	*sebastien.paris@lsis.org*	CNRS LSIS, USTV, Toulon, FR
Plumbley Mark D.	*mark.plumbley@eecs.qmul.ac.uk*	Queen Mary Univ. of London, London, UK
Pollack Gérald	*gerald.pollack@gmail.com*	McGill Univ., Montréal, CA
Potamitis Ilyas	*potamitis@staff.teicrete.gr*	Technological Educational Institute of Crete, Greece
Razak Khaleel	*khaleel.abdulrazak@ucr.edu*	UC Riverside, USA
Razik Joseph	*Joseph.Razik@univ-tln.fr*	CNRS LSIS, USTV, Toulon, FR
Rey Arnaud	*arnaud.rey@univ-amu.fr*	Aix-Marseille Univ., CNRS, FR
Rotschafer Sarah	*srots001@ucr.edu*	UC Riverside, USA
Shokoohi-Yekta Mohammad	*mshok002@ucr.edu*	UC Riverside, USA
Stathopoulos Vassilios	*v.stathopoulos@ucl.ac.uk*	Univ. College London, UK
Stowell Dan	*dan.stowell@qmul.ac.uk*	Queen Mary Univ. of London, London, UK
Straley Jan	*jan.straley@uas.alaska.edu*	Univ. of Alaska Southeast, Sitka, Alaska, USA
Tchernichovski Ofer	*otcherni@hunter.cuny.edu*	Hunter College - CUNY, NY, USA
Theunissen Frédéric E.	*theunissen@berkeley.edu*	Berkeley Univ., USA
Thode Aaron	*atthode@ucsd.edu*	Univ. of California San Diego, USA
Trone Marie	*mtrone@valenciacollege.edu*	Valencia College, Spain
Zakaria Jesin	*jzaka001@ucr.edu*	UC Riverside, USA
Zamora-Gutierrez Veronica	*vz211@cam.ac.uk*	Univ. of Cambridge Cambridge, CB2 3E, USA
Zhao Gang	*zhaogang@comp.nus.edu.sg*	National Univ. of Singapore, Singapore

Chapter 1

Introduction

Introduction

This book is the content of the 1st Big Bioacoustics Data [NIPS4B] that took place at Tahoe lake, Nevada, in december 2013, during the NIPS international conference. The 40 attendies provided further insights into the analysis of large scale bioacoustic data and modeling of animal sounds, not only from a neuro- perspective, but also by highly reinforcing the need to approach these unique signals within the machine learning community.

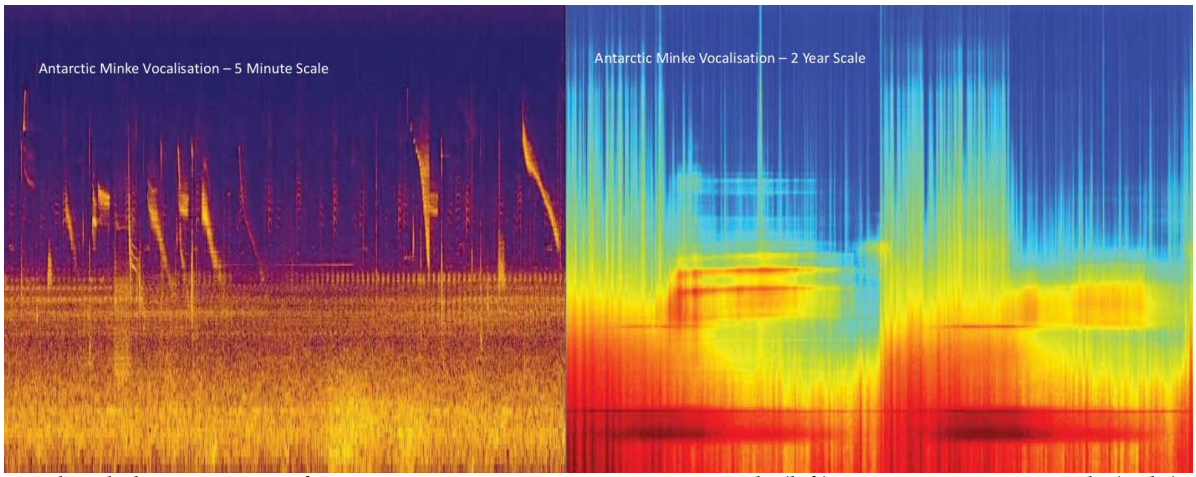
As a result both the bioacoustics community and the mainstream NIPS community met, leading to new collaborations: the communications ranged from the complexity of bioacoustics to scaled analyses, from understanding and monitoring bird song ontogeny, to cricket auditory neural functions, from use of sparse architectures for whale sound classification, to range estimation and bat tracking...

Although, in recent years, the majority of the existing applications lend themselves to advanced acoustic signal processing methodologies, our efforts are successfully integrating robust processing and machine learning algorithms for scaled analysis of these abundant recordings. Major issues such as data repositories and the need for standardizations within the bioacoustics field discussed and addressed.

We exchanged ideas on how to proceed in understanding bioacoustics to provide methods for biodiversity indexing, and to open a novel paradigm toward a Bioacoustic Turing Test: one might model animal communication before tackling the original Turing test for human being.

The scaled bioacoustic data science is a novel challenge for artificial intelligence that requiere new methods. For example Minke whales, observed all around the planet have been recorded by Kindermann's acoustic observatory at the ice shelf around Antarctica during 8 years. Big data scientists are today invited to look into that data using advanced methods to definitely new knowledge about this important species.

Similarly, large cabled submarine acoustic observatory deployments permit data to be acquired continuously, over long time periods. For examples, Neptune observatory in Canada, Antares or Nemo neutrino workshop on Neural Information Processing Scaled for observatories in Mediterranean sea are 'big data' challenges to the scientists. Automated analysis, including the classification of acoustic signals, event detection, data mining and machine to discover relationships among data streams are techniques which promise to aid scientists in discoveries in an otherwise overwhelming quantity of acoustic data as it is presented in this book.



Minke whale Fourier time-frequency representation, on 5 minutes scale (left) versus on two years scale (right), showing season effect and global frequency shift [from L. Kindermann 2013, in this book].

1.1 Objectives

Bioacoustic data science aims at analyzing and modeling animal sounds for neuroethology / biodiversity assessment. However, given the complexity of the collected data along with the different taxonomies of the different species and their environmental contexts, it requires original approaches. In recent years, the field of bioacoustics has received increasing attention due to its diverse potential benefits to science and society, and is steadily required by regulatory agencies as a tool for timely monitoring and mitigation of environmental impacts from human activities. The increased expectations from bioacoustic research have been coincident with a dramatic increase in the spatial, temporal and spectral scales of acoustic data collection efforts. One of the most promising strategies concerns neural information processing and advanced machine learning.

The features and biological significance of animal sounds, while constrained by the physics of sound production and propagation, have evolved through the processes of natural selection. Additional insights have been gained through analysis and attempts of modeling of animal sounds as related to critical life functions (e.g. communicating, mating, migrating, navigating, etc.), social context, and individual, species and population identification. These observations have led to both quantitative and qualitative advancements, as for example MRIs for monitoring bird song ontogeny. These yieled to new paradigms such as prococesses that underlie song learning and their modelisation. Although, the majority of the existing applications lend themselves to widely used, advanced acoustic signal processing methodologies, the field has yet to successfully integrate robust signal processing and machine learning algorithms, applied for example to bird, insect, or whale song identification, source localisation, (neural)modelisation of the biosonar of bats or dolphins...

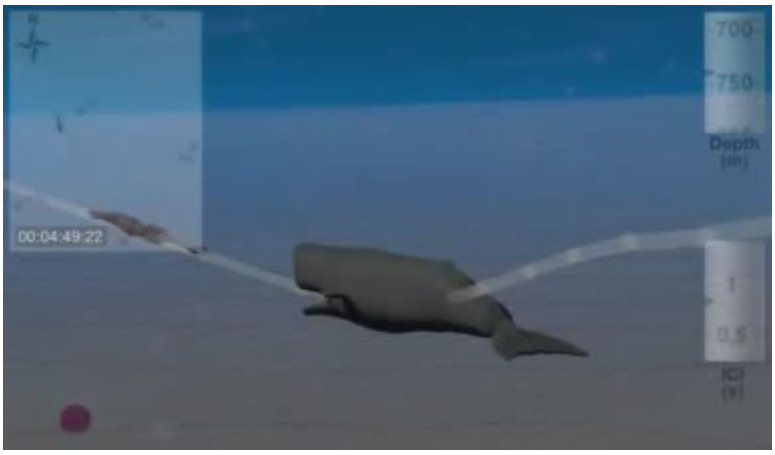


Figure: Sperm whale tracking demo[1], more informations here[2].

This NIPS4B workshop has helped to introduce and solidify an innovative computational framework in the field of bioacoustics by focusing on the principles of neural information processing in an inheretly hierarchical manner. State of the art machine learning algorithms have been explored in order to draw physiological parallels within bioacoustics, while an applicative framework has address classification tasks.
For example, new sparse feature representations have been pursued by using both shallow and deep architectures in order to model the underlying highly complex data distribution. Cost creation and hyper-parameter optimization in architectures such as Deep Belief Networks (DBN), Sparse Auto Encoders (SAE), Convolutional Networks (ConNet), Scattering transforms, ..., have provided insights in the analysis of these complex signals. Any interesting new learning technique for this type of bioacoustic signal is very welcome.
NIPS4B has encouraged interdisciplinary, scientific exchanges and foster collaborations among the workshop participants for the bioacoustic signal analysis and understanding of the auditory process. NIP4B aims at bringing together experts from the machine learning and computational auditory scene analysis fields

1 http://www.youtube.com/watch?v=0Szo3gdiTRk
2 http://glotin.univ-tln.fr/oncet/

with experts in the field of animal acoustic communication systems to promote, discuss and explore the use of machine learning techniques in bioacoustics for signal separation, classification, localisation,... It has concerned researchers in modeling the auditory cortex, neurophysiological process in perception and learning, machine listening, signal processing, and computer science to discuss these complementary perspectives on bioacoustics.

1.2 Song Bird Classification Challenge

Challenge 1: Bird Song Classification / Kaggle web site now available

This Bird NIPS4B competition asks participants to identify which of 87 sound classes of birds and their ecosystem are present into 1 000 continuous wild recordings (from different places in Provence France - nearly 2 hours of recordings, frequency sample = 44.1 kHz, SM2 system). The data is provided by the BIOTOPE[1] society (having the largest collection of wild recordings of birds in Europe). The training set matches the test set conditions.
This challenge is a more complex task than our previous one at ICML4B challenge[2] for which 77 teams participated - see proceedings at sabiod.org[3].
This enhanced challenge opens the 2nd of october. The metrics is the Area Under the Curve, as for our previous previous[4] challenge.

1/ SOUND FILES: WHOLE WAV FILES, TRAIN and TEST, 138 Mo[5]

2/ SUGGESTED FEATURES: we provide baseline features of these train and test .wav files, computing optimized MFCC for bird's sound representation, as distributed in ICML4B 2013 bird challenge : MEL FILTER CEPSTRA COEFFICIENTS (MFCC) of WHOLE TRAIN and TEST FILES (158 Mo)[6] The format is a matrix 17xN: 17 cepstral coefficients x N frames, frame size 11.6 ms, frame shift 3.9 ms, one line per frame. You may compute their speed and acceleration by simple line differences. These suggested features minimize the signal reconstruction error in average on bird species. The script which produced these MFCC is:MFCC SCRIPT for BIRD SOUND REPRESENTATION[7] (please cite if you use these features).

3/ LABELS:Here are the tables of the 87 classes to learn (.csv, xls, html)[8].
** This archive also includes the TRAINING LABELS of the 687 train files (.csv, xls, html)[9].
For some species we discriminate the song to the call (and to the drum). We also include some species living within with these birds: 7 insects and a batracian. Each of these 87 classes in this table are to be predicted in the 1000 test files. Some training files are empty (background noise only called 'empty class') to tune your model, this class is not to be predicted. The training set contains 687 files. Each species is represented by nearly 10 training files (within various context / other species).

4/ EXAMPLES: The test set is composed of 1000 files. All the species into the test set are in the training set. We give here two samples containing each two species: Sylvia cantillans (which is singing) and Sylvia melanocephala (which is calling)[10] .
Second sample: Sylvia cantillans (which is also singing) and Petronia petronia (which is calling)[11]

1 http://www.biotope.fr/
2 http://www.kaggle.com/c/the-icml-2013-bird-challenge/
3 http://sabiod.univ-tln.fr/ICML4B2013_proceedings.pdf
4 http://www.kaggle.com/c/the-icml-2013-bird-challenge/
5 http://sabiod.univ-tln.fr/nips4b/media/birds/NIPS4B_BIRD_CHALLENGE_TRAIN_TEST_WAV.tar.gz
6 http://sabiod.univ-tln.fr/nips4b/media/birds/NIPS4B_BIRD_CHALLENGE_TRAIN_TEST_MFCC.tar.gz
7 http://sis.univ-tln.fr/~glotin/Kaggle_BIRD_challenge_ICML4B_MFCCcomputation.m
8/9 http://sabiod.univ-tln.fr/nips4b/media/birds/NIPS4B_BIRD_CHALLENGE_TRAIN_LABELS.tar
10 http://sabiod.univ-tln.fr/nips4b/media/birds/nips4b2013_birds_file_0001.wav
11 http://sabiod.univ-tln.fr/nips4b/media/birds/nips4b2013_birds_file_0002.wav

Scaled bioacoustics is a new challenge that requiere new methods that will be discussed in this book. For example antarctic Minke whales are today observed all around the planet. Long term recordings from Kindermann's acoustic observatory at the ice shelf show its acoustic emission around Antarctica during years (see figure below). Tenth of thousands of hours of this sound have been recorded during the last 8 years. Big data scientists are today invited to look into that data using advanced methods to extract definitely new knowledge about this important species.

The large cabled submarine acoustic observatory deployments permit data to be acquired continuously, over long time periods. For examples, the current running ones are the Neptune observatory in Canada(see M.H. talk in this book), Antares or Nemo neutrino observatories in Mediterranean sea (see H.G.'s talk). This capability presents a “big data” challenge to the scientist using and accessing the data. Automated analysis, including the classification of acoustic signals, event detection, data mining and machine learning to discover relationships among data streams are techniques which promise to aid scientists in making discoveries in an otherwise overwhelming quantity of acoustic data as it will be prensented in this book.

Organizers:

•Pr. H. Glotin[1] - Institut Universitaire de France[2], CNRS LSIS[3] and USTV[4], FR
Email: glotin@univ-tln.fr
•O. Dufour[5] - CNRS LSIS, FR
•Dr. Y. Bas - BIOTOPE[6], FR

1.3 Whale Song Clustering Challenge

Challenge 2: Whale Song Processing

It is well documented that Humpback whales produce songs with a specific structure. We provide 26 minutes of a remarkable Humpback whale song recording produced at few meters distance from the whale in La Reunion - Indian Ocean, by our "Darewin" research group in 2013 (frequency sample = 44.1kHz, 32 bits, mono, wav, 130MB). Upload here 26 minutes of a remarkable Humpback whale song recording[7].

1 http://glotin.univ-tln.fr/
2 http://iuf.amue.fr/iuf/presentation
3 http://www.lsis.org/
4 http://www.univ-tln.fr/
5 http://dyni.univ-tln.fr/~odufour/
6 http://www.biotope.fr/
7 http://sabiod.univ-tln.fr/nips4b/media/NIPS4B_Humpback_Darewin_LaReunion_Jul_03_2013-001_26min.wav

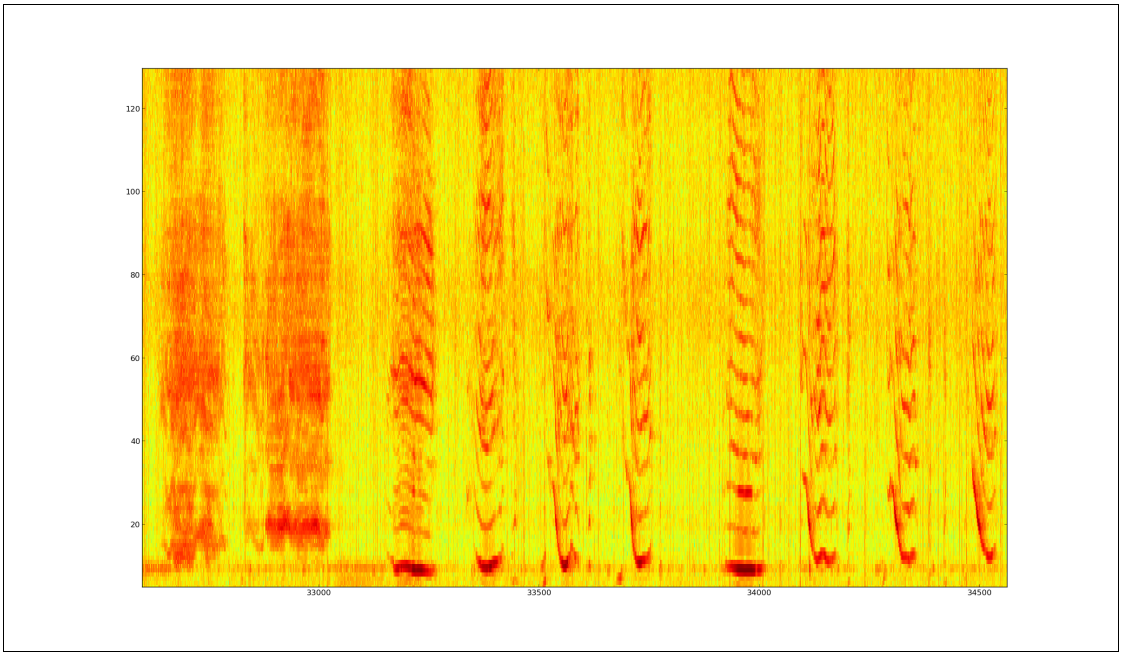


Figure: Spectrum of around 20 seconds of the given song of Humpback Whale (start from about 5'40 to 6'. Ordinata from 0 to 22.05 kHz, over 512 bins (fft on 1024 bins), frameshift of 10 ms.

We also give the usual Mel Filter Cepstrum Coefficients of this wav file (octave / matlab v6 format)[1]. The parameters of extraction of these MFCC are given here.[2].

For this challenge, you may propose any efficient representation of this song that helps to study its structure, discover and index its song units. You can find an interesting preliminary approach in: *Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. (2010) Subunit definition for humpback whale call classification*[3]*, int. journal Applied Acoustics, Elsevier, 11(71)*

The workshop allows discussions over the proposed representations (clustering, indexing, sequence modeling etc.). Your representation of this song file shall be sent to nips4b@gmail.com in usual format (.xml, .csv or .mat ...). The size (bytes) of your representation and its quality (MSE on the reconstructed signal of interest) are used to rank it.

Figure: Humpback Whale

Organizers:

Doh Yann (UTLN), Joseph Razik (UTLN) and Hervé Glotin (UTLN & IUF)

We thank Darewin for the recording.

1 http://sabiod.univ-tln.fr/nips4b/media/NIPS4B_Humpback_Darewin_LaReunion_Jul_03_2013-001_26min_1024_CORRECTED.mat

2 http://sabiod.univ-tln.fr/nips4b/media/NIPS4BparametersMFCChumpbacksongsample.txt

3 http://sabiod.univ-tln.fr/nips4b/media/Pace_etal_APAC2010.pdf

1.4 Neurosonar Analysis

Example of topic of interest (non restrictive) : BioSonar - new files to upload since 30th sept

One of the possible (non restrictive) topic of interest in this workshop is the biosonar analysis (mostly from bats or dolphins): Which neural processes underlie pulse train emission of biosonar? Are there pulse categories? What is the high / social information content of these sonar, and if any, at which scale? Which could be the efficient decompositions / self-learned representations for these complex sounds?
Some samples and paradigms of sonar sequences of wild dolphin and bats are given below, more are available, please ask to nips4b@gmail.com.

Here a .pdf with the summary of some dolphin data and suggested topic of interest.[1]
First we provide these clear recordings of Stenella dolphin sonar from the Port-Cros National Park, Côte d'Azur [USTV-PELAGOS DECAV project 2011-12]:

- file a[2] (5MB)
- file b[3](11MB).

The two next files contain sonar of another dolphin species, the biggest one, i.e. Physeter macrocephalus (15 meters, 40T): file c[4] (28MB), and file d[5] (high signal to noise ratio, recorded at Toulon in 2012-DECAV SABIOD, FS=48kHz, 55MB).

Nice 25 minutes of one Physeter have been recorded on 5 channels in Bahamas by NATO, and we have precisely computed the 4D positions of this whale [Glotin 2008][6] with its real animation on YouTube[7]. The whole recordings 25 minutes x 5 channels at 48kHz and the positions and references are in this archive[8] (500 MB) (here is one sample of 5 min[9]). You find a sparse coding representation of these clicks in [Paris et 2013][9].

1 http://sabiod.univ-tln.fr/nips4b/media/NIPS4B_Humpback_Darewin_LaReunion_Jul_03_2013-001_26min_1024_CORRECTED.mat
2 http://sabiod.univ-tln.fr/nips4b/media/DECAV_20110607_073535_v2_raccourcie_bateauenfond_v2.wav
3 http://sabiod.univ-tln.fr/nips4b/media/DECAV_20120916_174818_v2_raccourcie_propre2.wav
4 http://sabiod.univ-tln.fr/nips4b/media/DECAV_20121006_171343_v2_dauphin_cachalot_assez_propre_org.wav
5 http://sis.univ-tln.fr/~glotin/DECAV_20120917_135935.wav
6 http://sis.univ-tln.fr/~glotin/NIPS4B_MATERIAL/DATA_WAV_POSITIONS/BAHAMAS/GLOTIN_etal_Whale_Cocktail_Party_Int_JOURN_CANADIAN_ACOUSTICS_spring2008.pdf
7 http://www.youtube.com/watch?v=0Szo3gdiTRk
8 http://sis.univ-tln.fr/~glotin//NIPS4B_MATERIAL/DATA_WAV_POSITIONS/BAHAMAS_Physeter_4channels.tar.gz
9 http://sis.univ-tln.fr/~glotin/NIPS4B_MATERIAL/DATA_WAV_POSITIONS/BAHAMAS/HYDRO10/10S_ch5_10-15.wav
10 http://arxiv.org/pdf/1306.3058v1

Fig: The sonar sample given below is from Nicky, here with her calf, recorded at Shark Bay[1] Australia (cred. Giraudet 2013).

This Tursiops sonar sample[2] is from the wild dolphin called Nicky, 37 years old, visiting nearly daily Monkey Mia Bay (frequency sample 96kHz, 32 bits, with CR55 hydrophone of Cetacean Research). Here is its time-amplitude representation:

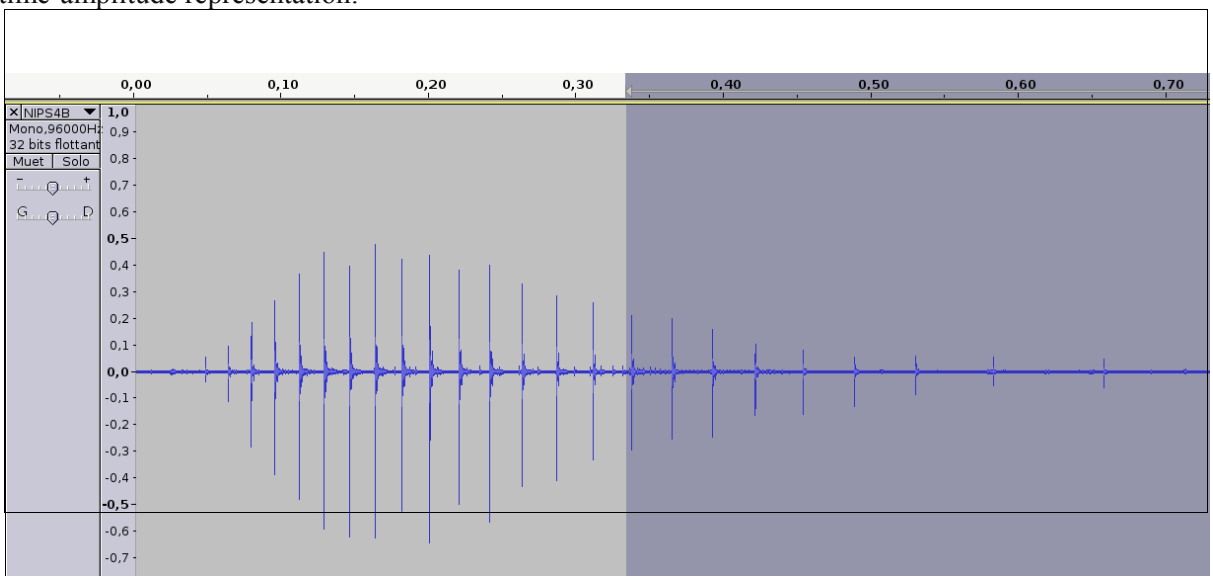


Fig: The time-amplitude representation of this Nicky's sonar short sample (0.7 sec).

Here we give a longer sequence of Nicky[3] (same FS=96kHz, 32 bits, 19MB), recorded at 2m from her noise. You may use Audacity[4] or GNU Octave[5] to read it.

1 http://www.monkeymiadolphins.org/
2 http://sabiod.univ-tln.fr/nips4b/media/NIPS4B_sonar_S1.wav
3 http://sabiod.univ-tln.fr/nips4b/media/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_32bits_after19min_nips4bfile_e.wav
4 http://audacity.sourceforge.net/
5 http://www.gnu.org/software/octave/

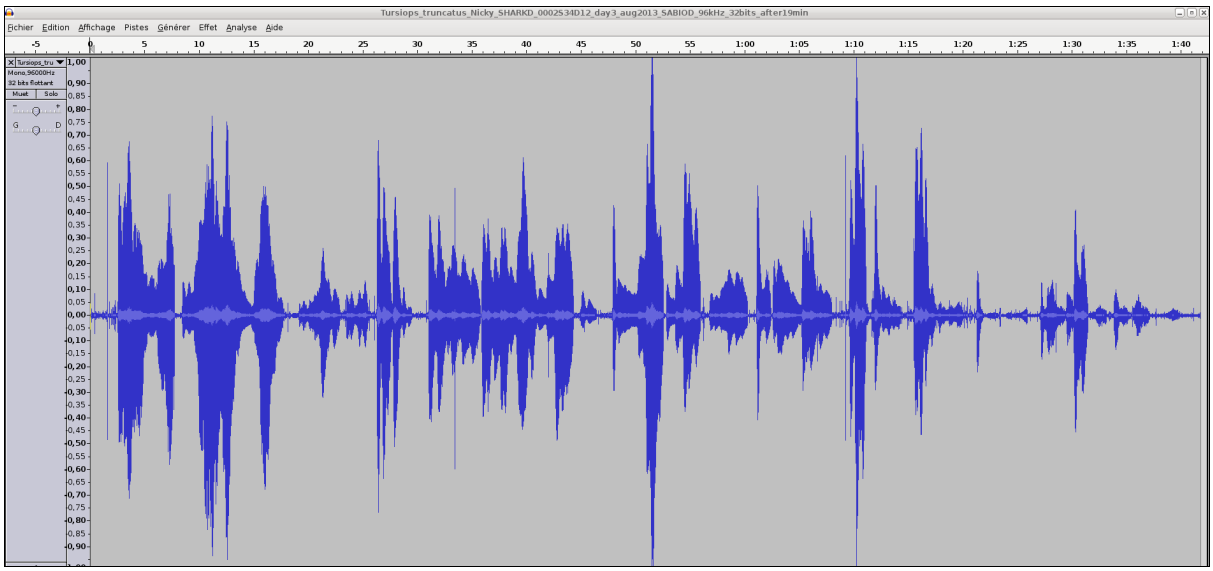


Fig: Longer time-amplitude of Nicky's sonar (100 sec.): same file e[1].

In [Ryabov 2011][2] it is shown that Tursiops dolphin are producing the packs of coherent and non-coherent broadband pulses. The waveform and spectrum of coherent pulses are invariable within a pack (see fig. below), but considerably varies from a pack to a pack. The waveform of each non-coherent pulse vary from a pulse to a pulse in each pack, therefore their spectrum also vary from a pulse to a pulse and have many extrema. It is very likely that the non-coherent pulses play a part of phonemes of a dolphin spoken language and the probing signals of dolphin's non-coherent coherent sonar. Efficient feature extraction and classification on sonar sequences are requiered for such studies.

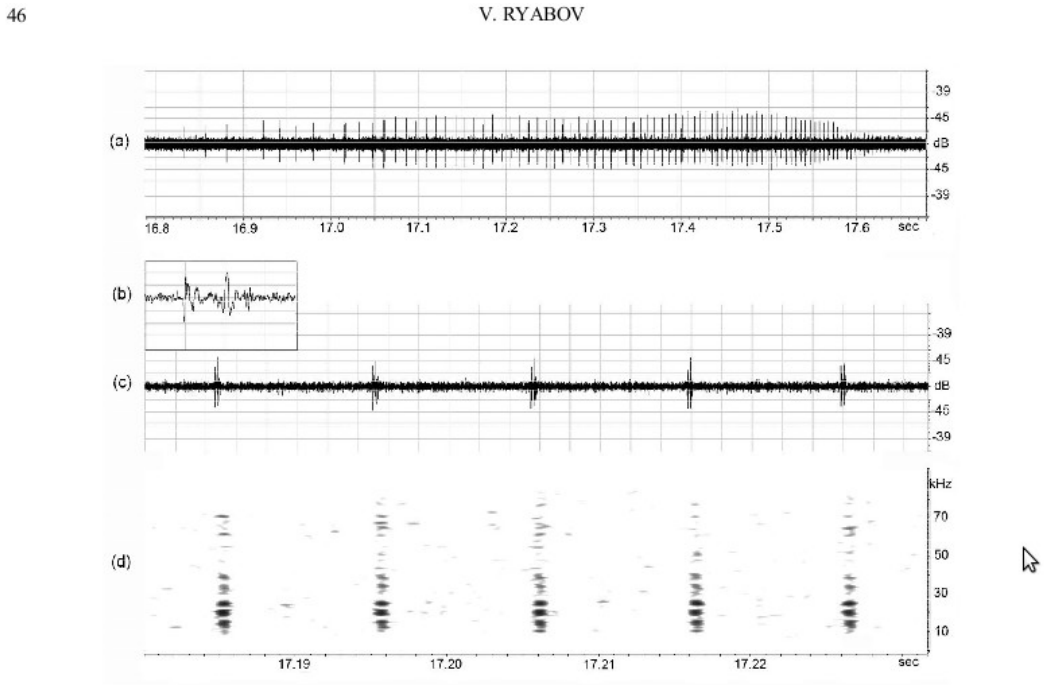


Figure 4. Example of the time dependence of interpulse interval in the pack of coherent pulses (a) that Yana produced from 16.8th up to 17.6th sec (n_4, Figure 2) and a magnification of the area from 17.185th up to 17.23th sec represented in the time (c) and frequency (d) domain. The single pulse (b) at the expanded time scale of 200 μsec/div. Along X-axis are pulses location on the time axis of Figure 2(a). Along Y-axis are SPL in dBs relatively 350 Pa and frequencies in kHz, respectively. The relative amplitudes scale of sonogram is the same as in Figure 3.

1 http://sabiod.univ-tln.fr/nips4b/media/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_32bits_after19min_nips4bfile_e.wav

2 http://www.scirp.org/journal/PaperDownload.aspx?paperID=7397

Sonar of Dolphin River

We give also Amazone dolphin river (Inia) recorded by David E. Bonnett (bonnettde@gmail.com)[1]: Inia, 2007, 96kHz FS, 600Mo[2], and Inia, 2009, 96kHz FS, 300Mo[3], and Inia, 2011, 96kHz FS, 300Mo[4] We also give some Indian River bottlenos dolphins record, recorded by M. Trone (DRS). Indian River Dolphin, 2013a, 96kHzFS, 358 Mo[5] and Indian River dolphin, 2013b, 500kHz FS, 1Go[6] From these recordings, you may try to learn features that may correlate with some morphological difference, here are some point of interest on river files.[7]

Similar paradigm applies to bats' sonar: [Kno 2012] shows that if bat echolocation is primarily used for orientation and foraging, it also holds great potential for social communication.

Fig: Bat's sonar share many properties with cetacean one's

1 http://sis.univ-tln.fr/~glotin/DRS_and_Microtrack_Recording_System_Descriptions.pdf
2 http://sabiod.univ-tln.fr/nips4b/media/Amazon_2007_96kHz.zip
3 http://sabiod.univ-tln.fr/nips4b/media/Amazon_2009_96kHz.zip
4 http://sabiod.univ-tln.fr/nips4b/media/Amazon_2011_96kHz.zip
5 http://sabiod.univ-tln.fr/nips4b/media/Indian_River_Lagoon_2013_96kHz.zip
6 http://sabiod.univ-tln.fr/nips4b/media/Indian_River_Lagoon_2013_500kHz.zip
7 http://sis.univ-tln.fr/~glotin/nips4b_dolphin_river_point_of_interests.pdf

- Myopterus bat's sonar sample (credit Cyberio)[1] (frequency sampling=250kHz, duration=4 seconds, 2MB)

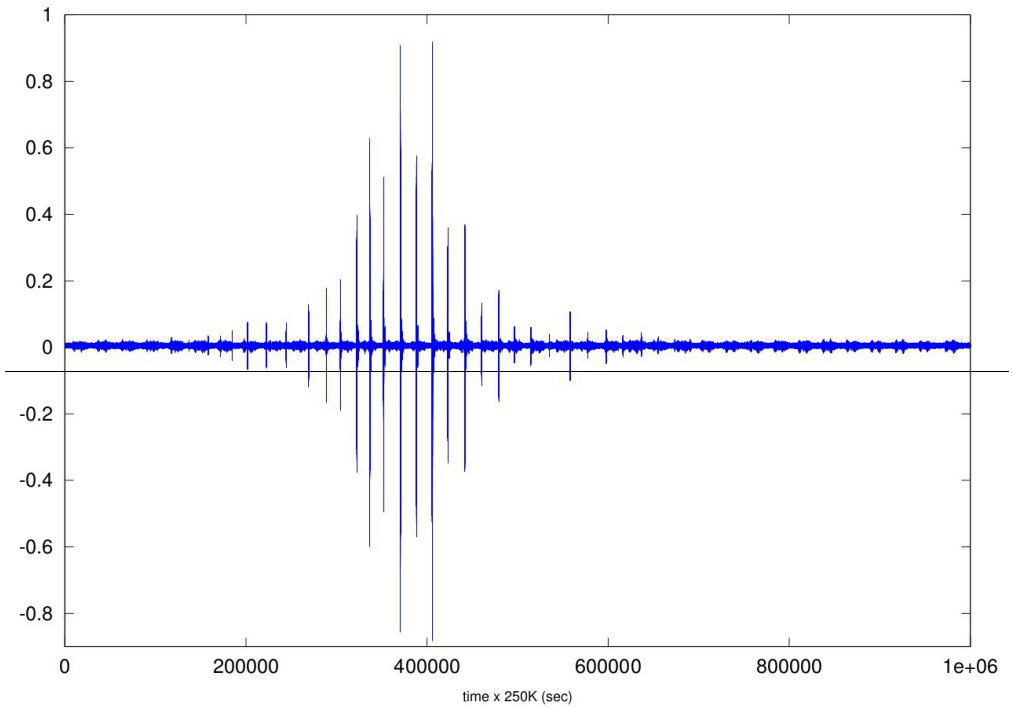


Fig: The time-amplitude representation of this bat sonar sample. The nearest point of approach (NPA) to the microphone corresponds to the highest amplitude (near the sample #400K). Before NPA, the bats flies in direction to the microphone, after NPA the bat emits in the opposite direction.

The communicative function of echolocation calls is still largely unstudied, especially in the wild. The vocal signatures encoding social information in echolocation calls has not been up to now well studied. The authors found pronounced vocal signatures encoding sex and individual identity : free- living males discriminate approaching male and female conspecifics solely based on their echolocation calls. Males always produced aggressive vocalizations when hearing male echolocation calls and courtship vocalizations when hearing female echolocation calls; hence, they responded with complex social vocalizations in the appropriate social context. Advanced statistics may reveal other dependences into biosonar sequences...

1 http://sabiod.univ-tln.fr/nips4b/media/MyopterusECORx_091612_223528COMP1.wav

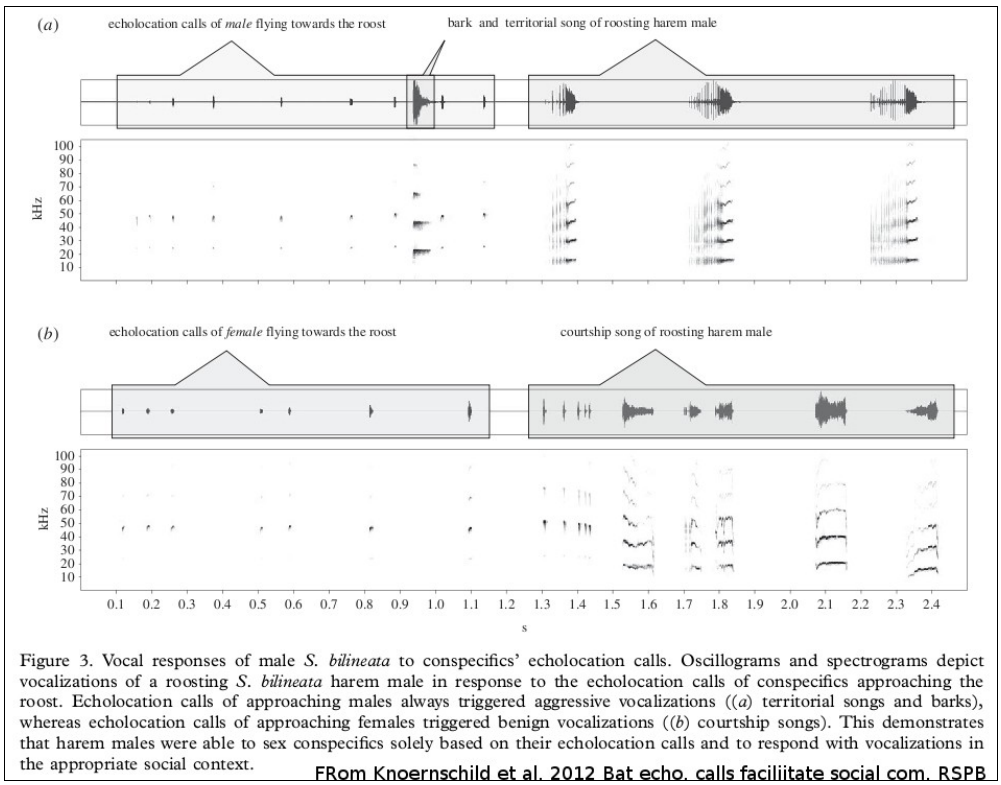


Figure 3. Vocal responses of male *S. bilineata* to conspecifics' echolocation calls. Oscillograms and spectrograms depict vocalizations of a roosting *S. bilineata* harem male in response to the echolocation calls of conspecifics approaching the roost. Echolocation calls of approaching males always triggered aggressive vocalizations ((*a*) territorial songs and barks), whereas echolocation calls of approaching females triggered benign vocalizations ((*b*) courtship songs). This demonstrates that harem males were able to sex conspecifics solely based on their echolocation calls and to respond with vocalizations in the appropriate social context.

FRom Knoernschild et al. 2012 Bat echo. calls faciliitate social com. RSPB

Fig: The sonar sequence dependencies illustration (from [Kno 2012]).

Organizers:

Marie Trone (Valencia Coll. USA) and Hervé Glotin (UTLN & IUF)
We thank Cyberio SA for the bat recordings, and MASTODONS SABIOD for its support for the Tursiops recordings

References NeuroSonar :

- [Ryabov 2011] Some Aspects of Analysis of Dolphins'Acoustical Signals, Ryabov (2011), Open Journal of Acoustics, 2011, 1, 41-54 doi:10.4236/oja.2011.12006 Published Online (http://www.SciRP.org/journal/oja)
- [Kno 2012] Knörnschild M, Jung K, Nagy M, Metz M, Kalko EKV (2012) Bat echolocation calls facilitate social communication. Proceedings of the Royal Society of London B 279: 4827-4835.
- [Glotin 2008] Glotin H., Caudal F., Giraudet P., 'WHALE COCKTAIL PARTY: REAL - TIME MULTIPLE TRACKING AND SIGNAL ANALYSES', int. Journal Canadian Acoustics, V.36(1), ISSN 0711-6659, DEMO @ sabiod.org
- [Paris 2013] S. Paris and al.(2013) Physeter catodon localization by sparse coding, ICML for Bioacoustics workshop, 2013
- [Benard 2010] Benard F., Giraudet P., Glotin H. (2010) 'Whale 3D monitoring using astrophysic NEMO ONDE 2m wide platform with state optimal filtering by Rao-Blackwell Monte Carlo data association', int. jour. App. Acoustics, (71)

1.5 Acknowledgements

We thank the members of the organizing committee H. Glotin, T. Artières, R. Balestriero, Y. Doh, M. Bartcus for their continuous effort for NIPS4B.

Chapter 2

Natural Neural Bioacoustic Learning

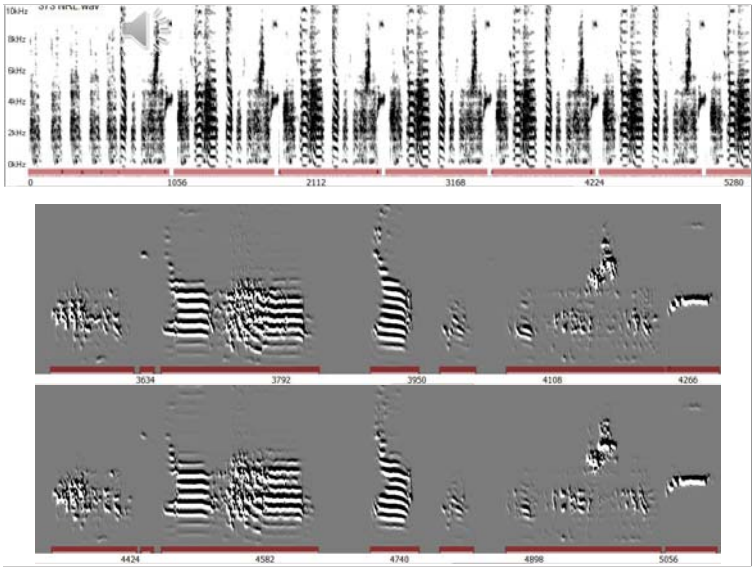
2.1 Physiological brain processes that underlie song learning

Ofer Tchernichovski- Hunter College - CUNY, NY, USA

Human language, as well as birdsong, relies on the ability to imitate vocal sounds and arrange them in new sequences. During developmental song learning, the songbird brain produces highly variable song patterns, which allow vocal exploration to guide learning. Tracking song development continuously show that exploratory variability is regulated in fine time scales, such that each song element becomes less variable independently when approaching the target (adult tutor) song. Therefore, multiple localized reinforcement-learning processes can explain how the bird learn to match specific song elements. However, we found that vocal exploration alone cannot explain how birds learn to match vocal combinatorial sequences. Combining an experimental approach in zebra finches with an analysis of natural development of vocal transitions in Bengalese finches and pre-lingual human infants, we found a common, stepwise pattern of acquiring vocal transitions across species. Results point to a common generative process that is conserved across species, suggesting that the long-noted gap between perceptual versus motor combinatorial capabilities in human infants may arise partly from the challenges in constructing new pairwise vocal transitions. Therefore, learning vocal sequences is likely to be constraint by a neuronal growth process, perhaps of establishing connections between representations of song gestures.

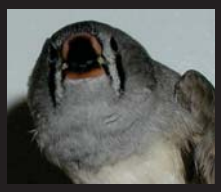
Neuronal mechanisms of vocal learning in songbirds and human

Ofer Tchernichovski
Dept. of Psychology, Hunter College
The City University of New York

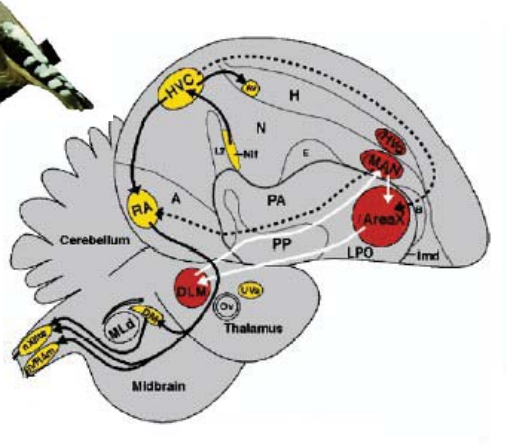


Song imitation

Developmental Song Learning

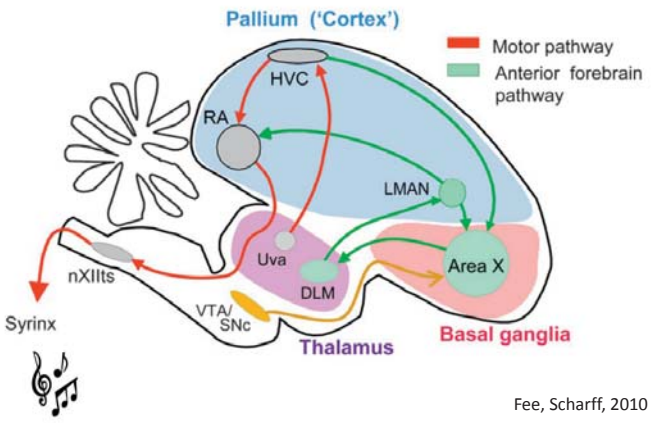


How birds sing



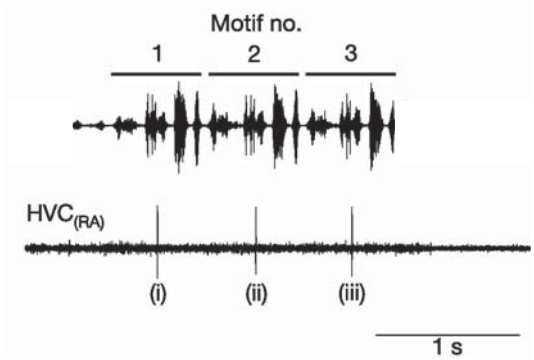
Jarvis & Nottebohm 1998

Songbird neuroanatomy



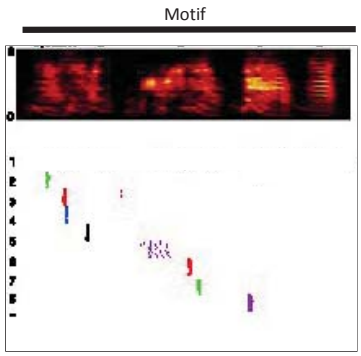
Fee, Scharff, 2010

Each HVC premotor neuron is a clock

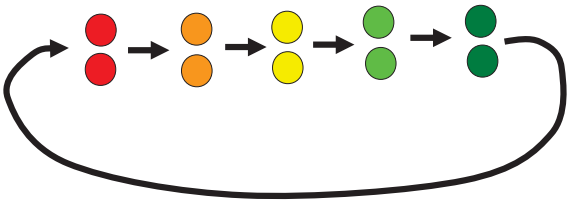


From: Hahnloser RHR, Kozhevnikov AA, Fee MS, NATURE 2002

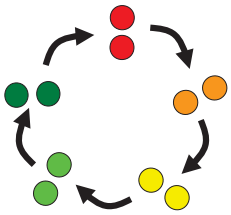
An array of clocks in HVC
Explicit representation of song time

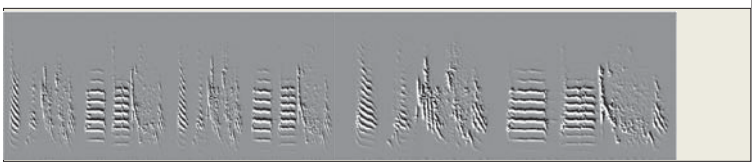


Hahnloser RHR, Kozhevnikov AA, Fee MS, NATURE 2002



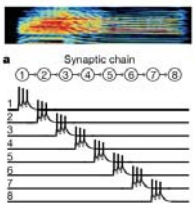
Here is our "music box"





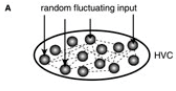
HVC synfire chain and how they may be learned

There good evidence for sysnfire chains of bursting neurons in HVC which may fire in sequence during a stereotyped syllable, e.g Long, Jin, Fee 2010

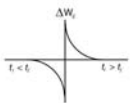


Such HVC synfire chains can be learned with (Fiete, Senn, Wang, Hanloser 2010)

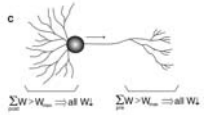
correlated external inputs



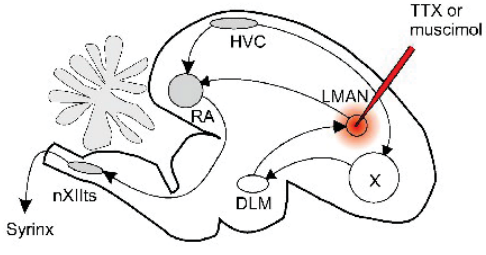
STDP



heterosynaptic competition

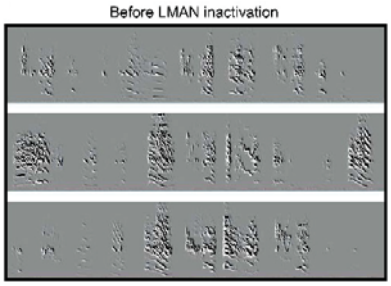


Variability in song production



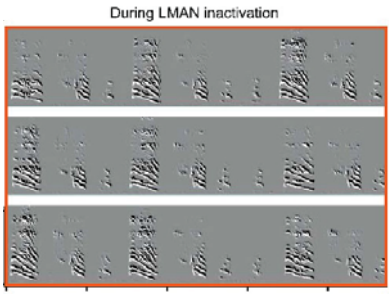
From: Bence P. Ölveczky, Aaron S. Andalman, Michale S. Fee, PLOS 2005

Learning by experimentation



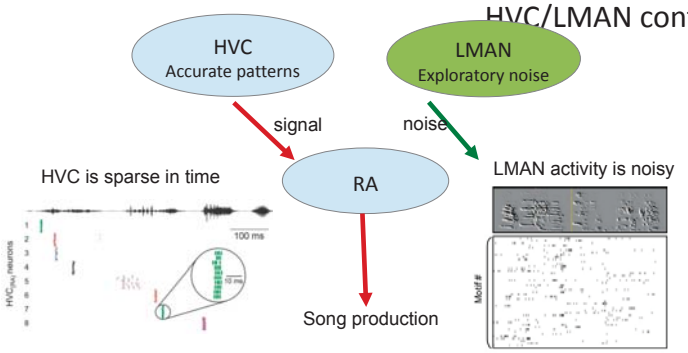
From: Bence P. Ölveczky, Aaron S. Andalman, Michale S. Fee, PLOS 2005

Learning by experimentation



From: Bence P. Ölveczky, Aaron S. Andalman, Michale S. Fee, PLOS 2005

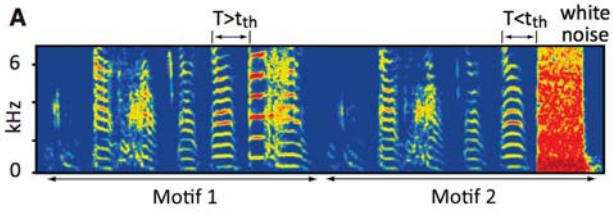
Common final p
HVC/LMAN contr



Fee, Goldberg, 2011

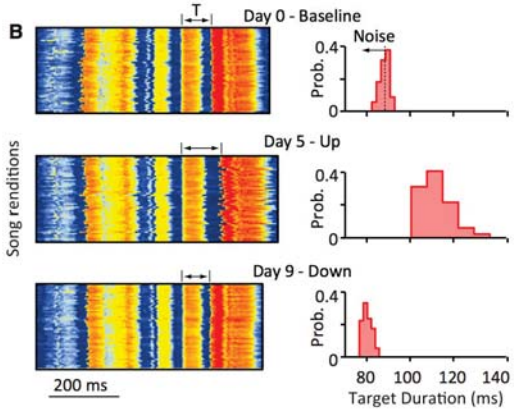
$$SNR = \frac{HVC}{LMAN}$$

Negative reinforcement during singing, targeting a specific svllable:

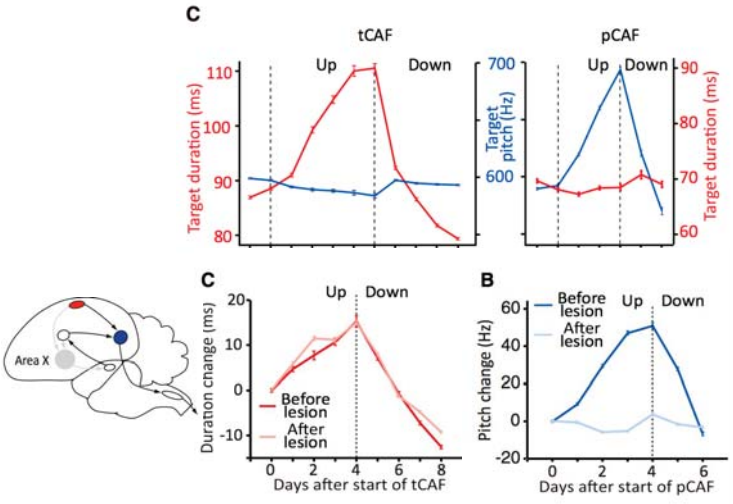


Ali … & Olveczky, Neuron, Dec 2013

The production variability in duration is assessable for



Ali … & Olveczky, Neuron, Dec 2013



Ali … & Olveczky, Neuron, Dec 2013

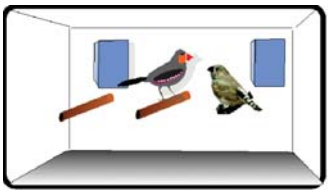
So, how to bird imitate their song?

- Several people think about vocal learning as an instance of reinforcement learning algorithm
- I will show evidence that birds control SNR adaptively as learning progresses: this will tell us at what time scales RLA operates
- And then I will show that at the level of acquiring combinatorial abilities – we have a problem…

Experimentally controlled song learning

days 7-30:
Young raised by their mother

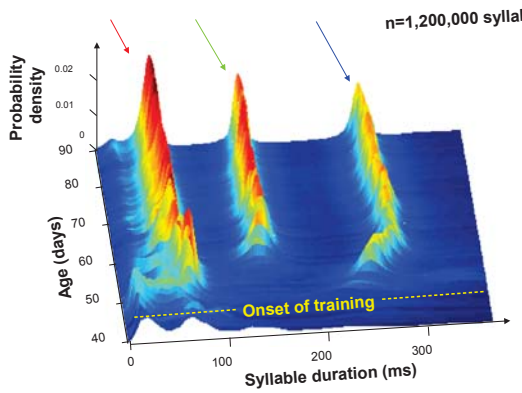
day 31-:
Social and acoustic isolation

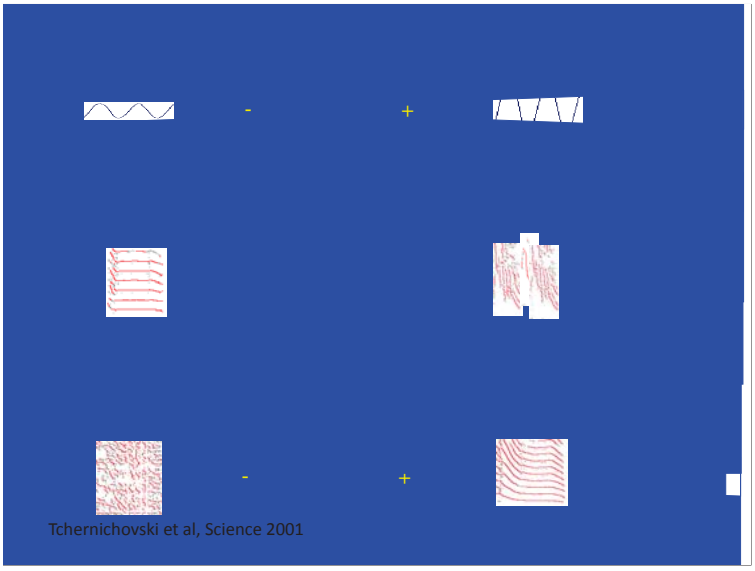


day 43: Start training

Sound Analysis Pro (GNU public license)
http://SoundAnalysisPro.com

Analysis of an entire song learning

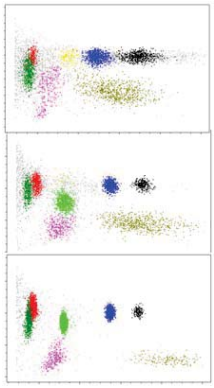




Start	Duration	Mean Amp	Mean Pitch	Mean Entropy	Mean FM	Mean Continuity
16454	76	0.216472968	769.9150391	-2.356431723	39.29466629	0.794104338
16571	54	0.52569139	687.6394043	-1.956387162	37.81315613	0.616944551
17000	58	0.135118335	864.5578613	-2.363121986	31.00643349	0.858065724
17189	51	0.124977574	752.3527222	-1.94250226	36.36558151	0.691144586
17761	58	0.144002378	1021.027527	-2.258356094	40.53672409	0.708231866
17873	47	0.066938281	1339.068604	-1.668018103	46.29984665	0.69986397
18051	38	0.066276349	1847.560913	-2.551876307	38.55633545	0.805839062
18092	81	0.200010121	2080.408936	-3.075473547	50.34065247	0.776402116
18219	66	0.335276693	858.1080933	-1.750756502	46.40740204	0.511499882
18536	69	0.261755675	890.3964233	-1.860459447	42.50422668	0.500995994
19446	46	0.15915972	993.3217773	-1.601477981	43.11263275	0.527124286
20405	51	0.193706796	800.2883911	-1.413753867	41.22149277	0.428571522
20644	65	0.24410592	802.0982666	-1.589150429	39.50386429	0.429761887
20729	61	0.166723967	901.6841431	-1.771348119	47.49161148	0.556119919
20847	51	0.198818251	852.6430664	-1.053611994	48.11198425	0.44106108
23287	68	0.178408563	784.8914185	-2.134843588	41.99195862	0.656920671
24243	70	0.185866207	990.8589478	-2.562700748	39.49663925	0.763919473

Show song development movies

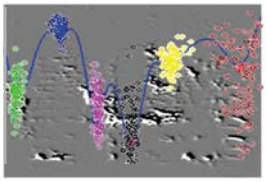
What's the scope of vocal exploration ?
Can the bird control exploratory noise locally?



Deregnaucourt et al, Nature 2005

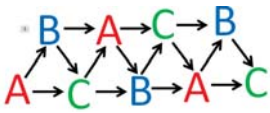
Is vocal exploration 'l controlled'?

Can the bird confine exploratory variability to parts of the song that need improvement?



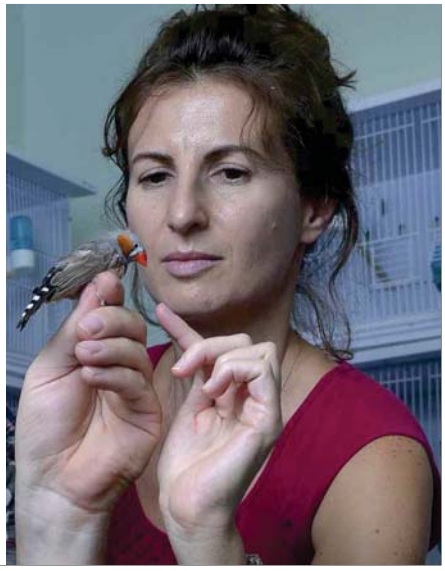
How is song syntax learned?

Can the bird explore at the syntax level?

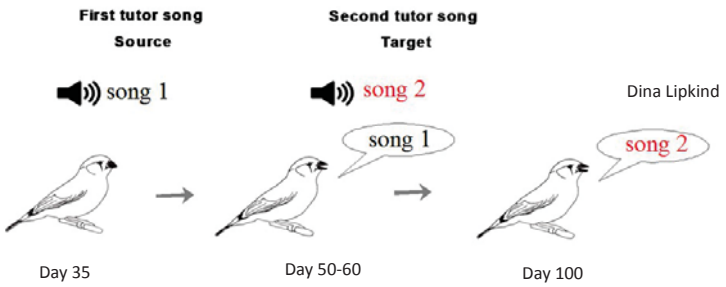


Altered-target training technique

Dina Lipkind



Altered target training



Design source and target songs to present the bird with a specific imitation task

Can the bird control vocal exploration locally?

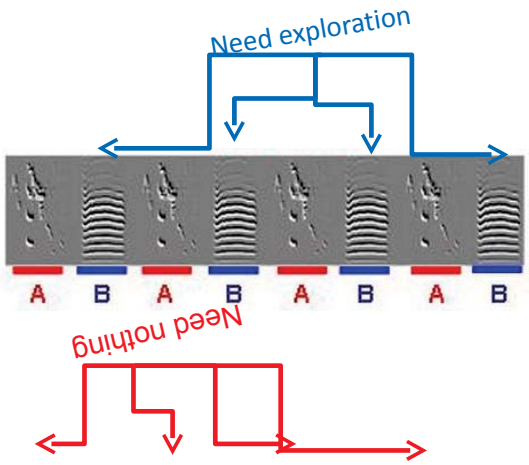
ered target training

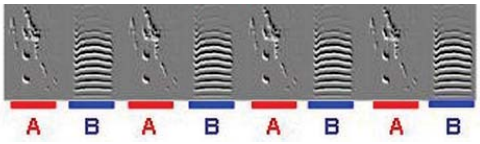
Dina Lipkind

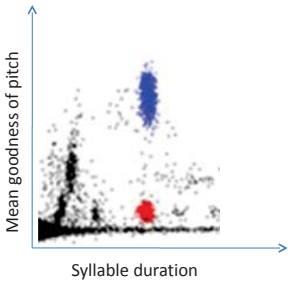
Conflicting demands

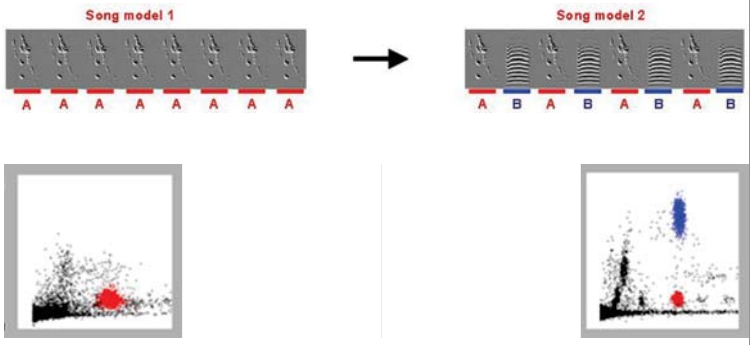
Primoz Ravbar

Dina Lipkind



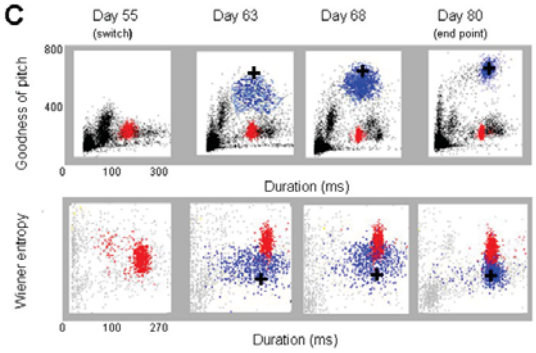




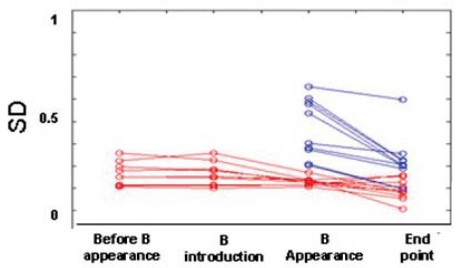


Day 55

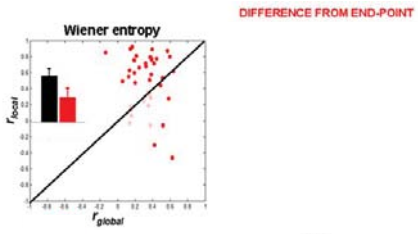
Day 90



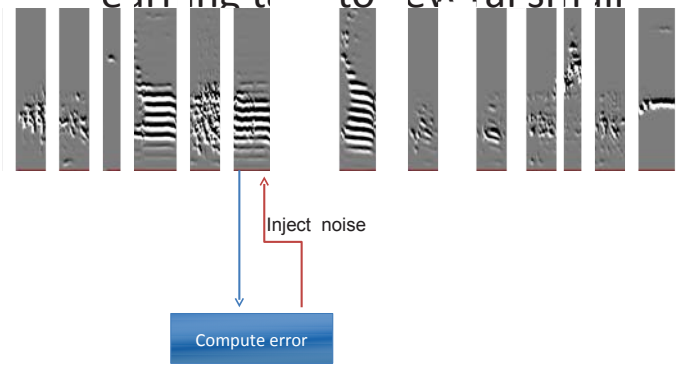
The high variance of the newly learned syllable does not leak to other sounds



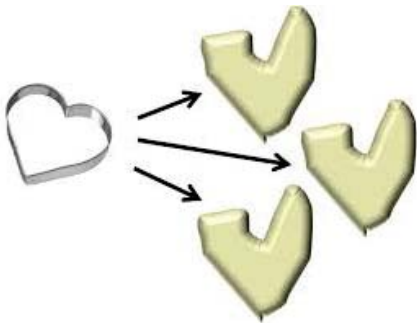
At fine time scales of vocal gestures variability correlates better with local similarity to the target song



An explanation: the bird partition the learning task to several small



But how is the song compared to the model template?



But how is the song compared to the model template?

training: ABC-ABC -> AC+B-AC+B

Learning:

ABC-ABC -> ABC+-ABC+ -> AC+B-AC+B

Conclusions

- Birds seem to be capable of locally controlling exploratory variability and confine vocal exploration to parts of the continuous actions that need improvement
- This may suggest that the continuous singing action is learned by virtually segmenting the song to short units (several ms) that can be learned independently
- In each segment, variability is gated by local error
- The bird can compare song elements to its model memory template out of context (not published yet)

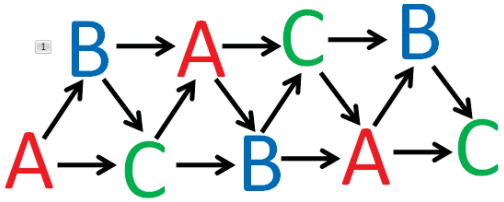
Ravbar et al, Journal of Neuroscience, 2012

What about sequence rearrangement?

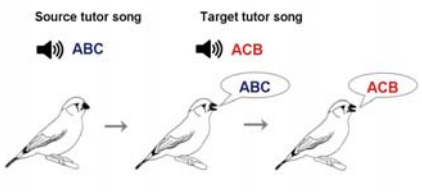
Dina Lipkind
Hunter College, CUNY

Gary Marcus
NYU

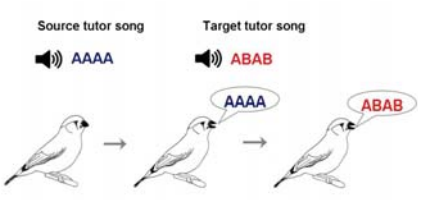
Doug Bemis
NYU



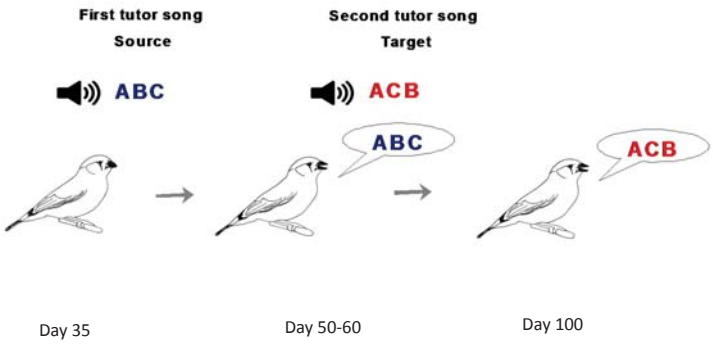
Rearrange syllable order
(permutation task)



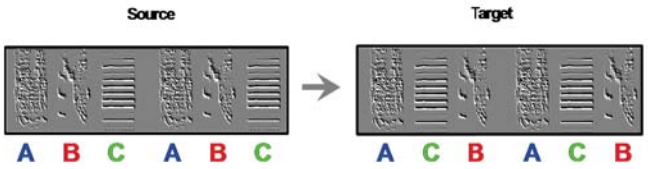
Insert a new syllable into
a string
(insertion task)



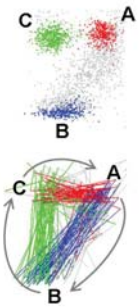
Rearrangement of syllables



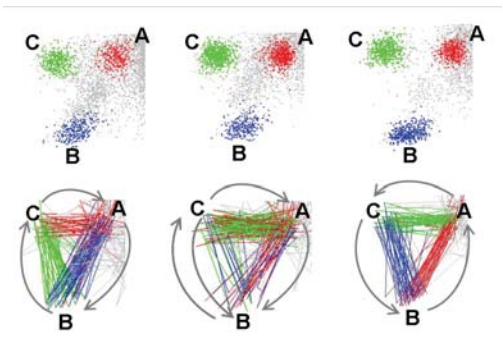
Day 35 Day 50-60 Day 100



Can the bird rearrange syllables at all?

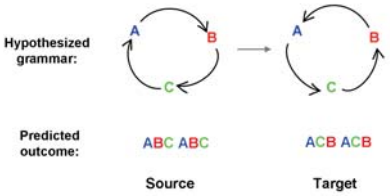


YES: in more than 90% of the cases

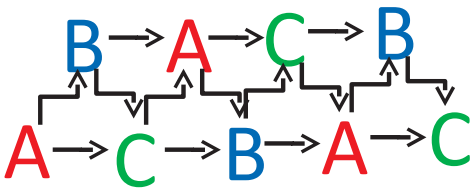


Hypothesis 1: Direct syllable swap:

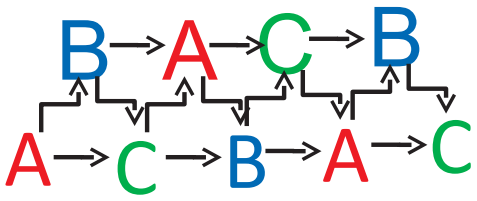
ABC ABC ABC

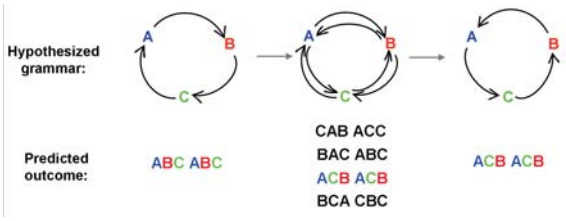


Hypothesis 2: Randomize & select

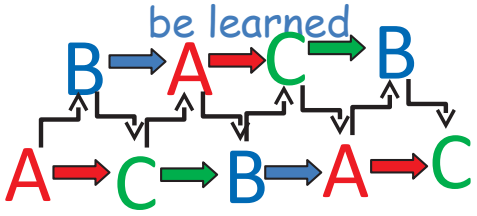


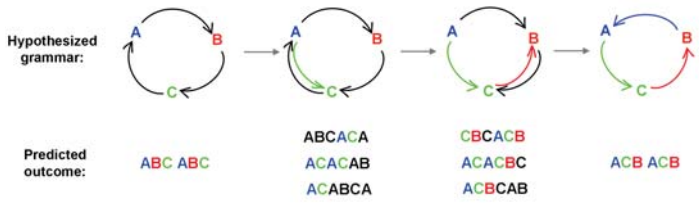
Hypothesis 2: Randomize & select





Hypothesis 3: New transitions must be learned

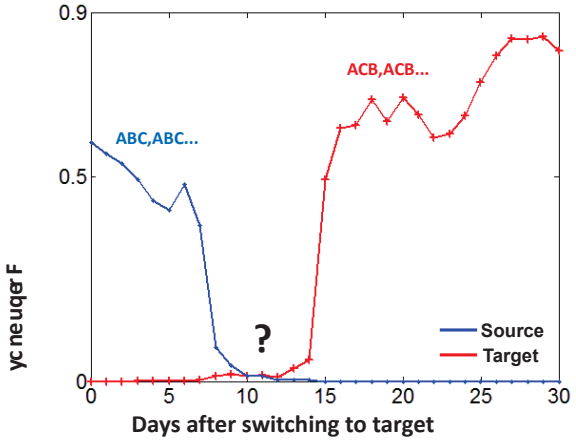




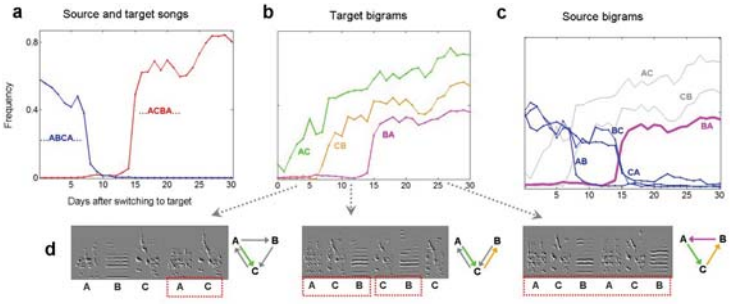
These three hypotheses are not necessarily the appropriate way of thinking about the problem:

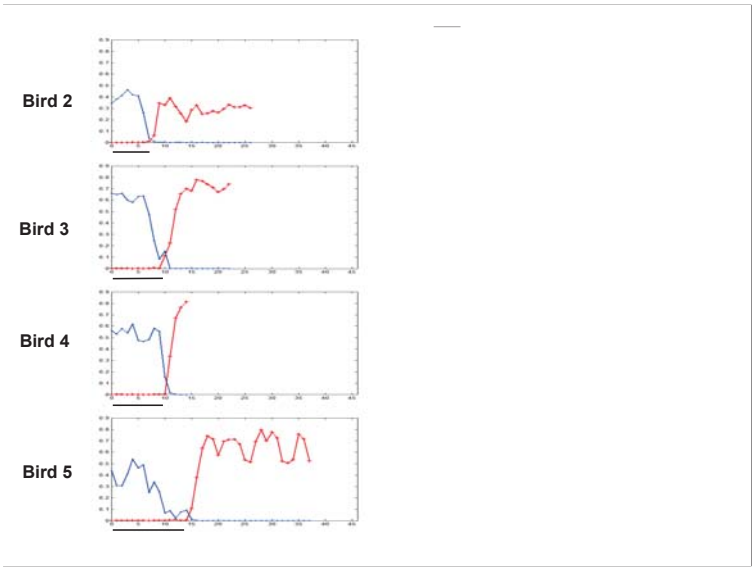
- **We do not know what are the units that the bird samples and learns**
- **The bird appears to accurately represent the frequencies of transitions**
- **Therefore, one should distinguish between the capability of performing a certain transition, and of adjusting its frequency to match a model**

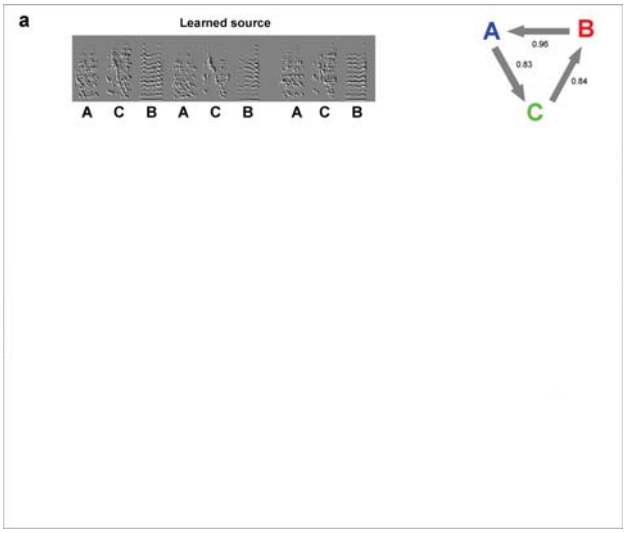
The transition from source to target syntax

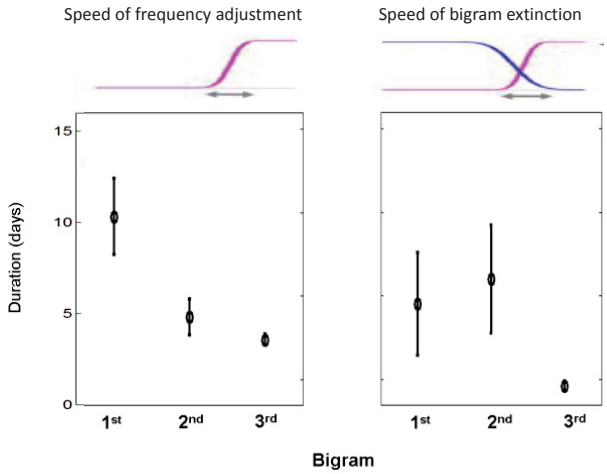


New syntax is gradually assembled from individual pair wise syllable transitions

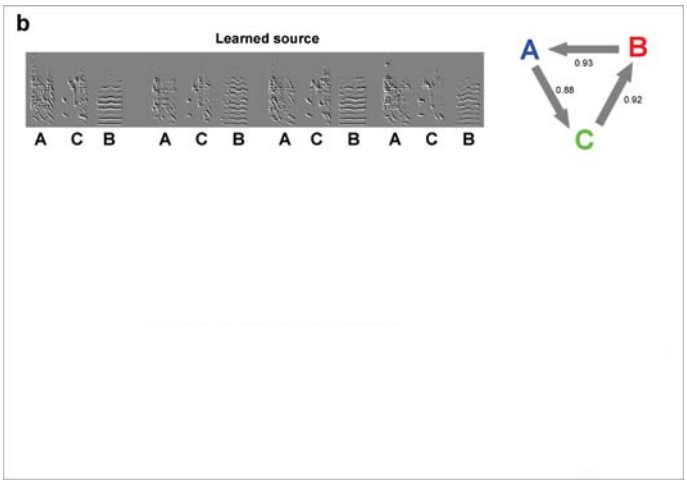


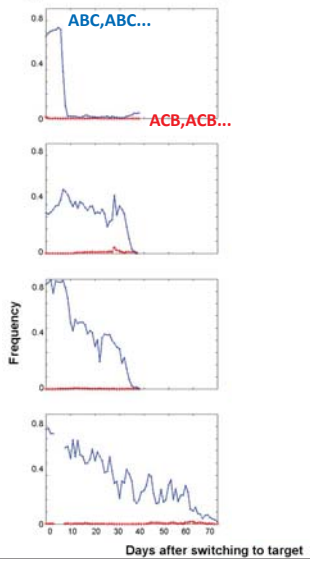


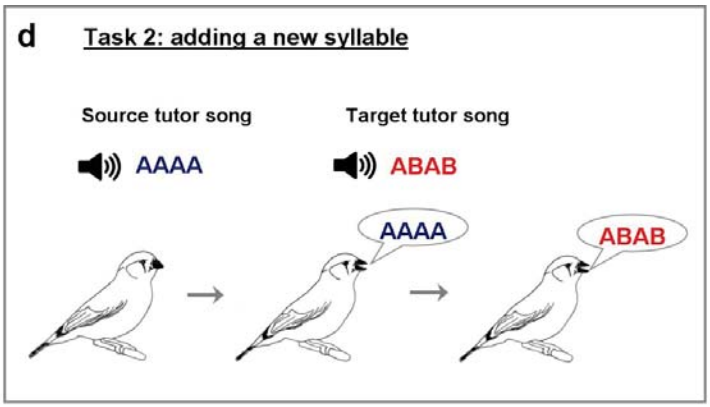




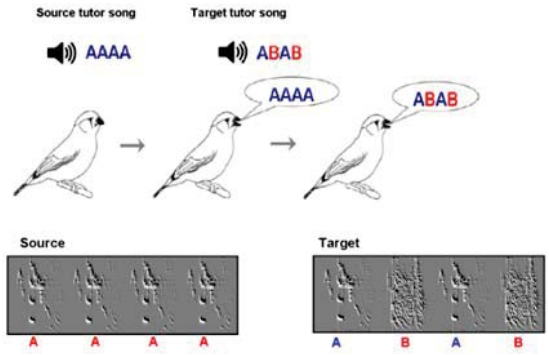
What about birds that failed to imitate the target?







Question: how AAAA✹ ABAB syntax should be learned?



Answer: In two stages

- **If A✹B is learned first:**

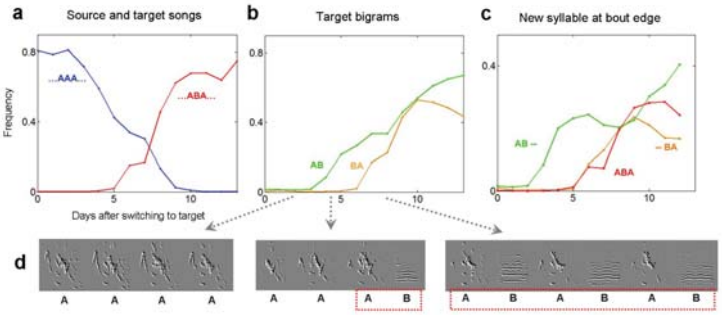
first - …AAAAB-stop {AB, but not BA}

then- …ABAB… {AB & BA}

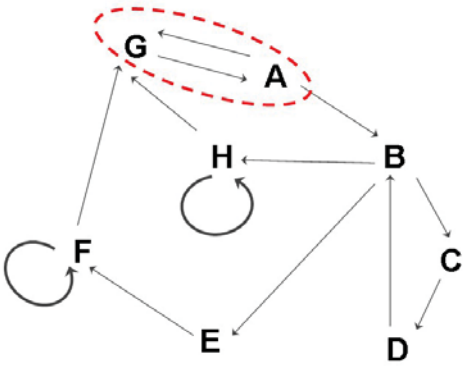
If B✹A is learned first:

First - stop-BAAAA… {BA, but not AB}

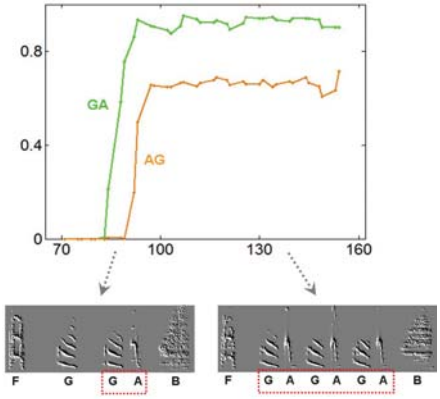
then: …ABAB… {AB & BA}

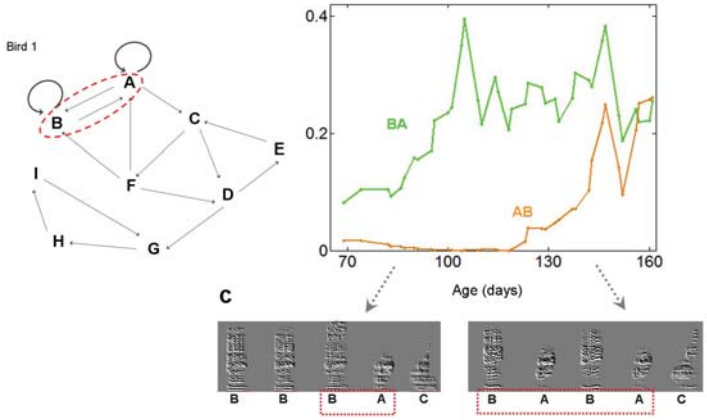


Bengalese finches song syntax



The ontogenetic origin of bi-directional transitions





What about early speech development?

Doug Bemis
NYU

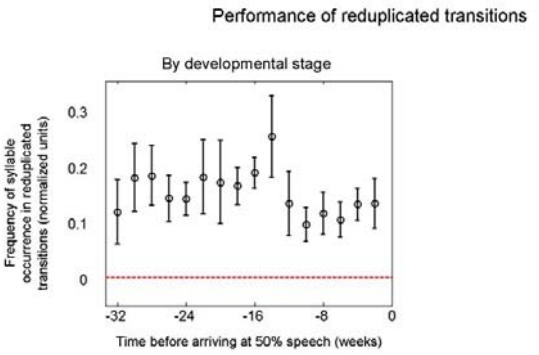
Gary Marcus
NYU

- Classical studies identified a transition from predominantly reduplicated to variegated babbling
- These results have not been reliably replicated in later studies

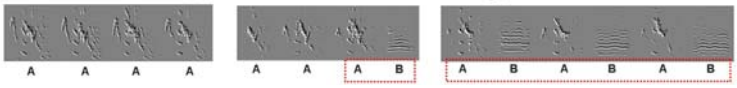
Hypothesis: babbling development is shaped by a stepwise process of acquiring transitions to and from a syllable

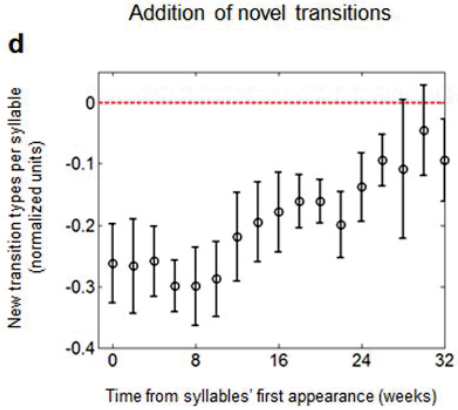
- Syllable types appear one by one during babbling development
- We therefore need to align the developmental data according to appearances of syllable type (time zero = the first time we see the syllable)

Stepwise transition to variegation



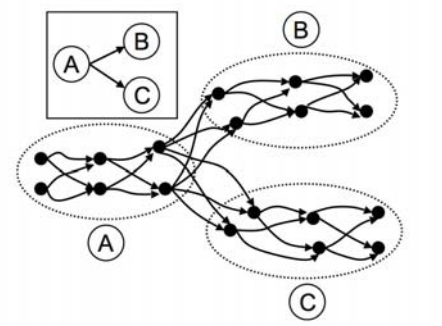
Edge effect:





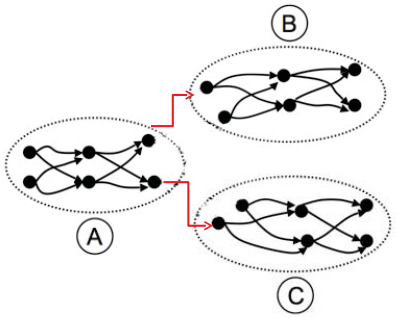
Summary

Head-to-tail connections of chains of neuronal activity (endpoint)

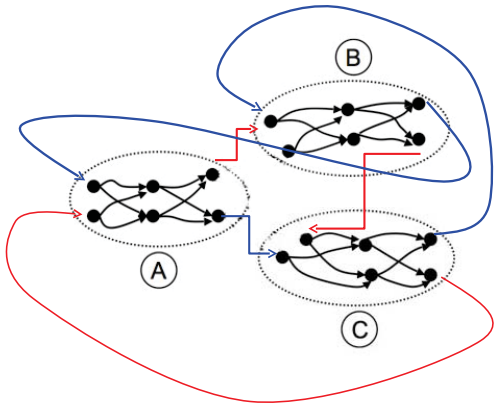


Dezhe Z. Jin: **Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC (Phys. Rev. 2009)**

The starting point may be sparse:



Rearranging is also sparse:



A simple model: A slow growth process of constructing chains

Adding elements:

A → B ⇒ A → B → C

Removing elements:

A → B → C ⇒ A → B

A simple model: A slow growth process of constructing chains

Swapping elements:

A → B → C ⛔ A → C → B

Inserting elements into chains:

A → A → A ⛔ A → B → A

Summary:

- In zebra finches: a balanced process of adding and removing transitions
- In Bengalese finches (complex song): more additions, resulting in branching chains
- In human infants: many (but slow) additions of vocal transitions, leading eventually to an all-to-all network

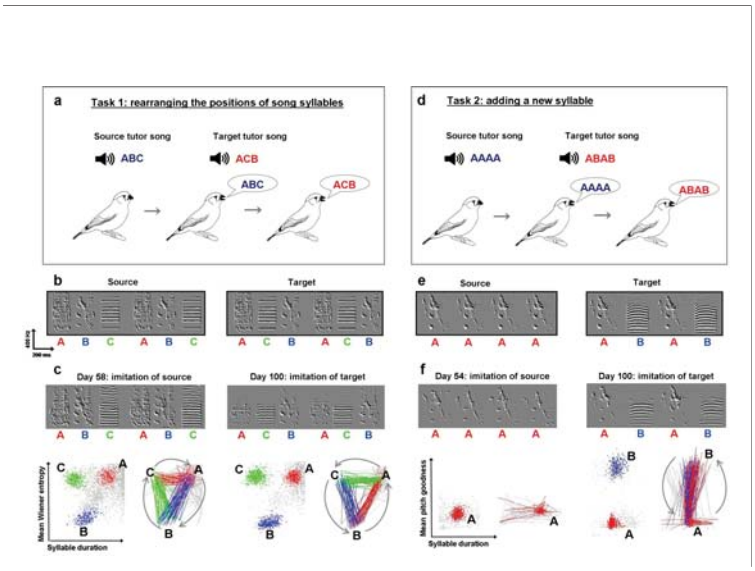

Dina Lipkind
Hunter College

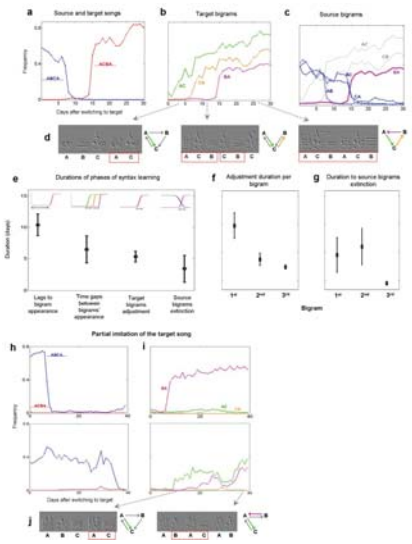

Primoz Ravbar
City College & GC

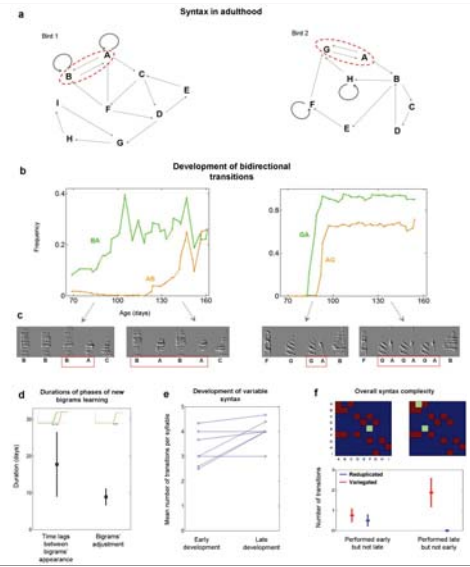

Doug Bemis
NYU

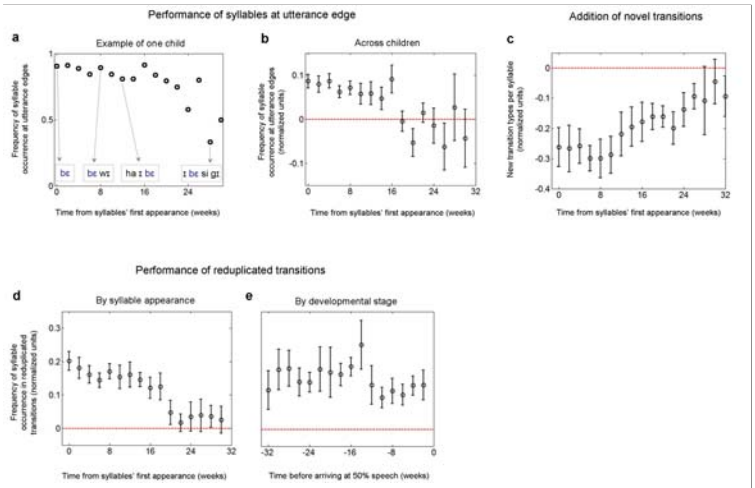

Gary Marcus
NYU

Supported by grants from NIH(NIDCD) & NSF

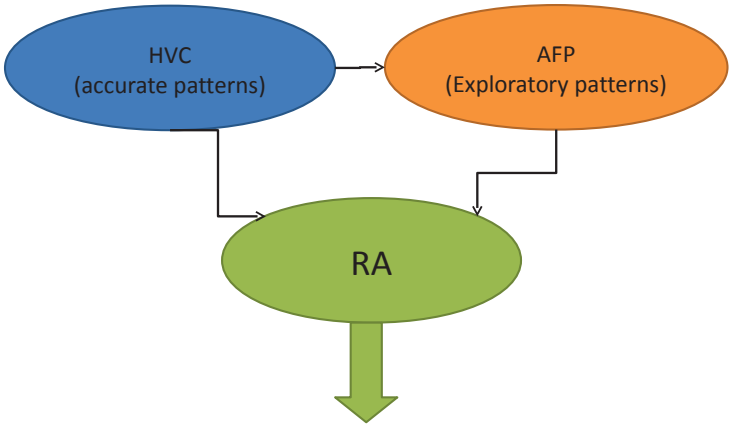








Common final path to RA ("motor cortex")



An example from speech development

- **Grapes = ANAVIM (Hebrew)**

First production: VIM-ANA
Then: ANA-VIM-ANA
Finally: ANA-VIM-~~ANA~~

Asymmetry between perceptual and production combinatorial capabilities

- A long delay between infants' precocious ability to perceive complex grammars, and their limited ability to produce vocal sequences

Gary Marcus
NYU

What are the evidence for combinatorial constraints in babbling?

- Classical studies identified a transition from predominantly reduplicated to variegated babbling
- These results have not been reliably replicated in later studies

The development of combinatorial capacity

- **H0: Vocal combinatorial capacity is the starting point of vocal learning**
- **H1: Vocal combinatorial capacity is gradually acquired via an elaborated generative process**

A comparative study:

Three species, whose vocal behavior span a broad range of combinatorial capabilities:

Zebra finch:
sings mostly linear sequences

Bengalese finch:
songs include branching sequences

Pre-lingual human infants:
an extraordinary vocal combinatorial capabilities

Lipkind et al, Nature 2013

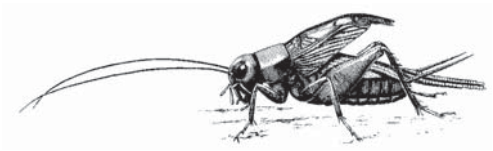
- In zebra finches we tested how they solve different types of combinatorial tasks
- In Bengalese finches, we explored the ontogenetic origin of combinatorial plasticity, looking at specific vocal transitions
- In human infants we examined statistically, how vocal diversification of thousands vocal transitions comes about

2.2 Neuroethology of hearing in crickets: embeded neural process to avoid bat

Gérald Pollack- McGill University, Montréal, CA

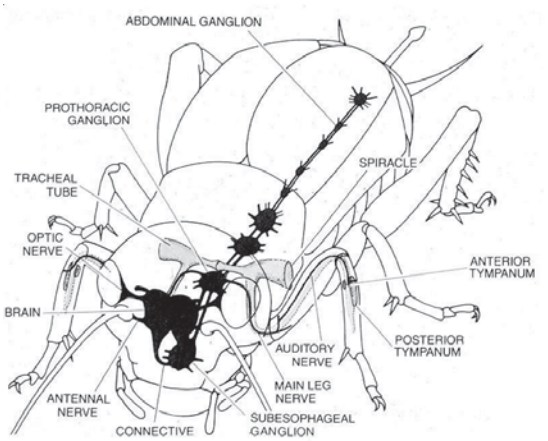
Many behavioral studies on crickets have identified the relationships between signal structure and behavioral effectiveness, and the neural basis for sound reception and analysis. We'll present the behavioral studies on signal recognition; relationships between stimulus structure and behavioral effectiveness; roles of sound frequency, stimulus temporal structure; positive and negative phonotaxis to cricket-like and bat-like signals, respectively Early auditory processing: separate channels for processing mate-attraction signals and predator-derived signals (ultrasound); temporal response properties of receptor neurons and first-order interneurons. And descending brain neurons: conveying the results of processing in the brain to motor centers that control behavior.

Bugs and bats: Neural analysis of behaviorally relevant sounds in crickets.

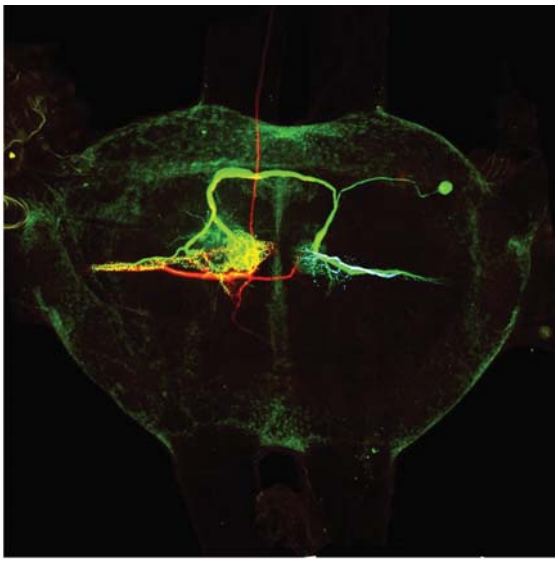


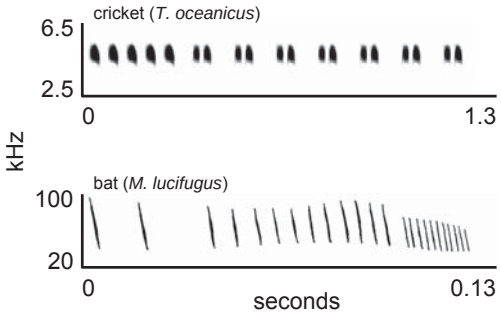
Drawing by Dan Otte

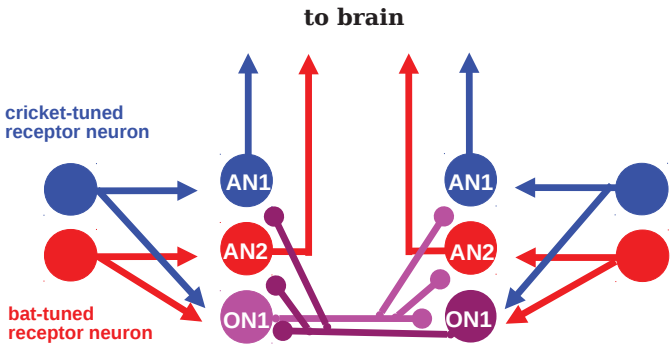
Gerald Pollack
Dept. of Biology, McGill University

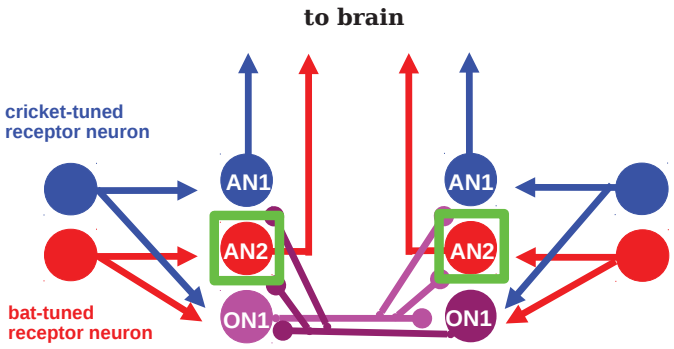


Huber & Thorson (1985) Scientific American 253:60-68

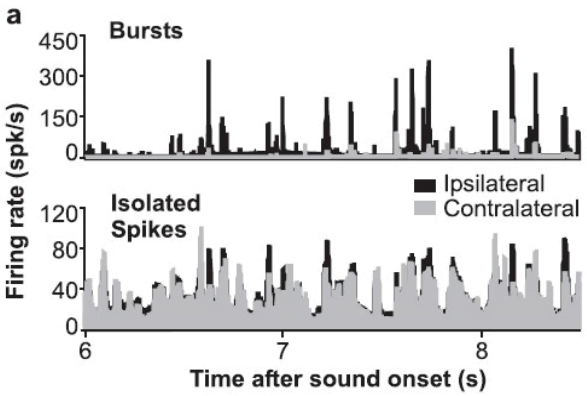






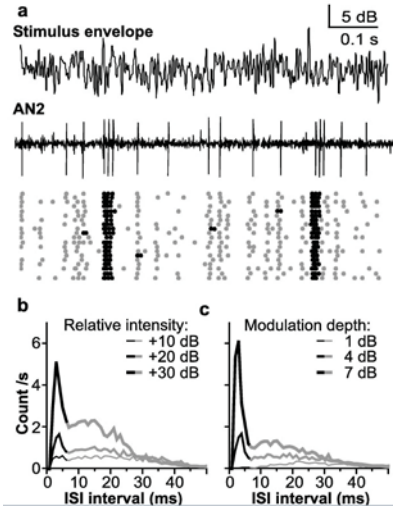


Bursts, and only bursts, code for sound direction



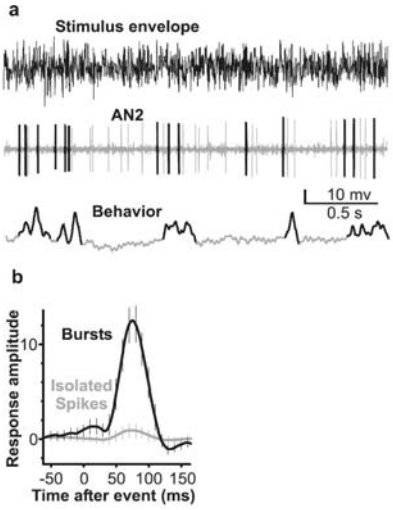
Marsat & Pollack (2006) J Neurosci 26:10542-7

AN2 produces bursts; brief episodes of high-rate firing



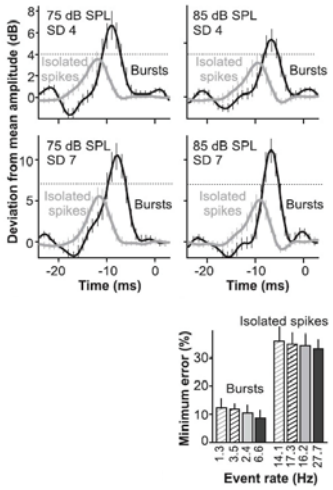
Marsat & Pollack (2006) J Neurosci 26:10542-7

Bursts, and only bursts, elicit steering responses

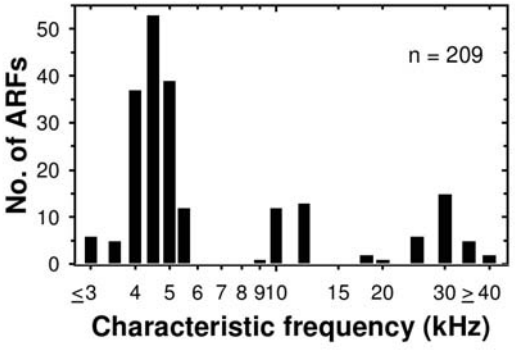


Marsat & Pollack (2006) J Neurosci 26:10542-7

Bursts accurately detect conspicuous increases in amplitude

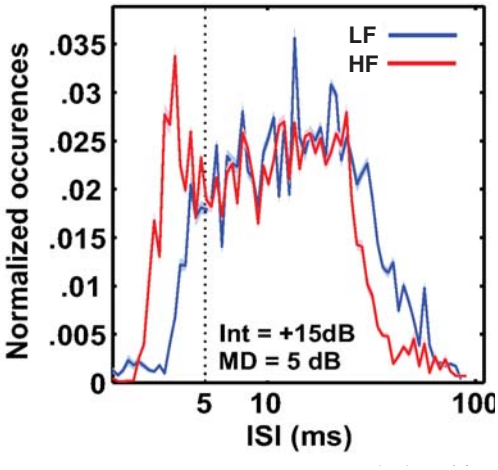


Distribution of best frequencies

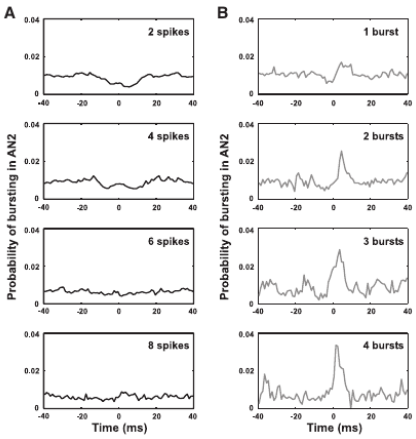


Imaizumi and Pollack (2001) JASA 109:1247-1260

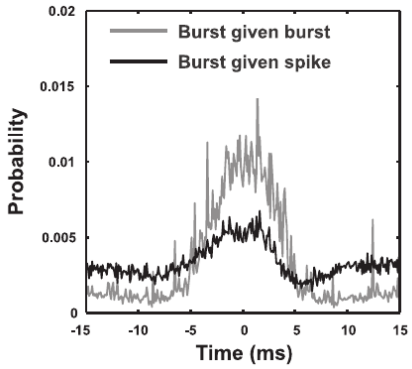
Bursting in HF-tuned receptors



"Distributed" bursts vs. real bursts



Synchronous receptor bursts



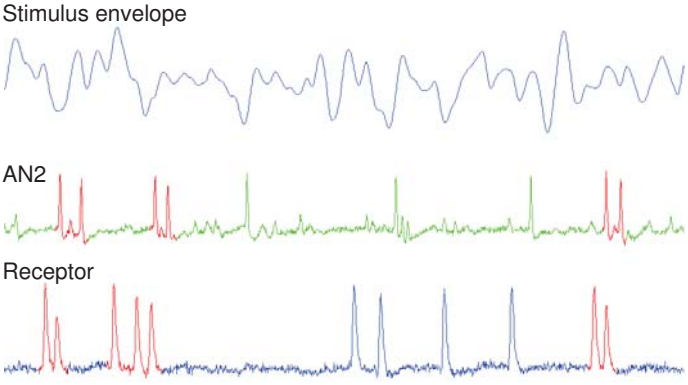
Some outstanding questions

Why do ultrasound-tuned receptors, but not cricket-tuned receptors, burst?

What are the synaptic mechanisms linking receptor-bursts and AN2 bursts?

What are the downstream effects of AN2 bursts in brain neurons?

Receptor bursts and AN2 bursts are near-coincident

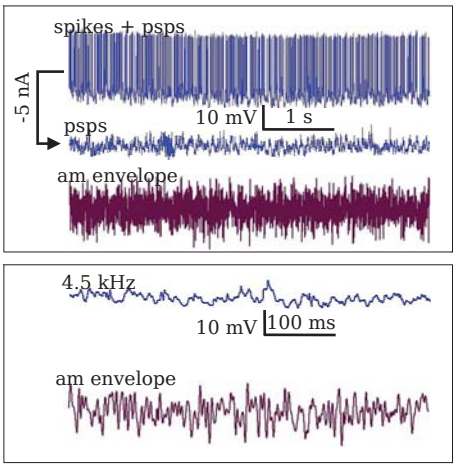


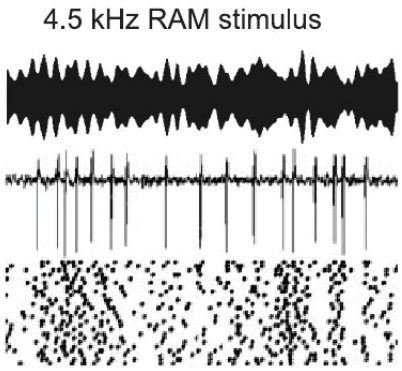
Gary Marsat

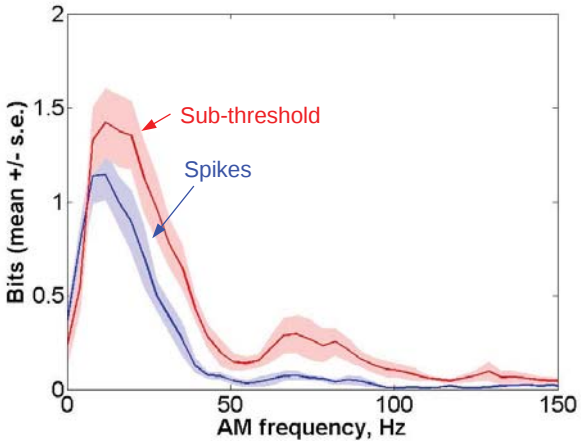
Patrick Sabourin

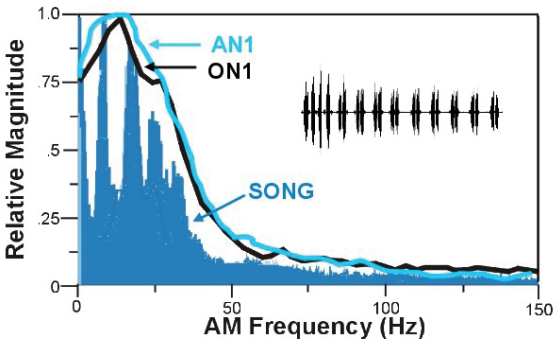
NSERC CRSNG
CIHR IRSC

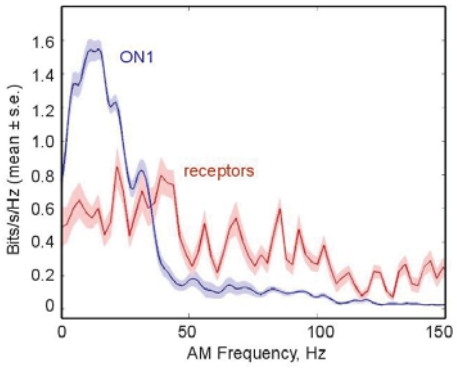
Frequency-specific temporal coding is apparent at the level of sub-threshold variations in membrane potential



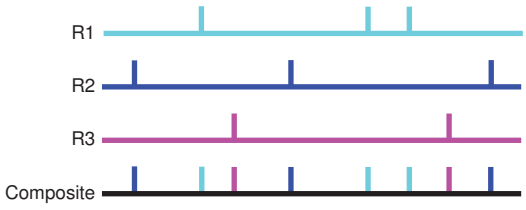




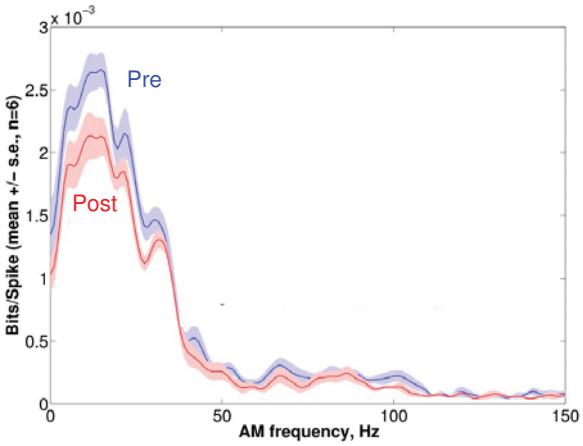


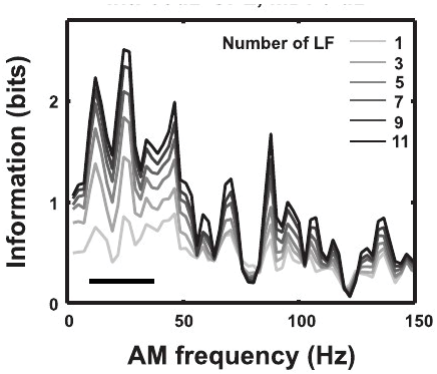


Modeling coding by receptor-neuron populations

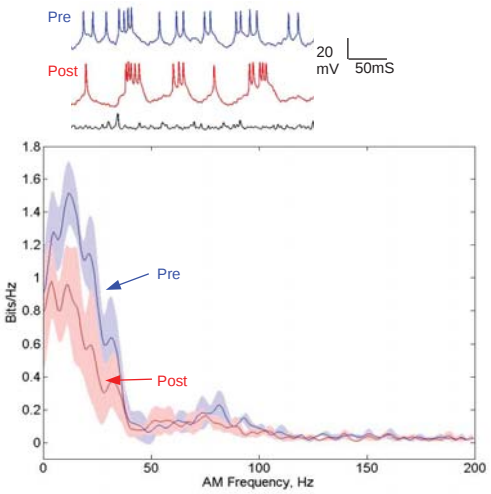


BAPTA

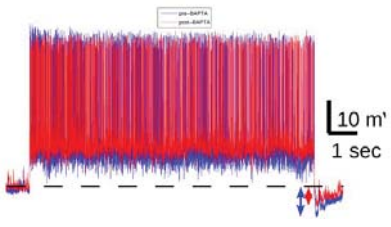




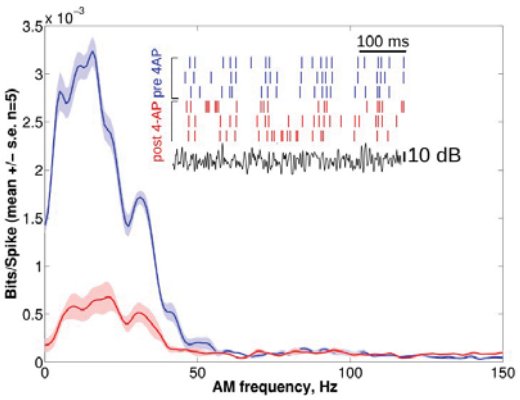
TEA



Ca-activatedK current?



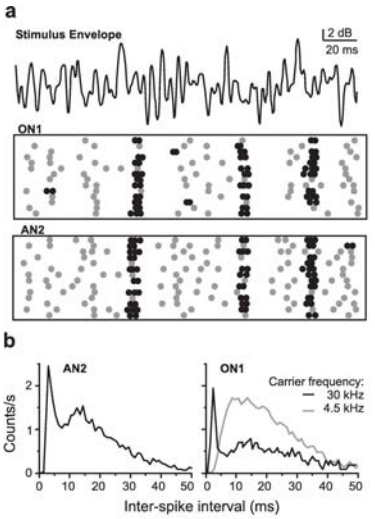
4-AP



Summary

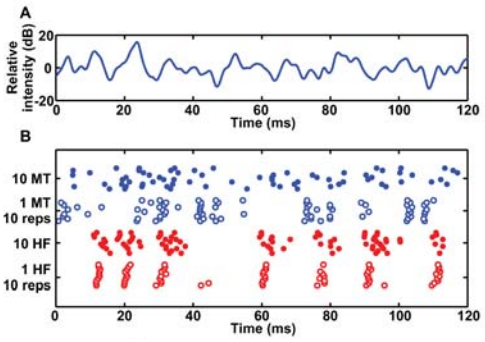
- Adaptive, frequency-specific temporal coding
  - For ultrasound: bursting is elicited by conspicuous increases in ultrasound amplitude
  - For cricket song: accurate encoding of species-typical rates of amplitude modulation
- In both cases, timing of receptor-neuron spikes seems to determine stimulus encoding by interneurons
- Intrinsic properties of interneurons appear to play little role (so far)

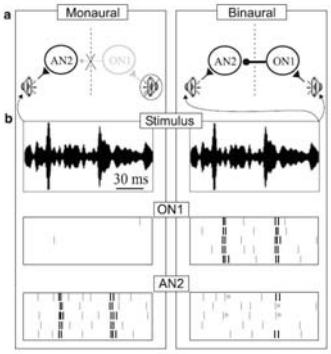
ON1 also produces bursts, but only for ultrasound stimuli



Marsat & Pollack (2007) J Comp Physiol

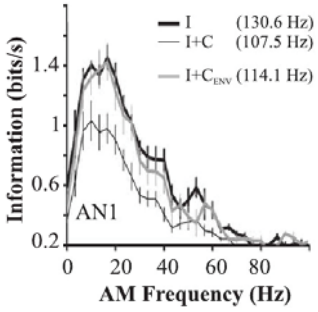
Low redundancy of responses of low-frequency receptors





* AN2 bursts eliminated by contralateral inhibition

Marsat & Pollack (2007) J Comp Physiol A 193:625-633



Marsat & Pollack (2005) J Neurosci 25:6137-6144

Chapter 3

Representation for Bioacoustics

3.1 Dynamic Time Warping and Gaussian Process Multinomial Probit Regression for Bat Call Identification

Vassilios Stathopoulos
Department of Statistical Science
University College London
London, WC1E 6BT
v.stathopoulos@ucl.ac.uk

Veronica Zamora-Gutierrez
Department of Zoology
University of Cambridge Cambridge, CB2 3E
vz211@cam.ac.uk

Kate Jones
Centre for Biodiversity and Environment Research
Dept. Genetics, Evolution and Environment
University College London
London, WC1E 6BT
kate.e.jones@ucl.ac.uk

Mark A. Girolami
Department of Statistical Science
University College London
London, WC1E 6BT
m.girolami@ucl.ac.uk

Abstract

We study the problem of identifying bat species from echolocation calls in order to build automated bioacoustic monitoring algorithms. We employ the Dynamic Time Warping algorithm which has been successfully applied for bird flight calls identification and show that classification performance is superior to hand crafted call shape parameters used in previous research. This highlights that generic bioacoustic software with good classification rates can be constructed with little domain knowledge. We conduct a study with field data of 21 bat species from the north and central Mexico using a multinomial probit regression model with Gaussian process prior and a full EP approximation of the posterior of latent function values. Results indicate high classification accuracy across almost all classes while misclassification rate across families of species is low highlighting the common evolutionary path of echolocation in bats.

1 Introduction

In many tropical ecosystems, bats are keystone species as they act as important pollinators, seed dispersal agents and regulators of insect populations [1]. In spite of their importance, most bat population studies in the tropics have been short term and the lack of long term bat monitoring programs is a result of their inherent difficulty. Bats produce unique sounds at frequencies that usually do not overlap with other species and most bat species have evolved species-specific echolocation calls [2, 3, 4]. However, their calls also show great interspecific variation and flexibility caused by habitat, geography, sex, age, etc. and in other cases there is a great overlap of call structures between species which makes species identification complicated [5, 6, 7]. Developing automatic identification tools would therefore assist in creating long term acoustic monitoring programs for biodiversity.

This work is a first step towards this direction. Our aim here is not to do an exhaustive comparison of methods but to show that using state of the art algorithms from the Machine Learning literature

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

and with no significant tunning or heavily engineered feature extraction methods good identification rates can be achieved.

In this study we use data of 21 species collected in North and Central Mexico and treat bat call identification as a supervised classification problem where a representative set of bat calls is used to train a classification model which is then applied to classify novel instances of bat calls. We employ a Multinomial probit regression model with Gaussian process prior [8] which can achieve good generalization capabilities with moderate to low numbers of training data. We also utilize a kernel representation of the data that directly compares the calls' spectrograms and thus it requires minor tunning.

2 Methodology

We approach bat call identification as a classification problem where the class response variables $y_n \in \{1, \dots C\}$ indicate the species id for the n^{th} call in the library and $\boldsymbol{x} \in \mathbb{R}^D$ is a D-dimensional vector representation of the call, e.g. features extracted from the call's spectrogram.Species' ids from all calls in the library are collected in a vector $\boldsymbol{y} = [y_1, \dots, y_N]$ and all call vector representations are collected in the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^T$. of size $N \times D$. In Section 2.1 we will define a probabilistic model for the conditional probability $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ where θ denotes a vector of unknown model parameters with an associated prior distribution $p(\boldsymbol{\theta})$. The id for a new call, y^*, with vector representation $\boldsymbol{x}^*$ is obtained by the class with highest probability from $p(y^*|\boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})$ where parameter estimates $\hat{\boldsymbol{\theta}}$ are obtained by maximizing the posterior distribution, i.e. $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})$.

2.1 Multinomial Probit Regression with GP prior

The probabilistic model assumes a *latent* function $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^C$ with *latent* values $\boldsymbol{f}(\boldsymbol{x}_n) = \boldsymbol{f}_n = [f_n^1, f_n^2, \dots, f_n^C]^T$ such that when transformed by a sigmoid-like function give the class probabilities $p(y_n|\boldsymbol{f}_n)$. Here we use a the multinomial probit function, Equation (1), which is convenient for deriving the EP approximation and Gibbs sampling [18, 8].

$$p(y_n|\boldsymbol{f}_n) = \int \mathcal{N}(u_n|0,1) \prod_{j=1, j\neq y_n}^{C} \Phi(u_n + f_n^{y_n} - f_n^j) du_n \tag{1}$$

For the *latent* function values we assume independent zero-mean Gaussian process priors for each class similar to [14]. Collecting *latent* function values for all calls and classes in $\boldsymbol{f} = [f_1^1, \dots, f_N^1, f_1^2, \dots, f_N^2 \dots, f_1^C, \dots, f_N^C]^T$ the GP prior is $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}(\boldsymbol{\theta}))$ where $\boldsymbol{K}(\boldsymbol{\theta})$ is a $CN \times CN$ block covariance matrix with block matrices $\boldsymbol{K}^1(\boldsymbol{\theta}), \dots, \boldsymbol{K}^C(\boldsymbol{\theta})$, each of size $N \times N$, on its diagonal. Elements $K_{i,j}^c$ define the prior covariance between the *latent* function values f_i^c, f_j^c governed by a covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$.

Optimising the unknown kernel parameters $\boldsymbol{\theta}$ involves computing and maximising the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{\theta}) \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{f}. \tag{2}$$

Making predictions for a new call, y_*, $\boldsymbol{x}_*$, involves two steps. First computing the distribution of the *latent* function values for the new call

$$p(\boldsymbol{f}_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \int p(\boldsymbol{f}_*|\boldsymbol{x}_*, \boldsymbol{X}, \hat{\boldsymbol{\theta}}) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}) d\boldsymbol{f} \tag{3}$$

and then computing the class probabilities using the multinomial probit function

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}) = \int p(y_*|\boldsymbol{f}_*) p(\boldsymbol{f}_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}) d\boldsymbol{f}_*. \tag{4}$$

2.2 Full EP approximation

Unfortunately exact inference is not possible and we have to either resort to numerical estimation through Markov Chain Monte Carlo or use approximate methods. Due to the large number of classes

(21 species in our data) in this work we consider the latter approach and use Expectation Propagation (EP) [19] to approximate the posterior of the *latent* function values $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta})$ in Equations (2) and (3) while for computing the integral in (4) we can again use the EP algorithm.

The EP method approximates the posterior using $q_{EP}(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}) \approx \frac{1}{Z_{EP}} p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) \prod_{n=1}^{N} \widetilde{t}_n(\boldsymbol{f_n}|\widetilde{Z}_n, \widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ where $\widetilde{t}_n(\boldsymbol{f_n}|\widetilde{Z}_n, \widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n) = \widetilde{Z}_n \mathcal{N}\left(\boldsymbol{f_n}|\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n\right)$ are local *likelihood* approximate terms with parameters $\widetilde{Z}_n, \widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n$. The approximation parameters are updated by first computing the *cavity* distribution $q_{-n}(\boldsymbol{f}_n) = q_{EP}(\boldsymbol{f}_n|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta})\widetilde{t}_n(\boldsymbol{f_n}|\widetilde{Z}_n, \widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)^{-1}$ and then matching them with the moments of the *tilted* distribution

$$\hat{q}(\boldsymbol{f}_n) = \hat{Z}_n^{-1} q_{-n}(\boldsymbol{f}_n) p(y_n|\boldsymbol{f}_n). \tag{5}$$

Unlike the binary probit case, where the *tilted* distribution (5) is univariate and thus its moments are easy to compute, the *tilted* distribution for the multinomial probit model is C-dimensional. Previous work on EP approximations for the multinomial probit model [16] further approximated the moments of the *tilted* distribution using the Laplace approximation which assumes that the distributions can be closely approximated my a multivariate normal.

In this work we show that a full EP algorithm can be derived by augmenting the *latent* function values $\boldsymbol{f}$ with the *auxiliary* variables u_n from Equation (1) and permuting both the augmented variables and the covariance matrix $\boldsymbol{K}(\boldsymbol{\theta})$. This results in the same algorithm as the *"nested"* EP approximation presented by [17], however this presentation clearly shows why a single iteration of the *inner* EP for the *tilted* distributions using the moments estimated from the previous iteration of the *outer* EP is enough for the algorithm to converge.

We introduce the new variables $\boldsymbol{w}$ which are formed by augmenting $\boldsymbol{f}$ with u_n and permuting such that $\boldsymbol{w} = [f_1^1, \ldots, f_1^C, u_1, f_2^1, \ldots, f_2^C, u_2, \ldots, f_N^1, \ldots, f_N^C, u_N]^T$. Similarly we augment the covariance matrix $\boldsymbol{K}(\boldsymbol{\theta})$ and permute accordingly such that the new covariance matrix $\boldsymbol{V}(\boldsymbol{\theta})$ is a $(C+1)N \times (C+1)N$ block matrix with blocks $\boldsymbol{V}(\boldsymbol{\theta})_{i,j} = \text{diag}([K_{i,j}^1, \ldots, K_{i,j}^C, 1]), \quad i, j \in \{1, \ldots, N\}$ of size $C+1 \times C+1$. Now we can write the posterior for $\boldsymbol{w}$ as

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{V}) \prod_{n=1}^{N} \prod_{j=1, j \neq y_n}^{C} \Phi(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) \tag{6}$$

where $\boldsymbol{w}_n = [f_n^1, \ldots, f_n^C, u_n]^T$ and $\boldsymbol{b}_{n,j} = [(\boldsymbol{e}_{y_n} - \boldsymbol{e}_j), 1]^T$ with $\boldsymbol{e}_j$ a C-dimensional vector of zeros and the j^{th} element set to 1.

The EP approximate posterior for $\boldsymbol{w}$ follows as

$$q_{ep}(\boldsymbol{w}) = Z_{ep}^{-1} \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{V}) \prod_{n=1}^{N} \prod_{j=1, j \neq y_n}^{C} \widetilde{t}_{n,j}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) \tag{7}$$

where $\widetilde{t}_{n,j}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) = \widetilde{Z}_{n,j}^{-1} \mathcal{N}\left(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}|\widetilde{\beta}_{n,j}, \widetilde{\alpha}_{n,j}\right)$ are the local approximate terms with parameters $\widetilde{Z}_{n,j}, \widetilde{\beta}_{n,j}, \widetilde{\alpha}_{n,j}$. This corresponds to an approximate posterior with $N(C-1)$ local approximation terms which have to be updated by matching their moments with the corresponding *tilted* distributions

$$\hat{q}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) = \hat{Z}_{n,j}^{-1} q_{-n,j}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) \Phi(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) \tag{8}$$

where $q_{-n,j}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) = q_{ep}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j}) \widetilde{t}_{n,j}(\boldsymbol{w}_n^T \boldsymbol{b}_{n,j})^{-1}$ are the cavity distributions. Calculating the moments for the *tilted* distribution can now be done analytically as Equation (8) resembles the *tilted* distribution of the probit model [14, 17].

2.3 Spectrogram Features

The vector representation $\boldsymbol{x}_n$ for each call is constructed by extracting call shape parameters from the call's spectrogram similar to [9]. The spectrogram of a call is calculated by using a hamming window of size 256 with 95% overlap and an FFT length of 512. The frequency range of the spectrogram is thresholded by removing frequencies below 5kHz and above 210kHz. An example

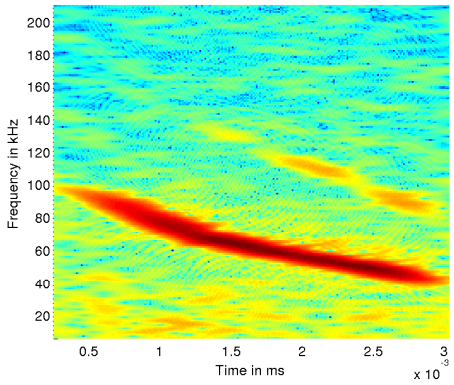


Figure 1: Example of a call's spectrogram. See text for details on spectrogram computation.

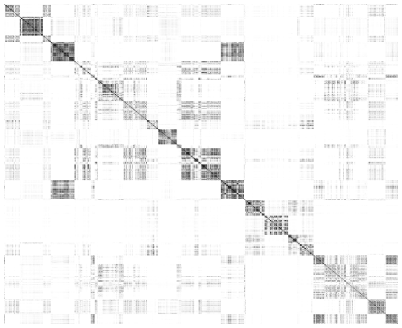


Figure 2: Class sorted optimal alignment scores using the DTW.

of a call's spectrogram is illustrated in Figure 1. In total 32 parameters are calculated including the call's duration in miliseconds, the highest and lowest frequencies of the call, its total frequency spread, the frequency with maximum amplitude, the frequencies at the start and end of the call etc. We do not give a full list of the call parameters here due to luck of space but a complete list can be found in [9].All 32 call parameters are concatenated in the vector $\boldsymbol{x}_n$ and a squared exponential kernel with individual length scales is used for the GP prior.

2.4 Dynamic Time Warping Kernel

Although extracting call shape parameters from the spectrogram of a call captures some of the call's characteristics and shape, there is still a lot of information that is discarded, e.g. harmonics. An alternative to characterising a call using predefined parameters is to directly utilise its spectrogram. However due to the differences in call duration the spectrograms will need to be normalised in order to have the same length using some form of interpolation. In this work we borrow ideas from speech recognition [11] and previous work on bird call classification [13] and employ the Dynamic Time Warping (DTW) kernel to directly compare two calls' spectrograms.

Given two calls i, j from the library and their spectrograms $\boldsymbol{S}_i, \boldsymbol{S}_j$, where $\boldsymbol{S}_i \in \mathbb{C}^{F \times W}$ with F being the number of frequency bands and T the number of windows, the dissimilarity matrix $\boldsymbol{D}^{i,j} \in \mathbb{R}^{W \times W}$ is constructed such that

$$\boldsymbol{D}^{i,j}(w, v) = 1 = \frac{\boldsymbol{S}_i(:, w)^T \boldsymbol{S}_j(:, v)}{\sqrt{\boldsymbol{S}_i(:, w)^T \boldsymbol{S}_i(:, w) \boldsymbol{S}_j(:, v)^T \boldsymbol{S}_j(:, v)^T}}. \tag{9}$$

DTW uses the dissimilarity matrix in order to stretch or expand spectrogram $\boldsymbol{S}_i$ over time in order to match $\boldsymbol{S}_j$ by calculating the optimal warping path with the smallest alignment cost, $c_{i,j}$, using dynamic programming. For each call we construct a vector representation $\boldsymbol{x}_n$ by computing the optimal warping paths with all N calls from the library and concatenating the alignment costs such that $\boldsymbol{x}_n = [c_{n,1}, \ldots, c_{n,N}]$. We then use the squared exponential covariance function for the covariance matrix of the GP classifier. Figure 2 shows the optimal alignment scores for the training data used in this study.

2.5 Multiple Kernel GP

GP classifiers allow for integrating information from different sources or different representations of the data by combining covariance functions. Although both representations discussed in the previous sections are extracted from a call's spectrogram, some of the call parameters used in Section 2.3 involve non-linear and complex transformations of the spectrograms by utilising prior knowledge of bat call shapes. Since such knowledge is important for bat call identification and is not present in the DTW representation we combine both kernels by a weighted sum and treat the weights as unknown

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

Table 1: Dataset statistics

Species	Samples	Calls
Family: *Emballonuridae*		
1 *Balantiopteryx plicata*	16	384
Family: *Molossidae*		
2 *Nyctinomops femorosaccus*	16	311
3 *Tadarida brasiliensis*	49	580
Family: *Mormoopidae*		
4 *Mormoops megalophylla*	10	135
5 *Pteronotus davyi*	8	106
6 *Pteronotus parnellii*	23	313
7 *Pteronotu personatus*	7	51
Family: *Phyllostomidae*		
8 *Artibeus jamaicensis*	11	82
9 *Desmodus rotundus*	6	38
10 *Leptonycteris yerbabuenae*	26	392
11 *Macrotus californicus*	6	53
12 *Sturnira ludovici*	12	71
Family: *Vespertilionidae*		
13 *Antrozous pallidus*	58	1937
14 *Eptesicus fuscus*	74	1589
15 *Idionycteris phyllotis*	6	177
16 *Lasiurus blossevillii*	10	90
17 *Lasiurus cinereus*	5	42
18 *Lasiurus xanthinus*	8	204
19 *Myotis volans*	8	140
20 *Myotis yumanensis*	5	89
21 *Pipistrellus hesperus*	85	2445

parameters. The kernel weights are jointly optimized along with the individual kernel parameters by maximizing the marginal likelihood.

3 Experimental Setup and Data

3.1 Data

Bat echolocation calls were recorded across North and Central Mexico. Live-trapped bats were measured and identified to species level using field keys [20, 21] and bat taxonomy followed in [22]. We constructed an echolocation call library by recording the calls of captured individuals using two different techniques: 1) bats were recorded while released from the hand about 6 to 10 m from the bat detector in open areas and away from vegetation, 2) bats were tight to a zip-line and recorded while flying along the zip flight path. The bat detector was set to manually record calls in real time, full spectrum at 500 KHz. Each recording consists of multiple calls from a single individual bat.

In total our dataset consists of 21 species, 449 individual bats and 8429 calls. Table 1 gives a summary of the dataset. Care must be taken when spiting the data to training and test sets during cross-validation in order to ensure that calls from the same individual bat recording are not in both sets. For that we split our dataset using recordings instead of calls. For species with less than 100 recordings we include as many calls as possible up to a maximum of 100 calls per species.

3.2 Experiments

We compare the classification accuracy of the multinomial probit regression with Gaussian process prior classifier using the three representations discussed in Sections 2.3-2.3. The values of the call shape parameters are normalised to have zero mean and one standard deviation by subtracting the mean and dividing by the standard deviation of the call shape parameters in the training set. For the 33 covariance function parameters, σ^2 and $\lambda_1, \ldots, \lambda_3 2$ we use independent Gamma priors with shape parameter 1.5 and scale parameter 10. For the DTW representation each call vector of optimal alignment costs is normalised to unit length and independent Gamma (1.5, 10) priors are used for the magnitude and length-scale covariance function parameters. The weights for the linear combination of the DTW and call shape kernel functions are restricted to be positive and sum to 1 and a flat Dirichlet prior is used.

3.3 Results

Table 2 compares the misclassification rate of the three methods. Results are averages of a 5-fold cross validation. We can see that the DTW representation is significantly better for characterising the species variations achieving a better classification accuracy. However, results can be improved by also considering information from the call shape parameters. Moreover, the optimised weights

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

for the kernel combination significantly favor the DTW covariance function with a weight of ≈ 0.8 in contrast to the call shape parameters with weight ≈ 0.2. If we fix the the weight parameters to equal values we obtain a classification error rate of 0.22 ± 0.031 highlighting the importance of the DTW kernel matrix.

The independent length scales allow us also to interpret the discriminatory power of the call shape parameters. In our experiments the frequency at the center of the duration of a call, the characteristic call frequency (Determined by finding the point in the final 40% of the call having the lowest slope or exhibiting the end of the main trend of the body of the call) as well as the start and end frequencies of the call have consistently obtained a small lengthscale parameter value indicating their importance in species discrimination. This coincides with expert knowledge on bat call shapes where these call shape parameters are extensively used for identifying species.

Table 2: Classification results, smaller values are better.

Method	Error rate	Std.
Call shape parameters	0.24	± 0.052
DTW	0.21	± 0.026
DTW + shape parameters	**0.20**	$\pm$**0.037**

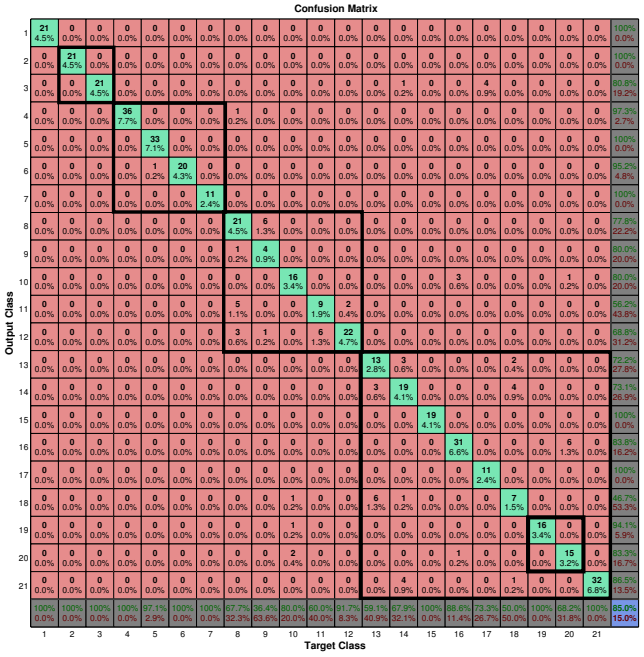


Figure 3: Confusion matrix of the best classification. Classes are in the same order and grouped as in Table 1

In Figure 3 the confusion matrix from the best classification results, 15% misclassification rate, are shown. There is an overall high accuracy for all classes with the exception of species *Lasiurus xanthinus*, class 18, which is often misclassified as *Antrozous pallidus*, class 13, which needs to be investigated further. In contrast, the very similar call shapes of the *Myotis* species are easily discriminated. Finally, misclassification rates are higher to within family species compared to species from other families indicating a common evolutionary path of bat echolocation.

4 Conclusions and Future Work

Previous works highlight the complexity to discriminate species from the Phyllostomidae family, while others recognized *Myotis* species hard to classify as well. The high accuracy obtained in this

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

study to separate species in the Phyllostomidae family from other families and its ability to discriminate between *Myotis* species sets the ground for a further development of an automatic identification tool for Mexican bats. Although only a small set of Mexican bat species was used in this study, it shows promising applications to a bigger set of species. Despite these limitations, the development of a national call library of full-spectrum calls together with the echolocation classification tool will set the foundations to establish a long-term National Bat Acoustic Monitoring Program. This is a feasible alternative for developing countries to create biodiversity monitoring programs and develop volunteer networks because they are easier and less costly to implement at broad scales and long term compared to other monitoring techniques.

References

[1] G. Jones, D. Jacobs, T. Kunz, M. Willig, and P. Racey. Carpe noctem: the importance of bats as bioindicators. *Endangered Species Research*, 8:93–115, 2009.

[2] M. B. Fenton and G. P. Bell. Recognition of Species of Insectivorous Bats by their Echolocation Calls. *Journal of Mammology*, 62(2):233–242, 1981.

[3] G. Jones and E. Teeling. The Evolution of echolocation in bats. *Trends in Ecology and Evolution*, 21:149–156, 2006.

[4] I. Ahlen and H. Baage. Use of ultrasound detectors for bat studies in Europe : experiences from field identification , surveys , and monitoring. *Acta Chiropterologica*, 1:137–150, 1999.

[5] M.K Obrist. Flexible bat echolocation: the influence of individual, habitat and conspecifics on sonar signal design. *Behavioral Ecology and Sociobiology*, 36:207–219, 1995.

[6] K. Murray, E. Britzke, and L. Robbins. Variation in search phase calls of bats. *Journal of Mammalogy*, 82:728–737, 2001.

[7] H. U. Schnitzler, C.F Moss, and A. Denzinger. From spatial orientation to food acquisition in echolocating bats. *Trends in Ecology and Evolution*, 18:386–394, 2003.

[8] Mark Girolami and Simon Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006. Probit regression; gaussian process, variational bayes, multi-class classification.

[9] Charlotte L. Walters, Robin Freeman, Alanna Collen, Christian Dietz, M. Brock Fenton, Gareth Jones, Martin K. Obrist, Sébastien J. Puechmaille, Thomas Sattler, Björn M. Siemers, Stuart Parsons, and Kate E. Jones. A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, 49(5):1064–1074, 2012.

[10] S. Parsons and G. Jones. Acoustic identication of twelve species of echolocating bat by discriminant function analysis and artifact neural networks. *Journal of Experimental Biology*, 203:2641–2656, 2000.

[11] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEE Transactions on Acoustics, Speech and Signal Processing*, 26:43–49, 1978.

[12] Hansheng Lei and Bingyu Sun. A study on the dynamic time warping in kernel machines. In *Signal-Image Technologies and Internet-Based System, 2007. SITIS '07. Third International IEEE Conference on*, pages 839–845, 2007.

[13] Theodoros Damoulas, Samuel Henry, Andrew Farnsworth, Michael Lanzone, and Carla Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications*, ICMLA '10, pages 424–429, Washington, DC, USA, 2010. IEEE Computer Society.

[14] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[15] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Research*, pages 1679–1704, 2005.

[16] Mark Girolami and Mingjun Zhong. Data integration for classification problems employing gaussian process priors. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 465–472. MIT Press, Cambridge, MA, 2007.

[17] Jaako Riihimaki, Pasi Jylanki, and Aki Vehtari. Nested expectation propagation for gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

[18] Matthias Seeger, Neil D. Lawrence, and Ralf Herbrich. Efficient nonparametric bayesian modelling with sparse gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tubingen, Germany, 2006.

[19] Thomas Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

[20] R. A. Medellín, H. Arita, and O. Sánchez. *Identificación de los murciélagos de México: Clave de campo*. Publicaciones Especiales. Asociación Mexicana de Mastozoología A. C., Mexico, DF, 2008.

[21] G. Ceballos and G. Oliva. *Los mamíferos silvestres de México*. CONABIO UNAM Fondo de Cultura Econmica, Mexico, DF, 2005.

[22] N. B. Simmons. Order Chiroptera. In *Wilson, D.E., Reeder, D.M. (eds.), Mammal Species of the World: A Taxonomic and Geographic Reference, third ed.*, pages 312–529. Johns Hopkins University Press, 2008.

3.2 Whale songs classification using Sparse Coding

Hervé Glotin, J. Razik, S. Paris, O. Adam and Y. Doh - USTV, Institut Universitaire de France, CNRS LSIS

The humpack whale songs relies on the ability of these whales to copy and recombine vocal sounds and to arrange them in new sequences, around many tropical sites all over the planet. We present the advantages of the sparse coding to represent these song sequences in order to track their structure and evolution, and to automatically recognize the area from where this song has been emitted. This representation may also help to understand learning processes that can explain how the whale build new songs. Demonstrations are conducted on true recordings (we thank C. Clark and O. Lammers for sharing some of their samples).

Sparse Coding for Scaled Bioacoustics

at NIPS 2013

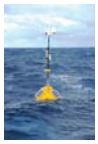
Glotin Hervé

Université de Toulon
Institut Universitaire de France (IUF)
Aix Marseille University
UMR CNRS LSIS – DYNI team
SABIOD MASTODONS CNRS Big Data project

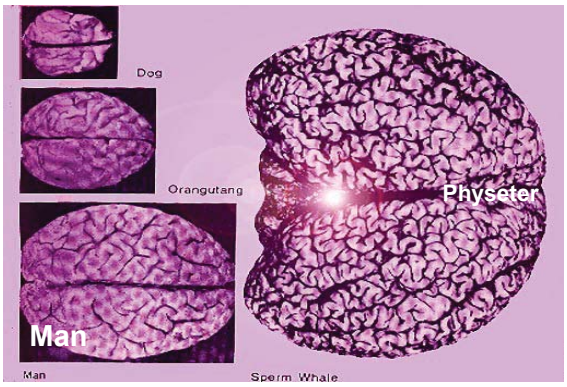
http://sabiod.univ-tln.fr

with Razik J., Paris S., Giraudet P., Prevot J.M., Doh Y., Abeille R., Bénard F., Monnin A, Chamroukhi F .

Prelude : 20 minutes far from univ. Toulon, live Physeter macrocephalus, Minke whale, dolphin, fin whale,...

Other Project (with O. ADAM)	location	technical properties	data Go/month	tasks
		Array of 6 hydrophones sampling frequency : 44 kHz or 192 kHz, coded by 16 or 24 bits	200 to 9000	Inventory of Humpback whales in the Indian Ocean Estimation of the trend of the size of the population Impact of the human activities Impact of the global change
	Guadeloupe 	Single hydrophone sampling frequency : 44 kHz or 192 kHz, coded by 16 or 24 bits	**200 to 2000**	Inventory of the different species of resident/non resident cetaceans Presence in the new sanctuary AGOA of marine mammals Impact of the human activities (touristic, fishermen, harbour)
	St Pierre et Miquelon 	2 single hydrophones sampling frequency: 32 kHz, coded by 16 bits	350	-Inventory of the different species of resident/non resident cetaceans -Migration routes / feeding areas
total			**800 Go to 11 To/month**	

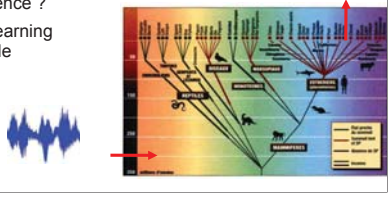
Neural Information Processing in Marine Mammals ?
Which representations ?



A challenge : Classification on large number of classes, with high class variability

- Usually organized (Hierarchical or Graph) through ontology
  - Examples : DMOZ (> 600 000) classes, Wikipedia, etc
  - Often multilabel
- **Quantitative change implies a qualitative change in methods**

- Problems
  - Which criterion to use for training ?
  - What to do to perform fast inference ?
  - Connexion to on-line learning, learning with imbalanced data, large scale learning, …



- New trend in machine learning and data mining

Thousands of classes

Large Scale Classification : which model ?

- **Approaches**
  - Flat (none relational information between classes e.g. one vs all classifiers)
    Accurate but slow at test time O(# classes)
  - Hierarchical (modeled on the taxonomy)
    Less accurate than flat methods but fast inference O(log(# classes))
  - In between methods
    Compromise wrt accuracy and inference time between the two extremes

[L Cai, T Hofmann, Hierarchical document categorization with support vector machines, CIKM 2004

K. Weinberger, O. Chapelle, Large Margin Taxonomy Embedding with an Application to Document Categorization, Neural Information Proc. Sys. NIPS 2008

J. Weston, S. Bengio and D. Grangier, Label Embedding Trees for Large Multi-Class Tasks, NIPS 2010

M. Cissé, T. Artières, P. Gallinari, Learning compact class codes for fast inference in large multi class classification, ECML 2012]

I. Introduction

II. Single hydrophone 'etho-acoustics'

III. 3D RT Single whale tracking by PA

IV. 3D RT Multiple whale tracking by PA

V. Classification by sparse coding

VI. Tracking by sparse coding

VII. Conclusion: Scaled passive Acoustic SABIOD

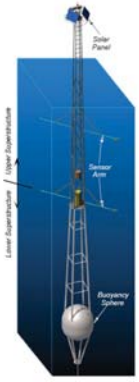
Predation and Moon effect ?

BOUSSOLE project

- Definition of the click detector and noise iltering
- Detection results
- Conclusion on ICI at new-moon versus full-moon

BOUSSOLE project

BOUSSOLE Buoy



periods of recording:

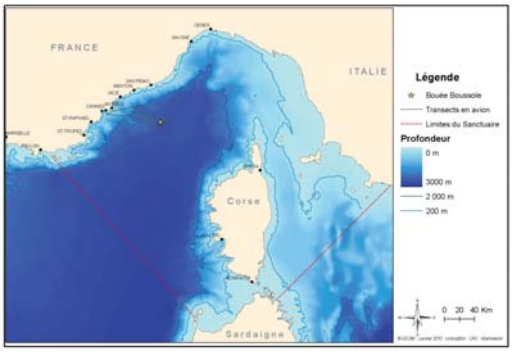
- 15 october to 9 december 2008
- 5 january au 2 march 2009
- 15 april au 15 juin 2009
- 16 july au 1 september 2009

5 minutes audio files recorded every quarter of an hour by the Ecological Acoustic Recorder (EAR, cf LAMMERS)

→ 6000 files of 19Mo

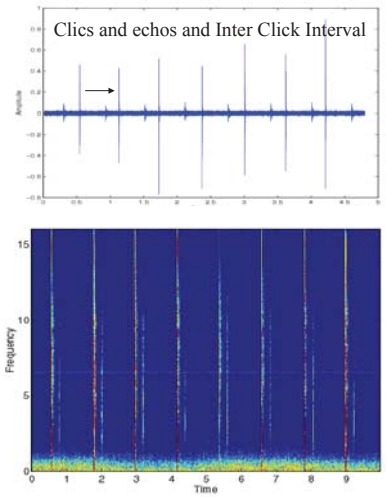
Fe = 32000Hz

18 monthes of recording south Antibes



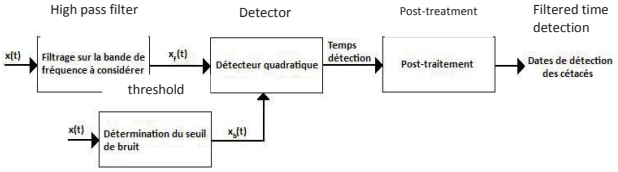
The Pelagos Sanctuary for Mediterranean Marine Mammals is a special marine protected area extending about 90.000 km2 in the north-western Mediterranean Sea between Italy, France and the Island of Sardinia, encompassing Corsica and the Archipelago Toscano.

An Inter-Click Interval study on Physeter catodon



Click detector and noise filter

For scaled processes, we designed a quadratic detector.
=> 2h30 to process one month of data.
followed by a filter in order to remove chain noises.

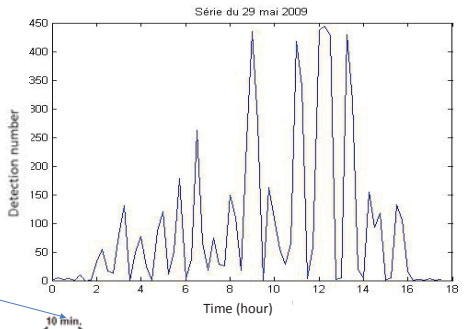


(high pass filter 5000Hz ; window length: 40 ms)

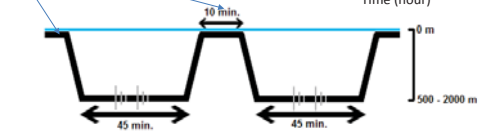
Detection results

Continuous detection on 15 hours of one Physeter !

Coherent detections profile according to the resting time at surface and diving periods :

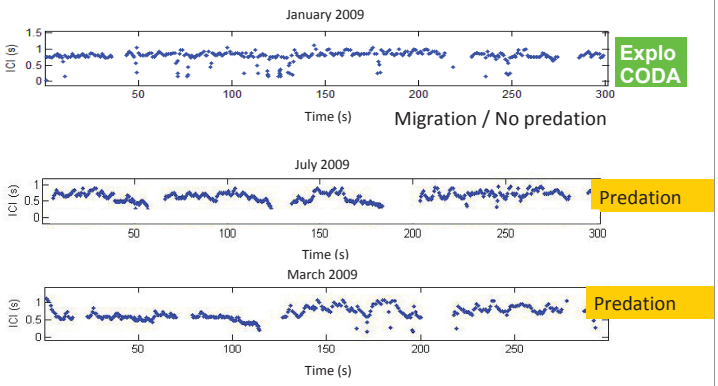


No acoustic activity



[Laran ...Glotin in Boussole Pelagos report 2008]

Inter-Clic Interval (ICI)



Predation behavior

(Bénard, Giraudet, Glotin 2007 - AUTEC Bahamas)

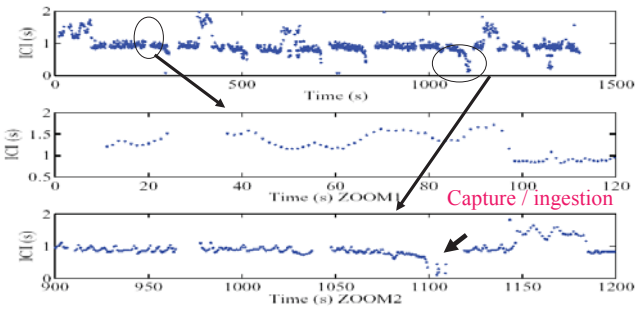
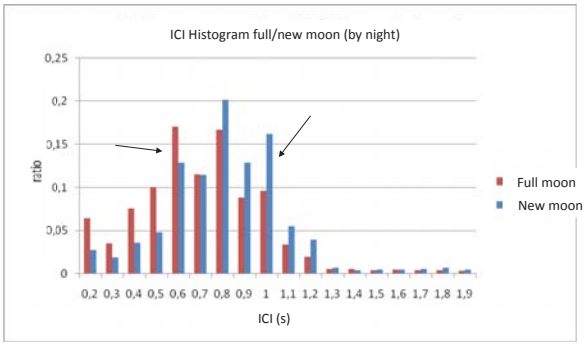


FIG. 3.11 – Intervalles inter-clics $ICI(t)$ estimés sur le set 2. En haut sur 1400s, en bas deux zoom. On observe des modulations qui correspondent au comportement de prédation. Dans le zoom de 900 à 1200s, on observe un creak à 1100s, suivi de l'ingestion d'une proie (silence).

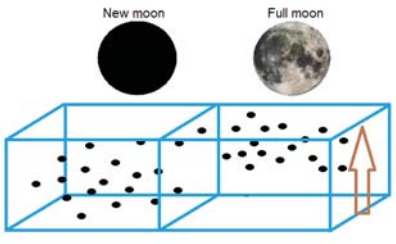
Statistics on ICI distribution new moon versus full moon



Again, Kolmogorov Smirnov test positive for $p < 0,01$.
nb : Recordings with more than 1 sperm whale are processed.

[Laran ...Glotin in Boussole Pelagos report 2008]

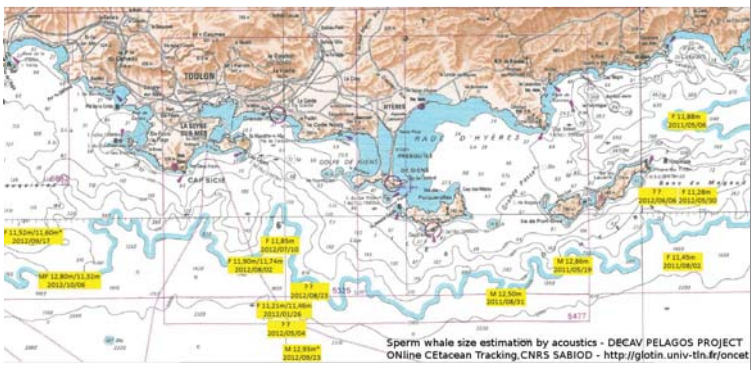
? : ICI(new-moon) >> ICI(full moon)



Interpretation : full moon light could result in a higher prey concentration at small depth water layers.
Thus, sperm whales are more often predating in this higher prey density at full moon, than at new moon.

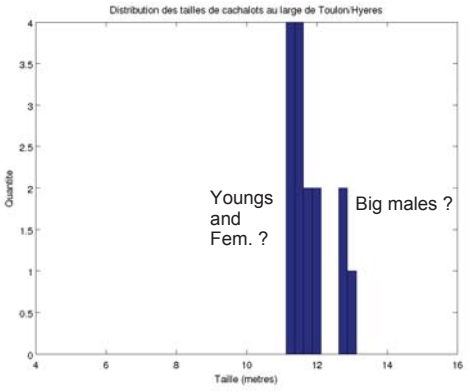
? Moon effect on social dialects ?

Inter Pulse Interval : proposition for robust IPI estimator
application on 2011-2012
DECAV PELAGOS project (each detection has its size est.)



2013 Abeille Phd, Coll G. Pavan

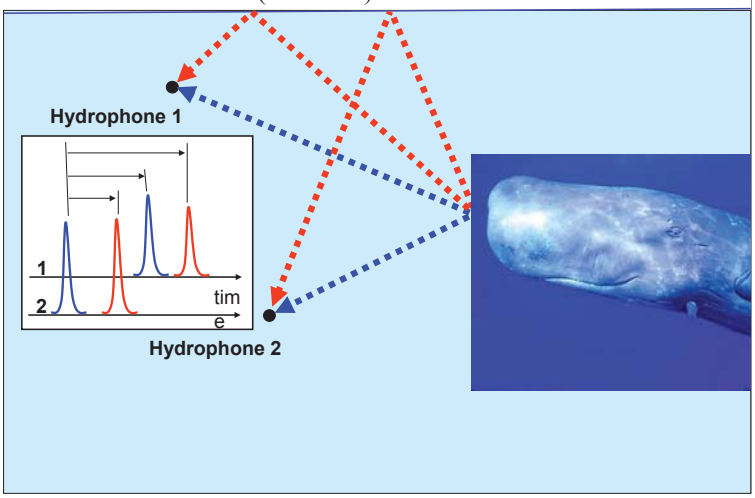
Allometric rules on IPI => Sizes distribution



[In Abeille, Glotin, Giraudet. IEEE OCOSS 2013]

TDOA *Time Delay of Arrival*

But multi-reflections (surface...) => directs + indirects

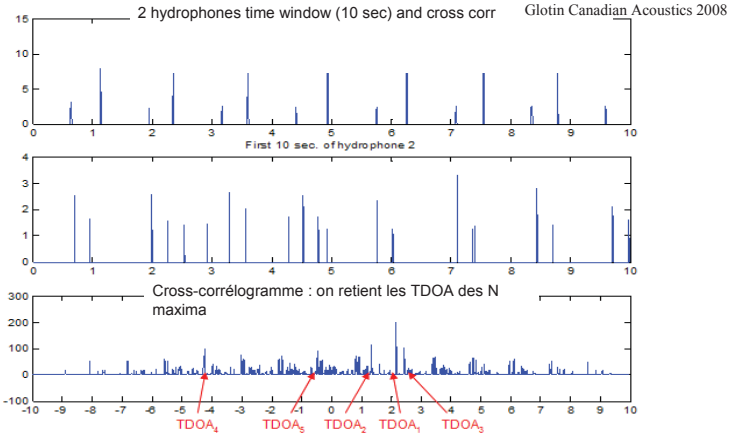


TDOA filtering ? Filtering the combinations between N local max

Criteria : MSE on the residual of TDOA transitivity system

[Glotin Giraudet Bénard patent 2007, PCT 2009-2011, USA, EU.]

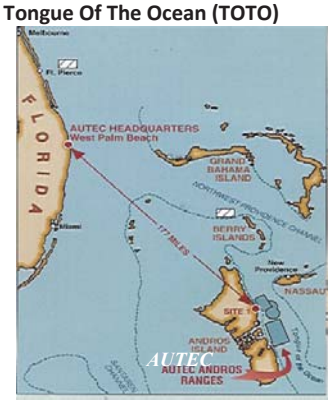
Glotin Canadian Acoustics 2008

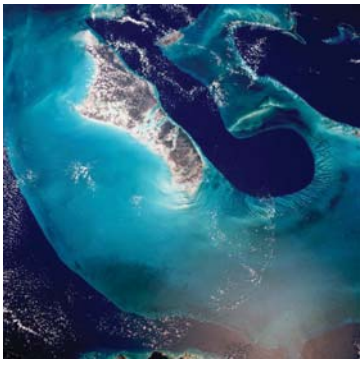


AUTEC NATO data set

Atlantic Undersea Test & Evaluation Center (AUTEC)

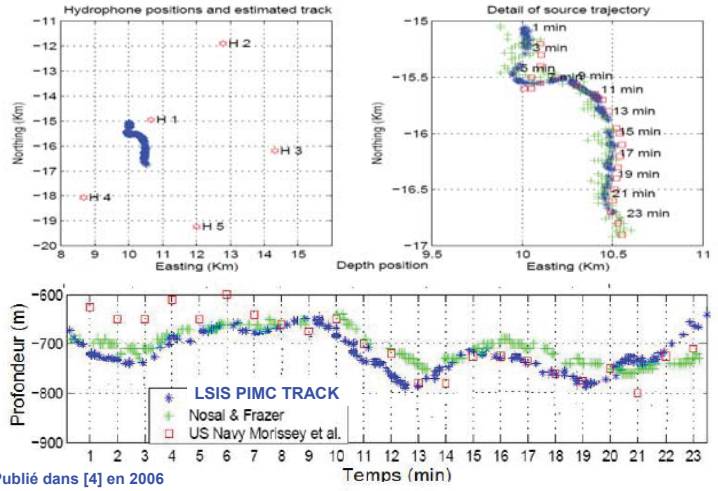
Tongue Of The Ocean (TOTO)





Glotin LSIS

RESULTS



Publié dans [4] en 2006

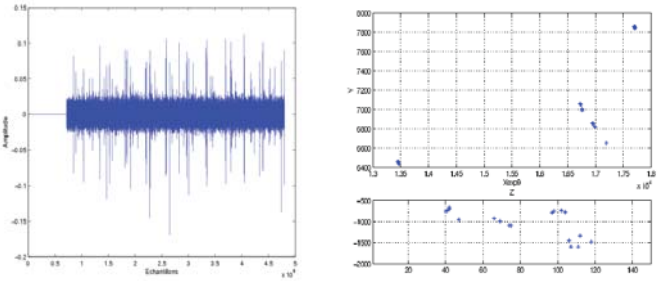
ONCET : Online Cetacean Tracking = Etho-acoutics ?

[Patent Glotin et al. Multiple whale tracking PCT USA,... 2008-2012
Glotin et al. Whale Cocktail Party, Canac Acoustics, 2008
Bénard Glotin, Neutrino whale tracking, Applied Acoustics 2011]

Online demo at http://sabiod.org
RANGE [500 to 5000 m] prec :15m

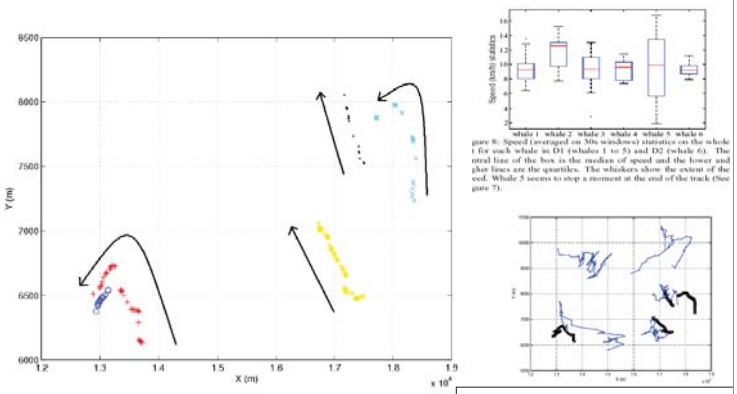
2nd Challenge : Simultaneous clicking whales...

– First results 2005 :



- 4 whales localization without TDOA selection ...

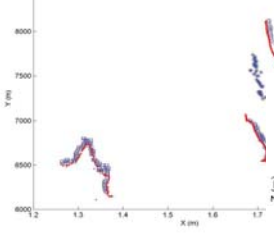
Second step 2006 : DYNI vs SOEST Hawaï
(abandon US Navy)



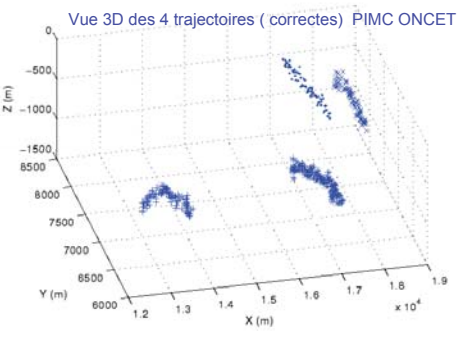
Dyni (BLACK) VS SOEST (Hawaï
Blue (V > 100km/h)

2009 : final ONCET model (Stochastic Adaptive Filtering)

PIMC-DYNI (rouge)
SOEST Hawaï contenant des fausses détections

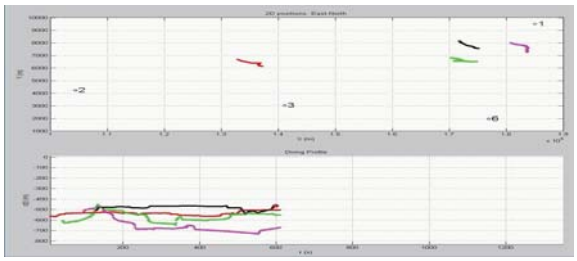


Vue 3D des 4 trajectoires (correctes) PIMC ONCET

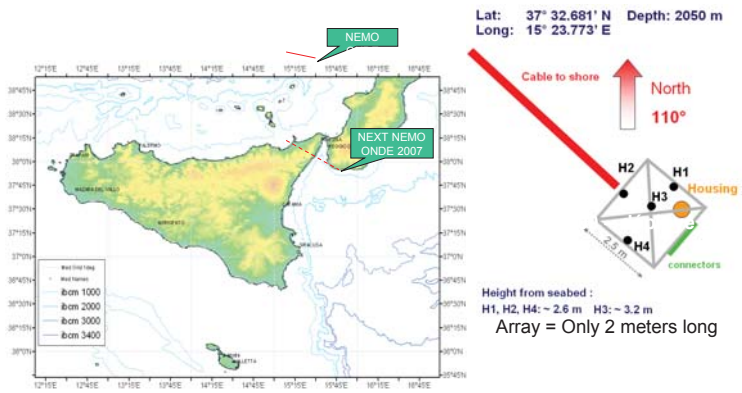


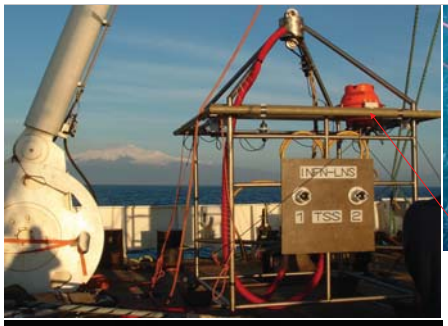
Demonstrations on line at :

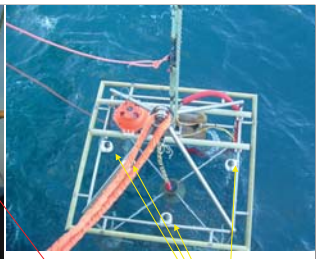
http://sabiod.univ-tln.fr/oncet



Astrophysics meets bioacoustics
Run 3D tracking on NEMO



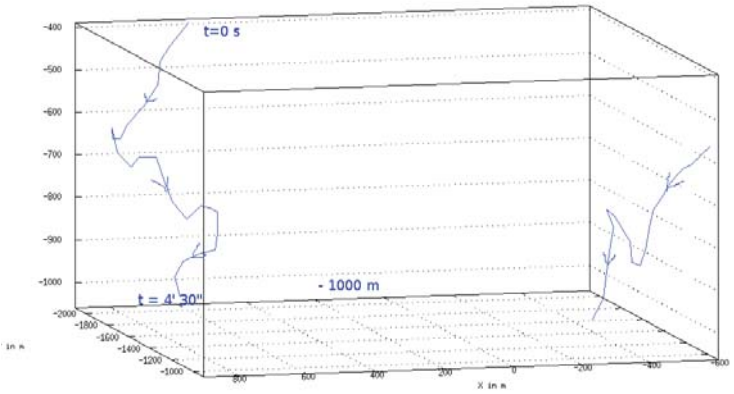




Once deployed on the sea floor the frame was connected to the optical cable by a ROV (Remotely Operated Vehicle)

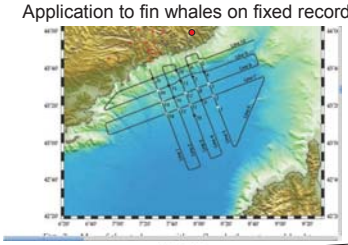
We thank INFN, NEMO, Ricobenne and G. Pavan for the record samples

LSIS results : 15 august 2005 15h00, Sicile Est :
2PC dive together from -400 m to -1000 m in 5 minutes



[Benard Glotin Applied Acoustics 2011]

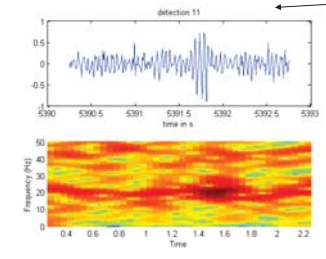
Application to fin whales on fixed recording in Toulon area...



Raw recording from a single hydrophone fixed on the OBS at the sea bottom ~2500 m

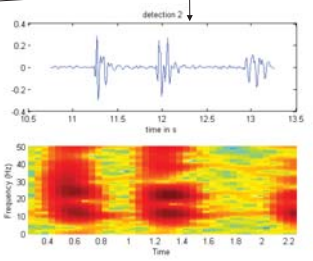
Sampled at 100 Hz

Detection processing



High spectral density around 20 Hz

Possibly rorqual sound emission



Wide band spectrum

Possibly sperm whale sound emission

I. Introduction

II. Single hydrophone 'ethoacoustics'

III. 3D RT Single whale tracking by PA

IV. 3D RT Multiple whale tracking by PA

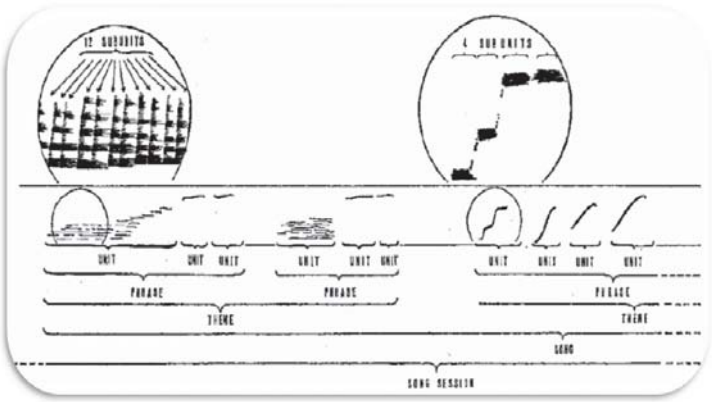
V. Classification by sparse coding

VI. Tracking by sparse coding

VII. Conclusion : Scaled Acoustic Methods SABIOD

Humpback whale song sparse coding : exploring song components

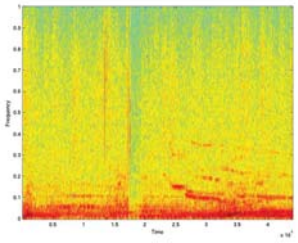
- Humpback songs are structured, but most of their decomposition algorithm are using a priori informations
- An usual way to determine recurrent pattern in a data flow in an unsupervised manner is to cluster the data. However the main drawback of k-means clustering is that the centroids of each cluster may not cover all the space and unfortunately not suit to the data.
- In this study we investigate the hypothesis of « subunit » and we propose a method to automatically identify these subunit components of the song.

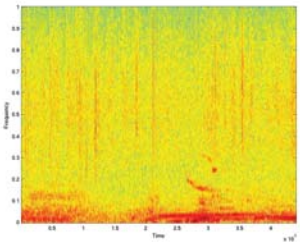


Humpack whale songs' structure : [Payne 1970]

Two song spectrograms : mostly unit of one second.

=> Unit patterns last nearly .1sec
=> The dictionnary is learned on 1 sec.frame window.





Sparse Coding ?

➔ Sparse Coding (SC) : unsupervised dictionnary generated from the complete data set

➔ SC may be more adapted to the differenciation of natural acoustic sources

➔ Development of methods for selecting and classifying relevant dictionnary atoms

➔ Applications :

- Supervised classification of whales
- Discovering spatio-temporal / 4D behavior patterns for (un)known species

Sparse Coding by Lasso and K-SVD

Why Sparse Coding ? : More discriminative
Better generalization for new data
Reduction of the reconstruction error

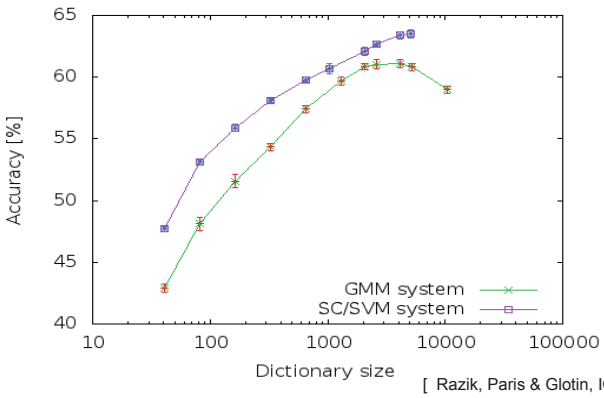
$$\arg\min_{D,C} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|^2 + \lambda\|\mathbf{c}_i\|_{\ell^1} \quad s.t. \quad \|\boldsymbol{c}_i\|_{\ell^1} = 1$$

and **Data compression**

– Each data vector x_i is expressed as a c_i linear

combination of a dictionary **D** of size **K** (only one in usual K-means)

SC *vs* state of the art :
SC improves automatic speech recognition



[Razik, Paris & Glotin, ICPRAM 2011]

Material

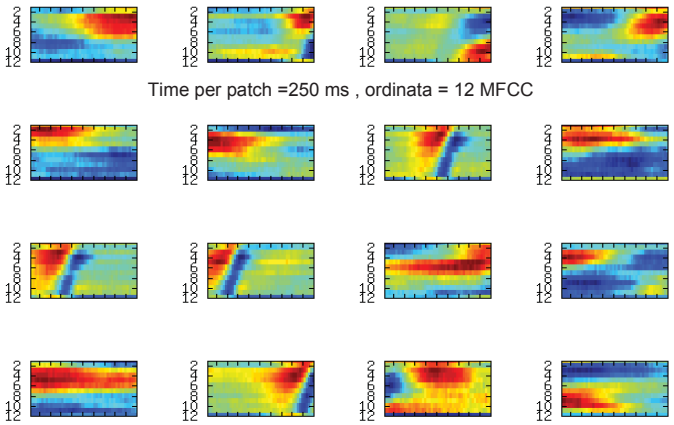
- Songs have been recorded at Hawaii (Lammers), Tonga (Clark), Madagascar (Adam & Doh), Reunion (Darewin), Guadeloupe (Adam), NewCaledonia (Glotin et al., Bachet et al.)...
- Each set contains clear song sequences of at least 10min, 44 kHz

Features extraction

- 13 MFCCs
  - 10 ms frameshift,
  - 32 ms frame length.
- N windows are concatenated to get the desired scale (e.g. N=25 for 250 ms).

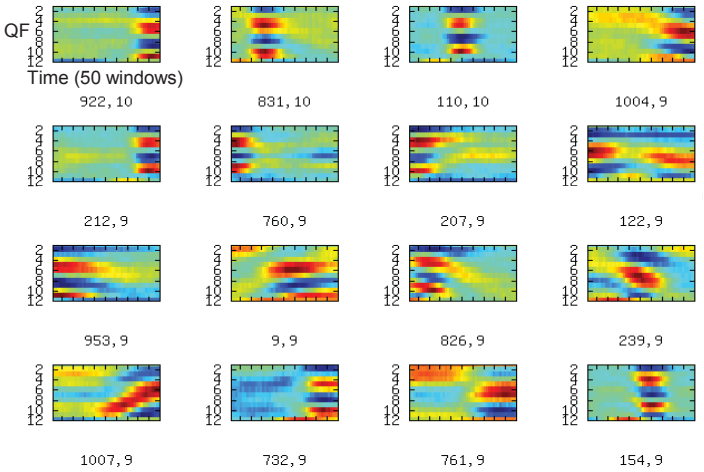
- On those vectors we :
  - Learn Unsupervised Dictionary : one codebook of 1024 words
  - Filter these words (units) according maximum quefrency movement (articulation).
  - Project the song sets using this codebook.

SPARSE MFCC CODE :
Samples of the 1024 words, some are coding sea noise

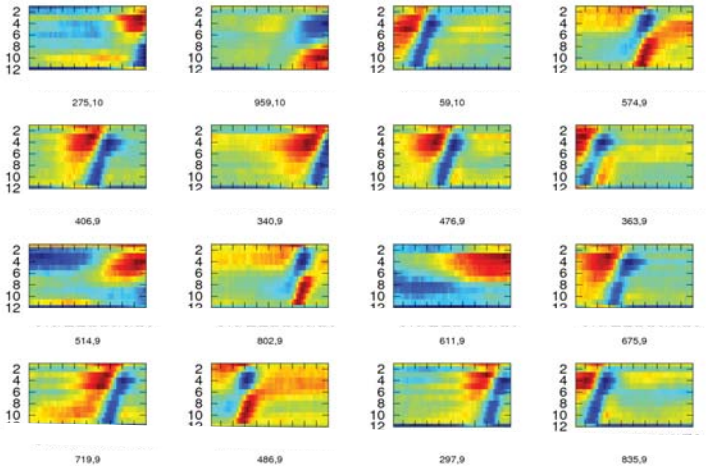


Time per patch =250 ms , ordinata = 12 MFCC

CODE SELECTION : the 16 most 'articulated' words (units)
criteria : selected by gabor filtering



The 16 most 'articulated' words
according to max variance in time and quefrency => whale ARTICULATIONS



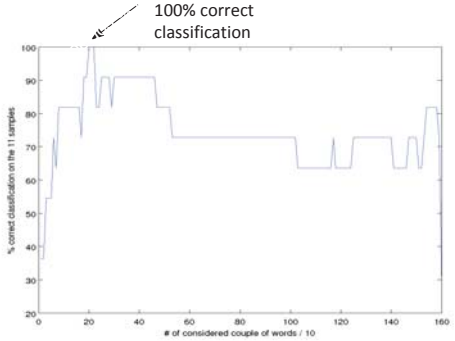
WHALE SONG CLASSIFICATION (preliminary results)

Song representation : filter 16 'most articulated codes' : C1...C16
Considering vector of N word couples = [. C(i,t)... C(j,t')...]
Build bigrams B(i,j) in a short time window (10 seconds)
Song = Histogram of bigram mean activity B(i,j) into time window 10 sec.
sim(Sp,Sq) = Cosine similarity between Sp and Sq representation

Validation :
Songs classification by sites

Material :
Hawaii
Tonga
Guadeloupe
Madagascar
Reunion

Task :
20 files of few min,
5 classes



Submitted to JASA Razik et al.

Conclusion

- We presented an algorithm to create by unsupervised dictionary learning a proto-lexicon of the song of the humpback whale.
- These representations are more generic than manual segmentation
- Different unit types have been learned on MFCC vectors.
- Long term units that are variously composing the songs from one year to another may be extracted **systematically**
  => WORLD SCALE BIOPOPULATION ANALYSIS

Perspectives

- Within this new song decomposition method, we find common patterns through time (subunits) and discover song differences considering long time units.
- By computing these features on different recordings through several years, we expect to find more efficiently stable patterns, and patterns hire between different whale groups.
- Our approach is naturally applicable to any marine mammals, and will be tested on dophin whistles.

Sparse coding for
Very fast TDOA
estimation
application to Hawaii

- 1. Objectives
- 2. Why and how to learn sparse dictionnary ?
- 3. Sparse matching of Minke whale boings in Hawaii
- 4. Time delay estimations
- 5. Tracking results
- 6. Conclusion and perspectives

Scaled Sparse Time Delay of Arrival estimations

on stereo recordings

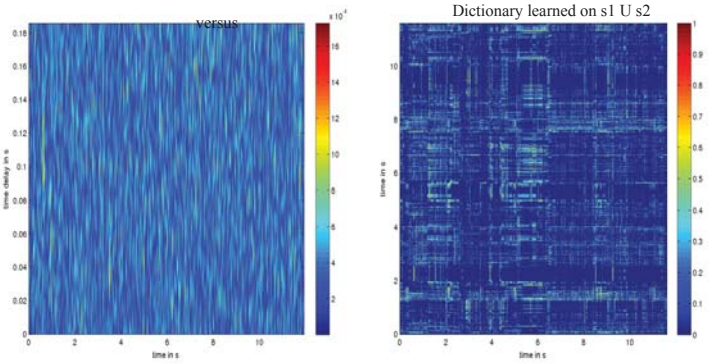
Humpack Whales, Madagascar recordings

CrossCorr(s1, s2)

versus

Cosinus(SC(s1) ,SC(s2))

Dictionary learned on s1 U s2

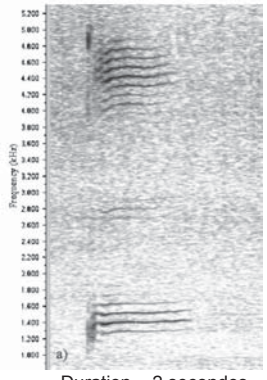


[Hervé GLOTIN - Joseph RAZIK - GIRAUDET Pascale - Sébastien PARIS - Frédéric BÉNARD Sparse coding for fast minke whale tracking with Hawaiian bottom mounted hydrophones" , *International Workshop on Detection, Classification, Localization & Density Estimation of Marine Mammals using Passive Acoustics, Portland, USA, supported by ONR Dpt of the Navy & Acoustical Society of America (ASA)*]

Objectives

- We propose in order to process efficient detection of minke whales (Balaenoptera acutorostrata), a sparse coding of their boings vocalizations.
- This sparse coding confers several advantages : it makes the structure in natural signals explicit and it represents complex data
- More generaly, l1-norm yields to robust Time Difference Of Arrival (TDOA). Recently Yuanqing Lin has described a l1-norm sparse Bayesian learning for acoustic blind channel identification and provides dramatic improvement dereverberation and TDOA estimation in reverberant environments compared to conventional methods.
- Therefore we compute the projection of a MFCC vector into a sparse coding representation, which allows good properties for similarity computation.

Why Sparse Coding ?



Duration ~ 2 secondes
Sample of boing (from Rankin and al.)

- Sparse coding minimizes the reconstruction error and allows good generalization for undetermined data.
- No need for any knowlege on the target (the boing) : the Sparse coding shall reconstruct in priority the frequent and high SNR events (e.g. the boings).
- We aim first to show that sparse coding will infer a simple boing matching process.
- Autocorrelation may give similar matching patterns, but our sparse vector representation will allow very fast cosine similarity computation

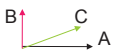
Features for minke boings detection

- 13 MFCCs
  - C0 to C12
  - 20 ms frameshift
  - 32 ms FFT
- 5 window length : 1/4, 1/2, 1, 2, 4 seconds
- Concatenation of the 5 vectors in one vector of 65 dimensions (non sparse).

Sparse projection of the MFCC

- On those 65 dim. vectors we :
  - Learn Unsupervised Dictionary : one common codebook of 1024 elements over the four hydrophones records and the three sets NN26, NN27, NN28 (4*30 minutes)
  - Project each hydrophones records using this codebook.
  - The resulting representation is very sparse : only 10% of the 1024 dimensions are non null.
  - This property allows relevant similarity measure between each projected signal window on different hydrophones.

Time Delay Estimation



The cosine similarity measure (def.)

cos(A,B) = (A .B) / (||A|| . ||B||),

here cos(A,B) = 0 < cos (A,C).

Higher the cosine is, the more the vectors are similar.

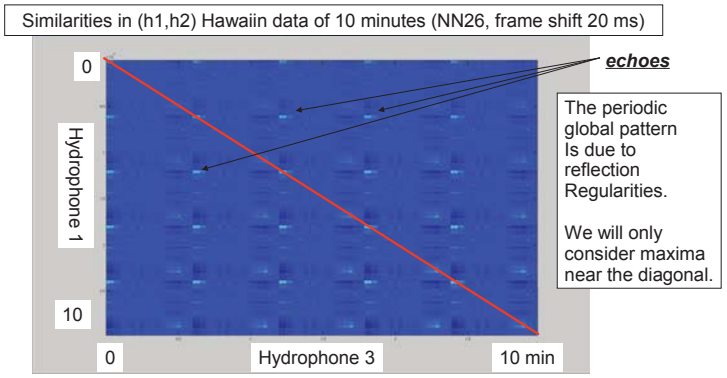
The multidimensional cosine between two hydrophones acoustic matrices, is very efficiently computed on parallel processing (much faster than correlation) :

allcosines(h1, h2) = (H1 * H2') / (norm(H1') * norm(H2)) ,

where Hi is the matrix of the 1024 by 10 minutes frames,
* is the matrix product,
norm(Hi) is the L2 norm of each frame vector of Hi.

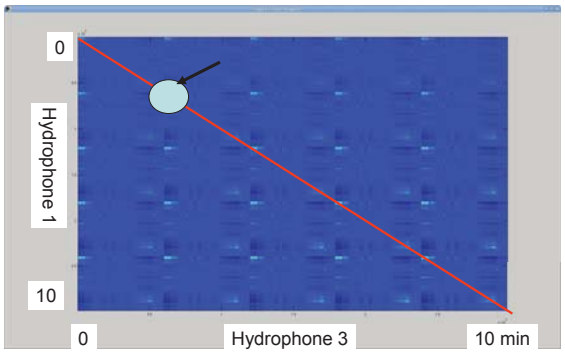
Time Delay Estimation

We compute the cosine between each vector pair from hi and hj
This representation allows a global analysis (far echoes…)
We figure out in red the 0 delay diagonal.

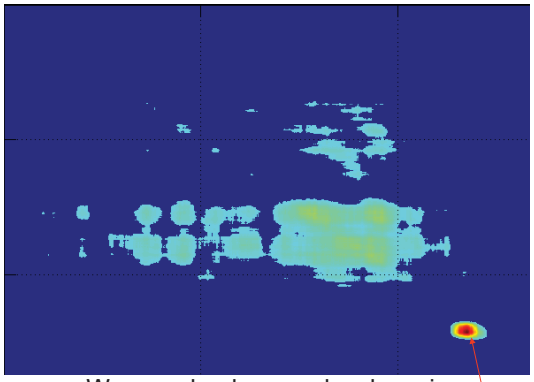


Time Delay Estimation

Zoom inside the map…

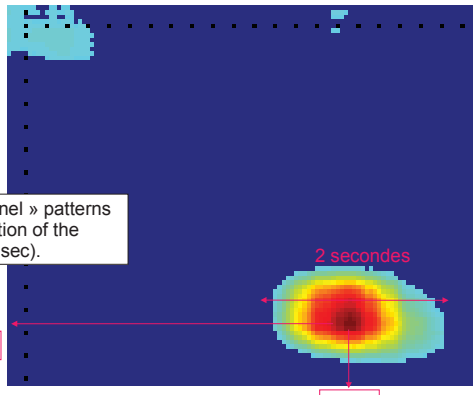


Zoom of this map between h1 and h3
zoom to 1 minute, after 5% superior quantile selection to remove the background noise



We see clearly some local maxima

Zoom of h1h3 map to 10 secondes



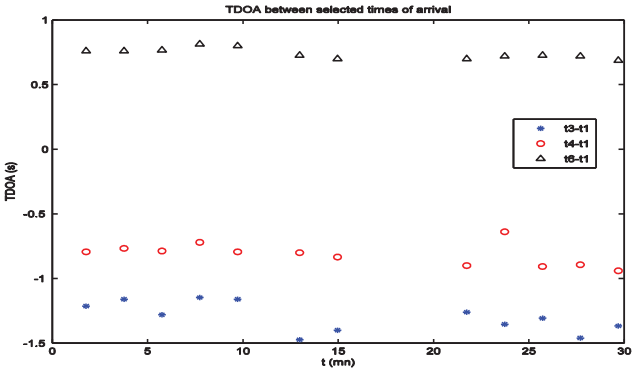
We get clear « kernel » patterns that have the duration of the boing sounds (= 2 sec).

The maximum of each kernel are measured iteratively to get the times on h1, h2

Then : **TDOA(h1,h2) = T(h1)-T(h3)**

Time Delays Of Arrival Estimations

We extract 14 TDOA over these 30 minutes, between h1,h3,h4,h6

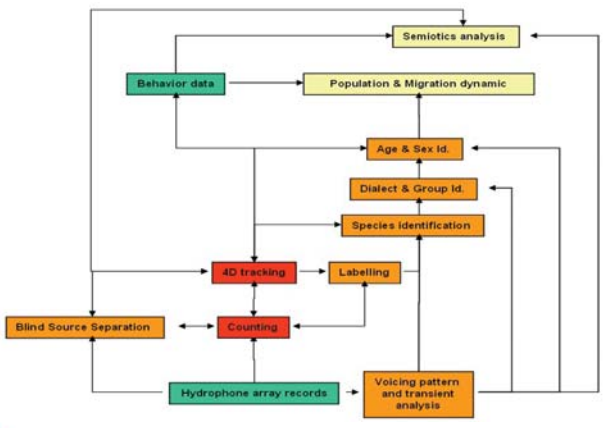


=> Coherent and regular variations

Conclusion

- We efficiently matched, through cosine of sparse projections, **and without any target knowledge**, the minke boing sounds.
- We got **clear boing detection** on hydrophone pairs. These TDOA generated straightforward **coherent track** with correct speed.
- Another set of TDOA has been detected (a second minke whale ?). We work further on that question.
- Perspectives : we will process our algorithm in the whole array and consider virtual hydrophones.

Global framework



LSIS PIMC framework since 2006 (Pole Mer PACA – SABIOD project Glotin http://sabiod.org)

- Ethoacoustic patterns extracted from SC statistics
- Biodiversity indexing may be easier based on SC
- 3D tracking may be accelerated using SC representations...

- Scaled detection and sparse decomposition

'Physeter catodon localization by sparse coding',
Paris, Glotin, Doh, Halkias, Razik, Workshop on Machine Learning for Bioacoustics, ICML4B, Atlanta 2013

- Whale localisation from SC

'Sparse coding for large scale bioacoustic similarity function',
Glotin, Razik, Paris, Halkias, POMA 19, 010015 (2013)
Report / paper available at http://sabiod.org

' Sparse coding for scaled bioacoustics: From Humpback whale songs evolution to forest soundscape analyses' H Glotin, J Sueur, T Artières, O Adam, J Razik, The Journal of the Acoustical Society of America 133 (5), 3311-3311

- Individual signature from SC of calls or transient

'Humpack sparse coding for group or individual identification',
Razik, Glotin, Paris, Adam, Doh, sub. in JASA

Example of scaled projects in SABIOD.ORG :
ONCET Online Cetacean Tracking,
Bombyx sono buoy,
ANTARES Neutrino & Bioacostics.

Tested on NEMO (see IV)
Wait for ANTARES data

Perspective : toward the Turing test

'Human' speech automatic processing was one of the first target of Artificial Intelligence

From the Turing test to a Bioacoustic Turing test ?
« Could an animal communicate with a computer ? »

If so, which patterns are conveying the information ?

How do animal learn them from so few samples ? (see Zebra Finch...)

Would this Bioacoustic Turing test be a mean to study causal inference in perception, and Artificial Intelligence ?

References 1/3 (please ask glotin@univ-tln.fr for copies if not found)

• Rankin S. and Barlow J., « Source of the North Pacific "boing" sound attributed to minke whales », J. Acoust. Soc. Am. 118-5, November 2005.

•Olshausen B., Field D., Sparse coding of sensory inputs, in Current Opinion in Neurobiology, 14 :481-487, 2004.

•Lewicki M., Efficient coding of natural sounds. Nat Neurosci, 5 :356-363, 2002.

•Yuanqing Lin, L1-Norm sparse bayesian learning, PhD univ. of Pennsylvania, 2008.

•Glotin H., F. Bénard, P. Giraudet, Whale Cocktail Party : a Real Time tracking of multiple whales. Canadian Acoustics Int. Journal, Vol. 36, p. 139-145, Mar 2008.

•Giraudet P. and H. Glotin Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely-spaced hydrophone array. Int. Jour. Applied Acoustics, Vol. 67, Issues 11-12, pp 1106-1117, Nov. 2006.

•Bénard F. Glotin H., Giraudet P, "Highly defined whale group tracking by passive acoustic Stochastic Matched Filter" , InTech, Advances in Sound Localization, fev 2011, available online :

• http://www.intechopen.com/articles/show/title/highly-defined-whale-group-tracking-by-passive-acoustic-stochastic-matched-filter

References 2/3

- Mairal « Sparse Coding », in. Proceedings of ERMITES summer school 2010, http://glotin.univ-tln.fr/ERMITES10
- Federica Pace, Fréderic Benard, Hervé Glotin, Olivier Adam, Paul White, "Subunit definition and analysis for humpback whale call classification", Applied acoustics, 71 2010, p 1107-1112
- Payne, R.S. and S. McVay, Songs of Humpback Whales. Science, 1971. 173(3997): p. 585-597.
- Mercado III, E. and A. Kuh. Classification of humpback whale vocalizations using a self-organizing neural network. in IEEE World Congress on Computational Intelligence. 1998.
- Rickwood, P. and A. Taylor, Methods for automatically analyzing humpback song units. The Journal of the Acoustical Society of America, 2008. 123(3): p. 1763-1772.

Reference 3/3

- Glotin H., F. Bénard, P. Giraudet, Whale Cocktail Party : a Real Time tracking of multiple whales. Canadian Acoustics Int. Journal, Vol. 36, p. 139-145, Mar 2008.
- Giraudet P., H. Glotin Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely-spaced hydrophone array. Int. Jour. Applied Acoustics, Elsevier Ed., Vol. 67, Issues 11-12, pp 1106-1117, Nov. 2006.
- Bénard, F, Glotin, H, Giraudet, P : Highly defined whale group tracking by passive acoustic Stochastic Matched Filter. Advances in Sound Localization Online Intech Book 2011 : http://www.intechopen.com/articles/show/title/highly-defined-whale-group-tracking-by-passive-acoustic-stochastic-matched-filter
- Bénard F., GLOTIN, CASTELLOTE, LARAN, LAMMERS "Passive acoustic monitoring in the Ligurian Sea" , 4th International Workshop on Detection, Classification and Localization of Marine Mammals using Passive Acoustics, 2009
- Bénard, H. Glotin, GIRAUDET P. "Whale 3D monitoring using astrophysic NEMO ONDE two meters wide platform with state optimal filtering by Rao-Blackwell Monte Carlo data association" , *in : Journal of Applied Acoustics, Vol. 71 (2010), pp. 994-999*, nov 2010
- Bénard, H. Glotin, "Automatic indexing and content analysis of whale recordings and XML representation" , *in : EURASIP Special Issue, Advances in Signal Processing for Maritime Applications, Vol. 2010 (2010), pp. 8*, 2010
- Bénard, H. Glotin, "WHALES LOCALIZATION USING a LARGE HYDROPHONE ARRAY: PERFORMANCE RELATIVE to CRAMER-RAO BOUNDS and CONFIDENCE REGIONS" , *in : Springer-Verlag, e-Business and Telecommunications*, sep 2009
- PACE F.,BENARD F., GLOTIN H., ADAM O. ,WHITE P. *"Automatic clustering of humpback whale songs for subunits sequence analyses" in Internat. Journal of Applied Acoustics, 2010*
- GLOTIN H., GIRAUDET P., CAUDAL F., BREVET international : *Procédé de trajectographie en temps réel de plusieurs cétacés par acoustique passive, institut National de la Propriété Intellectuelle, INPI, 2007. numéro 07/06162*, étendu PCT 2009 USA, Canada, Australie, Nouvelle Zeeland, Europe.
- Laran S., M. Castellote, F. Caudal, A. Monnin & H. Glotin, *Suivi par acoustique passive des cétacés au nord du sanctuaire*. Rapport de recherche du Parc National de Port Cros, 2009, 80 p
- Lelandais, F. Glotin, H., "Mallat's Matching Pursuit of sperm whale clicks in real-time using Daubechies 15 wavelets", in: New Trends for Environmental Monitoring Using Passive Systems, 2008 ISBN: 978-1-4244-2815-1,DOI 10.1109/PASSIVE.2008.4786977

Acknowledgments : **This work is supported by CNRS MASTODONS SABIOD and Institut Universitaire de France « scaled complex scene analysis project »**

Join us to Scaled Bioacoustic Plateform : SABIOD.ORG

SABIOD
Scaled Acoustic BIODiversity platform

Home
NEW: Bioacoustic Workshop @ NIPS, Nevada, dec. 2013
ICML 2013 Workshop
Data Samples
Online CEtacean Tracking
Team
Media
Contact

Bioacoustic signaling is a primary mode of communication and exploration for most of the animals. It enables quick load and transfer of information without any visible contact with the target, tackling the reduced visibility of deep forest (insect, frogs, birds, mammals...), cave or night activities (insects, bats), and/or the long distances like in ocean (krill, fishes, whales...). Bioacoustics is also one of the factors in optimizing natural selection, playing a significant role in signalling resource qualities to potential partners. The SABIOD project aims to detect, cluster, classify and index bioacoustic big data in various ecosystems, at different space and time scales, in order to reveal informations on the complex sensori-motor loop, and on the health of an ecosystem, yielding to new biodiversity insights.

NEWS:
- Int workshop Neural Information Scaled to Bioacoustics (joint to NIPS 2013) 10th dec - deadline ext. abstract 13th oct.
- IEEE ATSIP'14 last CFP - Special session on Bioacoustics Int. conf. on Ad. Tech. for Signal & Image Processing

3.3 Classification of Mysticete Sounds using Machine Learning Techniques

Halkias X., Paris S., Glotin H. - CNRS LSIS, USTV, Toulon, FR

Classification of mysticete sounds has long been a challenging task in the bioacoustics field. The diverse nature of the signals due to the inherent variations as well as the use of different recording apparatus and low Signal to Noise Ratio conditions, often lead to systems that are not able to generalize across different species and require either manual interaction or hyper-tuning in order to fit the underlying distributions. This talk presents a Restricted Boltzmann Machine (RBM) and a Sparse Auto-Encoder (SAE) in order to learn discriminative structure tokens for the different calls, which can then be used in a classification framework.

Classification of Mysticete Sounds using Sparse Architectures

LSIS/DYNI

12/10/2013

Outline

Overview

Goal
Automatic Classification of 5 different Mysticete species vocalizations.

1 Mysticete species: Blue whales, Humpback whales, Bowhead whales, Fin whales and Southern Right whales.
2 Applications
- Off-line analysis:
  - Big-data problem: Indexing/Archiving/Detection improvement of multiple, large recordings.
- On-line analysis:
  - Real-time monitoring: Navigation/Preservation (endangered species)/Migration patterns

Mysticete sounds I

Figure: Left to right: Southern Right, Humpback, Bowhead, Blue, Fin whale

Species	ROI Duration	F_s	Freq. Range
Southern Right Whale	130min	8kHz	50-1500Hz
Humpback Whale	55min	4kHz	150-1700Hz
Bowhead Whale	84min	4kHz	110-800Hz
Blue Whale	378min	100Hz	15-18Hz
Fin Whale	162min	100Hz	15-30Hz

Mysticete sounds II

Figure: Range of: Southern Right, Humpback, Bowhead, Blue, Fin whale

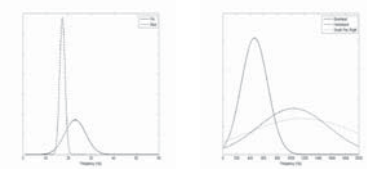


Figure: Ground truth distributions for mysticete species

Existing methodologies

- Region of Interest (ROI) detection
- Feature extraction
  - Auditory perception features, spectrograms, freaquency contours, cepstral coefficients etc.
- Classification
  - Single-species: Boosting of true positive rates (TPR) for single species detection/2 class problem species vs. noise
  - Multiple-species: Multiple species recognition in noisy environment/ multi-class problem
  - Ground truth limitations (experts), supervised approaches (SVM, ANN, RFT, HMM etc.)

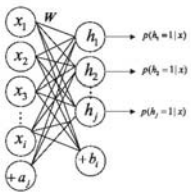
In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

Our approach

- Assume ROI detected
- Feature extraction
  - Learn features for the different species in an unsupervised way
  - Optimize the extracted features by using sparsity
- Classification
  - Classify features in a multi-class framework using a supervised approach

Feature extraction - RBM I

Architecture



Equations

$$p(\mathbf{x},\mathbf{h}) = \frac{e^{-E(\mathbf{x},\mathbf{h})}}{Z}$$

$$p(x_i = 1|\mathbf{h}) = \sigma(b_i + \sum_j h_j w_{ij})$$
$$p(h_j = 1|\mathbf{x}) = \sigma(a_j + \sum_i x_i w_{ij})$$

$$p(\mathbf{x}|\mathbf{h}) = \prod_i p(x_i|\mathbf{h})$$
$$p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$$

$$E_\theta(\mathbf{x},\mathbf{h}) = -\sum_{i=1}^{I}\sum_{j=1}^{J} x_i h_j w_{ij} - \sum_{i=1}^{I} b_i x_i - \sum_{j=1}^{J} a_j h_j$$

Feature Extraction - RBM II

- Training with Gradient descent (GD)
  $-\frac{\partial log p(\mathbf{x})}{\partial \theta} = \langle \frac{\partial E_\theta(\mathbf{x}^\mathbf{n},\mathbf{h})}{\partial \theta} \rangle_\mathbf{h} - \langle \frac{\partial E_\theta(\mathbf{x},\mathbf{h})}{\partial \theta} \rangle_{\mathbf{x},\mathbf{h}}$
- Update equations using Contrastive Divergence (CD)

$$\Delta \mathbf{w}_{\cdot j} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^\mathbf{n} p(h_j = 1|\mathbf{x}^\mathbf{n}) - \widetilde{\mathbf{x}}^\mathbf{n} p(h_j = 1|\widetilde{\mathbf{x}}^\mathbf{n})$$
$$\Delta b_i = \frac{1}{N}\sum_{n=1}^{N} p(x_i^n = 1|\mathbf{h}) - p(\widetilde{x}_i^n = 1|\mathbf{h})$$
$$\Delta a_j = \frac{1}{N}\sum_{l=1}^{N} p(h_j = 1|\mathbf{x}^\mathbf{n}) - p(h_j = 1|\widetilde{\mathbf{x}}^\mathbf{n})$$

Feature Extraction - RBM III

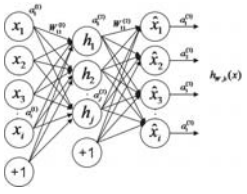
- Sparsity: Weight decay
  $J(\mathbf{W},\mathbf{b},\mathbf{a};\mathbf{x}) = -log p(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{W}\|_2^2$
  $\Delta \mathbf{w}_{\cdot j} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^\mathbf{n} p(h_j = 1|\mathbf{x}^\mathbf{n}) - \widetilde{\mathbf{x}}^\mathbf{n} p(h_j = 1|\widetilde{\mathbf{x}}^\mathbf{n}) + \lambda \mathbf{w}_{\cdot j}$
- Gaussian input units

$$E_\theta(\mathbf{x},\mathbf{h}) = -\sum_{i=1}^{I}\frac{x_i}{\sigma_i}\sum_{j=1}^{J} h_j w_{ij} - \sum_{i=1}^{I}\frac{(x_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^{J} a_j h_j$$

  - Normalize data to have zero mean and unit variance
  - Allow for real valued signals i.e. natural scenes, cepstral coefficients etc.

Feature extraction - SAE I

Architecture



Equations

$$\alpha_i^{(l)} = \sigma(z_i^{(l)})$$
$$z_i^{(l)} = \sum_j W_{ij}^{(l)} \alpha_j^{(l)} + b_i^{(l)}$$

$$\sigma(z^{(3)}) = W^{(2)}\alpha^{(2)} + b^{(2)}$$

$$h_{W,b}(x) = \alpha^{(3)} = \sigma(z^{(3)})$$

Feature Extraction - SAE II

- Training with Gradient descent (GD)
  $J(\mathbf{W},\mathbf{b};\mathbf{x}) = \frac{1}{2}\|h_{W,b}(\mathbf{x}) - \mathbf{x}\|^2$
- Update equations

$$\Delta W_{ij}^{(l)} = \frac{1}{N}\sum_{n=1}^{N} \nabla_{W^{(l)}} J(\mathbf{W^{(l)}}, \mathbf{b^{(l)}}; \mathbf{x^{(n)}})$$

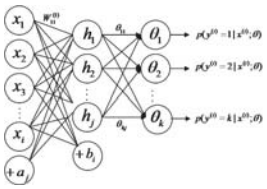
  - Compute gradients via backpropagation
    $\nabla_{W^{(l)}} J(\mathbf{W},\mathbf{b};\mathbf{x}) = \delta^{(l+1)}(\alpha^{(\mathbf{l})})^T$
    $\nabla_{b^{(l)}} J(\mathbf{W},\mathbf{b};\mathbf{x}) = \delta^{(l+1)}$

Feature Extraction - SAE III

- Sparsity: Weight decay
  $J(\mathbf{W},\mathbf{b};\mathbf{x}) = \frac{1}{2}\|h_{W,b}(\mathbf{x}) - \mathbf{x}\|^2 + \frac{\lambda}{2}\|\mathbf{W}\|_2^2$
  $\Delta W_{ij}^{(l)} = \frac{1}{N}\sum_{n=1}^{N} \nabla_{W^{(l)}} J(\mathbf{W^{(l)}}, \mathbf{b^{(l)}}; \mathbf{x^{(n)}})] + \lambda W_{ij}^{(l)}$
- Sparsity: Kullback-Lieibler Divergence
  $J_{sparse} = J + \beta \sum_{j=1}^{J_2} KL(\varrho||\hat{\varrho}_j)$
  - Penalize activations based on a constant/constrain hidden units
  - Incorporate into backpropagation for gradient computation

Classification - Softmax Regressor

Architecture



Equations

$$\mathbf{h_\theta(x)} = \frac{1}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \\ \vdots \\ e^{\theta_k^T x} \end{bmatrix}$$

$$J(\boldsymbol{\theta}) = -\frac{1}{N}[\sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{1}\{y^{(n)} = j\} log \frac{e^{\boldsymbol{\theta_j}^T x^{(n)}}}{\sum_{l=1}^{k} e^{\boldsymbol{\theta_l}^T x^{(n)}}}] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{N}\theta_{ij}^2$$

Overview

- Data (5 Mysticete species) and ground truth from Mobysound.org[1]
- Prior information: Classifying within a frequency range
  - Classifying with a noise class
  - Classifying without a noise class
- No prior information: Classifying all species
  - Classifying with a noise class
  - Classifying without a noise class

[1] http://www.mobysound.org/

Data and pre-processing

- Assume extraction of Regions of Interest (ROI)
  - ROI: Detected box areas in spectrogram that include calls
  - Common ground truth format
- Extract 5000 random patches from ROI and 5000 from noise
- Normalize patches: zero mean, unit variance
- Feature vector: 1600x1 scaled and normalized patch

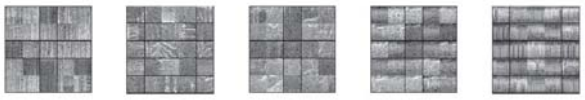
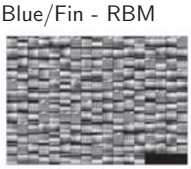


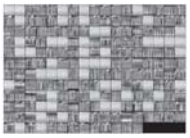
Figure: Sample patches(x-axis, y-axis: bin number). Left to right: Southern Right, Humpback, Bowhead, Blue and Fin.

Blue vs. Fin Whale

Blue/Fin - RBM



Blue/Fin - SAE



SAE/RBM(%) - Actual Value

Predicted Value	Blue	Fin
Blue	93.42/96.25	2.63/1.20
Fin	6.58/3.75	97.37/98.80

Table: Confusion matrix for Blue and Fin whale using the SAE and RBM architectures

Blue Whale vs. Fin Whale vs. Noise

SAE/RBM(%) - Actual Value

Predicted Value	Blue	Fin	Noise
Blue	92.85/93.67	1.66/1.66	2.52/2.71
Fin	1.50/1.19	90.10/90.69	3.20/2.92
Noise	5.65/5.14	8.24/7.65	94.28/94.37

Table: Confusion matrix for Blue whale, Fin whale and noise using the SAE and RBM architectures

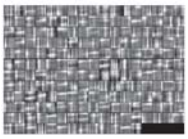
Overall results: Blue and Fin whales

Blue/Fin	
Model	Classification Accuracy
SAE	95.40%
RBM	97.50%
Blue/Fin/Noise	
Model	Classification Accuracy
SAE	92.88%
RBM	93.26%

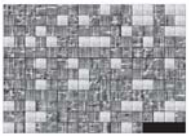
Table: Classification accuracies for species within the same frequency range with and without a noise class using the SAE and RBM architectures

Bowhead vs. Humpback vs. Southern Right Whale

Bowhead/Humpback/South.R - RBM



Bowhead/Humpback/South.R - SAE



SAE/RBM(%) - Actual Value

Predicted Value	South. Right	Humpback	Bowhead
South. Right	83.70/85.26	11.00/9.63	4.74/7.69
Humpback	7.99/6.20	61.20/58.09	15.95/14.88
Bowhead	8.31/8.54	27.80/32.28	79.31/77.43

Bowhead vs. Humpback vs. Southern Right Whale vs. Noise

SAE/RBM(%) - Actual Value

Pr. Value	S. Right	Humpback	Bowhead	Noise
S. Right	77.89/75.69	6.58/5.62	2.49/2.58	3.02/2.62
Humpback	4.23/2.34	28.84/27.77	4.03/4.14	4.72/4.95
Bowhead	1.39/0.81	3.82/3.13	13.68/13.19	4.37/4.47
Noise	16.49/21.16	60.76/63.48	79.80/80.09	87.89/87.95

Table: Confusion matrix for Southern Right whale, Humback whale, Bowhead whale and noise using the SAE and RBM architectures

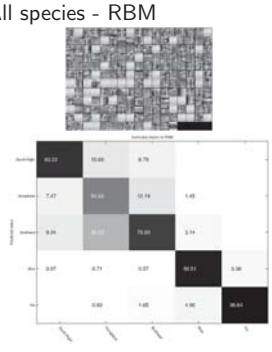
Overall results: Bowhead vs. Humpback vs. Southern Right Whale vs. Noise

Bowhead/Humpback/Southern Right	
Model	Classification Accuracy
SAE	74.75%
RBM	73.30%
Bowhead/Humpback/Southern Right/Noise	
Model	Classification Accuracy
SAE	63.73%
RBM	63.16%

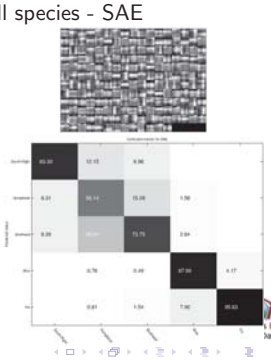
Table: Classification accuracies for species within the same frequency range with and without a noise class using the SAE and RBM architectures

All species

All species - RBM

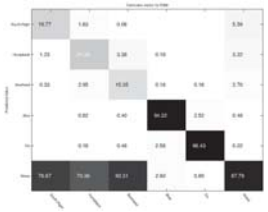


All species - SAE

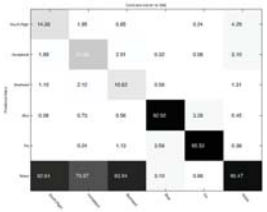


All species and noise

All species and noise - RBM



All species and noise - SAE



Overall results: All species and Noise

Bowhead/Humpback/Southern Right/Blue/Fin	
Model	Classification Accuracy
SAE	79.54%
RBM	80.68%
Bowhead/Humpback/Southern Right/Blue/Fin/Noise	
Model	Classification Accuracy
SAE	69.40%
RBM	68.90%

Table: Classification accuracies

Parameter tuning

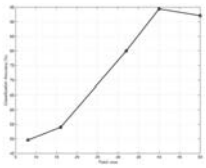
Model	Parameter	Value
SAE/RBM/Softmax	Weight-decay λ	0.003
SAE	Sparsity ϱ	0.05
SAE	Sparsity weight β	3
SAE/RBM	Hidden Units	200
SAE/Softmax	GD Iterations	300
RBM	CG Iterations	500

Table: Values of hyper-parameters.

Important variables

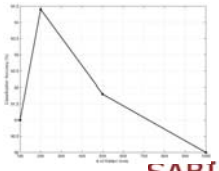
Patch size

- Input space dimensionality



Hidden units

- Feature space dimensionality



Closing thoughts

- Automatic classification for mysticete sounds for indexing/labeling
  - Multi-class task for species recognition
- Semi-supervised approach
  - RBM/SAE unsupervised methodology for feature extraction
  - Sparse constraints for discriminative features
  - Features identify salient call structures
  - Extension for both detection and classification

Thank you
QUESTIONS?

References

S. David,N. Mesgarani, and S. Shamma, Estimating sparse spectro-temporal receptive fields with natural stimuli, Network: Computation in Neural Systems,volume 18, pages 191–212,2007.

P. J. Dugan, A. N. Rice, I. Urazghildiiev, and C. W. Clark, North atlantic right whale acoustic signal processing: Part i. comparison of machine learning algorithms, in Proceedings of the IEEE Applications and Technology Conference, pages 1–6, Long Island, USA, 2010.

Y. Ma and K. Chen, A time-frequency perceptual feature for classification of marine mammal sounds, in 9th International Conference on Signal Processing ICSP, pages 2820–2823, 2008.

References II

- M. A. Roch, M. S. Soldevilla, R. Hoenigman, S. M. Wiggins, and J. H. Hilderbrand, Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes, Journal of Canadian Acoustics, volume 36, pages 41–47, 2008.
- D. K. Mellinger, A comparison of methods for detetcting right whale calls, Journal of Canadian Acoustics, volume 32, pages 55–65, 2004.
- M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, Efficient learning of sparse representations with an energy-based model, in Advances in Neural Information Processing Systems (NIPS), edited by B. Schölkopf, J. Platt, and T. Hoffman, pages 1137–1144, publisher MIT Press, Cambridge, MA, 2007.

Chapter 4

Advanced ANN

4.1 ConvNets & DNN for Bioacoustics

Yann LeCun - New York University, USA

Intelligent perceptual tasks such as audition require the construction of good internal representations. Theoretical and empirical evidence suggest that the perceptual world is best represented by a multi-stage hierarchy in which features in successive stages are increasingly global, invariant, and abstract. An important challenge for Machine Learning is to devise "deep learning" methods for multi-stage architecture that can automatically learn good feature hierarchies from labeled and unlabeled data. A class of such methods that combine unsupervised sparse coding, and supervised refinement will be described. We demonstrate the use of these deep learning methods to train convolutional networks (ConvNets). ConvNets are biologically-inspired architectures consisting of multiple stages of filter banks, interspersed with non-linear operations, and spatial pooling operations, analogous to the simple cells and complex cells in the mammalian auditory cortex. A number of applications will be shown.

The full talk is available on line at http://sabiod.univ-tln.fr/nips4b/

4.2 Mapping functional equations to the topology of networks yields a natural interpolation method for time series data

Lars Kindermann
Alfred Wegener Institute
for Polar and Marine Research
Bremerhaven, Germany
lars.kindermann@awi.de

Achim Lewandowski
Austrian Research Institute
for Artificial Intelligence
Vienna, Austria
achim@oefai.at

Abstract

Typically machine learning methods attempt to construct from some limited amount of data a more general model which extends the range of application beyond the available examples. Many methods specifically attempt to be purely data driven, assuming, that everything is contained in the data. On the other hand, there often exists additional abstract knowledge about the system to be modeled, but there is no obvious method how to combine these two domains. We propose the calculus of functional equations as an appropriate language to describe many relations in a way that is more general than a typical parameterized model, but allows to be more specific about the setting than using an universal approximation scheme like neural networks. Symmetries, conservation laws, and concepts like determinism can be expressed this way. Many of these functional equations can be translated into specific network structures and topologies, which will constrain the possible input-output relations of the network to the solution space of the equations. This results in less data that is necessary for training and may lead to more general results, too, that can be derived from the model. As an example, a natural method for inter- or extrapolation of time series is derived, which does not use any fixed interpolation scheme but is automatically constructed from the knowledge/assumption that the data series is generated by an underlying deterministic dynamical system.

1 Introduction

To interpolate data which is sampled in finite, discrete time steps into a continuous signal e.g. for resampling, normally a model has to be introduced for this purpose, like linear interpolation, splines, etc. In this paper we attempt to derive a natural method of interpolation, where the correct model is derived from the data itself, using some general assumptions about the underlying process. Applying the formalism of generalized iteration, iteration semigroups and iterative roots from the mathematical branch of functional equations, we attempt to characterize a method to determine if such a natural interpolation for a given time series exists and give a method for it's calculation, a formal one for linear autoregressive time series and a neural network approximation for the general nonlinear case.

Let x_t be an auto regressive time series: $x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-n}) + \varepsilon_t$. We will not deal here with finding f, i.e. predicting the time series, instead we assume f is already known or already approximately derived from the given data. We will attempt to embed the discrete series of x_t, $t = 0, 1, 2, \ldots$ into a continuous function $x(t)$, $t \in R^+$. To clarify the idea we present the method at first for the case that the timeseries is generated totally deterministically ($\varepsilon_t = 0$) by an underlying autonomous dynamical system. Later we will consider the influences of additional external inputs and noise.

The time evolution of any autonomous dynamical systems is represented by a solution of the translation equation [1],

$$\Phi(x_0, t_1 + t_2) = \Phi(\Phi(x_0, t_1), t_2) \quad (1)$$

where x_0 is an state vector representing an initial condition and t_1, t_2 are arbitrary time intervals. For continuous time dynamical systems this equation holds for every positive t. If we assume that the given time series is a discrete time sampling of an underlying continuously evolving signal, we have to solve (1) under the conditions $\Phi(x, 0) = x$ and $\Phi(x, 1) = f(x)$, where f is the discrete time mapping represented by the data. (Without loss of generality we can assume the sampling rate of the discrete time data to be one, which will result in a nice and very intuitive formalism.)

To double the sampling rate for example, (1) becomes $f(x) = \Phi\left(\Phi\left(x, \frac{1}{2}\right), \frac{1}{2}\right)$.

Substituting $\varphi(x) \equiv \Phi\left(x, \frac{1}{2}\right)$ we get $\varphi(\varphi(x)) = f(x)$, the functional equation for the iterative root of the mapping f [3].

By introducing the formal notation $\Phi(x, t) \equiv f^t(x)$ the connection to iteration theory becomes clearly visible: Time evolution of discrete time systems can be regarded as the iteration of a time step mapping function (iterative map) and this concept extends to continuous time dynamical systems by means of generalized or continuous iteration, allowing for non-integer iteration counts. The following mathematical problems appear [3,4]:

- For a given function f, does there exist the iteration semigroup f^t?,
- Is the solution unique?
- How to calculate it explicitly or numerically.

To apply this theory, usually x has to be a complete state vector of the dynamical system. This means that f has to be a function of the last state only: $x_t = f(x_{t-1})$. When f also depends on earlier values of the time series $x_{t-2}, \ldots, x_{t-n}$, there must be some hidden variables. In order to obtain a self-mapping we introduce the function $F: R^n \to R^n$ which maps the vector

$\vec{x}_{t-1} = \left[x_{t-1}, \ldots, x_{t-n}\right]$ to $\vec{x}_t = \left[x_t, x_{t-1}, \ldots, x_{t-(n-1)}\right]$

with $x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-n})$. Except for the first element this is a trivial time shift operation, each element of $\vec{x}$ is just replaced by its successor. But because F is a self mapping within a n-dimensional space now, time development can be calculated by iterating F and we can try to find the generalized iteration with non-integer iteration counts to find a time continuos embedding F^t, the continuous iteration semigroup of F and extract a function $x(t)$ from this [2].

2 Linear Case

The idea is best demonstrated for the linear case, where it's application simplifies and unifies several problems. For a *linear* autoregessive time series AR(n) model with $x_t = \sum_{k=1}^{n} a_k x_{t-k}$, the mapping F can be written as a square matrix $F = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

with the coefficients a_k in the first row and the lower subdiagonal filled with ones. Then we can compute $\vec{x}_t = F \cdot \vec{x}_{t-1}$ and the discrete time evolution of the system can now be calculated using the matrix powers $\vec{x}_{t+n} = F^n \cdot \vec{x}_{t-1}$.

This autoregressive system is called linear embeddable if the matrix power F^t exists also for all real $t \in R^+$. This is the case if F can be decomposed into $F = S \cdot A \cdot S^{-1}$ with A being a diagonal matrix consisting of the eigenvalues λ_i of F and S being an invertible square matrix which columns are the eigenvectors of F. Additionally all λ_i must be non-negative to have a linear and *real* embedding, otherwise we will get a *complex* embedding.

Then we can obtain $F^t = S \cdot A^t \cdot S^{-1}$ with $A^t = \begin{bmatrix} \lambda_1{}^t & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \lambda_n{}^t \end{bmatrix}$.

Now we have a continuous function $\vec{x}(t) = F^t \cdot \vec{x}_0$ and the interpolation of the original time series $x(t)$ consists of the first element of $\vec{x}$.

In case there is also a constant term, i.e. the mean is not zero, $x_t = \sum_{k=1}^{n} a_k x_{t-k} + \boldsymbol{b}$, we just have to append a constant one to all the vectors $\vec{x}_t = \left[x_t, x_{t-1}, \dots, x_{t-(n-1)}, \mathbf{1}\right]$ and take $F = \begin{bmatrix} a_1 & a_2 & \dots & a_n & \boldsymbol{b} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$.

A special case is $F^{1/2}$, the square root of a matrix, which solves the matrix equation $F^{1/2} \cdot F^{1/2} = F$. It resembles the iterative root of linear functions and corresponds to a doubling of the sampling rate.

A few lines of Maple code can automate this procedure both for symbolic and numeric expressions. A sample worksheet is available at the authors web page.

3 Examples

We will now provide some simple examples to demonstrate this formalism.

3.1 One dimensional linear case

The time series given by some x_0 and $x_t = 2x_{t-1}$ simply doubles every time step. The natural interpolation we immediately get by applying the former formalism in the trivial one dimensional case is $x(t) = 2^t x_0$, which is of course exactly what we expect: exponential growth. But a little change makes the problem much more difficult: If $x_t = ax_{t-1} + b$ we expect a mixture of constant and exponential growth, but what is the exact continuous law?

Take $\vec{x}_t = [x_t, 1]$ then the series is generated by $F = \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}$. We get immediately by eigenvalue decomposition $\vec{x}(t) = F^t \vec{x}_0 = SA^t S^{-1} \vec{x}_0 = \begin{bmatrix} 1 & 1 \\ \frac{1-a}{b} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & a^t \end{bmatrix} \begin{bmatrix} 0 & \frac{b}{1-a} \\ 1 & \frac{b}{a-1} \end{bmatrix} \begin{bmatrix} x_0 \\ 1 \end{bmatrix}$

and for the first component $x(t) = a^t x_0 + b\frac{a^t - 1}{a-1}$ (which equals $ax_0 + b$ for $t \to 1$).

We don't have to consider about stability or stationarity of the AR(1) model here but note that to obtain a completely real valued function $x(t)$, a has to be positive. Later we will discuss about the meaning of such cases with complex embeddings, but for short it means that there is no one-dimensional continuous time dynamical system that can generate such timeseries. In the linear case this should be clear because a negative a implies oscillatory behavior of $x(t)$. This means, some initial condition x_0 won't be enough to determine the continuation of the trajectory, it could be on the rising or falling slope. The underlying dynamical system needs to have one more hidden dimension to allow embedding. The other dimension can be represented by the imaginary part of $x(t)$ which will vanish at all integer times t. But taking only the real part will still result in a valid interpolation of the given series, the observable of the system.

This is such an embedding of the AR(2) process $x_t = -\frac{1}{2}x_{t-1} + \frac{1}{3}x_{t-2} + \frac{1}{2}$ with $x_0 = x_1 = 1$.

Circles mark the time series x_t, the left graph shows the real part of $x(t)$, our *natural interpolation*, the imaginary part is on the right.

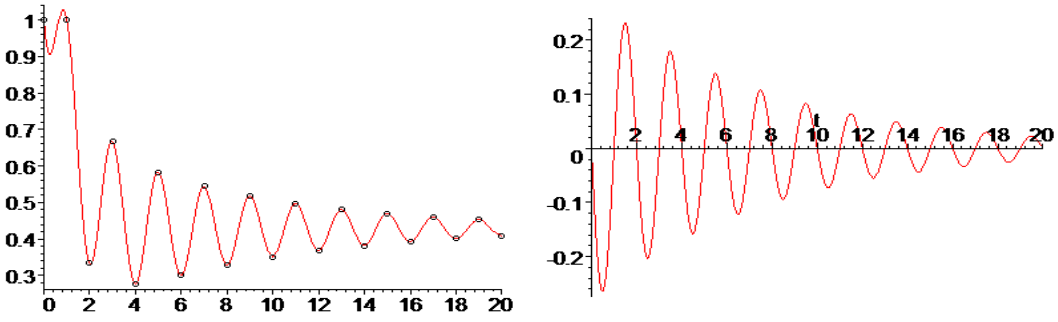


Figure 1: Embedding of an AR(2) process

.

3.2 Two dimensional linear case

The well known Fibonacci series $x_0 = 0$, $x_1 = 1$, $x_t = x_{t-1} + x_{t-2}$ can be generated in this manner by $F = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ and $\vec{x}_1 = [1, 0]$. By eigenvalue decomposition of F we get

$$\vec{x}_{t+1} = F^t \vec{x}_1 = SA^t S^{-1} \vec{x}_1 = \begin{bmatrix} \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^t & 0 \\ 0 & \left(\frac{1-\sqrt{5}}{2}\right)^t \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{1}{2} - \frac{1}{2\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{1}{2} + \frac{1}{2\sqrt{5}} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

which turns out to evaluate exactly to Binet's famous formula for the Fibonacci series in the first component $x_t = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^t - \left(\frac{1-\sqrt{5}}{2}\right)^t\right]$ [5].

Because the second eigenvalue is negative, a *real* linear continuous time embedding does not exist and $x(t)$ takes complex values on non-integer x. Figure 2 shows real and complex part of $x(t)$.

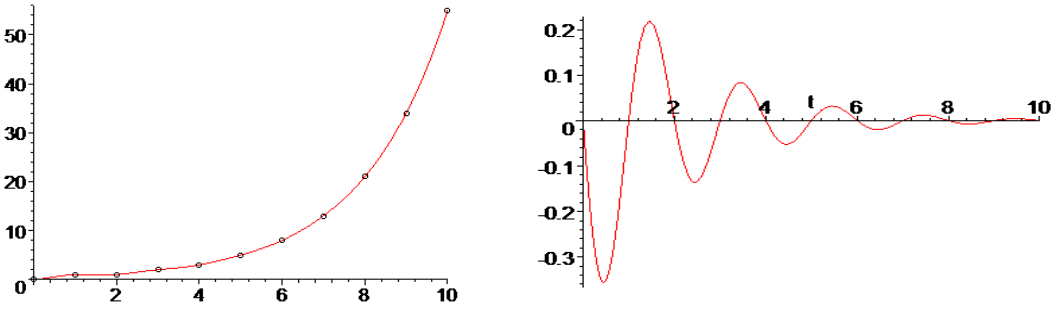


Figure 2: Embedding of the Fibonacci series.

4 The Nonlinear Case

So far we considered only linear dependencies of the past f which could easily be mapped to matrix expressions. The problem becomes much more complicated if we allow for arbitrary f. Even for one dimension this cannot be solved analytically in the general case, so we use neural networks to compute approximations for fractional iterates of arbitrary functions [7].

4.1 One dimensional nonlinear systems

The general solution of the real valued translation equation $\Phi(x_0, t_1 + t_2) = \Phi(\Phi(x_0, t_1), t_2)$ $\Phi\colon R \times R \to R$ with Φ being continuous and strictly monotonic in x and t is given by $\Phi(x, t) = \varphi^{-1}(\varphi(x) + t)$. If the discrete time mapping $\Phi(x, 1) = f(x)$ is given, this is Abel's functional equation $f(x) = \varphi^{-1}(\varphi(x) + 1)$. If f has a fix point this can further be transformed to Schröder's functional equation $f(x) = \varphi^{-1}(c\varphi(x))$, the eigenvalue problem for nonlinear functions. In those cases when either of these equations can be solved for the unknown function φ, it is immediately possible to obtain the embedding by $f^t(x) = \varphi^{-1}(\varphi(x) + t)$ [2].

This can be easily solved for example in the linear case: If we take $x_t = ax_{t-1}$ and initial value x_0 the time step mapping is $f(x) = ax$, we get the Abel type functional equation $ax = \varphi^{-1}(\varphi(x) + 1)$ or $\varphi(ax) = \varphi(x) + 1$, which is solved by $\varphi(x) = \log_a x$. So we get for the continuous embedding the expected result again, exponential growth
$f^t(x) = a^{\log_a x_0 + t} = x_0 a^t$.

However, this analytical method is limited to a small selection of functions and it can be shown that there exist embeddings for a much wider range of mappings which cannot be calculated analytically yet. Furthermore the theory so far is developed mainly for real or complex valued functions, solving Abel or Schröder type functional equations in higher dimensions is currently for the general case beyond reach.

But simple neural networks can be used to find precise approximations for those embeddings [7,8]. The basic idea is to use a MLP with a special topology which approximates $f(x)$. To compute the *fractional iterate* $f^{m/n}$, we use a network that consists of n subnetworks in a row with *pairwise identical weight matrices*. The use of special training algorithm allows to perform the function approximation with the whole network and keep the subnets identical at the same time [9]. The fractional iterate of the function can be read out after the *m*-th subnet.

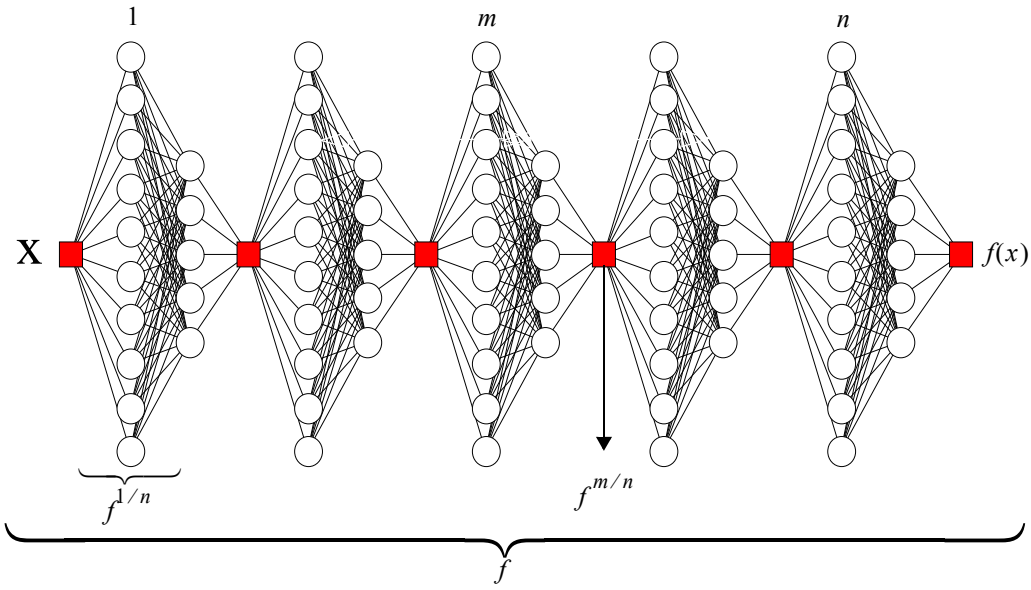


Figure 3: MLP for computing fractional iterates.

4.2 Multidimensional nonlinear system

We took a time series of yearly snapshots from a discrete non linear Lotka-Volterra type predator - prey system (x = hare, y = lynx) as training data.

$$x_{t+1} = (1 + a - b\, y_t)\, x_t$$

$$y_{t+1} = (1 - c + d\, x_t)\, y_t$$

From these samples only we calculated the monthly population by use of a neural network based method to compute iterative roots and fractional iterates with a pseudo newton algorithm [9].

This figure shows the yearly training data as circles and the interpolated monthly values. Additionally the forecasted values for the next 12 months are shown, together with the true value after one year which was excluded from model fitting.

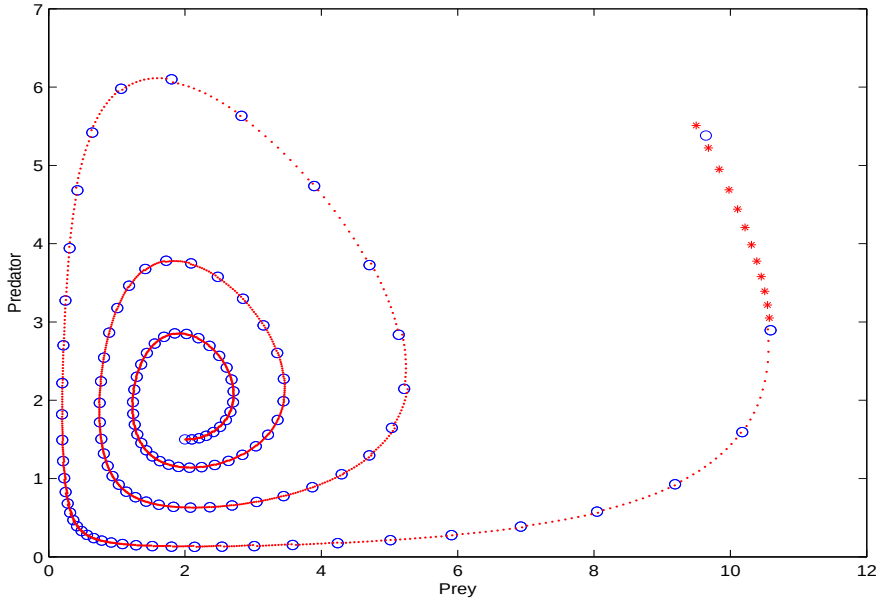


Figure 4: Embedding of a Volterra type system.

The given method provides a natural way to estimate not only the values over a year, but also to extrapolate arbitrarily smooth into the future.

5 Discussion and Outlook

The method demonstrates that there is a close relation between prediction and interpolation. A necessary condition for the existence of a natural interpolation of a time series is predictability. If there are random influences and we require that the values $x(t)$ coincide with x_t for integer t, we can still use the embedding function to get a near fit and add an additional interpolation method for the residuals $x_t - x(t)$. This has again to be selected freely of course, but it minimizes the amount of arbitrariness involved in interpolating.

Another problem are impossible embeddings. Take $x_{t+1} = 4\lambda x_t(1-x_t)$, the iterated logistic map, which is a favorite textbook example for the emergence of chaotic behavior within a simple dynamical system. However, this is a discrete time system, so the question should arise naturally if it is possible to embed the x_t into continuous trajectories $x(t)$ which now obey the functional equation $x(t+1) = 4\lambda x(t)(1-x(t))$ for any non-integer t. Or even more general, is there *any* continuous time system that takes the same values at integer times? Iteration theory proofs that the answer is no if $\lambda > 3/2$ [6], but this could be expected also by the theorem of Poincare-Bendixon, which implies that chaotic behavior is impossible in continuous time systems of less than three dimensions. To obtain a continuous embedding of this series, we had to introduce some hidden dimension, like allowing complex values for *x*, in iteration theory these generalized solutions are called phantom roots of functions [1]. In neural networks this could be accomplished by introducing additional hidden nodes, allowing to address the general embedding problem.

Acknowledgements

Part of the work was conducted at the RIKEN Brain Science Institute, Wako-shi, Japan.
The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education Science and Culture.

References

1. G. Targonski: *Topics in Iteration Theory.* Vandenhoeck und Ruprecht, Göttingen (1981)
2. M.C. Zdun: *Continuous iteration semigroups.* Boll. Un. Mat. Ital. 14 A (1977) 65-70
3. M. Kuczma, B. Choczewski & R. Ger: *Iterative Functional Equations.* Cambridge University Press, Cambridge (1990)
4. K. Baron & W. Jarczyk: *Recent results on functional equations in a single variable, perspectives and open problems.* Aequationes Math. 61. (2001), 1-48
5. R.L. Graham, D.E. Knuth & O. Patashnik: *Concrete Mathematics.* Addison-Wesley, Massachusetts (1994)
6. R.E. Rice, B. Schweizer & A. Sklar: *When is f(f(z)) = az^2+bz+c for all complex z?* Amer. Math. Monthly 87 (1980) 252-263
7. L. Kindermann: *Computing Iterative Roots with Neural Networks*. Proc. Fifth Conf. Neural Information Processing, ICONIP (1998) Vol. 2:713-715
8. E. Castillo, A. Cobo, J.M Gutiérrez & R.E Pruneda: *Functional Networks with Applications. A Neural-Based Paradigm*. Kluwer Academic Publishers, Boston/Dordrecht/London (1999)
9. L. Kindermann & A. Lewandowski: *A Comparison of Different Neural Methods for Solving Iterative Roots.* Proc. Seventh Int'l Conf. on Neural Information Processing, ICONIP, Taejon (2000) 565-569

Chapter 5

Learning to Track by Passive Acoustics

5.1 Mono-Channel Spectral Attenuation modeled by Hierarchical Neural Net Estimates Hydrophone-Whale Distance

Doh Y.
DYNI team
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13284 Marseille, France
Université de Toulon,CNRS, LSIS, UMR 7296, 83957 La Garde, France
yanndoh.m2@gmail.com

Glotin H.
IUF, Institut Universitaire de France
103 Bd. Saint-Michel, 75005 PARIS - France
DYNI team
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13284 Marseille, France
Université de Toulon,CNRS, LSIS, UMR 7296, 83957 La Garde, France
glotin@univ-tln.fr

Razik J.
DYNI team
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13284 Marseille, France
Université de Toulon,CNRS, LSIS, UMR 7296, 83957 La Garde, France
Joseph.Razik@univ-tln.fr

Paris S.
DYNI team
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13284 Marseille, France
Université de Toulon,CNRS, LSIS, UMR 7296, 83957 La Garde, France
sebastien.paris@lsis.org

Abstract

We aim to allow whale monitoring and anti-collision system using single hydrophone. We then propose a new model to estimate the range from wideband signals such as clicks emitted by odontocetes. We demonstrate that it is possible to link the intrinsec distorsion of the signal with the distance of the acoustic path. We provide different models to establish the relationships between the signal energy and propagation distance. We deal with different energy scales: the global received energy of the signal E_0, the frequency bands energy and frequency bin energy. We then demonstrate that intermediate prediction of the whale orientation enhances the distance estimation, yieding to only 6 % of relative error rate.

1 Introduction

Passive acoustics is one of the best ways to enhance the knowledge of marine mammals emitting sounds trough various tasks: detection, classification, localization and density estimation. 3D whales localisation is mostly achieved using hydrophone arrays. Although these methods have been em-

ployed successfully for this task on mysticetes [6] and odontocetes [1, 3] with a high level of accuracy, they require the use of heavy and expensive hardware. Despite of the information loss, using a single, light and cheap hydrophone device, quick to deploy, could provide the necessary data to satisfy certain applications such as mobile listenning point and anti-collision system where a simple range estimation is sufficient. Therefore, in this paper, we decided to focus on single-hydrophone methods. Theoretically, it is possible to use the virtual hydrophone framework and the acoustic property of the water column in order to estimate the position of the whale [4, 7, 5]. This technique involves the acquisition of the direct acoustic path, the bottom reflection and the surface reflection. However, in practice, we often observe only a subset of the needed information. The above mentioned constraints led us to find a new model of range estimation applied to wideband signals such as clicks emitted by odontocetes. Specifically, this paper applies the proposed model on sperm whale recordings. In order to focus on range estimating in this first approach, and to avoid source separation problems, we choose to deal with single animal recordings.

It is well known that sound attenuation depends on frequency and propagation distance [8, 9]. The attenuation impacts the total energy and generates a distortion of frequency representation in the emitted signal that we can link to the distance between source and receiver. Madsen et al. in 2002, put forward the variation of the centroid with the distance et suggest a low pass effect provided by the acoustic propagation [18].

In this paper, we try to establish the expression of the relationship between the signal energy and propagation distance by an empirical model based on a neural network and by the theoretical model Inter-Frequency Attenuation (IFA) [10, 11].

Real data and their associated ground truth [3, 12, 13, 14], will allow us to develop this model by optimizing a few but important parameters. Also, an independent partition of the data will be useful to test the ability and limits of the proposed estimator.

2 Motivation

2.1 The relationship between the received signal and loss by transmission

Our goal is to extract information regarding the propagation distance, from the observed signal from a unique hydrophone. The relationship between Transmission Loss (TL) and distance, as provided by the passive sonar equation, is not well adapted to bioacoustics signal applications. Solving this equation depends mainly on the signal's power at the origin Source Level (SL). Emitting with a variable sound level, a sperm whale is not a constant acoustic energy source. Moreover, we must also take into account other variability factors like the animal's size, the Inter-Click Interval (ICI) and Inter-Pulse Interval (IPI), diving depth [18] or the directionality of the animal relative to the hydrophone position [2, 17, 20, 21].

First we introduce the expression of the received energy E at the distance r from the source, as a function of the energy source level E_{SL} and TL [15, 7] for a given frequency in dB. We consider the simple framework of omnidirectionnal spherical source:

$$E(r, f) = E_{SL}(f) - TL(r, f), \tag{1}$$

where the transmission loss TL can be decomposed by

$$TL(r, f) = 20 \log_{10}(r) + \alpha(f) r \ , \tag{2}$$

where r is the propagation distance (in m), f is the frequency (in Hz) and α is the frequency attenuation coefficient ($dB.m^{-1}$). The first term of Eq. (2) is due to loss by geometric divergence of a spherical wave while the second term represents the frequency attenuation because of interactions between the wave and the medium.

In a first approximation, we can assume that the loss by divergence is predominant on frequency attenuation and TL does not depend on frequency which allows to consider the total energy E_0 of the signal.

The problem $r = F(E_0)$ remains unsolved without a theoretical or statistical model of the energy source level. This function can be approximated and learned by a neural [24] network in particular a MLP, since MLPs are universal function approximators and will be described in a further section. It will be the basic model LER (for Loss Estimation Regression). But this function is empirical and very dependant on the data used to learn the model. Thus, it has motivated us to find a theoretical relationship only based on the frequency attenuation which could work without any knowledge of the total SL. This model imposes to consider not only the total energy but also the detailled frequency composition.

2.2 Data driven IFA regressions

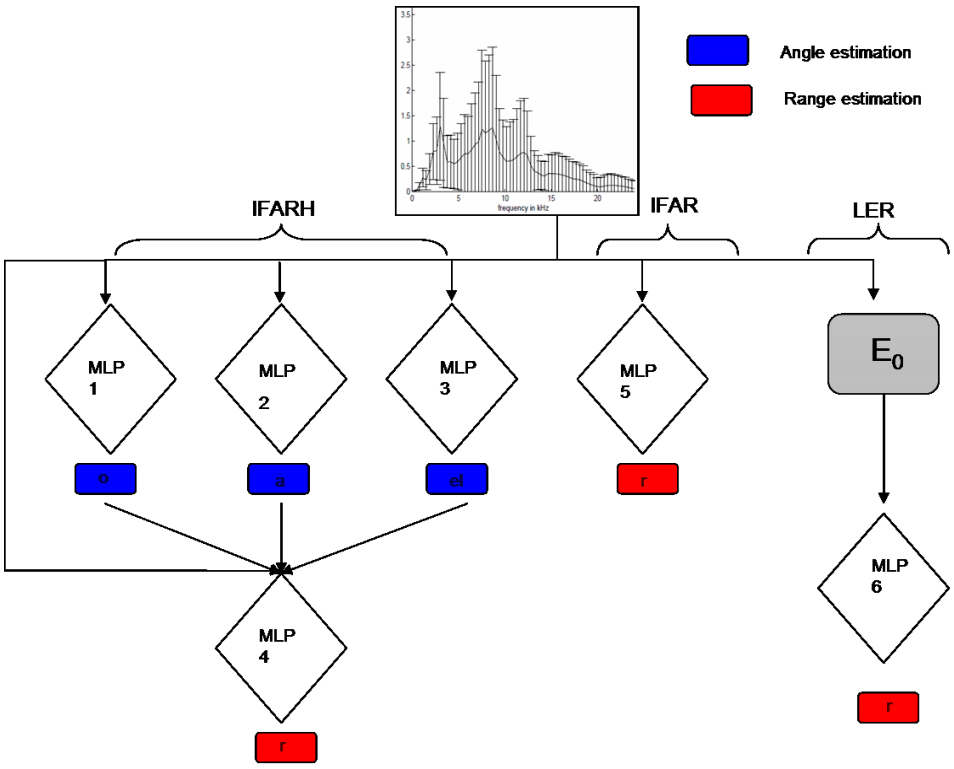


Figure 1: Representation of different neural network we used. LER, IFAR, IFARH

Based on the simplification done, we expect to have limited results using the theoretical model of the IFA. We also used a neural network algorithm [24] to learn a regression model to fit the relation between spectral informations and distance estimations such as :

$$\hat{r} = h\left(|X(f;r)|;\mathbf{W}\right), \tag{3}$$

where $\mathbf{W}$ represents the matrix of weights provided by the training step.

We trained a MLP to a learn a empirical relationship between spectrum and radial distance. Spectrum bins $|X_j(k\Delta F)|$ will be used as input to the network and ground truth radial distance d_j as the output. As presented in previous sections regarding the IFAT estimators, we will work on $N = 256$ samples. It will be the Model $IFAR$. The used MLP is comprised of two hidden layers which are composed by $2N + 1$ units. The MLP attributes the optimal weights describing the regression between spectrums bins and propagation distance.

We know that information on the animal directionality can be extracted from the click spectrum and its energy. Therefore we propose the model $IFARH$ in order to enhance the model by associating in "waterfall" 2 models. A first one learns a regression estimating the 3 angles describing the postion of the animal: off axis o, azimut a and elevation el. A second one learn a regression function given the estimated angles $\hat{o}$,$\hat{a}$ and $\hat{el}$ such as :

$$\hat{r} = h(|X(f;r)|;\hat{o};\hat{a};\hat{el};\mathbf{W}). \tag{4}$$

2.3 Proposition of a theoretical Model : Inter Frequency Attenuation (IFAT)

The proposed Theoretical Inter Frequency Attenuation (IFAT) [10] model aims to extract information from the source distance by taking advantage of the energy ratio between two frequency bands of the emitted signal. the derivation of the attenuation laws allowed us to establish the following relationship :

$$r(B_1, B_2) = \frac{10\log_{10}\left(\frac{E_1}{E_2}\right)}{\int_{F_1'}^{F2'} \alpha(f)df - \int_{F1}^{F_2} \alpha(f)df}. \tag{5}$$

In this expression, r is the acoustic propagation distance, $B_1 = [F_1, F_2]$ and $B_2 = [F_1', F_2']$ are the frequency band involved, F_1' E_1 and E_2 the energy of band 1 and 2. In this expression r does not depend on loss by divergence and energy at the origin, but it only depends on frequency attenuation.

3 Material

In this section, we present our dataset which is extracted from the Bahamas dataset distributed by AUTEC at the second DCL workshop in Monaco 2005. It consists of five hydrophones deployed off the Bahamas Island, and a total of 25 minutes of recording of one sperm whale with a sample rate of 48 KHz. The trajectory computed by LSIS/DYNI (Fig. 2) [3, 14] is similar to the one resulting by different methods by the scientific community [22, 23], and it will be considered as the ground truth.

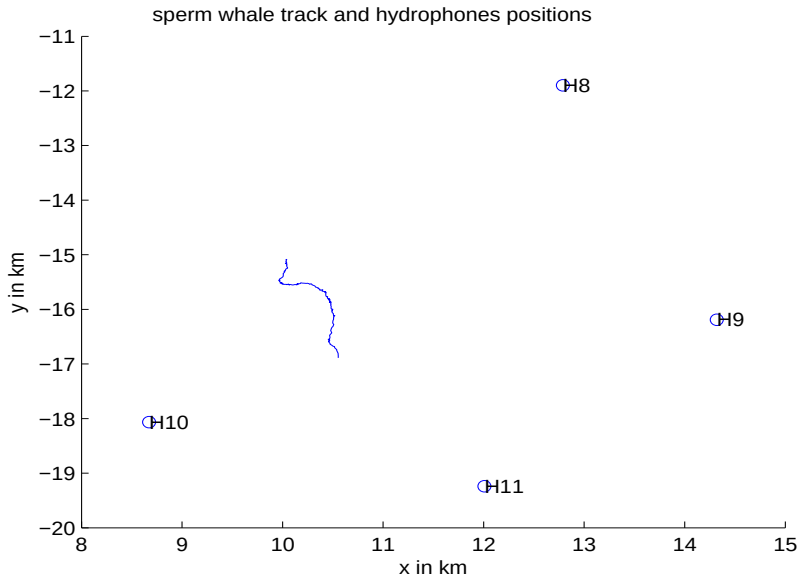


Figure 2: The 2D trajectory (in $x - y$ plan) of the single sperm whale observed during 25 min (LSIS/DYNI [3]) and corresponding hydrophone's positions. The whale goes to south east. Supplemental material with the animated 3D tracking of this whale is available at http://glotin.univ-tln.fr/oncet and http://www.youtube.com/watch?v=0Szo3gdiTRk. We also give there in supplemental material with the file containing the $[x, y, z, t]$ whale positions, and the $[az, el, offaxis, t]$ files for each hydrophone H8.. H11

	H11	H10	H9	H8
hydrophone depth	-1522 m	-1361 m	-1553 m	-1556 m
mean off axis	36 degrees	60 degrees	74 degrees	125 degrees
mean azimut	29 degrees	56 degrees	73 degrees	129 degrees
mean elevation	-14 degrees	-15 degrees	-14 degrees	-13 degrees
mean distance	3937 m	2900 m	4150 m	4716 m
distance std deviation	452 m	242 m	234 m	283 m

Table 1: Mean ground truth distance, azimut and hydrophone depth

Considering the four hydrophones H8, H9, H10, H11, the range of distance between source and receiver is from 2500 m to 5500 m. The precise angle associated to the ground truth trajectory, has been calculated. The mean values are mentionned in Tab. 1.

We then divide the data in 2 partitions. Partition 1 will be used for the development and parameter optimization step. Partition 2 will be dedicated to estimation and predictions.

4 Results

4.1 IFAR and IFARH IFA estimation using MLP model

4.1.1 Training and development

The training step is employed on partition 1 of the data. The temporal order of the data is not taken into account when applied as the input of the MLP. During the development stage, we optimized only one parameter, which is the number of system iterations for the training session (early stopping) related to the quality of prediction.

Predictions are generated from partition 2 of the data and we assume independence with the training data. We were surprised to observe that the MLP learned the relationship between spectrum and propagation distance quite quickly. 300 iterations are sufficient to obtain satisfactory results for our prediction. In order to avoid over-fitting and to extract the most general predictor we keep the number of iterations low.

4.1.2 Distance prediction

In this section, we propose a temporal MLP prediction following LER, $IFAR$ and $IFARH$ on all hydrophone in the same data subset as section IV.A.2 for the IFAT estimator. It has been computed using the MLP for regression on the spectrum.

In Fig. 3 we can compare the fidelity of prediction and ground truth (results summarized in Tab. 3 . Firstly, the prediction given by $IFAR$ and $IFARH$ is better than LER one. It demonstrates the usefullness of considering spectrum inter bins and not only the global energy of the signal. LER shows some transitions which lead to a recess that others models seem to control. However, the bias seems to increase in the sections where the azimut is the lowest (20 degrees). This behavior can be explained by a lack of an on-axis configuration during the training session. This assumption means that the regression law is different between an on-axis and off-axis configuration. It may also be caused by the different frequency structure of pulses (P or PJ) in a click following the receiver position [16].

Then, $IFARH$ mean error is similar to $IFAR$. However, the dispersion of the predicted distances seems enhanced by the use of an intermediate MLP predicting the position angles. The Results of angle estimation are not presented directly in this paper. We noted that azimut prediction was more precise than elevation prediction which could suggest that the spectrum shape is more dependent on azimut.

4.2 Estimation of the radial distance from the theoretical model IFAT

We represent the final temporal estimation $\widehat{\boldsymbol{r}}^*$ of radial distance between hydrophone 11 and the sperm whale. The computation of the estimators has been implemented on Partition 2 of the recordings (test set) according to F_P^* and N_{best}^* learned on the train set data (see previous section).

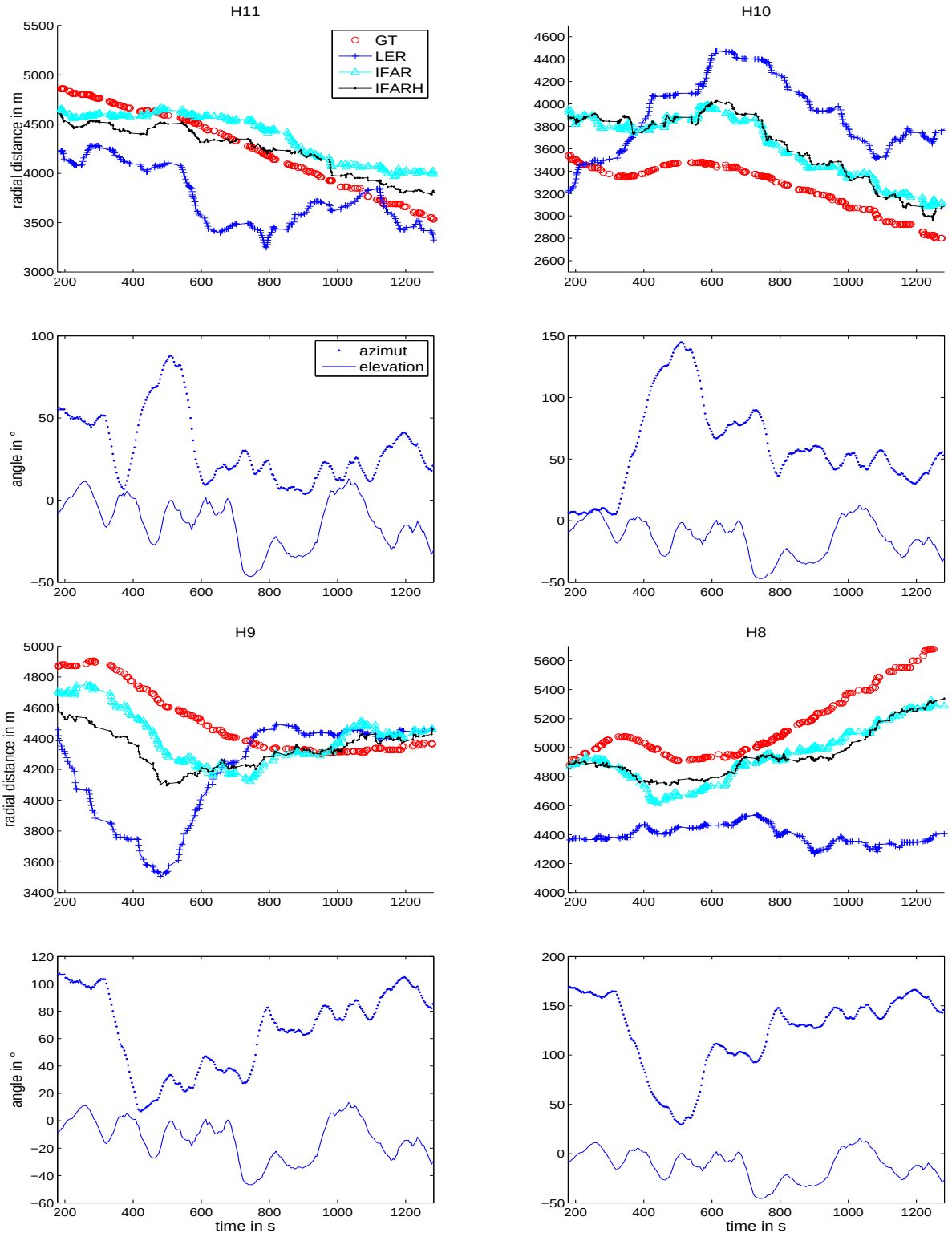


Figure 3: Temporal prediction of radial distance compared to ground truth using model LER, IFAR, IFARH.

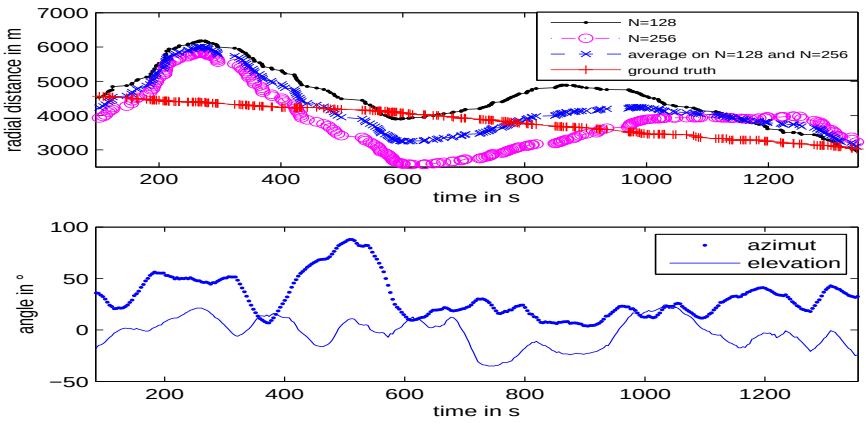


Figure 4: An example of final IFAT temporal estimation of radial distance compared to ground truth radial distance and azimut evolution on H11. fixed parameters : $N_{best} = 4$, $F^*_{P_{128}} = 8.5$kHz and $F^*_{P_{256}} = 9$kHz.

In Fig. 4 we see that both curves have almost the same dispersion. Using $N = 128$ samples, the distances are overestimated, while underestimated using $N = 256$ samples.

The significant dispersion of our estimation cannot be due to the range variation. It can be related to the animal's off-axis variation, meaning that IFAT does not cancel the animal's directionality effects. Most of the errors of IFAT using $N = 128$ samples seem to accumulate on sections where the azimut is high ($> 20\,\mathrm{deg}$). When the animal is supposed to be on-axis, we observe that the estimation is converging to the ground truth.

On the other hand, for the $N = 256$ samples, the lowest error seems to match with high azimut, and it results to a better estimation on high off-axis configurations. Thus the average between $N = 128$ and $N = 256$ samples is computed to provide a uniform behavior relative to the azimut possibilities.

Finally, we also observe that the radial distance is decreasing which is in agreement with the directionality of the whale on the ground truth. IFAT correctly affirms that the whale is traveling towards the hydrophone.

5 discussion

According to the results (tables 2, 3) IFAT errors tend to vary between the different hydrophones, but the IFAT model and the neural network led to common conclusions. The performance of $IFAR$ and $IFARH$ confirms the usefullness to consider spectrum inter bin property and not only the total energy which is more sensitve to natural and voluntary variations of the Source Level.

The IFAT, IFAR, IFARH and LER models provide us with differences regarding the azimut. With the IFAT model, the performance of the estimator seems to be more impacted by the azimut configuration. Since we developed the model without the consideration of the animal's directionality, a future step may be the inclusion of this variable in the estimator expression.

The MLP demonstrates the relationship between spectrum and propagation distance in this data set on only 300 iterations. We demonstrated that the IFAT model could estimate distance with about 15 % of mean relative error, while the MLP IFAR or IFARH reduces it to 6 %.

As IFAT produces locale range estimates, a particle filtering process [19] could be efficiently added after IFAT in order to produce more reliable and complex estimates.

In the case of multiple emitting whale and a monohydrophone recording, IFA would play an important role in order to cluster the clicks and thus estimate the number of emitting whales. Morever, in the case of multiple hydrophones, IFA can help in the localisation of each whale. The IFAR and

	H11	H10	H9	H8	MEAN
IFAT model $N = 128$	21	16	14	41	23
IFA model $N = 256$	20	21	9	16	16.5
average on model $N = 128$ and $N = 256$	17	16	11	28	18
mean total relative error	19	17	11	28	-

Table 2: Absolute mean relative error of estimated distance in % for all hydrophones with IFAT theoretical model estimators

	H11	H10	H9	H8	MEAN
LER mean	14 ± 6	20 ± 10	8 ± 9	15 ± 5	14.2 ± 7.5
$IFAT$ mean	20 ± 10	21 ± 13	9 ± 7	16 ± 14.0	16.5 ± 11
$IFAR$ mean	4 ± 2	14 ± 4	3 ± 2	4 ± 2	6.2 ± 2.5
$IFARH$ mean	4 ± 2.5	11 ± 3	4 ± 3.5	4 ± 2	5.75 ± 2.75

Table 3: Absolute mean relative error of estimated distance by with MLP: IFAR/IFARH and standard deviation (in %)

IFARH models trained on a data set, can be applied on another recording set for similar species and hydrophones. Also one can model and run IFA for other species using biosonar, like bats.

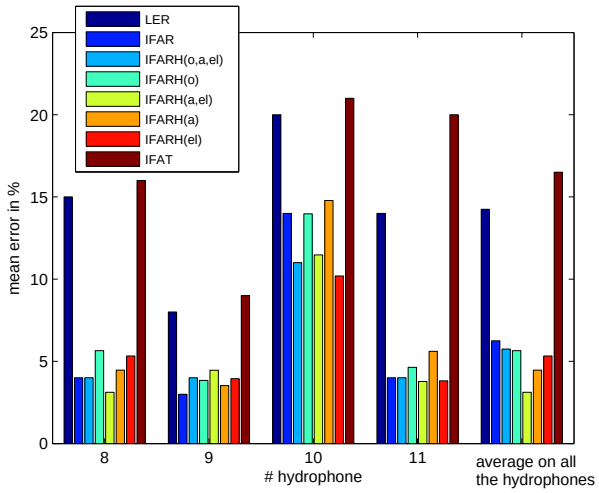


Figure 5: mean relative error between predictions and groundtruth. Different IFARH model versions have been tested : IFARH(o,a,el), IFARH(a,el), IFARH(o), IFARH(a), IFARH(el).

Acknowledgments

The authors gratefully acknowledge the contribution of the Provence Alpes Cote d Azur region, Université Toulon Var, CESIGMA company and the anonymous referees.

References

[1] E.-M. Nosal and L. Frazer, “Sperm whale three-dimensional track, swim orientation, beam pattern, and click levels observed on bottom-mounted hydrophones”, The Journal of the Acoustical Society of America **122(4)**, 1969–1978 (2007).

[2] B. Mohl, M. Wahlberg and P. -T Frazer, “The monopulsed nature of sperm whale clicks”, The Journal of the Acoustical Society of America **114(2)**, 1143–1154 (2003)

[3] P. Giraudet and H. Glotin, “Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely-spaced hydrophone array”, Applied Acoustics **67**, 1106–1117 (2006).

[4] N. Josso, “Characterizing the underwater environment using moving sources by opportunity”, Ph.D. thesis, Université de Grenoble (2010).

[5] X. Mouy and D. Hannay, “Tracking of Pacific walruses in the Chukchi Sea using a single hydrophone”, The Journal of the Acoustical Society of America **131(2)**, 1349–1358 (2012)

[6] S. W. Martin, T. Norris, E. M. Nosal, D. K. Mellinger,R. P. Morrissey and S. Jarvis “Automatic localization of individual Hawaiian minke whales from boing vocalization”, The Journal of the Acoustical Society of America **129**, 2506 (2011)

[7] W. Au and M. Hastings, *Principles of marine bioacoustics* (Springer Science + business Media, LLC, New York), 680p (2008).

[8] C. Leroy, “Sound attenuation between 200 and 10000 cps mesured along single paths”, Technical Report 43, Saclant ASW Research Center (1965).

[9] W. Thorp, “Analytic description of the low frequency attenuation coefficient”, The Journal of the Acoustical Society of America **42(1)**, 270 (1967).

[10] H. Glotin, Y. Doh, R. Abeille, and A. Monnin, “Physeter distance estimation using sub-band leroy transmission loss model”, in *5th Internationnal Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals using Passive Acoustics*, 49 (2011).

[11] Y. Doh, “A model of Inter Frequency Attenuation (IFA) in order to estimate the distance source receiver”, Master’s thesis, Universités d’Aix Marseille/Centrale Marseille (2011).

[12] F. Bénard and H. Glotin, “Whales localization using a large array : performance relative to cramer-rao bounds and confidence regions”, in *e-Business and Telecommunications*, 294–306 (Springer - Verlag, Berlin Heidelberg) (2009).

[13] H. Glotin, F. Bénard, and P. Giraudet, “Whales cocktail party: a real-time tracking of multiple whales”, International Journal Canadian Acoustics **36**, 139–145 (2008).

[14] F. Bénard, H. Glotin, and P. Giraudet, “Highly defined whale group tracking by passive acoustic stochastic matched filter”, in *Advances in Sound Localization*, chapter 28, 527–544 (InTech, Rijeka, Croatia) (2011).

[15] C. Viala, “Real time inversion in geoacoustic of wideband signals in deep water”, Ph.D. thesis, Université de Toulon et du Var (2007).

[16] C. Laplanche, “Studies by passive acoustics of the hunting behaviour of the sperm whale”, Ph.D. thesis, Université Paris XII Val-de-Marne (2005).

[17] B. Mohl, M. Wahlberg, P. Madsen, LA. Miller and A. Surlykke, “Sperm whale clicks, directionality and source level revisited”, J. Acoust. Soc. Am **107**, 638–648 (2000).

[18] P. Madsen, R. Payne, N. U. Kristiansen, M. Walhberg, I. Kerr and B. Mohl, “Sperm whale sound production studied with ultrasound time/depth-recording tags”, The Journal of Experimental Biology **205**, 1899–1906 (2002)

[19] M. Sanjeev Arulampalam, S. Maskell and N. Gordon “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking”, IEEE Transactions on Signal Processing **50**, 174–188 (2002)

[20] W. M. X. Zimmer, P. L. Tyack, M. P. Johnson and P. T. Madsen , “Three-dimensional beam pattern of regular sperm whale clicks confirms bent-horn hypothesis”, J. Acoust. Soc. Am **117(3)**, 1473–1485 (2004).

[21] W. M. X. Zimmer, P. T. Madsen, V. Teloni, M. P. Johnson and P. L. Tyack , “Off-axis effects on the multipulse structure of sperm whale usual clicks with implication for sound production”, J. Acoust. Soc. Am **118(5)**, 3337–3345 (2005).

[22] R. Morrissey, J. Ward, N. DiMarzio, S. Jarvis, and D. Moretti., “Passive acoustic detection and localization of sperm whales (physeter macrocephalus) in the tongue of the ocean”, Applied Acoustics **67**, 1091–1105 (2006).

[23] E.-M. Nosal and L. Frazer, “Track of a sperm whale from delays between direct and surface-reflected clicks”, Applied Acoustics **67**, 1187–1201 (2006).

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (Wiley-Interscience, New York), 654p (2000).

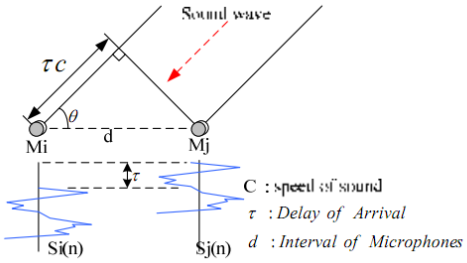
5.2 Physeter Localization: Sparse coding & Fisher vectors

S. Paris, H. Glotin, Y. Doh, J. Razik
USTV, Institut Universitaire de France, CNRS LSIS

We present in this work some range and azimuth estimators from spermwhale's clicks based on machine learning technics such as sparse coding and fisher vectors. The main idea is to use data collected from long-range array field to learn offline an acoustic model (animal sound production and sound propagation). These raw estimates of range/azimuth can feed a non-linear filter to track the animal.

1 Basic motivations

Most of efficient cetacean localisation systems are based on the Time Delay Of Arrival (TDOA) technic [NF06, BG09].

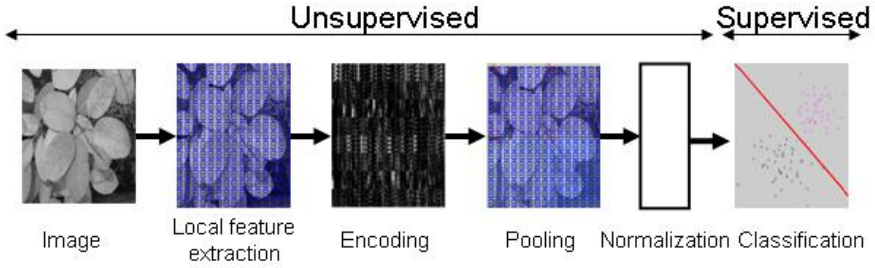


Long base hydrophones'array offers precise localization but, They represent a fixed, centralized and expensive solution.

Question : It is possible to obtain a decentralized and cheapmono-hydrophone localization procedure for collision detection/cetacean watching systems ?
Most surely yes, thanks to dataset collected by TDOA/DTAG systems and machine learning technics.
This work introduces how to obtain two rough estimates of range r and azimuth az by sparse coding method.
Both can feed a non-linear lter to localize precisely the spermwhale.

2 Global feature extraction pipeline

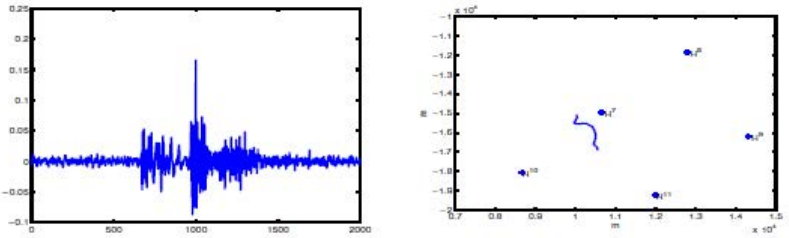
In speech processing, computer vision [BBLP10], the popular pipeline local features extraction-encoder-pooler gives a global representation robust to signal intraclass variations



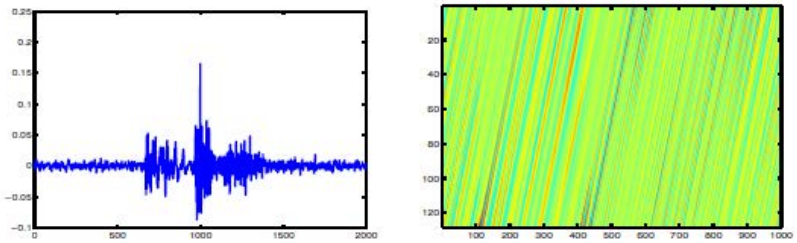
Mimesis of cortex architecture. We will employ the same kind of architecture by changing the classification module by a regression one

3 Local feature

- Let's denote by $\boldsymbol{C} \triangleq \{\boldsymbol{C}^j\}$, $j = 1, \ldots, H$ the collection of detected clicks associated with the j^{th} hydrophone of the array composed by H hydrophones.



- $\boldsymbol{C}^j \triangleq \{\boldsymbol{c}_i^j\}$, $i = 1, \ldots, N^j$ where $\boldsymbol{c}_i^j \in \mathbb{R}^n$ is the i^{th} click of the j^{th} hydrophone.
- Total number of detected clicks : $N = \sum_{i=1}^{H} N^j$
- **local features** : signal patches of $p \leq n$ samples (typically $p = 128$) extracted by sliding windows and denoted by $\boldsymbol{z}_{i,l}^j \in \mathbb{R}^p$. Local features are ℓ_2 normalized



- $\forall \boldsymbol{c}_i^j$, L local patches $\boldsymbol{Z}_i^j \triangleq \{\boldsymbol{z}_{i,l}^j\}$, $l = 1, \ldots, L$ equally spaced of $\lceil \frac{n}{L} \rceil$ samples are collected
- Local patches associated with the j^{th} hydrophone : $\boldsymbol{Z}^j \triangleq \{\boldsymbol{Z}_i^j\}$, $i = 1, \ldots, N^j$. $\boldsymbol{Z} \triangleq \{\boldsymbol{Z}^j\}$ is denoting all the local patches matrix

4 Local features encoded by sparse coding

Sparse coding overview

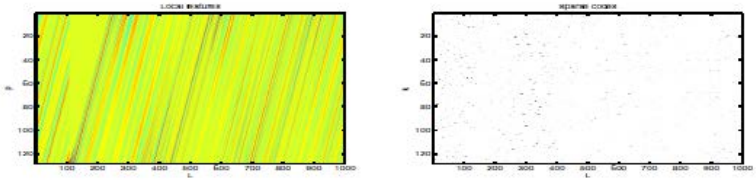
- Each local patch $\boldsymbol{z} \subset \boldsymbol{Z}$ are linearly encoded *via* $\boldsymbol{\alpha} \in \mathbb{R}^k$: $\boldsymbol{z} \approx \boldsymbol{D}\boldsymbol{\alpha}$ where $\boldsymbol{D} \triangleq [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k] \in \mathbb{R}^{p \times k}$ is a pre-trained dictionary such as $\boldsymbol{d}_j^T \boldsymbol{d}_j = 1$
- $\boldsymbol{\alpha}$'s are retrieved by solving the following LASSO problem [Tib94]

$$l_{SC}(\boldsymbol{\alpha}|\boldsymbol{z};\boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

- $\Longrightarrow$ Can be solved efficiently with LARS solver [MBPS]
- In practice, we choose $k \geq p$ to expand information in a higher dimension space
- Thanks to sparsity constraints, we can hope that data of interest will lie on a low-rank manifold much more easy to discriminate than in the original feature space

Pooling parse codes

- Aggregate sparse codes to have a global representation more robust to local deformations
- Let define $\boldsymbol{v}^j \in \mathbb{R}^L$, $j = 1, \ldots, k$ as the j^{th} row vector of $\boldsymbol{V} \triangleq \{\boldsymbol{\alpha}_i\}$, $i = 1, \ldots, L$



- ℓ_μ-norm pooling :

$$f_n(\boldsymbol{v};\mu) = \left(\sum_{m=1}^{L} |v_m|^\mu\right)^{\frac{1}{\mu}} \quad s.t.\ \mu \neq 0$$

- $\mu \to 1 \Leftrightarrow$ sum-pooling while $\mu \to \infty \Leftrightarrow$ max-pooling

Dictionary learning

- Dictionary is trained on a subset of $M \leq N$ local features

$$\mathcal{R}_M(\boldsymbol{V}, \boldsymbol{D}) \triangleq \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\|\boldsymbol{z}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$$
$$s.t.\ \boldsymbol{d}_j^T\boldsymbol{d}_j = 1$$

- Not jointly convex problem $\rightarrow$ alternating method :

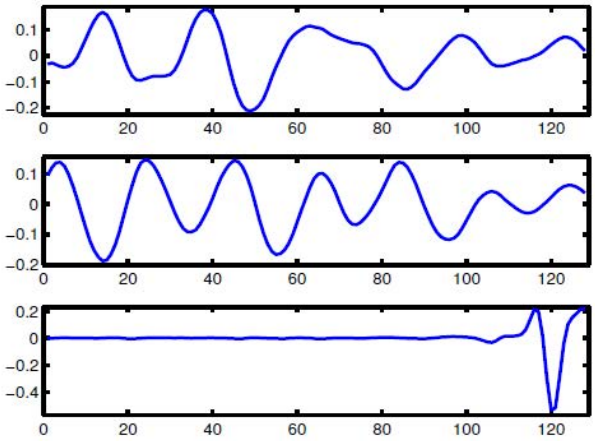
$$\mathcal{R}_M(\boldsymbol{V}|\hat{\boldsymbol{D}}) \triangleq \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\|\boldsymbol{z}_i - \hat{\boldsymbol{D}}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1,$$

solved in parallel by LASSO (LARS, feature sign, etc...) and

$$\mathcal{R}_M(\boldsymbol{D}|\hat{\boldsymbol{V}}) \triangleq \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\|\boldsymbol{z}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}}_i\|_2^2 \ \ s.t.\ \boldsymbol{d}_j^T\boldsymbol{d}_j = 1,$$

solved online (block coordinate descent [MBPS], etc...)

- Example of trained "atoms"



5 Local features encoded by Fisher vectors

Fisher vectors overview

- Let $u_{\boldsymbol{\theta}}$ a pdf which models $\boldsymbol{Z}$, *eg.* a GMM
- Fisher vectors are defined by the normalized gradient vector

$$\mathcal{G}_{\boldsymbol{\theta}}^{\boldsymbol{Z}} = Q_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(u_{\boldsymbol{\theta}}(\boldsymbol{Z})) = \sum_{l=1}^{L} Q_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\log(u_{\boldsymbol{\theta}}(\boldsymbol{z}_l))$$

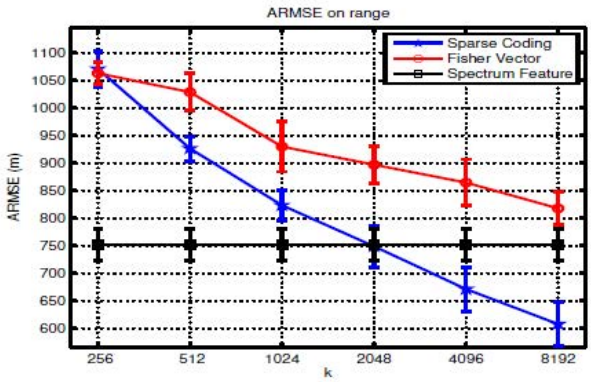
- If $u_{\boldsymbol{\theta}} = \sum_{b=1}^{B} w_b p(z; \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2)$ a GMM, with
  $\boldsymbol{\theta} = \{w_b, \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b^2\}, b = 1, \ldots, B$
- Fisher vectors [PSM10] are defined by average-pooling of local gradients :

$$\begin{cases} \mathcal{G}_{\boldsymbol{\mu}_b}^{\boldsymbol{Z}} &= \frac{1}{L}\sum_{b=1}^{B}\frac{\gamma_l(b)}{\sqrt{w_b}}\left(\frac{\boldsymbol{z}_l - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b}\right) \\ \mathcal{G}_{\boldsymbol{\sigma}_k}^{\boldsymbol{Z}} &= \frac{1}{L}\sum_{b=1}^{B}\frac{\gamma_l(b)}{\sqrt{w_b}}\frac{1}{\sqrt{2}}\left[\frac{(\boldsymbol{z}_l - \boldsymbol{\mu}_b)^2}{\boldsymbol{\sigma}_b^2} - 1\right] \end{cases}$$

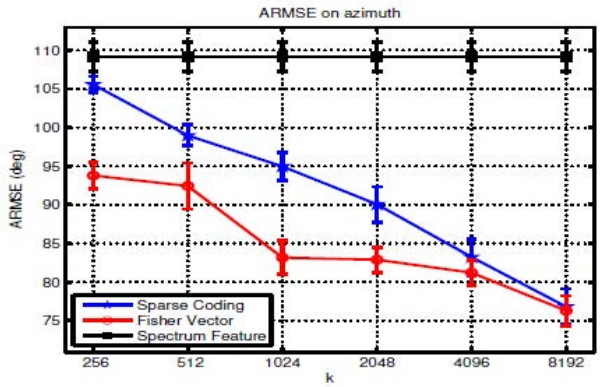
- Total size of FV is $2p.B$

6 Experimental results

- Bahamas2 dataset with $N = 6134$ clicks
- Local features extraction setup : $n = 2000$, $p = 128$, $L = 1000$
- Dictionary learning setup : $M = 400000$ sampled patches, $\lambda = 0.2$, $k \in \{128, 256, 512, 1024, 2048, 4096\}$
- For FV, whitening preprocessing *via* PCA of local features
- GMM learning setup : $M = 400000$ sampled patches, $b \in \{1, 2, 4, 8, 16, 32\}$
- Logistic regression for both r and az estimates, C optimized by CV
- $10-$CV with $70\%/30\%$ in train/test splits
- Average RMSE scores
- Range estimator performances vs k



- For Fisher vectors, the corresponding number of Gaussian is $b \in \{1, 2, 4, 8, 16, 32\}$.
- Azimuth estimator performances vs k



7 Conclusions & Perspectives

- Two rough estimators for range and azimuth by sparse coding and fisher vector framework have been presented
- No specific pre-processing required. Working directly on the click signal
- Promising results in mono-hydrophone configuration
- More efficient local features can be investigated (MFCC, Scattering,...)
- Deep learning (more than 1 layer) as general feature extractor
- Non-linear filtering for a precise localization (EKF, UKF, PF, ect,...)

References

Y-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, *Learning mid-level features for recognition*, CVPR' 10, 2010.

Frédéric Bénard and Hervé Glotin, *Whales localization using a large array : performance relative to cramer-rao bounds and confidence regions*, e-Business and Telecommunications, Springer - Verlag, Berlin Heidelberg, september 2009, pp. 294–306.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, *Online dictionary learning for sparse coding*, ICML '09.

E.-M. Nosal and L. Frazer, *Track of a sperm whale from delays between direct and surface-reflected clicks*, Applied Acoustics **67** (2006), 1187–1201.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink, *Improving the fisher kernel for large-scale image classification*, European Conference on Computer Vision, 2010, pp. 143–156.

Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1994), 267–288.

5.3 Range-depth tracking of multiple sperm whales over large distances using a two-element vertical array and rhythmic properties of click-trains

Delphine Mathias
Gipsa-Lab, Image-Signal Department
Grenoble-INP, France
delphine.mathias@gmail.com

Aaron Thode
Marine Physical Laboratory
Scripps Institution of Oceanography
University of California San Diego, USA

Jan Straley
University of Alaska Southeast,
Sitka, Alaska, USA

Russ Andrews
School of Fisheries and Ocean Sciences
University of Alaska Fairbanks, USA

Olivier Le Bot
Gipsa-Lab, Image-Signal Department
Grenoble-INP, France

Cédric Gervaise
Gipsa-Lab, Image-Signal Department
Grenoble-INP, France

Jérôme Mars
Gipsa-Lab, Image-Signal Department
Grenoble-INP, France

Abstract

Sperm whales (*Physeter macrocephalus*) have followed fishing vessels off the Alaskan coast for decades, in order to remove sablefish (”depredate”) from longlines. The Southeast Alaska Sperm Whale Avoidance Project (SEASWAP) has found that whales respond to distinctive acoustic cues made by hauling fishing vessels, as well as to marker buoys on the surface. Between 15-17 August 2010 a simple two-element vertical array was deployed off the continental slope of Southeast Alaska in 1200 m water depth. The array was attached to a longline fishing buoyline at 300 m depth, close to the sound-speed minimum of the deep-water profile. The buoyline also served as a depredation decoy, attracting seven sperm whales to the area. One animal was tagged with both a LIMPET dive depth-transmitting satellite and bioacoustic B-probe tag. Both tag datasets were used as an independent check of a passive acoustic scheme for tracking the whale in depth and range, which exploited the elevation angles and relative arrival times of multiple ray paths recorded on the array. The localization approach doesnt require knowledge of the local bottom bathymetry. Numerical propagation models yielded accurate locations up to at least 35 km range at Beaufort sea state 3. Ongoing work includes combining the arrival angle information with an algorithm developed by Le Bot *et al.* [1] that uses the rhythmic properties of odontocete click trains to separate interleaved click trains. This approach will improve our localization capabilities in presence of multiple sperm whales. [1] In order to achieve better separation of interleaved click trains it is possible to use machine learning

based algorithms. This new concept is based on finding useful information hidden in a large database. This useful information can then be represented by a sparse subspace. The first step of the approach is to extract informative features with a new detector proposed by Dadouchi *et al.* [2]. Once the dictionary of features is learned, any signal of this considered dataset can be approximated sparsely. By reducing the dimensional space, the sparse representation has the advantage to provide an optimally representation of the data. [Work supported by the North Pacific Research Board, the Alaska SeaLife Center, ONR, NOAA and ANR-12-ASTR-0021-03 "MER CALME"]

1 Introduction

In recent years, passive acoustic methods have become increasingly widespread for monitoring the general assessment of marine environments [3]-[5] and expanding knowledge about marine mammals vocalization repertoire, distribution and habitat characterization. In the past decade, considerable efforts have been made for this purpose using a combination of ocean science, signal processing, statistics and computational (algorithms) science [6].

This paper is concerned with the localization and tracking of sperm whales using a vertical array comprising only two hydrophones. Indeed, passive acoustic monitoring has become an important tool to study sperm whale behavior in the Gulf Of Alaska and their interaction with longlining fisheries [7]-[8]. Each click event generated by a sperm whale can arrive on a hydrophone via multiple ray paths. In this paper, the ray path that arrives first on a hydrophone will be called the primary path, and other ray path arrivals that arise from the same click event are called secondary paths, or multipath.

Most methods developed for localizing marine mammals use wide-baseline hydrophone arrays and the time-difference-of-arrival (TDOA) of a sound on pairs of hydrophones [9]-[11]. Methods are often based on ray-trace acoustic propagation modeling and exploit multipath arrival information from recorded sperm whale clicks. The algorithm compares the arrival pattern from a sperm whale click to range and depth dependent modeled arrival patterns in order to estimate whale location. The technique can account for waveguide propagation physics like interaction with range-dependent bathymetry and ray refraction. Tieman *et al.* [12] managed to track a sperm whale in three dimensions using only one acoustic sensor and a model of the azimuthally dependent bathymetry.

When multiple whales are simultaneously clicking, the biggest challenge is to arrange clicks into separate click-trains corresponding to individual whales, and then classify clicks as primary paths and multipaths. In the past decade several authors proposed algorithms for separating multipaths from the primary click-train, either on single hydrophones [13] two-hydrophone arrays [14] or wide-baseline acoustic arrays [15]. These algorithms exploit the slowly varying multipath structure of individual whales or the slowly varying features of clicks within a train (such as waveform, power). Recently, a few papers discussed how sparse coding can be an effective technique for solving the multiple-marine mammal tracking problem [16]-[18]. Sparse coding seems to be a promising alternative to usual time-frequency feature analysis.

To our knowledge the long-range tracking of multiple whales on a single deployment has not been performed yet. For many applications, the deployment of several hydrophones is impractical and too expensive. Here we discuss how a two-element vertical was used to track the range of multiple whales up to a 35 km range over a 3-day period and how the method could be automated using rhythmic properties of click-trains and sparse coding.

2 Semi-Automated tracking of multiple whales

A two-element vertical array deployed at the sound speed minimum was used to track sperm whales in the Gulf of Alaska between 15 and 17 August 2010. The vertical arrival angles and relative arrival times of multiple refracted and surface-reflected ray paths contain enough information for range-depth tracking without knowledge of the bottom bathymetry. A ray-tracing program was used to model the acoustic travel times from each candidate source location, using a measured sound speed profile. By comparing modeled and measured time lags and vertical angles, an ambiguity

surface was created, displaying the best-fit whale position. A tagged sperm was tracked up to 35 km range under Beaufort 3 conditions, using satellite tag data to independently verify tracking estimates. The technique also permitted to measure the drift of multiple whales away from the vertical array. The method and results are described in detail in Mathias *et al.* [8].

However we were not able to automate the tracking process in the presence of as much as six whales simultaneously vocalizing. Techniques described above to separate click-trains such as the cross-correlation or a rhythmic analysis failed in our case, because of the high number of multipaths received at the hydrophones produced by whales at various ranges.

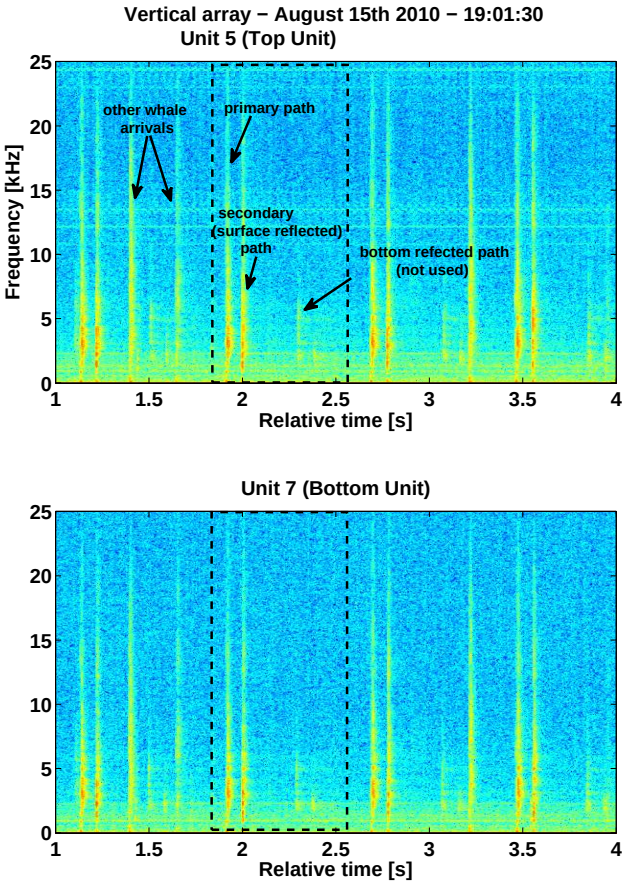


Figure 1: Spectrogram of multipaths produced by two sperm whales on 15 August 2010 at 19:01:30 and recorded on vertical array top and bottom hydrophones

3 Towards better localization of sperm whales using rhythmic properties of click-trains and sparse coding

Localizing a sperm whale using a two-element vertical array requires measuring the relative arrival times of at least two ray paths (the primary path and a multipath) on both hydrophones [8]. Therefore, we need to de-interleave click-trains and associate a primary path click-train with a multipath click-train. Two approaches seem promising for performing this task when three whales or more are vocalizing simultaneously. The first approach takes advantage of the two hydrophone array configuration and use the arrival angle information as an additional source of information for grouping sperm whale clicks into trains associated with a given whale and propagation path. Therefore

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

on-going work includes combining the arrival angle information with an algorithm developed by Le Bot *et al.* [1] that uses the rhythmic properties of odontocete click-trains to separate interleaved click-trains. The algorithm only uses the time of arrival of each click and a complex- autocorrelation function to compute a histogram that exhibits peaks at inter-click intervals (ICI) corresponding to the interleaved click trains, while suppressing harmonics due to ICI multiples. This complex autocorrelation is calculated in a window sliding along the click train leading to a time-ICI representation, which is thresholded to detect the different interleaved click trains. This sequential search could use some complementary features such as the click arrival angle, level or its frequency content.

The second approach is based on sparse coding [19]-[20] and recent publications on the application of this technique on marine mammal sounds [16]-[18]. We propose that a sparse transform of the clicks in the time-frequency domain can help determine the stable components between multipaths belonging to an individual and a given propagation path. In order to reduce the signal dimension for more efficient computation, sets of Mel Frequency Cepstral Coefficients (MFCC) can be computed and a dictionary of features can be generated. Any click detected on the hydrophone can therefore be represented in this space of reduced dimension. The similary between each projected click can be computed using the cosine similarity measure for example. Glotin et al. 2013 showed that this technique worked for tracking the sounds produced by the same minke whale during 30 minutes. It is also possible to work directly on the spectrogram to select areas of interest. A specific algorithm by Dadouchi *et al.* [2] has been developed to detect click and whistles. Based on a two-stage methodology, this algorithm estimates the instantaneous frequency law of non-linear frequency modulations under several constraints (high resolution estimation, ability to cope with multiple overlapping and/or close signals in the time-frequency plane). The first step of the methodology is applied on the square modulus of any linear time-frequency representation, and aims at detecting the time-frequency support of the signals of interest under probabilistic models. A Chi-squared model is used to do the detection of time-frequency bins hosting signal, a Poisson model for the gathering of detected bins into regions of interest (RoIs). Once the RoIs are detected, a high resolution estimator using local polynomial frequency law estimation and phase continuity criteria is used to link local approximation to get a whole estimate of the instantaneous frequency law.

Acknowledgments

This work was supported by the North Pacific Research Board, the Alaska SeaLife Center, ONR, NOAA and ANR-12-ASTR-0021-03 "MER CALME". We thank the Scaled Acoustic Biodiversity SABIOD MASTODONS CNRS project for providing travel funds.]

References

[1] O. Le Bot, J. Bonnel, J. Mars and C. Gervaise (2013). Odontocete click train deinterleaving using a single hydrophone and rhythm analysis, Proceedings of Meetings on Acoustics (19), 010019.

[2] F. Dadouchi, C. Gervaise, C. Iona, J. Huillery and J.I. Mars (2013). Automated segmentation of linear time-frequency representation of marine mammal sounds", The Journal of the Acoustical Society of America, 134(3), 2546-2555.

[3] O. M. Lammers, R.E. Brainard, W.W.L. Au, T.A. Mooney and K. Wong (2008). An Ecological Acoustic Recorder (EAR) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats, The Journal of the Acoustical Society of America 123, 1720-1728.

[4] M. N Anagnostou, J.A. Nystuen, E.N. Anagnostou, A. Papadopoulos, and V. Lykousis (2011). Passive aquatic listener (PAL): An adoptive underwater acoustic recording system for the marine environment, In Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment.

[5] M. André, M. Van Der Schaar, S. Zaugg, L. Houégnigan, A.M. Sénchez, and J.V Castell (2011). Listening to the Deep: Live monitoring of ocean noise and cetacean acoustic signals", Marine pollution bulletin, 63(1), 18-26.

[6] DCL book (2013). Detection, Classification and Localization of Marine Mammals using passive acoustics, 2003- 2013: 10 years of international research, edited by O. Adam and F. Samaran.

[7] D. Mathias, A.M. Thode , J. Straley, J. Calambokidis, G.S. Schorr and K. Folkert, K. (2012). Acoustic and diving behavior of sperm whales (Physeter macrocephalus) during natural and depredation foraging in the Gulf of Alaska, The Journal of the Acoustical Society of America 132(1).

[8] D. Mathias, A.M. Thode , J. Straley, and R.D. Andrews (2013). Acoustic tracking of sperm whales in the Gulf of Alaska using a two-element vertical array and tags, The Journal of the Acoustical Society of America 134 (3).

[9] E.M. Nosal and L.N. Frazer (2006). Track of a sperm whale from delays between direct and surface reflected clicks, Applied Acoustics 67, 11871201.

[10] C.O. Tiemann, M.B. Porter and L.N. Frazer (2004). Localization of marine mammals near Hawaii using an acoustic propagation model, Journal of the Acoustical Society of America 115(6), 2834-2843.

[11] H. Glotin, F. Caudal and P. Giraudet (2008). "Whale cocktail party: real-time multiple tracking and signal analyses. Canadian Acoustics", 36(1), 139-145.

[12] C.O. Tiemann, A. Thode, A., J. Straley, K. Folkert and V. OConnell (2006). Three-dimensional localization of sperm whales using a single hydrophone, Journal of the Acoustical Society of America 120(4), 2355-2365.

[13] P.M. Baggenstoss (2011). Separation of sperm whale click-trains for multipath rejection, The Journal of the Acoustical Society of America 129(6), 35983609.

[14] R. Bahl, T. Ura, T. Fukuchi, Towards identification of sperm whales from their vocalizations, The Journal of Scientific and Industrial Research 54, 409-413 (2002).

[15] E.M. Nosal (2013). "Methods for tracking multiple marine mammals with wide-baseline passive acoustic arrays", The Journal of the Acoustical Society of America, 134, 2383.

[16] Y. Doh, J. Razik, S. Paris, O. Adam and H. Glotin (2013). "Décomposition et analyse par codage parcimonieux des chants de cétacés", Traitement du Signal, in press.

[17] H. Glotin, J. Razik, S. Paris and X. Halkias (2013). "Sparse coding for large scale bioacoustic similarity function improved by multiscale scattering, Proceedings of Meetings on Acoustics Vol.19.

[18] S. Paris, Y. Doh, H. Glotin, X. Halkias and J. Razik (2013). Physeter catodon localization by sparse coding, ICML 2013 conference, 6pp.

[19] Q. Barthélemy, C. Gouy-Pailler, Y. Isaac, A. Souloumiac, A. Larue and J. I. Mars (2013). Multivariate temporal dictionary learning for EEG, Journal of Neurosciences Methods, in press.

[20] Q. Barthélemy, A. Larue, A. Mayoue, D. Mercier and J.I. Mars (2012). Shift and 2D Rotation Invariant Sparse coding for multivariate signals, IEEE Trans. Signal Processing 60(4), 1597-1611.

5.4 Optimization of Levenberg–Marquardt 3D biosonartracking

Ales Mishchenko LSIS-DYNI *ales.mishchenko@univ-tln.fr*
Pascale Giraudet dpt biology *giraudet@univ-tln.fr*
Hervé Glotin LSIS-DYNI *glotin@univ-tln.fr*
Toulon University, 83957 La Garde, France

Abstract

This paper presents a method for tracking of unknown calling and moving animals, such as bats or whales, using an arbitrary system of microphones, organized in a 3D structure. The time differences of sound arrival (TDOAs) result in a set of distances to pairs of microphones. However, in presence of echoes propagation speed non-uniformities, and background noises, the resulting set of distances may be non-consistent. We propose a method for optimization of TDOA data used in 3D trajectory reconstruction. Our method provides faster and smoother animal trajectory reconstruction than the state of the art .

1 Introduction

The location of an acoustic source using time differece of sound arrival (TDOAs) to multiple microphones has many military, bioacoustic and surveillance applications. Tracking of wildlife movements has been widely studied since the 1960s [1], [2]. For the majority of these studies, an operator supervises received signal strength while changing the orientation of a directional receiving antenna. The problem of localizing an acoustic source from Time-Differences-of-Arrival (TDOAs) is received recently a lot of scientific attention with a number of different solution approaches (the correspondent review can be found in [3]). However, most of these methods assume ideal conditions, such as known and constant propagation speed, reliable only under controlled conditions where the air temperature can be monitored. The effects of a wrongly assumed propagation speed is surveyed in [8]. One of the most successful approach to overcome the problem of variable propagation speed (and subsequent non-robustness of tracking), echoes, etc, in the field of bioacoustics is [6].

2 The algorithm overview

Due to the difference in sounds, emitted by different species, as well as difference in antennae geometry and media of sound-propagation, it is possible to use the animal call structure to detect and process the calls of a particular animal. In accordance with this, we perform the optimization of parameters for adaptation of TDOA calculation for a particular animal. First, we adapt the resulting algorithm of TDOA calculation to a variety of bat recordings, including optimisation of time-step, crosscorrelation- window, number of maximums and persentage of TDOA to filter with coherence condition formulas. Due to the difference in sounds, emitted by different species, as well as difference in antenae geometry and media of sound-propagation, it is necessary to adapt the resulting algorithm of TDOA calculation to a variety of bat recordings. For each recording we automatically optimize time-steps, crosscorrelation-windows, number of maximums and persentages of TDOA to filter with coherence condition formulas. To take this into account, for each recording, we automatically optimize time-steps, crosscorrelation-windows, number of maximums and persentages of TDOA to filter with coherence condition formulas. In this, 1st step, our algorithm subdivides these segments into subsegments, contaning, on average, a single distinct animal sound (click in case of whales and bats). This sundivisions facilitates the

detections of the given animal and cross-correlations between microphones in order to find the TDOAs between them.

Second, for each pair of sensors, their arrived signals are cross-correlated (between pairs of different microphoe-recordings), followed by extraction the maximum correlation values for TDOAs calculation, as well as calculation of TDOAs errors, described in the section 3 below. The segments of multimicrophone recording (in our case, recordings from 4 microphones with length of 6seconds are used) are used for this TDOAs calculation. The Figure 1 shows the typical geometries of antennas used for TDOA calculations.

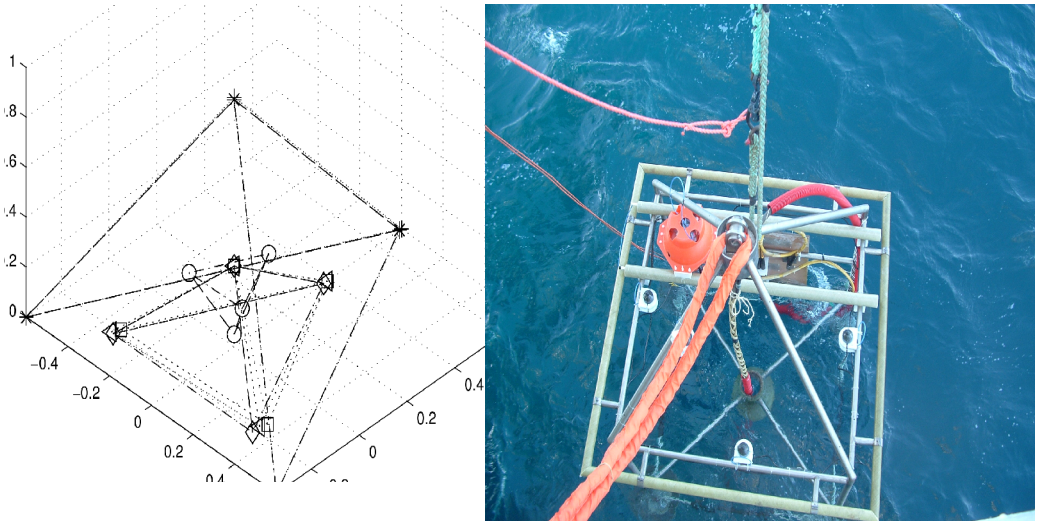


Figure 1: The antennas geometries used for TDOA calculations: Bats recordings(left) and whales recordings(right)

In this, 2nd step of our algorithm refines these TDOAs according to the formulas, described in the section 2. These TDOA-refinement-formulas allow to calculate the amount of the overall TDOA-calculation-error increase/decrease, depending on all possible shifts of detection-vectors in time. The best shifts of detection-vectors, providing the minimum overall TDOA-calculation-error, are used to find the most probable TDOAs, as described in details in the following section 2.

Material. In case of bats tracking we used 4 microphones, organized in a tetrahedron with 1,6m edge, shown in a Figure 1. The typical recording for one of bat-microphones is shown in a Figure 2.

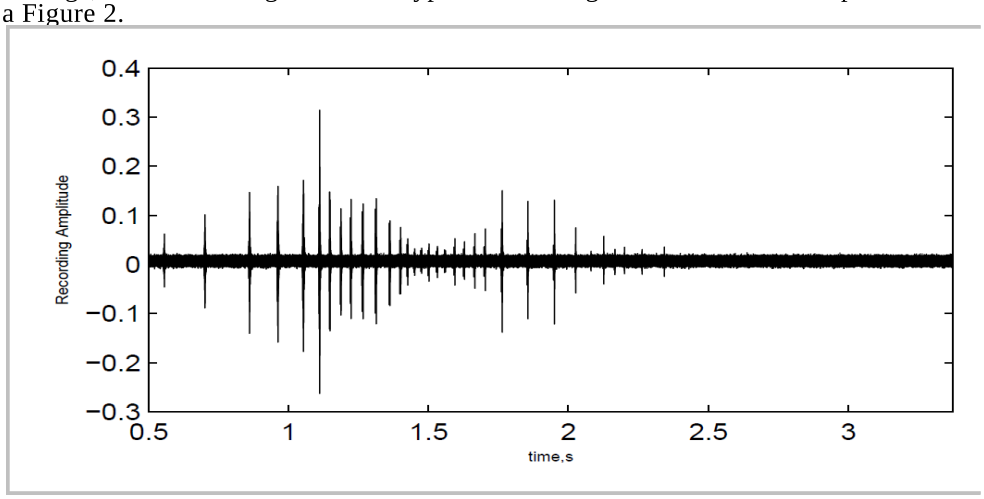


Figure 2: The flying bat recording ”ECOR2116”

For whale tracking the recording system was an hydrophone array fixed to the bottom of the ocean at the Bahamas Data, distributed in 2005 and also to the NIPS4B workshop.

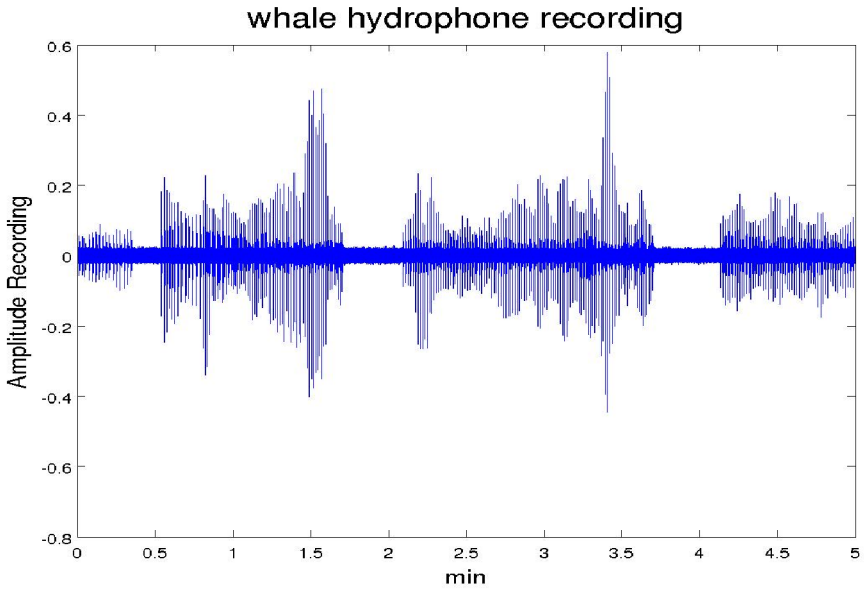


Figure 3: The Physeter macrocephalus whale recording.

3 TDOA refinement

Correlation of sounds, arrived to pair of sensors results in cross-correlation-functions, showing shifts at which signals are more similar (best shifts) or more different (worst shifts). TDOAs between two microphones are calculated as peaks of correlation of recordings of these microphones.

TDOAs by coherence formulas. The refinement of TDOAs by coherence formulas is used to filter the set of TDOA by application of the following coherence equations, developed in [4-6] (4 sensors case) :

$$\left| TDOA_{ij}^{m} + TDOA_{jk}^{m} - TDOA_{ik}^{m} \right| = |\varepsilon(m,i,j,k)| < \varepsilon$$

$$\left| TDOA_{ij}^{m} + TDOA_{jh}^{m} - TDOA_{ih}^{m} \right| = |\varepsilon(m,i,j,h)| < \varepsilon$$

$$\left| TDOA_{ik}^{m} + TDOA_{kh}^{m} - TDOA_{ih}^{m} \right| = |\varepsilon(m,i,k,h)| < \varepsilon$$

$$\left| TDOA_{jk}^{m} + TDOA_{kh}^{m} - TDOA_{jh}^{m} \right| = |\varepsilon(m,j,k,h)| < \varepsilon$$

Here indices i,j,k,h are microphone numbers, whereas index m is the rank of the maximum. ε is the maximum possible overall system error,depending on microphone sensitivity, noise level, sound quality, etc. This way the equations will serve as a filters for TDOAs, selecting the coherent TDOA, producing smooth TDOA and correct tracking. depending on microphone sensitivity, noise level, sound quality, etc. This way the equations will serve as a filters for TDOAs, selecting the coherent TDOA, producing smooth TDOA and correct tracking.

Using many maximuims together with coherence equations allows to significantly reduce the error of TDOA calculation.

The Figure 4 shows comparison between state of the artand our method, in 10 different time-moments from recording ”ECOR2116”, presented in the following section 4.

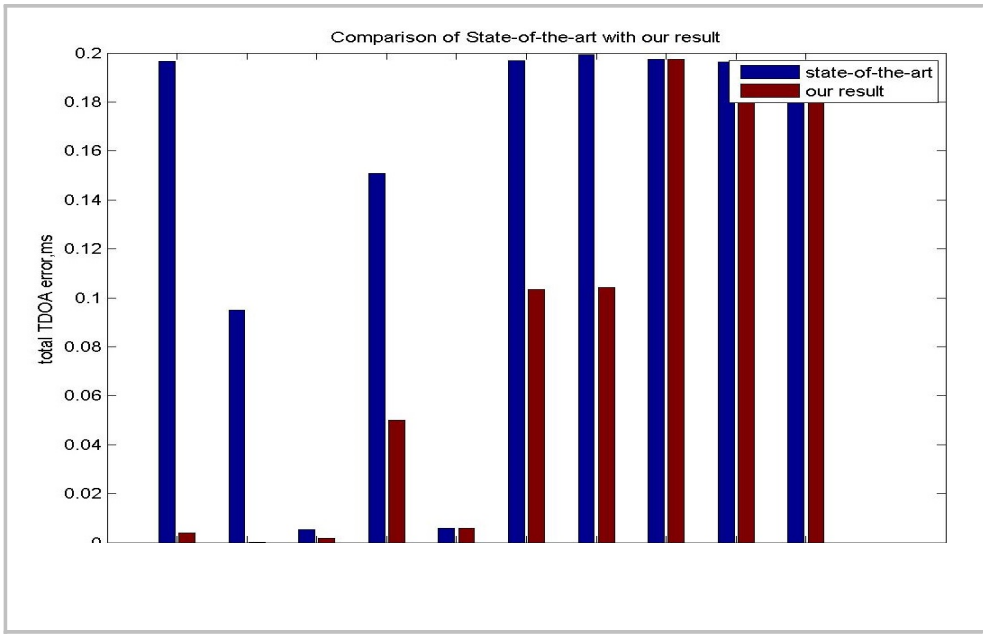


Figure 4: The TDOA calculation errors for flying bat recording ”ECOR2116” in 10 different time-moments.

Figure 5 shows the error distribution as a joint-histogram for the state of the art (using 1 TDOA correlation maximum) and our method, operating with 1-6 TDOA correlation maximums.

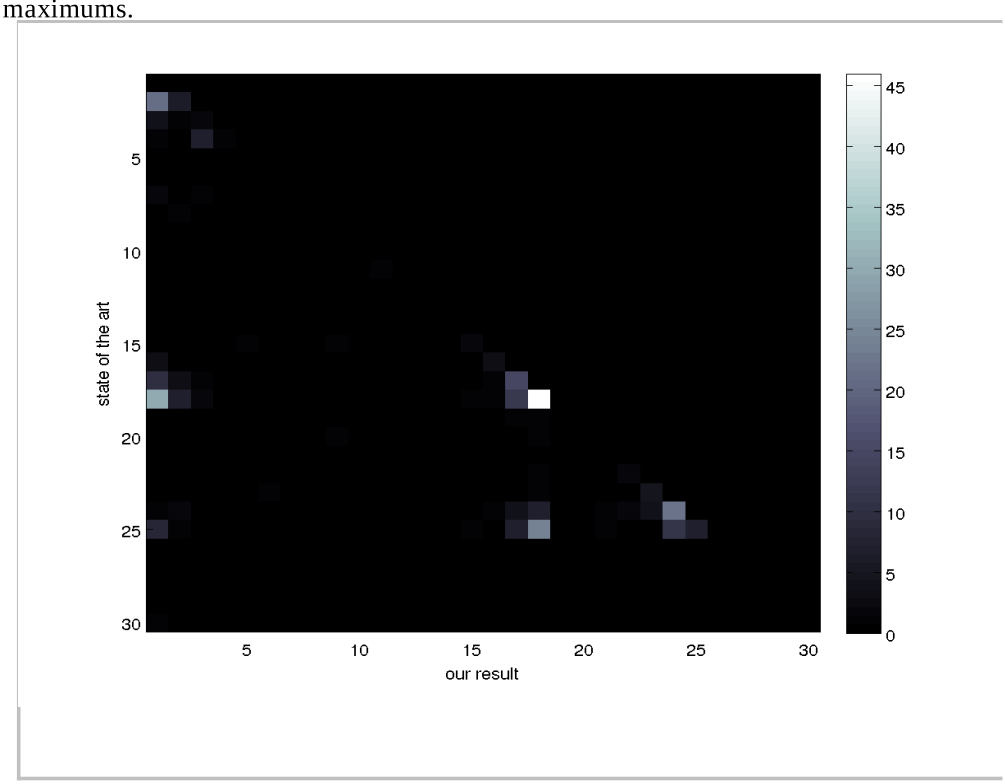


Figure 5: The statistics of TDOA calculation errors for all recordings. The non-symmetry of an error histogram with respect to diagonal (state of the art methods provide more errors) illustrates advantage of our method.

4 Results of TDOA calculation and refinement

The Figures 6 and 7 show comparison between state of the art and our method, using the same 10 different time-moments of recording ”ECOR2116”, presented in the Figure 5.

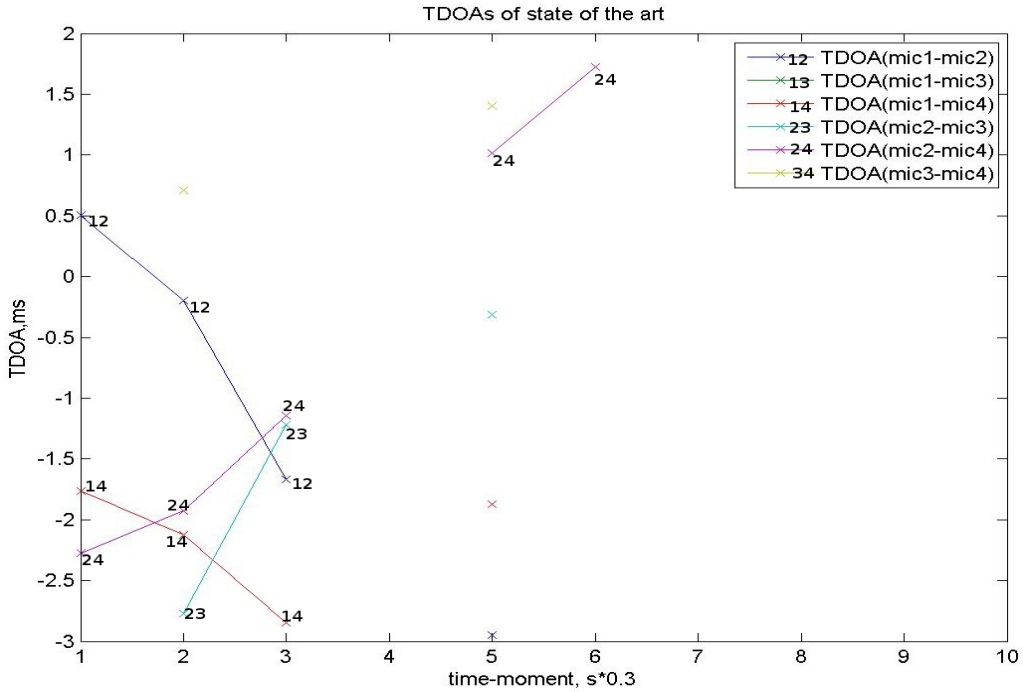


Figure 6: TDOA-tracking trajectory for the flying bat recording ”ECOR2116” , resulting from the state-of-the-art single-maximum TDOA calculation method.

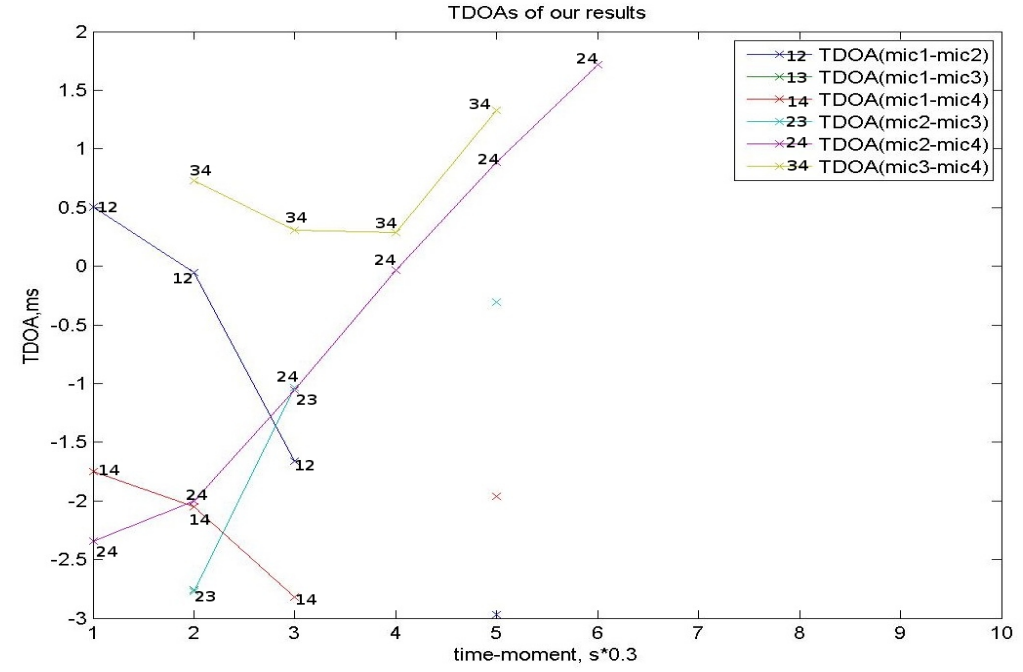


Figure 7: TDOA-tracking trajectory for the flying bat recording ”ECOR2116” , resulting from the our multiple-maximum TDOA calculation method.

As it can be seen, our method is able to fill the gaps of TDOA-trajectory, left by the state-of-the-art TDOA calculation methods. The advantage of usage of our patent (used with just 4 maximums) over baseline (1 maximum) is smoothness and completeness of the TDOA-trajectorie, resulting in smoothness and completenes of the distance-to-microphones difference trajectories, necessary to robust trajectory reconstruction. For the trajectory reconstruction we are using Levenberg solver and the following figure 8 shows the advantages of our patent in reconstruction of trajectory over state-of-the-art methods.

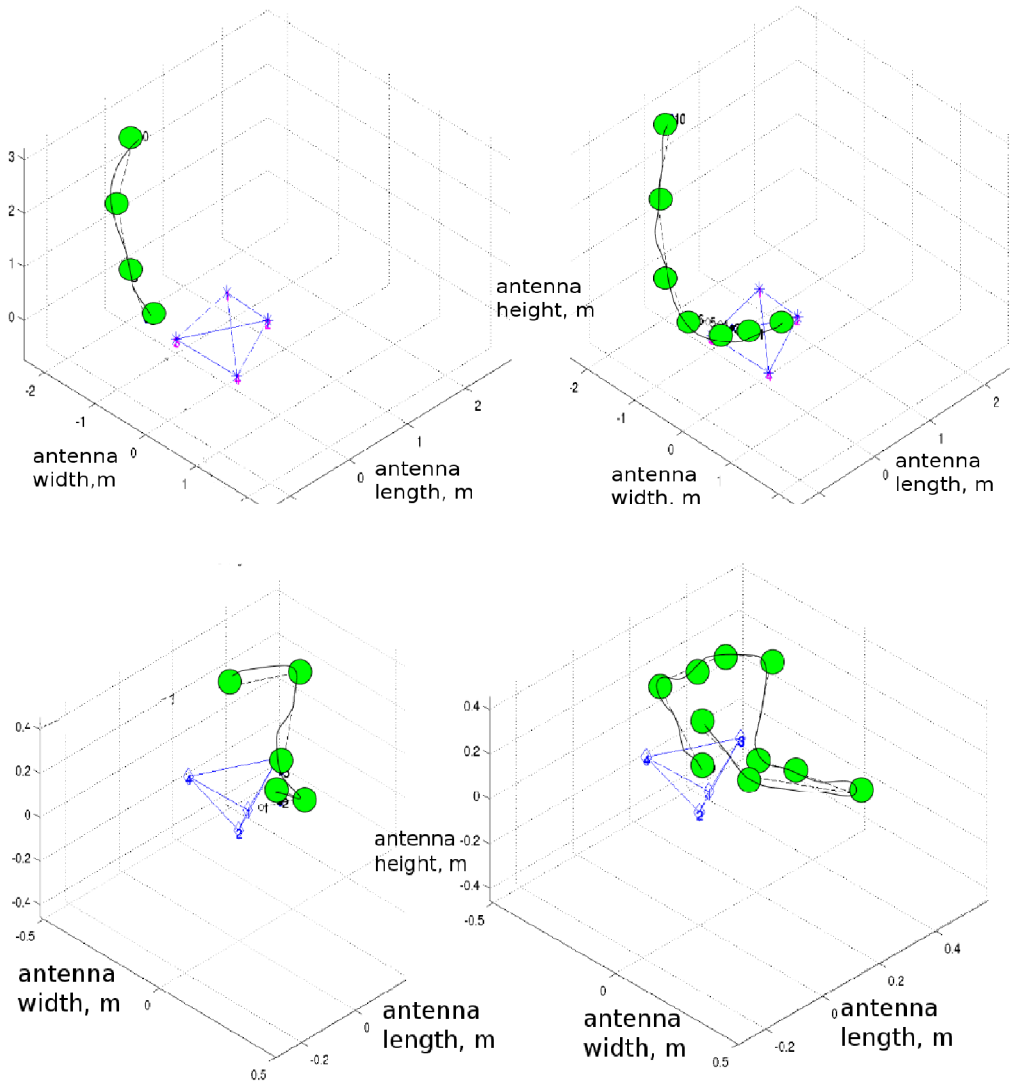


Figure 8: Comparison of reconstructed trajectories of the flying bat, resulting from the state-of-the-art single-maximum TDOA calculation method (left) and from the our multiple-maximum TDOA calculation method (right).

The snapshot of the tracking whales results is shown in a Figure 9.

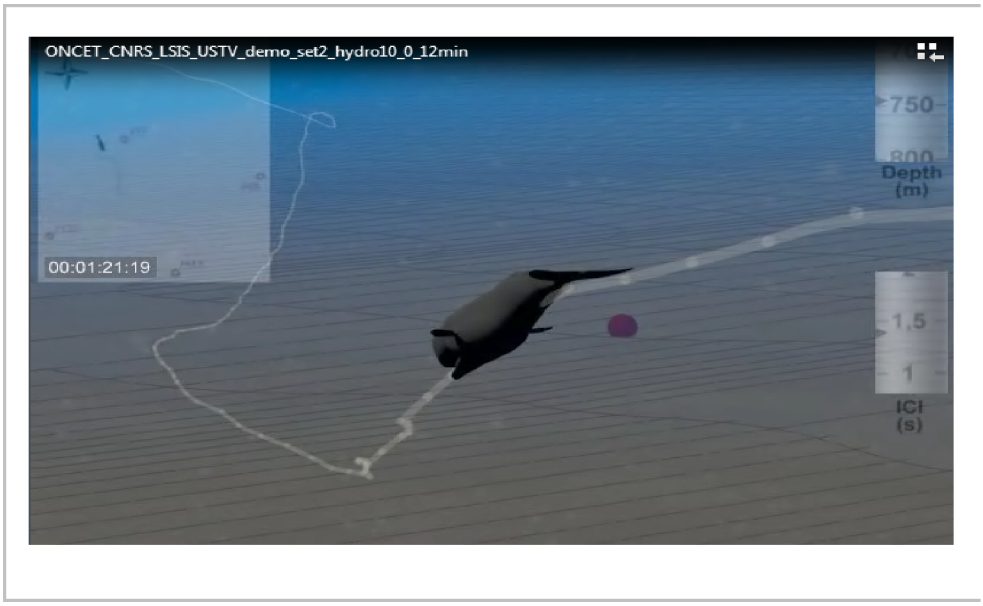


Figure 9. The snapshot of the video of the whale track computed in [5] (available online at http://sabiod.univ-tln.fr/tv).

5 Results of calculation-time, needed for reconstruction of trajectory

As it was shown above, our method provides more exact set of TDOAs, leading to more exact distance-to-microphones difference trajectories. Moreover, refinment of a TDOAs set by the formulas, described in section 3, leads to a smaller set of TDOAs. These advantages of our method lead to the acceleration of Levenberg–Marquardt solver in trajectory reconstruction. The following Figure 11 shows the comparison for the different choices of sound-velocity.

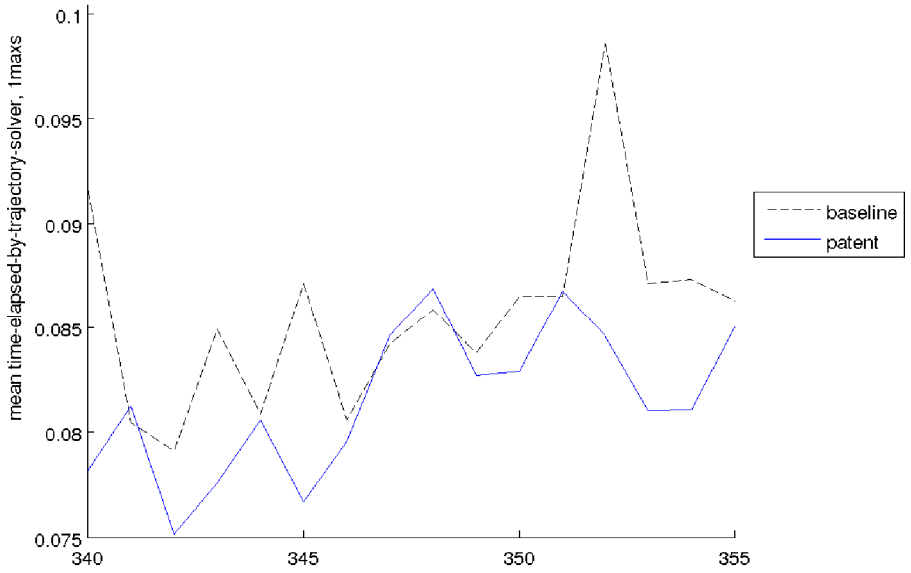


Figure 11: Comparison of a trajectory-reconstruction time, used by Levenberg–Marquardt solver in case of the baseline non-refined TDOA calculation method and our patented TDOA refinment method.

6 Conclusion and future work

We have shown that our patent on multi-maximum TDOA calculation and refinement show advantage over state-of-the-art methods.

It provides higher number and accuracy of reconstructed trajectory points, and shows less calculation-time for the LM solving due to enhanced inputs.

Our future plans are construction of the neural network for detection and classification of animal call recordings in big amount of recorded data. As it was shown in [9-11], this neural network can be constructed as the convolutional neural network (CNN), receiving the microphones recordings for detection and classification. The CNN were sucessfully applied not only for images, but for time-domain, as well as adapted to to process and detect many specific time-series patterns [10,11].

Acknowledgments

This work was done with the support of SABIOD project and SATT sud-est company. We also thank Cyberio company for bat recordings.

References

[1] C.D. LeMunyan, W. White, E. Nybert, Design of a miniature radio transmitter for use in animal studies, J. Wildl. Mgmt., vol. 23(1), pp. 107-110, 1959.

[2] W.W. Cochran, R.D. Lord, A radio tracking system for wild animals, J. Wildl. Mgmt., vol. 27(1), pp. 9-24, 1963.

[3] P. Stoica and J. Li. Lecture notes - source localization from range-difference measurements. Signal Processing Magazine, IEEE, 23(6):6366, November 2006.

[4] Giraudet P., Glotin H., Real-time 3D tracking of whales by echo-robust precise TDOA estimates with a widely-spaced hydrophone array. Int. Jour. Applied Acoustics, Elsevier Ed., Vol. 67, Issues 11-12, pp 1106-1117, Nov. 2006.

[5] Glotin H., Caudal F., Giraudet P., Whales cocktail party: a real-time tracking of multiple whales, in Internat Journal Canadian Acoustics, V 36, p139-145, 2008, ISSN 0711-6659.

[6] Glotin H., Giraudet P., Caudal F., Patent, Real time multiple whale tracking by passive acoustics, 2007. no 07/06162, Europe, extension 2009 USA.

[7] Bénard F., Glotin H., Giraudet P., Whale 3D monitoring using astrophysic NEMO ONDE two meters wide platform with state optimal filtering by Rao-Blackwell Monte Carlo data association, in Journal of Applied Acoustics, Vol. 71 (2010), pp. 994-999

[8] P. Annibale and R. Rabenstein. Accuracy of time- difference-of-arrival based source localization algorithms under temperature variations. In Proc. of 4th Int. Symposium on Communications, Control and Signal Processing, (ISCCSP), Li massol, Cyprus. IEEE, 2010.

[9] O. Abdel-Hamid, L. Deng, and D. Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In INTERSPEECH, 2013.

[10] Y. Bengio and Y. Lecun. Convolutional networks for images, speech, and time-series, 1995.

[11] Cecotti,H.,Graser,A. Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:33 , Issue: 3)

5.5 Data driven approaches for identifying information bearing features in communication calls.

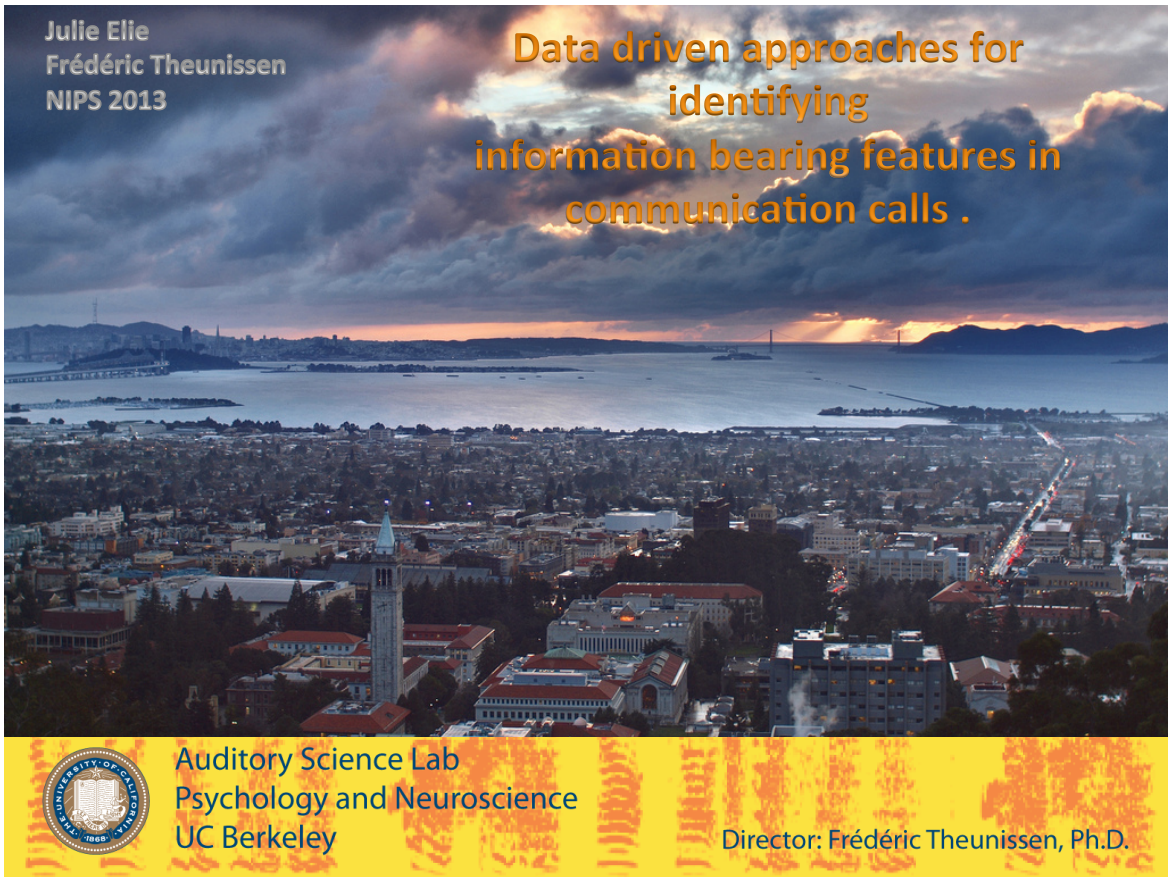
Julie E. Elie and Frédéric E. Theunissen. UC Berkeley. Dept of Psychology and Helen Wills Neuroscience Institute.

Bioacousticians have traditionally investigated the acoustical nature of information bearing features in communication calls by describing sounds using a small number of acoustical parameters that appear particularly salient (e.g. the mean frequency, duration, spectral balance). These measures are then used as parameters for linear discriminant analyses (LDA) or other supervised learning approaches to investigate what acoustic parameters drive sound categorization. This classical approach is computationally efficient and yields results that are easily interpretable. However this approach can also be limited by the *a priori* choice of the putative information bearing features: as long as the sound representation is not complete, one will not be able to determine whether the correct information bearing features are identified and, thus, whether the actual amount of information present in the calls (measured for example as the quality of a discrimination test) is correctly estimated.

To address this issue, we have adopted a data driven approach. In the traditional approach, specific acoustical parameters are chosen for two reasons: for dimensionality reduction and for the implementation of a non-linear transformation that could be required for linear discriminant approaches to effectively discriminate among sound categories. These two steps, however, can be implemented without *a priori* assumptions or loss of information. In our approach, our non-linear transformation is an invertible spectrographic representation of the sound. Then before using a classifier, the dimension of this high-dimensional representation is reduced using principal component analysis (PCA). For this approach to work, the spectrograms of the sounds must be aligned and cross-validation techniques that evaluate both the effect of the number of PCs in the PCA and the number of parameters in the classifier must be implemented to prevent over-fitting. In our approach alignment of spectrograms was based on the cross-correlation between amplitude envelopes. We then used proven cross-validation techniques to prevent over-fitting: bootstrap for LDA and the Random Forest algorithm for non-linear tree classifiers. Finally, we compared the classification performance of the two classifiers on this sparse spectrographic representation of the sound (PCA on spectrogram) with those obtained for two other feature spaces: the Mel frequency cepstral coefficients (MFCC) and the modulation power spectrum (MPS).

These approaches were tested for the analysis of calls in a unique database of 1275 zebra finch communication calls obtained in our laboratory. This database includes many exemplars of all the call types in the zebra finch repertoire for a large number of male and female birds. Our algorithms were used to determine the discriminability of vocal types. The PCA on spectrogram yielded the best feature space: these features are both easy to interpret and yield higher performance of classifiers. The best results were obtained using the Random Forest algorithm and a PCA spectrographic representation using 50 PCs; we show that the 11 distinct call types in the zebra finch repertoire, irrespective of the identity of the vocalizing bird, could be classified with 83.1% of accuracy. This classification

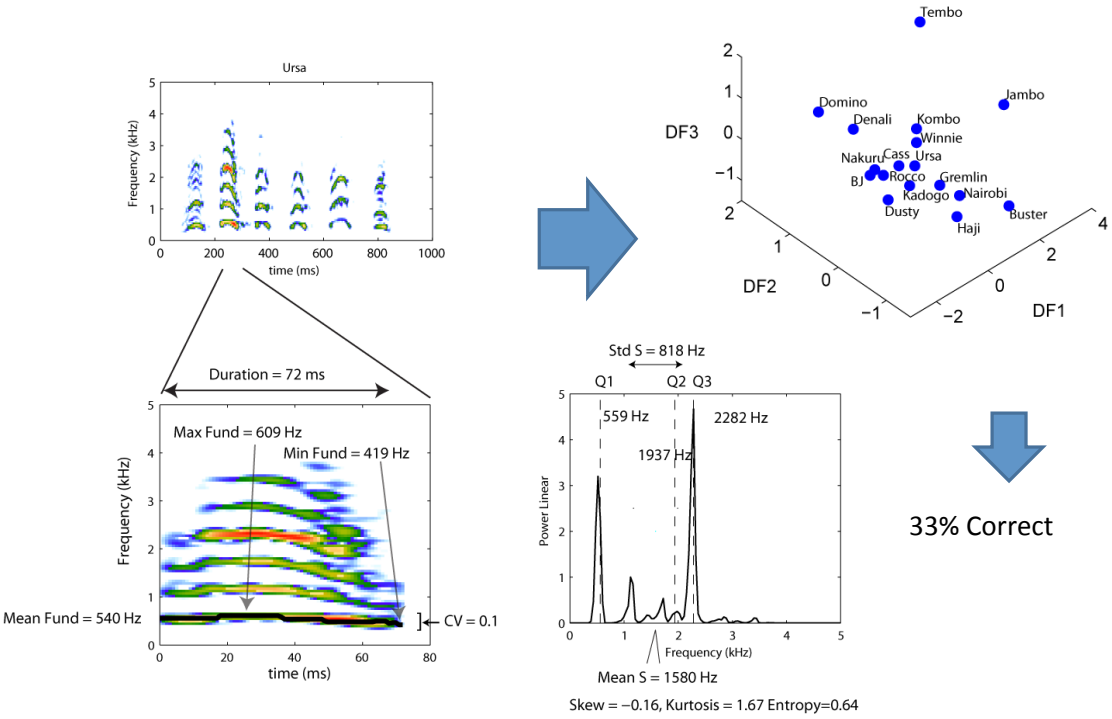
performance is significantly higher that what one could obtain from any of the two other sound representations tested with the Random Forest algorithm (MFCC, 53.1% of accuracy; MPS, 63.5% of accuracy). Besides, the Random Forest yielded better classification performance than LDA, irrespective of the feature space used. In conclusion, the data driven algorithm using the DFA on the spectrogram showed superior results and we propose that it could be used both for investigating behavioral and neural mechanisms of sound discrimination and for the vocalization based identification of species in ecological or environmental studies.



Motivation

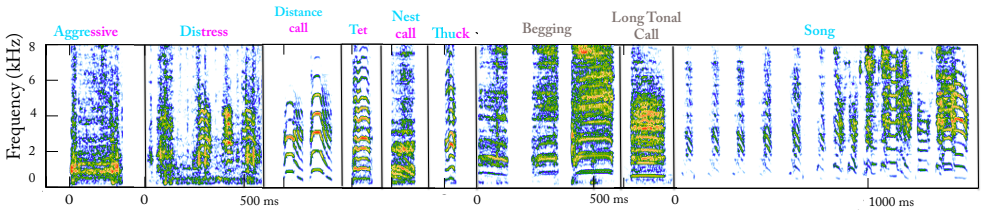
- Classical approaches in bio-acoustics and sound analyses ~ Using simple (ad hoc) features.
  - Hyena Vocalizations
- Perception and Neural Representation.

Individual Signature in the Hyena giggle sounds.



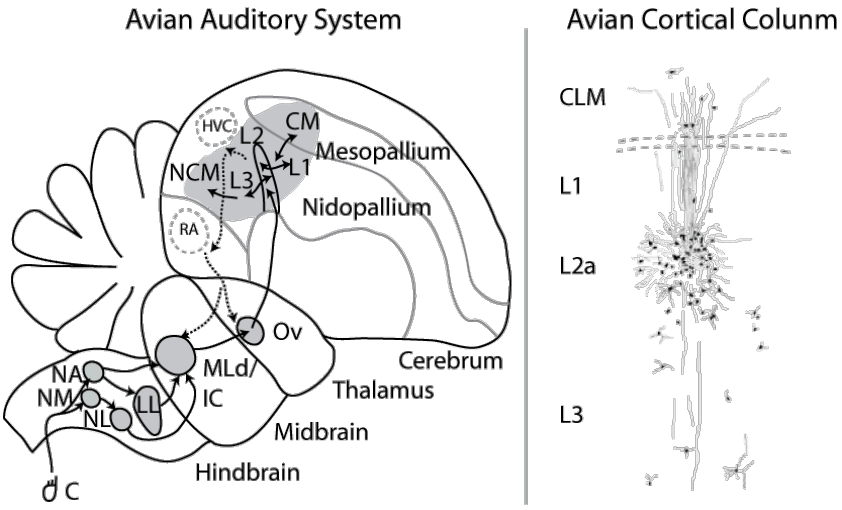
Mathevon et al, BMC Ecology 2010

Zebra Finch Complete Repertoire is Complex.

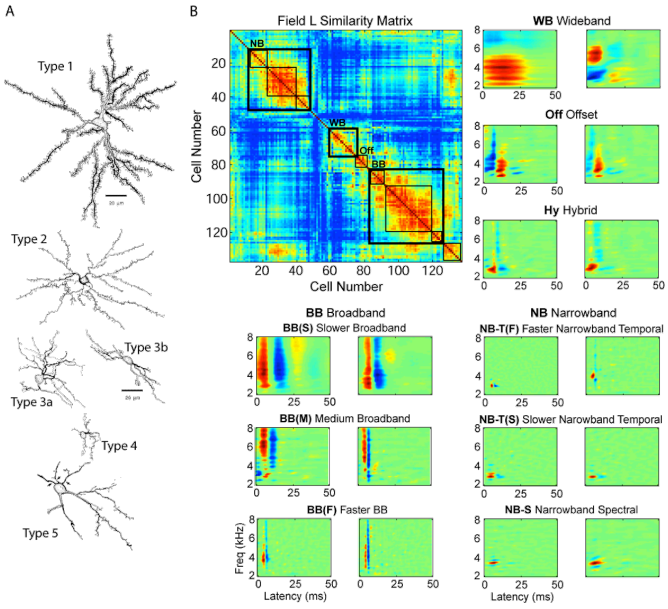


We need an automatic (unsupervised) feature extractor…

Auditory System of Song Birds



Functional Clusters



We would like to relate the features to neural representation

Woolley SM et al, J. Neurosci 2009

A UNIQUE DATA BASE OF THE COMPLETE VOCAL REPERTOIRE OF THE ZEBRA FINCH.

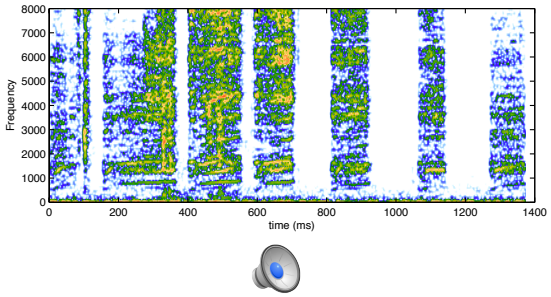
The Zebra Finch Vocal Repertoire: A quick tour.
Categorization of WHAT from the behavioral context

- Needy: Chick Calls
  - Begging
  - Long Tonal
- Affiliative: Social Contact
  - Whine
  - Nest
  - Tet
  - Distance
  - Song
- Non-Afiliative: Alarm – Distress – Aggression
  - Tuck
  - Thuck
  - Distress
  - Aggressive

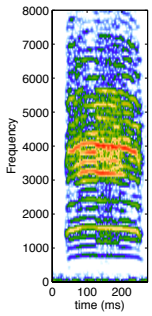
Julie Elie

Chicks (Juveniles): Begging and Long Tonal Call

Begging

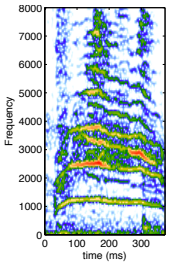


Long Tonal Call

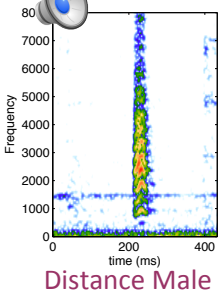


Affiliative Calls
Making and Preserving Social Bonds

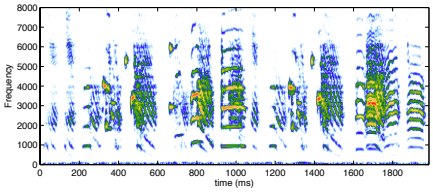
Whine



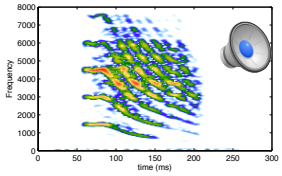
Tet



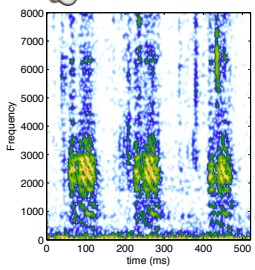
Male Song



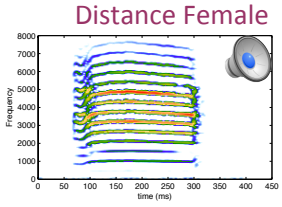
Distance Male



Nest

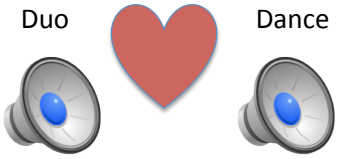


Distance Female



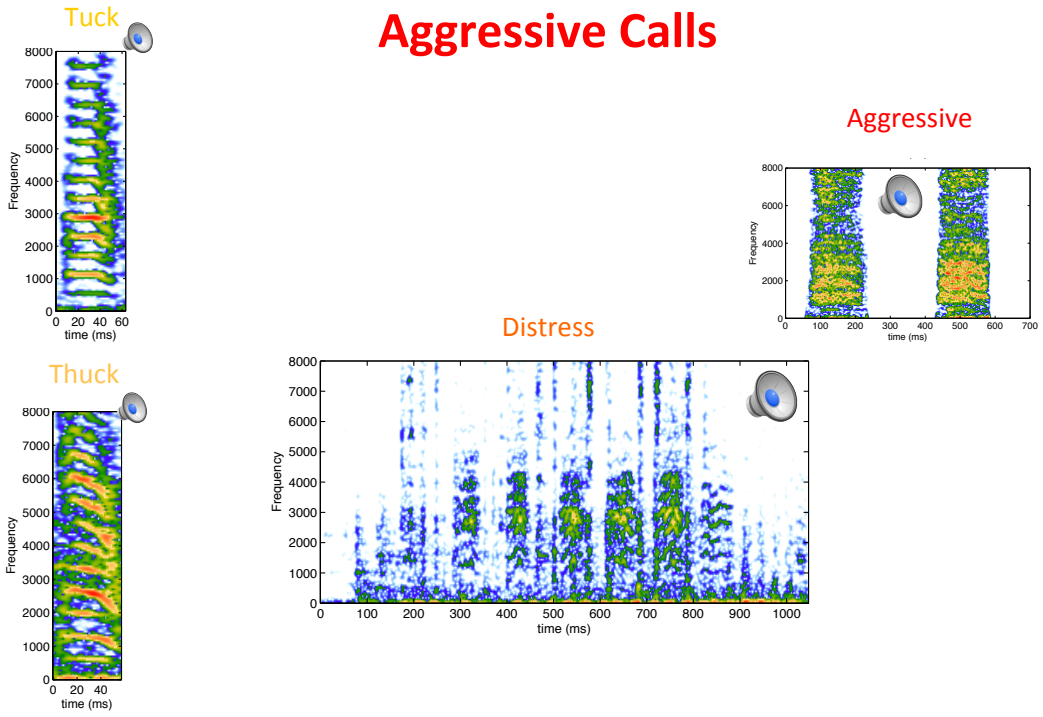
Duo

Dance



Alarm Calls → Distress Calls

Aggressive Calls



CLASSIFYING VOCALIZATION TYPES.

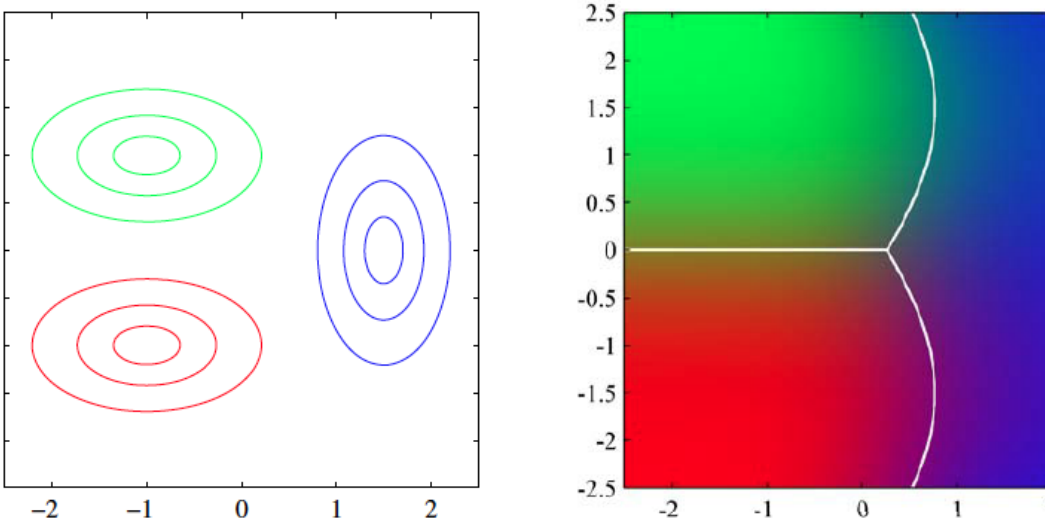
Three Feature Spaces

- Spectrogram + PCA
- Mel Frequency Cepstral Coefficients
- Modulation Power Spectrum

Two Classifiers

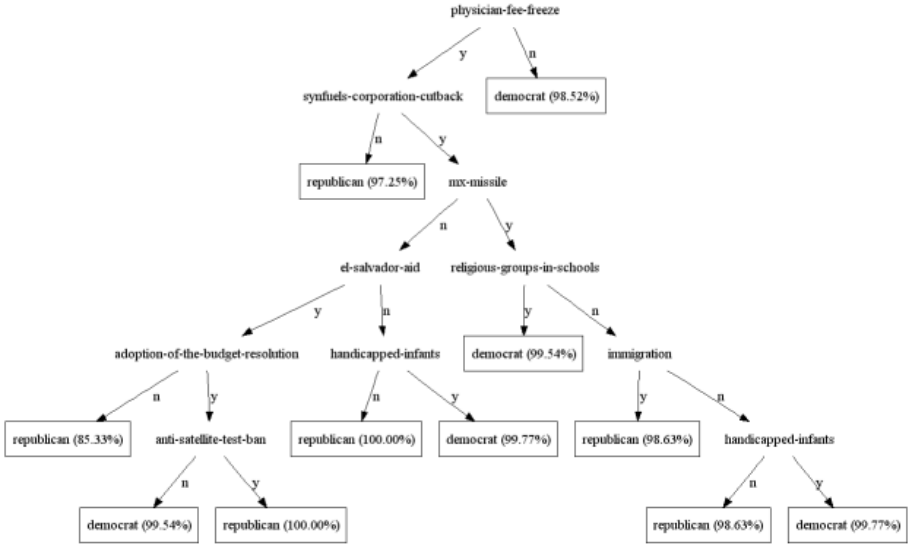
- Fisher Linear Discriminant Analysis
- Random Forest

Fisher Discriminant Analysis

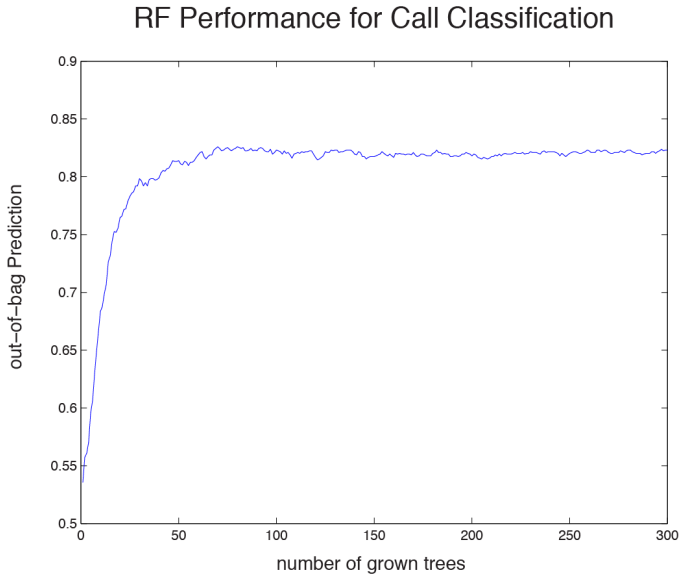


Relax the isotropic assumption and get quadratic decision boundaries

Tree Classifiers

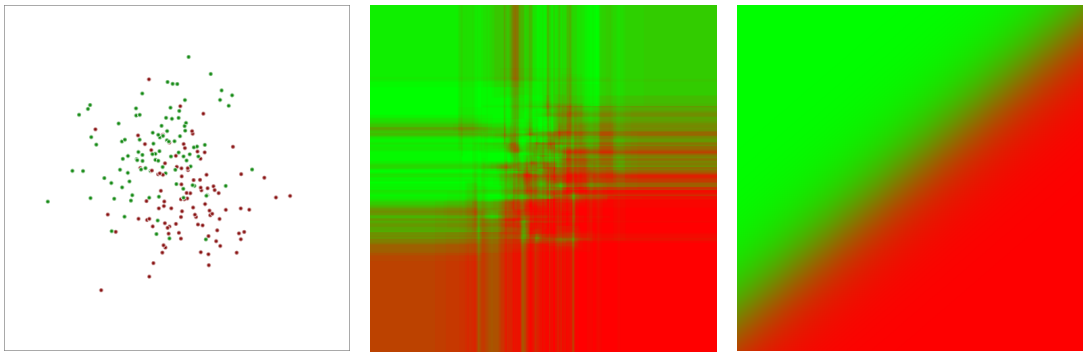


Random Forest = A forest of tree classifiers

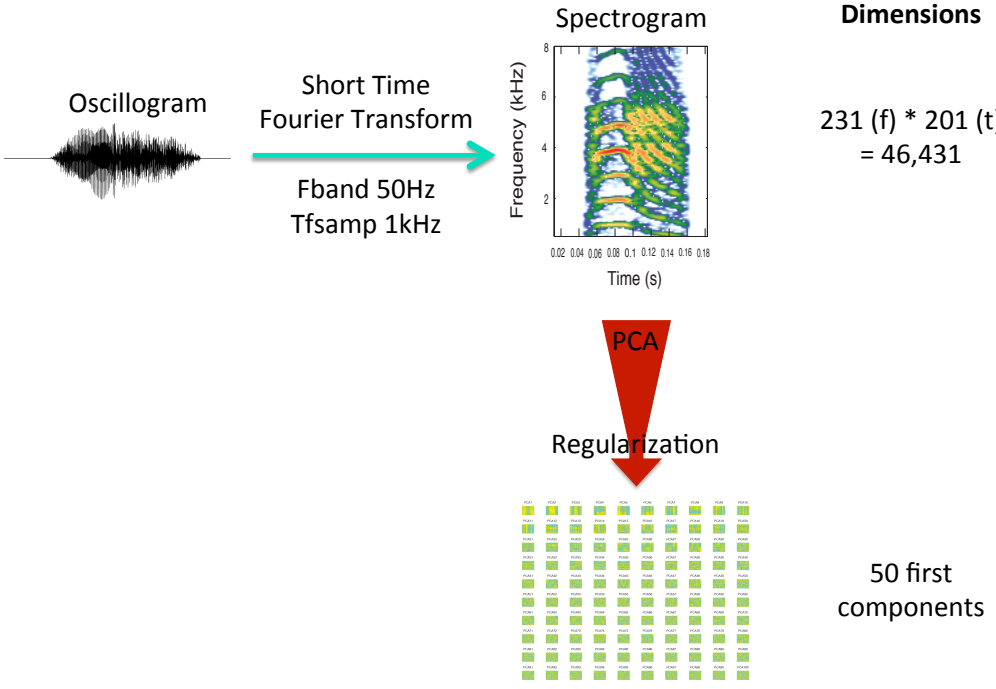


Random Forest

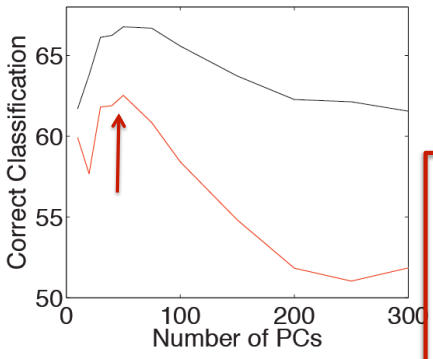
Data | Result from Random Forest | Logistic Regression

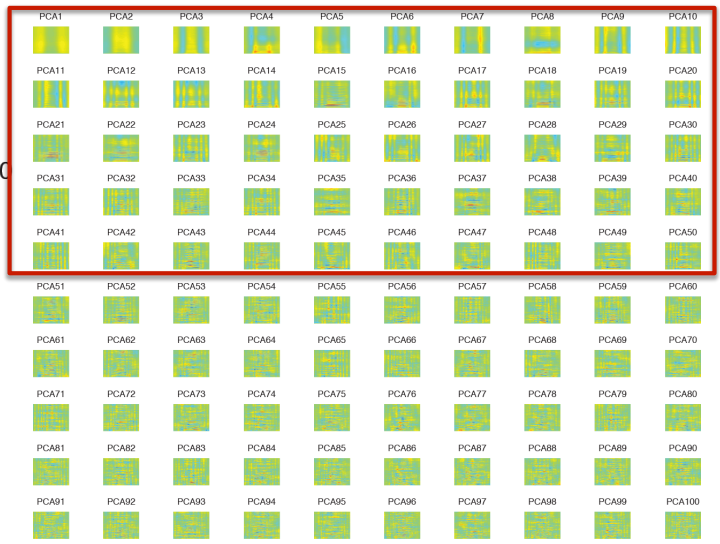


Spectrogram + PCA based features



PCA - Regularization



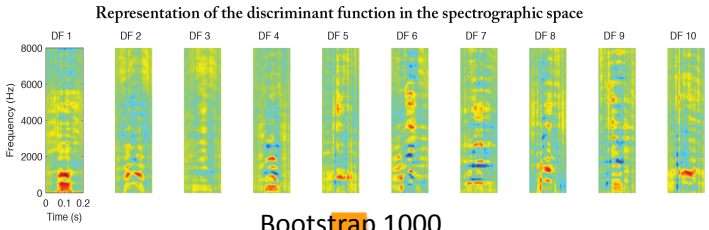


Dimensions:
50 first
components

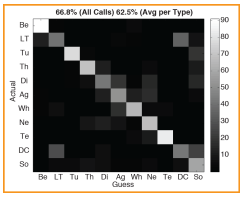
(Stectrogram+PCA) and DFA

Training data set
(6/7 of the data)

DFA



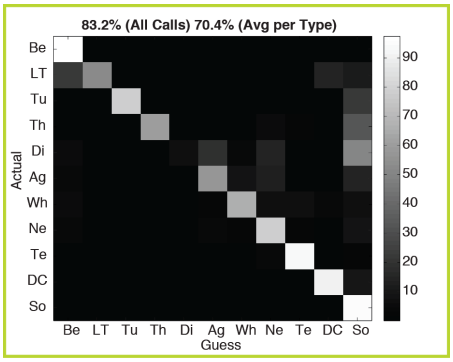
Bootstrap 1000
(cross-validation)



Correct classification:
66.8% (All Calls)
62.5% (Avg per Type)

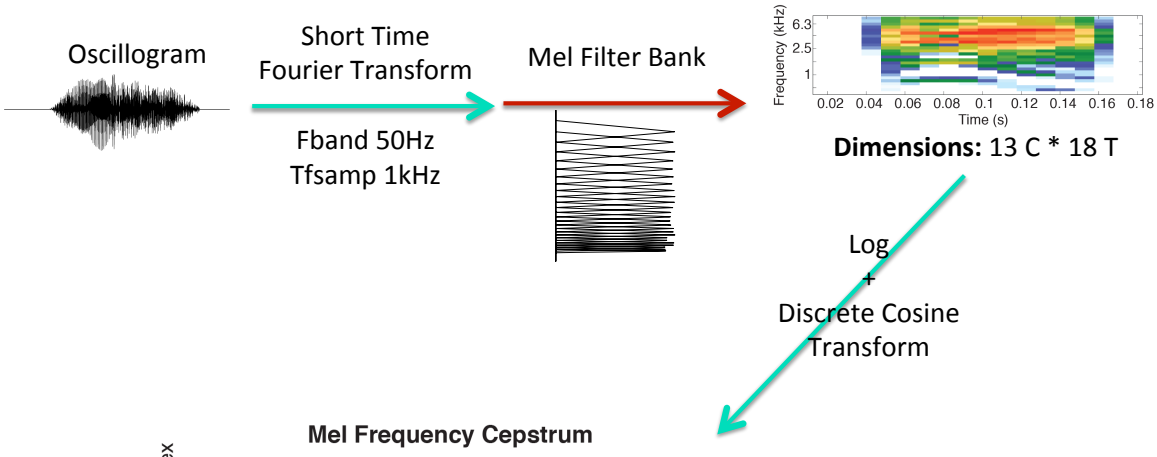
Spectrogram+PCA and RF

Tree Classifier
+
random forest



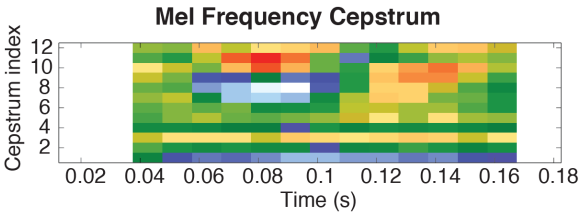
Correct classification:
83.2% (All Calls)
70.4% (Avg per Type)

Mel Frequency Cepstrum Coefficients



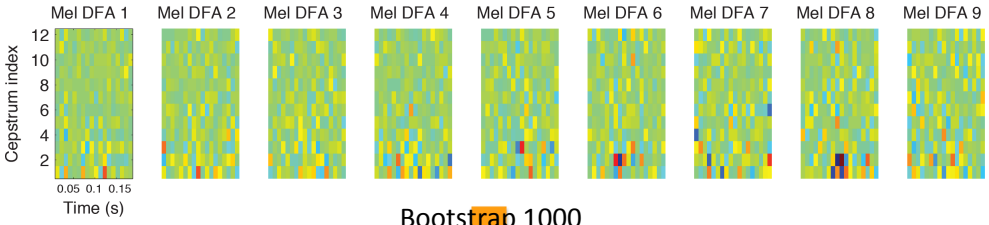
Dimensions: 13 C * 18 T

MFCC + DFA

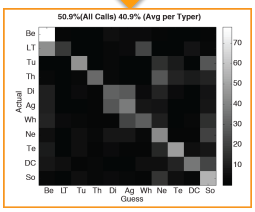


Training data set
(6/7 of the data)

DFA

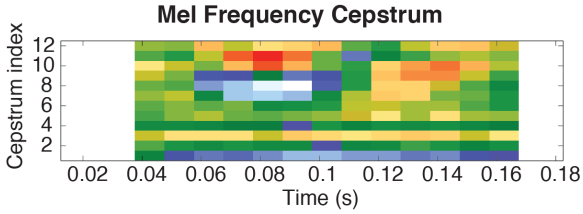


Bootstrap 1000
(cross-validation)

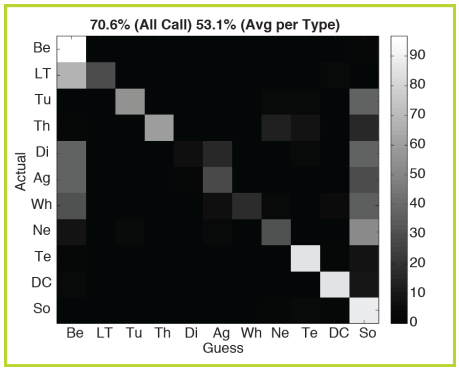


Correct classification:
50.9% (All Calls)
40.9% (Avg per Type)

MFCC + RF

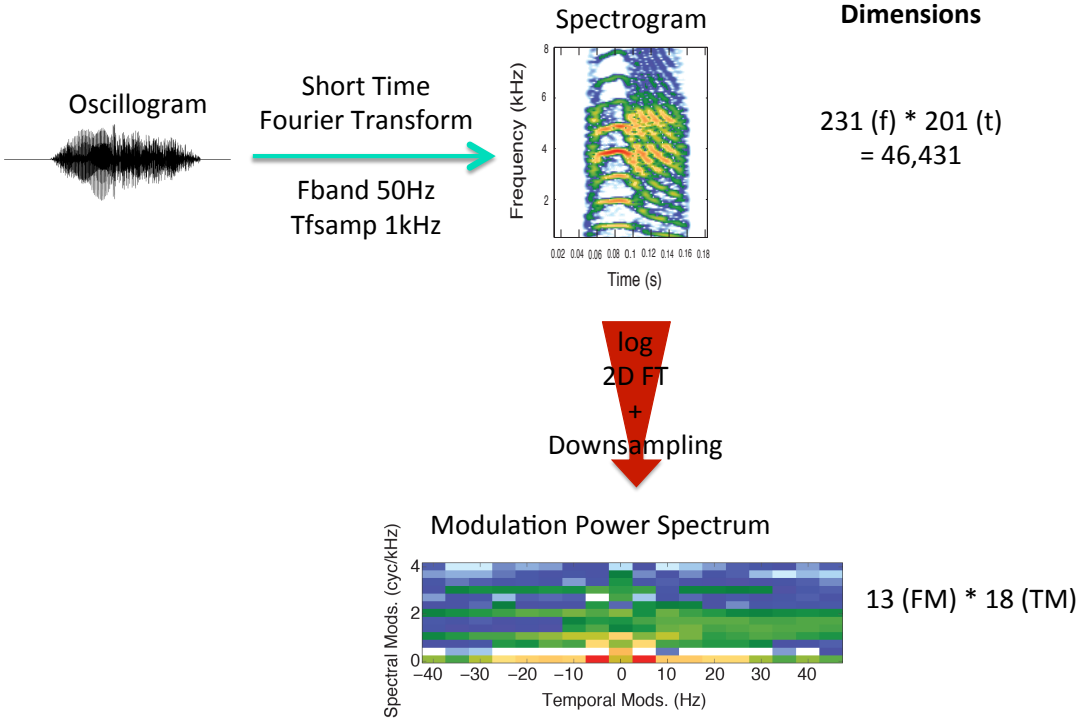


Tree Classifier
+
random forest

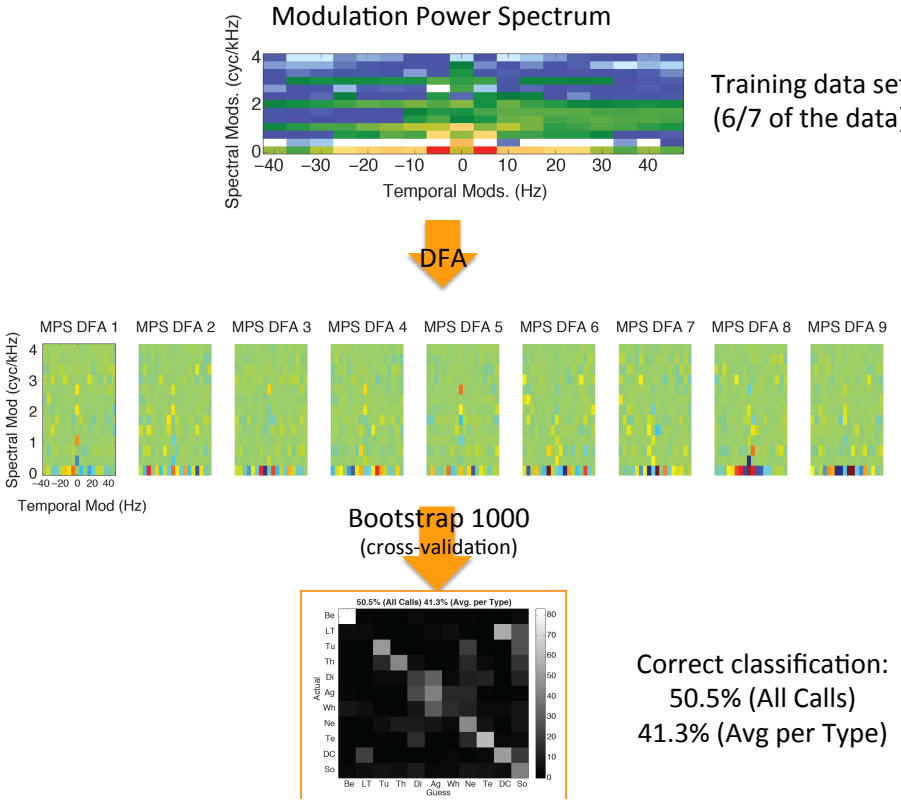


Correct classification:
70.6% (All Calls)
53.1% (Avg per Type)

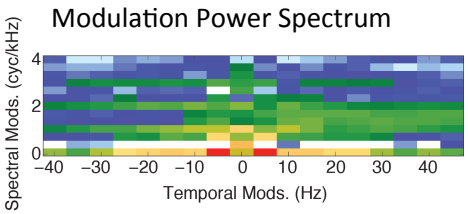
Modulation Power Spectrum (MPS)



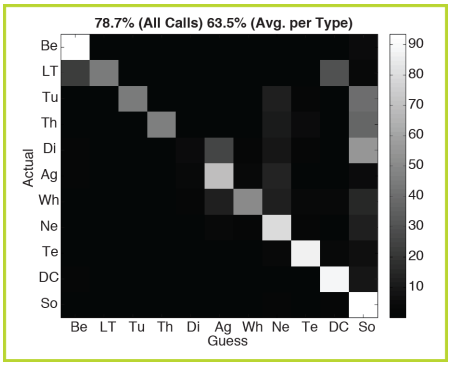
MPS + DFA



MPS + RF

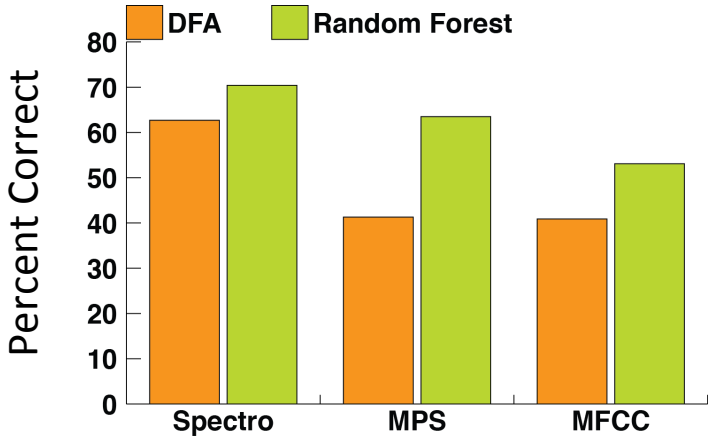


Tree Classifier
+
random forest



Correct classification:
78.7% (All Calls)
63.5% (Avg per Type)

Summary



Feature Space

Conclusions

- PCA on spectrogram yielded the “best” feature space (among 3):
  - Higher classification performance
  - Ease of interpretation
  - Room for improvement (Sparse, ICA)
- MPS might be better than MFCC.
- Random Forest is an efficient classifier.
- Zebra finches have a large repertoire of calls that can be categorized based on behavioral context and acoustics.

Chapter 6

Non Human Speech Processing

6.1 Gabor Scalogram Reveals Formants in High-Frequency Dolphin Clicks

Marie Trone
Valencia College
1800 Denn John Lane
Kissimee, FL 34744
mtrone@valenciacollege.edu

Randall Balestriero
Université de Toulon
av. de l'Université
La Garde, France
randallbalestriero@gmail.com

Hervé Glotin
Institut Universitaire de France
Bt St Michel Paris
& Université de Toulon
glotin@univ-tln.fr

Abstract

Toothed whales (suborder: Odontoceti) produce high-frequency clicks for navigation, and possibly communication. Determining the power spectrum within a click may help differentiate clicks and their possible communicative functions. The short time Fourier transform (STFT) is characterized by a time-frequency trade-off, resulting in difficulty in ascertaining local energy maxima within a short-duration click. We propose to use Gabor wavelet decomposition to get better local energy maxima contrast instead of the Fourier STFT. Click data collected from bottlenose dolphins (*Tursiops sp.*) sampled at 96 kHz and 500 kHz were analyzed using both the STFT and Gabor scalogram. The resulting scalograms were visually inspected. While the STFT spectrograms did not portray the regions of local energy maxima within each click clearly, the Gabor scalogram displayed distinct bands of local energy maxima with respect to frequency. Consecutive clicks that contained regions of higher acoustic energies at approximately the same frequency were defined as formants. Possible 'phonetic units' composed of these formants were subsequently identified. However, the function of these formants and possible phonemes remains speculative. This preliminary study demonstrates the need for scaled algorithms capable of analyzing high-frequency recordings, which may be essential in order to gain a deeper understanding of cetacean communication. Future studies should sample odontocetes with a minimum sampling rate of 500 kHz, or higher. Gabor scalogram analyses could then be used in conjunction with other algorithms to explore correlations between formant frequencies, frequency bandwidths of the entire click that contains the formant, the quantity of formants, and inter-click-intervals in order to discern the possible functions of dolphin formants.

1. Introduction

Cetacean acoustics research is currently expanding due to recent advances in available technology, in conjunction with decreasing costs of equipment. The capability to record the higher frequencies associated with click trains of many cetaceans of the suborder Odontoceti permits more complete acoustic assessments. However, the copious amount of digitally recorded data produced requires an interdisciplinary approach to develop innovative algorithms and procedures to process, reduce and analyze the resultant complex data sets.

Detailed information concerning Odontoceti click acoustics has been derived from studying animals in human care. The ability to finely manipulate variables under controlled conditions has yielded information regarding the timing, frequency and amplitude of clicks. On-axis click signals occur when the receiving animal or hydrophone is positioned directly in front of the signaling dolphin. Off-axis clicks are characterized by lower frequencies and amplitudes when compared to analogous on-axis clicks [R1]. Furthermore, off-axis click trains emitted by multiple animals can produce interference resulting in decreased signal amplitude [R1]. Finally, it has been suggested that the morphology of the cetacean head may cause the signal frequency to decrease as a function of the angle from the beam axis, functioning like a low-pass filter [R2]. However, higher frequency signals have narrower beam patterns than low frequency signals. As a result, high frequency signals are more resistant to signal distortion due to off-axis propagation [R2].

Traditionally the short time Fourier transform (STFT) has been used to analyze the time, frequency and amplitude parameters of an acoustic signal. However, the STFT employs a sliding window that is associated with a time/frequency trade-off, such that shorter STFT windows portray better time resolutions but poorer frequency resolutions in the resulting spectrograms, and vice versa [R2]. Similarly, scalogram analyses also produce time, frequency and amplitude spectrograms. However, these algorithms use wavelet transforms, and consequently improve knowledge of where frequency components are occurring in time [R3]-[R5]. This capacity is valuable when analyzing click trains that occur in short bursts. Furthermore, scalogram analyses are able to detect more clicks than the STFT when the signal to noise ratio is low [R4]-[R5]. Finally, the relatively simple scalogram computations facilitate the prompt acquisition of results [R3]. Thus, scalogram analyses may be beneficial to many real-world applications, such as acoustically detecting multiple sperm whales using a single hydrophone [R3].

Narrow-band click trains characterized by regular, incremental changes in energy, frequency and inter-pulse intervals are known as coherent pulses [R6]. Non-coherent pulses occur when the consecutive, broad-band pulsed sounds within a click train are characterized by wide fluctuations in energy, frequency and inter-pulse intervals [R6]. Dolphins residing in nondescript pools under human care have been recorded to alternately produce these non-coherent pulses [R6]. The function of these irregular, non-coherent pulses is not currently known. Ryabov [R6] has suggested that these pulses may function as phonemes, which are the smallest acoustic units that constitute a language.

Collectively, human languages consist of approximately 371 phonemes characterized by frequencies that fall between 20 and 20,000 Hz [R6]. Phonemes are depicted on spectrograms as formants and scalogram analyses have been used to detect formants in human speech [R7]. However, cetaceans possess far more extensive hearing ranges and acoustic repertoires than humans. Indeed, bottlenose dolphins (*Tursiops truncatus*) produce sounds that vary between 200 to over 500,000 Hz, leading one researcher to propose that these animals are capable of producing at least 3,000 phonemes consisting of non-coherent pulses [R6].

Studying free-ranging cetaceans requires the ability to utilize parameters that are not impacted by being recorded off-axis and by interference from multiple animals producing click trains in the same time interval. High-frequency acoustic signals are more resistant to distortion resulting from off-axis propagation [R2]. Thus, scalogram analysis may be a more appropriate tool in cetacean acoustic investigations due to its superiority in determining where frequency components are occurring in time [R3]-[R5]. Furthermore, scalogram analyses have been used to detect formants in human speech [R7]. Thus, we propose to analyze bottlenose dolphin click trains using both STFT and scalogram analyses to compare the efficiency of these two algorithms. Subsequently, we examine the resulting spectrograms for possible formants.

2. Material

Acoustic signals from two species of bottlenose dolphins (*Tursiops sp.*) were collected in their natural environments.

2.1 Indian River Lagoon, Florida, *Tursiops truncatus*

A group of three bottlenose dolphins were recorded in July of 2013, in the Indian River Lagoon, Florida. A Cetacean Research Technology CR-3 Hydrophone, a Reson EC 6061 preamp, and an IOTech Personal Daq/3000 Series digital acquisition system were used to obtain these recordings. A 500 kHz sampling rate at 16 bit resolution was employed during recordings yielding usable data up to 250 kHz. DaqView Data Acquisition software was used to convert and store data in the .wav format. These dolphins were attracted to the vicinity of the recording vessel by the fishing activities of the boat's captain. This .wav file is available in the NIPS4B website at

http://sabiod.univ-tln.fr/nips4b/media/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE.wav and is used under the copyright given in the NIPS4B website.

2.2 Western Australia, *Tursiops aduncus*

A 37-year-old female *Tursiops aduncus* dolphin was recorded in Monkey Mia-Shark Bay, Western

Australia in August 2013 during a mission of the SABIOD project, using a Cetacean Research Technology CR 55 hydrophone and a TASCAM audio digital recorder. The recordings were made with a 96 kHz sampling rate and 24 bit resolution. The recording is available in the NIPS4B website under the NeuroSonar session:

http://sabiod.univ-tln.fr/nips4b/media/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_32bits_after19min_nips4bfile_e.wav

3. Methods

3.1 Short term Fourier transform (STFT)

The *T. truncatus* and *T. aduncus* audio recordings were divided into 0.4 second and 0.7 second segments respectively. Subsequently, 168 *T. truncatus* segments and 149 *T. aduncus* segments were analyzed using STFT, with half-overlapping windows of 512 ms durations. The resultant spectrograms show the local power spectrum of the signal over time.

3.2 Gabor scalogram transform

The same files that were analyzed using the STFT were also analyzed using a Gabor scalogram transform with the ScatNet toolkit [R8] to produce scalograms. The T coefficient was set to 32, the Q coefficient was set to 16, and the J coefficient was set to 80. Again, the resultant scalograms show the local power spectrum of the signal over time.

It should be noted that, unlike STFT spectrograms, center frequencies of the wavelets are quantized in a geometric progression. Henceforth, the Y-axis of the scalogram is naturally logarithmic. Additionally, the temporal duration of the corresponding spectrograms and scalograms are the same. However, the scalogram X-axis is similarly transformed using the ScatNet toolkit, and in actuality reflects the time scale portrayed in the spectrogram [R8].

3.3 Formant detection

Following procedures outline by Jemma et al. [R7], formants were identified in the spectrograms and scalograms. Within each click, frequencies of highest amplitude were identified by visual inspection. Consecutive clicks that contained regions of higher acoustic energies at approximately the same frequency were identified as formants. Each audio .wav file was analyzed using the STFT and the Gabor scalogram transform, allowing a comparison of these two representations for formant tracking.

4. Results

The STFT spectrograms did not portray the regions of local energy maxima within each click clearly. Instead, the energy appeared to be equally distributed among the various frequencies that demarked the click. On the contrary, the Gabor scalogram layer 1 displayed distinct bands of local energy maxima with respect to frequency. For ease of comparison, Figure 1 depicts the Gabor scalogram layers 1, 2 and 3 of the scattering decomposition, as well as the STFT spectrogram from a segment of the Florida *T. truncatus* recordings.

We continued our exploration of dolphin acoustics by focusing our attention on layer 1 of each scalogram. Bands of high amplitude sharing the same frequency on adjacent clicks were connected with red lines (see Figure 2). Following human speech terminology [R7], we identified these clusters of local energy maxima as dolphin formants.

Moreover, we labeled the formants depicted in Figure 2 as phoneme units A, B, C and D. Phoneme B seemed to be composed of phoneme A plus an additional formant. Similarly, phoneme D appeared to be composed of phoneme C plus an additional formant. Furthermore, phoneme C seemed to be a shift upward of phoneme B. Additionally, three individual clicks contained local energy maxima, but could not be connected to either adjacent click. These clicks occurred at 200, 270 and 300 ms.

The original scalograms and spectrogram for this file are available at the following URL:
http://sabiod.univ-tln.fr/pimc/TURSIOPS_FLORIDA_norformants/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE_J80_Q16_T32/part6_983041_windowsnb2_T32_Q16_J80.png

The following Gabor scalograms derived from the *T. truncatus* audio file display distinct formants. These formants cannot be ascertained from inspection of the corresponding STFT spectrograms, which can be accessed at the following URLs for comparison:

http://sabiod.univ-tln.fr/pimc/TURSIOPS_FLORIDA_norformants/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE_J80_Q16_T32/part9_1572865_windowsnb2_T32_Q16_J80.png

http://sabiod.univ-tln.fr/pimc/TURSIOPS_FLORIDA_norformants/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE_J80_Q16_T32/part65_12582913_windowsnb2_T32_Q16_J80.png

http://sabiod.univ-tln.fr/pimc/TURSIOPS_FLORIDA_norformants/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE_J80_Q16_T32/part74_14352385_windowsnb2_T32_Q16_J80.png

Finally, a directory with links to the 168 corresponding scalograms/spectrogram spectra derived from our *T. truncatus* audio recordings, each 0.4 seconds in duration, can be accessed at the following URL:
http://sabiod.univ-tln.fr/pimc/TURSIOPS_FLORIDA_norformants/NIPS2_TURSIOPS_20_2013_mosquito_lagoon_florida_TRONE_J80_Q_16_T32/

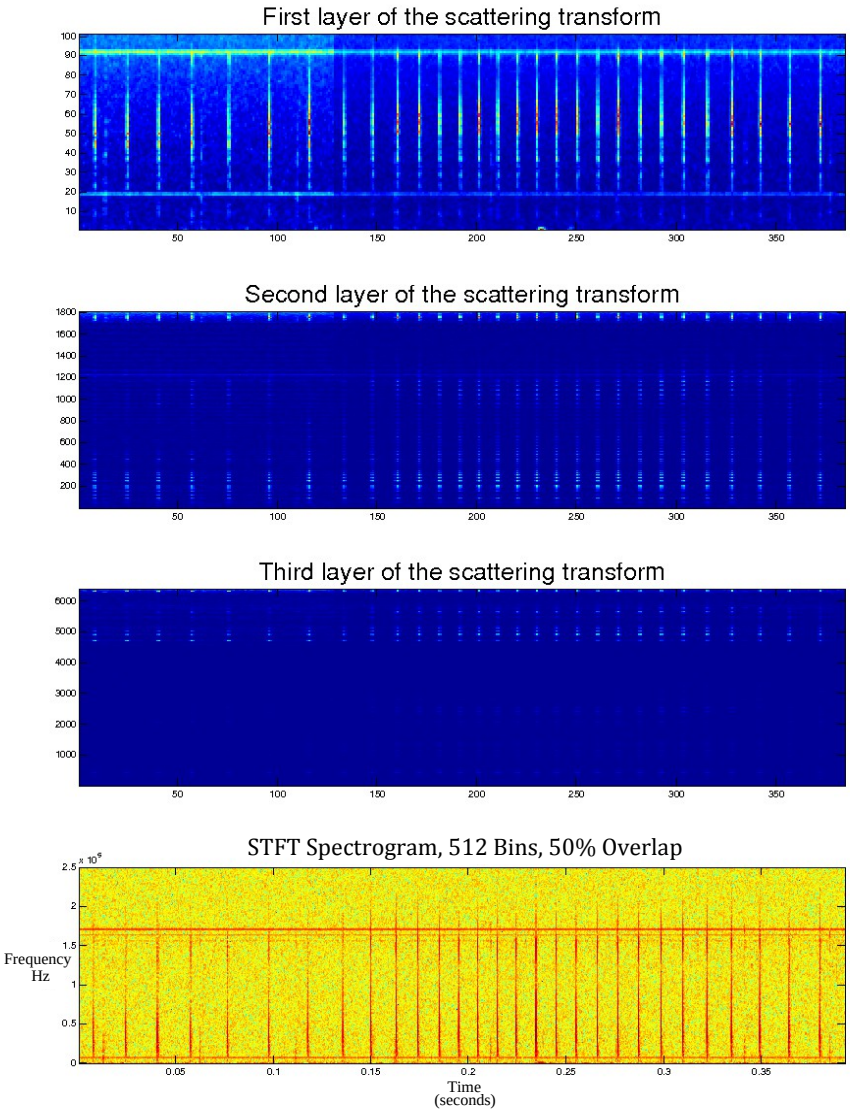


Figure 1: The scalograms (layers 1, 2 and 3) of the scattering representation of Gabor scalogram and the STFT spectrogram, from a segment of the Florida T. truncatus file, 500 kHz sampling rate, 0.4 seconds, coefficients T=32, Q=16, J=80.

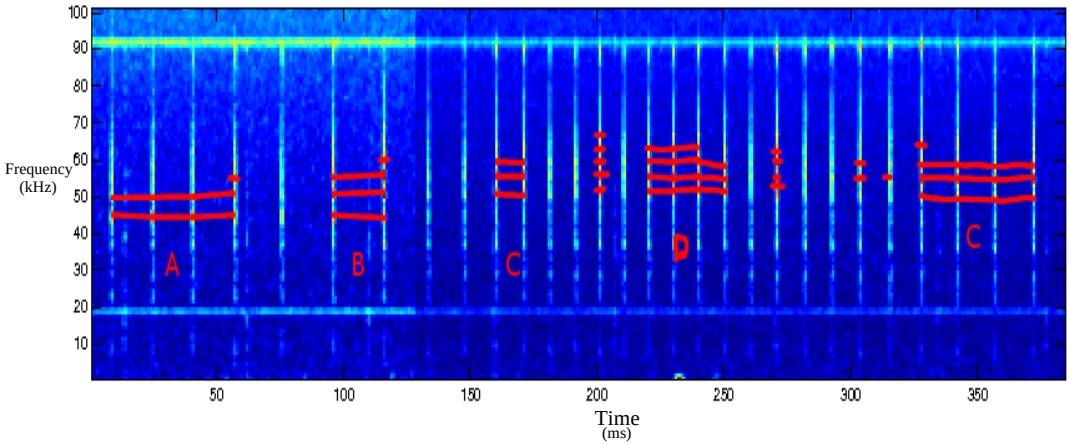


Figure 2: Scalogram layer 1 from the previous figure displaying the joint formant nodes. Bands of high amplitude sharing the same frequency on adjacent clicks have been connected with red lines. Phoneme units have been labeled as A, B, C and D. Total duration 0.4 seconds (Fs : 500 kHz).

In addition, scalograms and spectrograms derived from the audio recordings of *T. aduncus* in Western Australia are also available. Each of the 149 segments depicts 0.7 seconds of audio recordings. These corresponding scalograms/spectrogram spectra can be accessed from a file directory using the following URL:

http://sabiod.univ-tln.fr/pimc/NICKY_norformants/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_J80_Q16_T32/

Figure 3 depicts the scalograms and spectrogram derived from the first *T. aduncus* file listed below. Similar to the previous example, the first layer of the scalogram highlights the frequencies containing the most energy per click, whereas the energy appears to be evenly distributed among the various frequencies on the spectrogram. Moreover, when the energy maxima are connected with red lines, two phoneme units appear, which we have labeled as D_1, D_2 and E. D_1 contains two clicks, while the two phoneme units labeled D_2 consist of a single click each. Eight clicks make up phoneme unit E.

These clicks which were recorded with a 96 kHz sampling rate are characterized by a frequency bandwidth of approximately 35 kHz. All of these clicks portray their highest energies at or very near the 48 kHz ceiling of the spectrogram. This is in contrast to the clicks sampled at 500 kHz, which are typified by a frequency bandwidth of approximately 160 kHz (see Figures 1 and 3).

We have posted four Gabor scalograms derived from the *T. aduncus* audio recordings that display distinct formants. These formants cannot be ascertained from inspection of the corresponding STFT spectrograms, which can be accessed at the following URLs for comparison:

http://sabiod.univ-tln.fr/pimc/NICKY_norformants/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_J80_Q16_T32/part115_7471105_windowsnb1_T32_Q16_J80.png

http://sabiod.univ-tln.fr/pimc/NICKY_norformants/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_J80_Q16_T32/part5_262145_windowsnb1_T32_Q16_J80.png

http://sabiod.univ-tln.fr/pimc/NICKY_norformants/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_J80_Q16_T32/part73_4718593_windowsnb1_T32_Q16_J80.png

http://sabiod.univ-tln.fr/pimc/NICKY_norformants/Tursiops_truncatus_Nicky_SHARKD_0002S34D12_day3_aug2013_SABIOD_96kHz_J80_Q16_T32/part131_8519681_windowsnb1_T32_Q16_J80.png

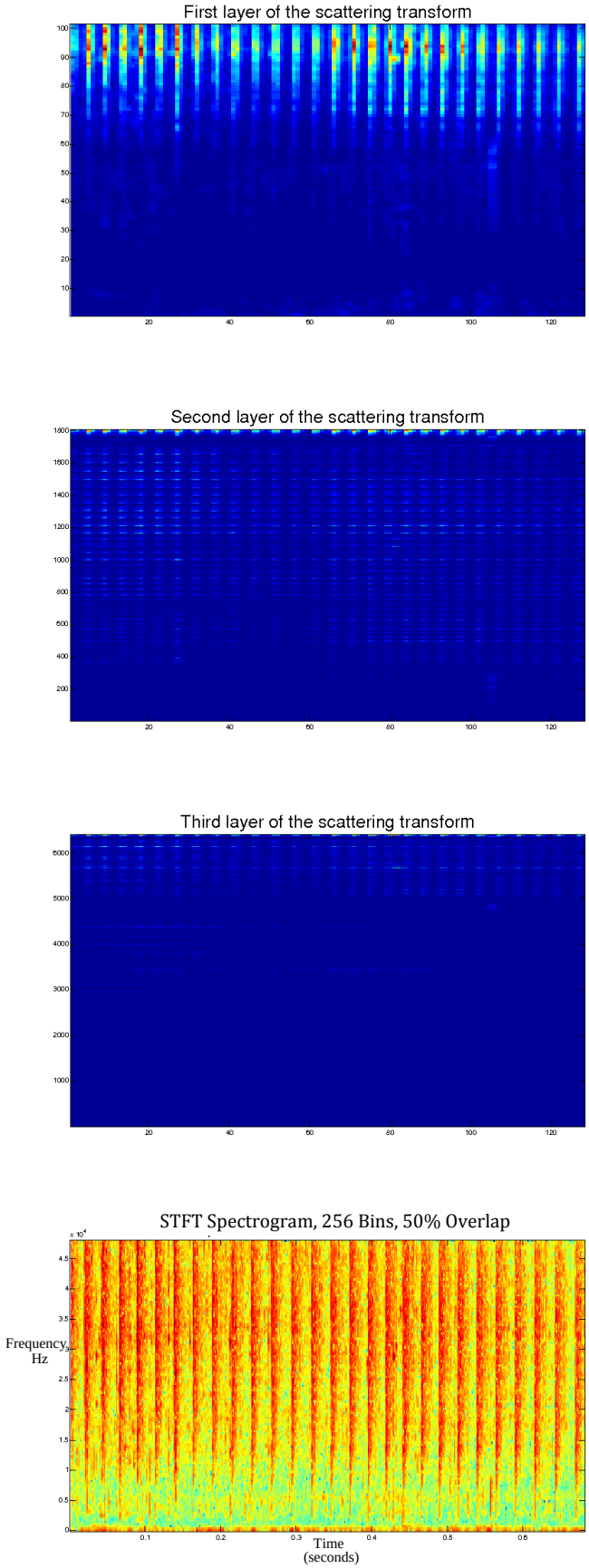


Figure 3: The scalogram (layers 1, 2 and 3) of the scattering representation of Gabor scalogram and the STFT spectrogram, from a segment of the Western Australia T. aduncus file, 96 kHz sampling rate, 0.7 seconds, Coefficients T=32, Q=16, J=80.

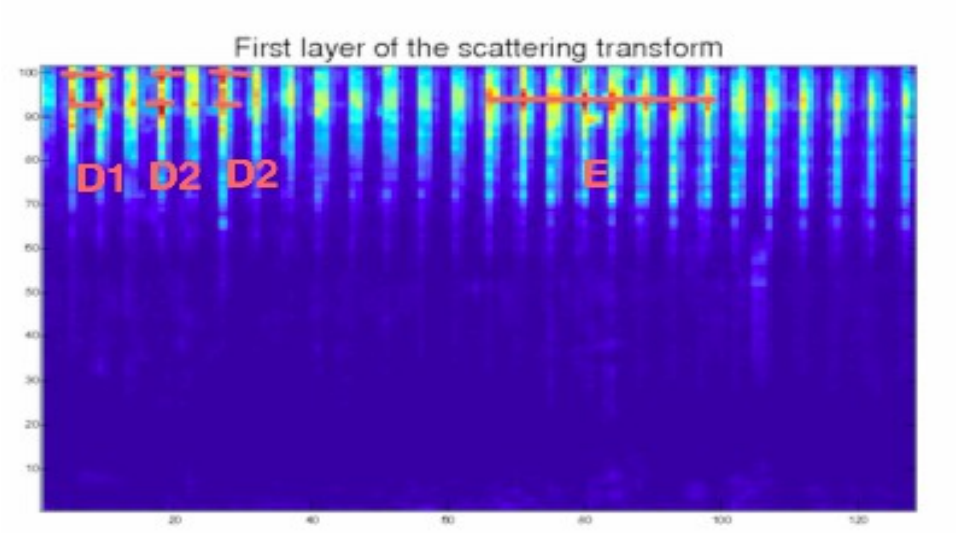


Figure 4: Scalogram layer 1 from the previous figure displaying the linked formant nodes. Bands of high amplitude sharing the same frequency on adjacent clicks have been connected with red lines. Phoneme units have been labeled as D_1, D_2 and E. Total duration 0.7 seconds (Fs : 96kHz).

5. Discussion and Conclusion

This exploratory study compared the quality of information obtained from the traditional STFT with that derived from a Gabor scalogram wavelet transform using high frequency, broad-band dolphin clicks. The timing of the clicks demonstrated strong correspondence between the two spectra and the Gabor scalogram did not detect more clicks. This result is contrary to those obtained by Adam [R4] and Lopatka et al.[R5]. Perhaps the signal to noise ratio was great enough that the STFT was as efficient as the Gabor scalogram.

However, the scalograms portrayed distinct local energy maxima at specific frequencies, whereas energy bands in the STFT spectrograms were generally indistinct. The explanation for this effect is two-fold. First, STFT analyses that portray strong time resolutions do so at the expense of poor frequency resolutions [R2]. Second, while the Gabor kernel employed in these scalogram analyses is efficient at capturing some energy in layers 2 and 3, it is extremely efficient at enhancing the bands of local energy maxima with respect to frequency in layer 1 [R9].

The scalograms at layer 1 reveal evident local energy maxima, reminiscent of formants found in human speech spectra. Thus, we propose that bands of high amplitude sharing the same frequency on adjacent clicks be connected to form dolphin formants (see Figures 2 and 4). As in human speech spectra, patterns are then ascertained. We demonstrate four formant patterns in Figure 2, and labeled each as a 'phoneme' unit. Moreover, three individual clicks contained unique energy maxima at different frequencies, and each could be considered a phoneme as well. Using the same methodology, 2 phoneme units were found in the scalogram depicted in Figure 4. We subdivided one phoneme unit into different varieties based upon the number of clicks within the phoneme unit.

Most likely there were additional high frequency formants in the clicks displayed in Figure 4. Frequencies higher than 48 kHz could not be detected given the 96 kHz sampling rate. All of these clicks were characterized with the local energy maxima in the top 10% of the spectrogram. Moreover, these *T. aduncus* clicks were demarked by a frequency bandwidth of approximately 35 kHz, while the *T. truncatus* clicks were typified by a bandwidth of approximately 160 kHz, even though all of these clicks had their lowest energies in the 10-20 kHz range. Indeed, most of the clicks that we have inspected that were sampled at 500 kHz have energies commonly ranging up to 165 kHz and higher. Thus, these differences in bandwidth can be attributed to sampling procedures. We therefore conclude that acoustic sampling should be conducted with a minimum sampling rate of 500 kHz when determining phoneme units in order to include all of the possible formants in the analyses.

It may be that these formants function in social communication, as suggested by Ryabov [R6].

Indeed, other researchers have suggested that cetaceans utilize pulsed signals for communication as well as navigation. For example, sperm whale clicks have been associated with various social contexts, including reunions, separations, contact calls, and in response to unusual underwater sounds [R10]. Additionally, pulsed signals have been associated with agnostic interactions, aggressive behavior, discipline, and excitement among Atlantic spotted dolphins (*Stenella frontalis*) and bottlenose dolphins (*T. truncatus*) [R11]. Finally, harbor porpoises (*Phocoena phocoena*) utilize stereotyped, narrow-band high frequency clicks among conspecifics during aggressive interactions [R12].

Currently, the significance of the local energy maxima identified in our scalogram analyses remains speculative. Although we are suggesting that they may function as communicative phonemes, many researchers will be skeptical. Indeed, some proclaim that humans will not be able to fully appreciate cetacean communication due to the difficulty and improbability of identifying the basic, communicative unit, or phoneme [R13].

Other critics might claim that these formants would only be valid if the clicks were recorded on-axis. However, high frequency signals are more resistant than low-frequency signals to the amplitude and frequency weakening caused by off-axis propagation [R2]. Such distortion has been demonstrated for 115 kHz beluga whale (*Delphinapterus leucas*) clicks and for 122 kHz bottlenose dolphin clicks [R14]. Nevertheless, our data demonstrate bottlenose dolphins produce clicks of even higher frequencies. Indeed, Figure 1 displays clicks with energy in the 200 kHz range. Furthermore, some of the clicks that we have recorded demonstrate high amplitudes at 250 kHz, suggesting a ceiling effect due to the limitations of our recording system (Figure 5). Finally, bottlenose dolphins have been reported to produce clicks ranging between 400 and 500 kHz [R15]-[R16]. To our knowledge, the degree to which these high frequency signals degrade due to off-axis propagation has yet to be determined.

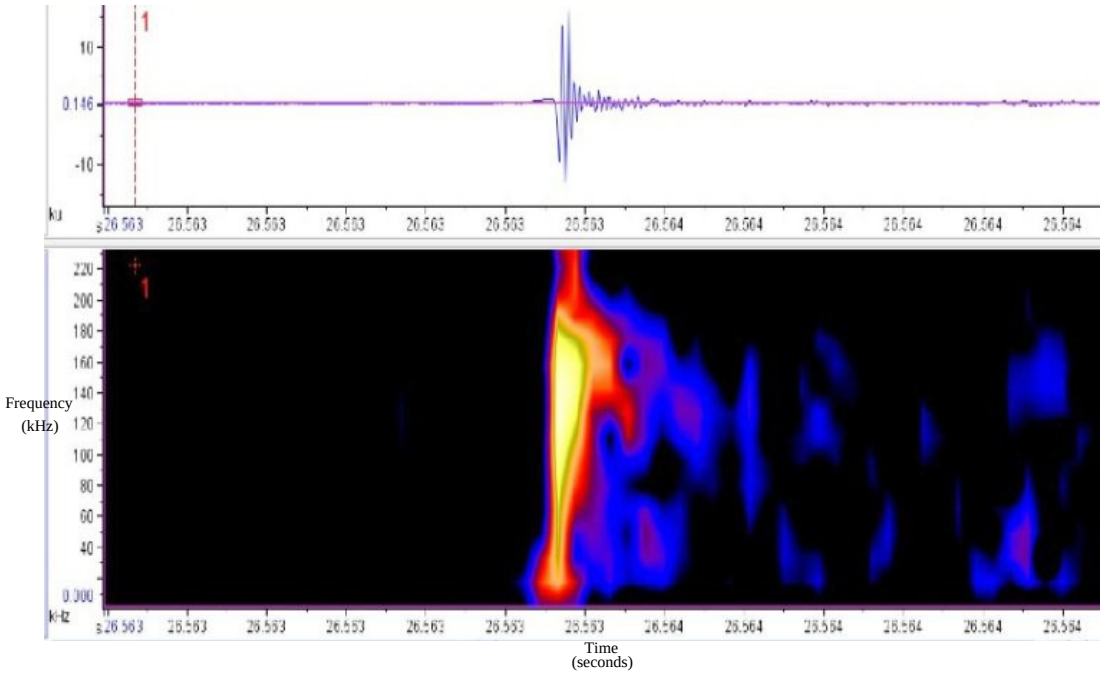


Figure 5. An FFT spectrogram of a bottlenose dolphin click that was recorded with a 500 kHz sampling rate, suggesting that information exists above 250 kHz. The click was analyzed with Raven Pro 1.5 software.

Furthermore, cetaceans live in a three-dimensional world, where most clicks are most likely received off-axis by conspecifics. Given that high frequency pulsed signals have been associated with certain social situations, it is highly probable that off-axis degradation does not significantly impede conspecific interpretation of these acoustic signals. Indeed, cetaceans may have evolved to produce and perceive high frequency signals precisely because these signals are resistant to off-axis distortion, despite the fact that high frequency signals attenuate more readily than low frequency signals.

In conclusion, we suggest that Gabor scalogram transforms outperform STFT analyses of cetacean acoustics. Scalogram analyses are not subject to the time/frequency distortion trade-off that is characteristic of the STFT. Thus, parameters that can be reliably ascertained through scalograms

include formant frequency; frequency bandwidth of the entire click that contains the formant; quantity of clicks; and inter-click intervals. Exploring patterns associated with these parameters may expand our understanding of cetacean communication and evolution, possibly facilitating conservation efforts.

Future studies should utilize recording equipment capable of recording up to 1 MHz in order to fully document the acoustic repertoire of cetaceans. Pattern recognition algorithms should be developed to automate the tedious task of identifying formants within audio recordings containing thousands of clicks. Furthermore, algorithms that are capable of ascertaining formant frequencies, frequency bandwidths of the entire click that contains the formant, the quantity of formants, and inter-click-intervals are essential in order to discern the possible functions of dolphin formants. Once patterns have been identified, playback studies could be utilized to determine the role of such patterns.

Acknowledgments

We thank David E. Bonnett for advice, access to the 500 kHz recording equipment and assistance in acquiring the *T. truncatus* audio files. We thank Vincent Lostanlen, Phd student at DATA team in DIENS laboratory ENS Paris for his collaboration. We thank the Big Data MASTODONS MI CNRS project for its support with the SABIOD research program (http://sabiod.org).

References

[R1] Au, W.L. (1993). *The Sonar of Dolphins*. Springer-Verlag, New York

[R2] Au, W.L. & Hastings M.C. (2008). *Principles of Marine Bioacoustics*. Springer, New York

[R3] Lelandais, F. & Glotin, H. (2008). Mallat's Matching Pursuit of sperm whale clicks in real-time using Daubechies 15 wavelets. In *New Trends for Environmental Monitoring Using Passive Systems*. IEEE conf. Passive *DOI*:10.1109/PASSIVE.2008.4786977

[R4] Adam, O. (2006). Advantages of the Hilbert Huang transform for marine mammals signals analysis. *Journal of the Acoustical Society of America*, 120(5), 2965–2973

[R5] Lopatka, M., Adam, O., Laplanche, C., Zarzycki, J. & Motsch, J.F. (2005). An attractive alternative for sperm whale click detection using the wavelet transform in comparison to the Fourier spectrogram. Aquatic Mammals, 31(4), 463-467. DOI: 10.1578/AM.31.4.2005.463

[R6] Ryabov, V. (2011). Some Aspects of Analysis of Dolphins' Acoustical Signals. *Open Journal of Acoustics*, 1, 41-54. DOI:10.4236/oja.2011.12006 Published Online (http://www.SciRP.org/journal/oja)

[R7] Jemma, I., Ouni, K., Laprie, Y., Ouni, S. & Haton, J.P. (2013). A new automatic formant tracking approach based on scalogram maxima detection using complex wavelets. Int. Conf. on Control, Engineering & Information Technology, Sousse, Tunisia

[R8] Mallat et al. (2013). The Scanet Toolkit, http://www.di.ens.fr/data/software/

[R9] Mallat S. (2013). Personnal communication

[R10] Watkins, W.A. and Schevill, W.E. (1977). Sperm whale codas. *Journal of the Acoustical Society of America*, 62, 1485-1490

[R11] Herzing, D.L. (1996). Vocalizations & associated underwater behavior of free-ranging Atlantic spotted dolphins, *Stenella frontalis* & bottlenose dolphins, *Tursiops truncatus*. *Aquatic Mammals*, 22(2), 61-79

[R12] Clausen, K.T., Wahlberg, M., Beedholm, K., Deruiter, S. and Madsen, P.T. (2011). Click communication in harbor porpoises *Phocoena phocoena*. *Bioacoustics: The International Journal of Animal Sound and its Recording*, 20(1), 1-28. DOI:10.1080/09524622.2011.9753630

[R13] Kuczaj, S. (2013). Why we will never be able to speak with dolphins. Public presentation at Eckerd College, St. Petersburg

[R14] Au, W.W.L., Penner, R.H., and Turl, C.W. (1987). Propagation of beluga echolocation signals. *Journal of the Acoustical Society of America*, 83, 807-813.

[R15] Lemerande, T.J. (2002). Transmitting beam patterns of the Atlantic bottlenose dolphin (*Tursiops truncatus*): Investigations in the existence & use of higher frequency components found in echolocation signals. Master's Thesis. Naval Postgraduate School, Monterey, CA. 148 p.

[R16] Toland, R.W. (1998). High frequency components in bottlenose dolphin echolocation signals. Master's Thesis. Naval Postgraduate School, Monterey, CA. 83 p.

6.2 Supervised Classification of Baboon Vocalizations

Maxime Janvier[*], **Radu Horaud**
INRIA Grenoble Rhône-Alpes
Grenoble, France
maxime.janvier@inria.fr
radu.horaud@inria.fr

Laurent Girin, Frédéric Berthommier, Louis-Jean Böe
GIPSA-lab
Grenoble-Alpes University, CNRS
Grenoble, France
laurent.girin@gipsa-lab.fr
frederic.berthommier@gipsa-lab.fr
louis-jean.boe@gipsa-lab.fr

Caralyn Kemp, Arnaud Rey
Laboratoire de Psychologie Cognitive
and Brain and Language Research Institute
Aix-Marseille University, CNRS
Marseille, France
caralyn@kemputer.com.au
arnaud.rey@univ-amu.fr

Thierry Legou
Laboratoire Parole et Langage
and Brain and Language Research Institute
Aix-Marseille University, CNRS
Marseille, France
thierry.legou@lpl-aix.fr

Abstract

This paper addresses automatic classification of baboon vocalizations. We considered six classes of sounds emitted by *Papio papio* baboons, and report the results of supervised classification carried out with different signal representations (audio features), classifiers, combinations and settings. Results show that up to 94.1% of correct recognition of pre-segmented elementary segments of vocalizations can be obtained using Mel-Frequency Cepstral Coefficients representation and Support Vector Machines classifiers. Results for other configurations are also presented and discussed, and a possible extension to the "Sound-spotting" problem, i.e. online joint detection and classification of a vocalization from a continuous audio stream is illustrated and discussed.

1 Introduction

Nonhuman primates produce a relatively limited variety of species-specific vocalizations in response to particular social events [1]. Until recently, classifying these vocalizations has been performed by ear and by time-consuming manual analysis [2]. Several researchers conducted various analyses, making comparison between studies, as well as between species, difficult [3]. The automatic classification of vocalizations can assist the field of primate communication in a multitude of ways: firstly, it can be used to assist and complement the classification made by experts. For example, it can be used to assess the relevance of different sets of acoustic features for the characterization of the different sound categories. Secondly, automatic classification of sounds in such a context can be exploited by audio/video recording systems dedicated to ethological studies or environment preservation. For example, the detection of relevant sounds emitted by the animals under study may indicate a scene of interest and trigger the video recording, thereby avoiding useless data storage and power consumption. Understanding the differences between the broad vocal classifications (i.e., comparison of a grunt to a scream) will better improve the fine-tuning of these analyses required for graded vocal calls and the differences in vocal production by different individuals for the same sound. In this work, we consider different supervised analyses for the classification of baboon vocalizations, which, to our knowledge, is the first study of its kind.

[*]M. Janvier is funded by the "Direction Générale de l'Armement" (DGA) included in the French Ministry of Defence.

In this paper we consider six categories of baboon vocalizations. We report the results obtained with the use of different audio signal representations and supervised classification methods to characterize and recognize these vocalizations. To this end, we tested different spectral features computed based on the usual short-term sliding window approach, e.g., Mel Frequency Cepstral Coefficients (MFCC). We propose to introduce a sparse subset of coefficients characterizing the harmonicity of the vocalizations, since, as opposed to (human) speech, the range of the fundamental frequency is quite different across the baboon sound categories. As for the classifiers, we used hidden Markov models (HMMs) [4] to model the dynamic evolution of the spectral patterns within each sound category. We also tested k-Nearest Neighbors (KNN) classifiers, Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) [5], [6] with different configurations and appropriate preprocessing of the data (especially for time alignment of feature vector sequences). Note that most of the presented experiments concern isolated sounds that were manually pre-segmented, but we also discuss and illustrate the feasibility of the extension of our system(s) to the "soundspotting" problem, i.e. online joint automatic detection (i.e. segmentation) and classification of vocalizations from a continuous audio stream.

The paper is organized as follows: Section 2 describes the data that were used for this study; Sections 3 and 4 present respectively the different features and classifiers that were used; Experimental results are presented in Section 5 and conclusions are drawn in Section 6.

2 Data

We recorded the vocal behavior of *Papio papio* Guinea baboons housed at the Rousset-sur-Arc CNRS primate center, France. The vocalizations of sixteen baboons (13 females, 3 males; aged between 2 and 27 years at the start of recording) were considered for this study. Fourteen of the baboons were housed as part of a larger group in a 25 $\times$ 30 m outdoor enclosure connected by wire tunnels to indoor housing (6 $\times$ 4 m) used at night. The other baboons were housed separately in a 4.7 $\times$ 6.4m outdoor enclosures connected to indoor housing (2 $\times$ 4m). All groups had visual and auditory contact with each other. The monkeys could be identified by their individual physical characteristics and by number tags on a chain around their neck. Once daily feeding (fruits, vegetables and monkey chows) occurred at 5PM; water was provided ad libitum. See [7] for a more detailed description of the research facilities at the Rousset-sur-Arc CNRS primate center. We used opportunistic sampling techniques to record spontaneous vocalizations produced in response to social events and to stimuli occurring naturally within the baboons' environment. The presence of the recorders and their equipment did not disturb the baboons from their natural daily activities. Recording took place between 8:00 and 21:00 (except 17:00-18:00 due to the baboons being fed at this time) between September 2012 and June 2013. Recording was conducted at a distance from the baboons of $<$ 2m to 20m, with the greater distances suitable only for the long-distance vocalizations. A digital Zoom Handy Recorder H4n (Zoom, Japan: 44.1kHz sampling frequency, 16-bit resolution, mono) with a Me66 Sennheiser directional microphone (Sennheiser Electronic KG, Germany; with windscreen) was used to record the vocalizations. This is a super cardioid microphone with a high sensitivity (50 mV/Pa $\pm$ 2.5dB) and a wide (40Hz–20000Hz) and flat ($\pm$ 2.5dB) frequency response. As the vocalizations were recorded outdoors, environmental sounds at different noise levels may have interfered with the sounds at the focus of the recordings.

From continuous audio streams, individual "homogeneous" sequences of vocalizations (i.e. a series of sounds of the same class) were first manually extracted by an expert for analysis. Those sequences were further manually segmented into elementary sounds that were labelled to be submitted to our classifiers. Six vocalization types were considered in the present study: barks, grunts, copulation grunts (denoted "Cops" throughout the rest of the paper for concision), screams, wahoos, and yaks. In total, the number of sounds per classification was: 110 barks, 130 copulation grunts, 384 grunts, 119 screams, 64 wahoos, and 336 yaks. Original sequences were used to illustrate the feasibility of the "Sound-spotting" task (see Sections 4.4 and 5.4).

3 Features

This Section presents the audio features used in this study. Although we consider here vocalization *elements*, i.e. elementary sounds that can be part of a series of longer vocalizations, and that have been previously segmented, those elementary sounds can be of variable length. Moreover, they can

be more or less stationary (and in general, they are rather non stationary). Therefore, from these elementary sounds, we first extracted *time sequences* of *feature vectors* computed using a short-term sliding window (for instance, a 30ms-Hamming window with 50% overlap). This approach is familiar in speech processing, as well as in audio processing in general (e.g. for the analysis of domestic or environmental sounds), and we inspire from those fields. Also, the features that we use have been largely presented in the related literature [8, 9], and, thus, we present them only briefly.

Mel-Frequency Cepstral Coefficients: MFCCs [10] are cepstral coefficients that represent the envelope of the short-term spectrum on a perceptive mel-frequency scale. Those coefficients are computed as the discrete cosine transform (DCT) of the logarithm of FFT power coefficients passed through a mel-filter bank (e.g. 40 log-spaced bands in the range 300Hz-10kHz; the bandwidth and number of bands can vary; see Section 5). The first coefficient was omitted since it represents the absolute energy of the signal frame and not the spectral shape, and the 1st and 2nd derivatives are added optionnally (depending on experiment).

Average Spectral Features: We tested a series of features that represent average properties of the Short-Term Fourier Transform (STFT) spectrum. The *Spectral Roll-off* is the cut-off frequency below which 99% the spectral energy is contained. The *Spectral Moments* characterize the overall shape of the spectrum using n-order moments of frequency bin weighted by spectral magnitude. We tested the 4 first moments. The *Spectral Slope / Decrease* represents the global amount of decreasing of the spectral amplitude. The *Spectral Flatness* of the magnitude spectrum is given by the ratio between its arithmetic and geometric mean. Finally, the *Spectral Flux / Correlation* measure the average variation between two consecutive spectra.

F_0 and Harmonicity Index: The above-mentioned MFCCs (resp. the ASF) are coefficients that characterize the spectral envelope (resp. the global shape of the spectrum) on a perceptive (resp. linear) frequency scale. MFCCs are widely used in Automatic Speech Recognition (ASR) systems [10] since the spectral envelope characterizes the different speech sounds through the effect of the speaker's vocal tract, while cutting loose of speech sound dependence on fundamental frequency F_0. This is a desirable property for ASR, in order to limit speech variablity across speakers and utterances. In contrast, in the present context of baboon vocalizations, we think that the F_0 *range* can be a discriminative feature since it varies much between some of the considered classes. Therefore, we propose to test the F_0 value (also extracted on a short-term basis) as an audio feature. We also tested the *harmonicity index*, which is the ratio between the second maximum of the signal (short-term) autocorrelation function (which is also used to detect F_0) and the maximum which is obtained at lag zero. The harmonicity index provides some simple confidence measure of the F_0 value.

Feature post-processing: The successive feature vectors of a sound can be further processed to produce different final features, which will feed the classifiers. In particular, the feature vector sequences are generally of different lengths, whereas some of the tested classifiers (KNN, GMMs and SVMs; see Section 4) are designed to process fixed-size vectors (or fixed-size sequences of vectors reorganized as vectors). Therefore, it is necessary adress the problem of time normalization. In the present study, we consider two simple forms of time normalization. The first one consists of *averaging* the vectors in the time dimension over the entire acoustic event. Therefore, the feature vector sequence is replaced with a single mean feature vector (the standard deviation can also be used). The second form regards the *interpolation* of the feature vector sequence to the class' average duration, using basic (e.g. spline) interpolation/resampling techniques. Note that the GMM-T and HMMs classifiers are fed directly with the original feature vector sequence and do not need time normalization (HMMs are specifically designed to model dynamic sound representations). Note finally that the final representation may consist of the (row-wise) concatenation of different features. This is a particular case of information fusion for classification (see Section 4.3).

Implementation The MFCC and ASF features have been computed with the Python/C++ toolbox YAAFE [9]. The F_0 and harmonicity index analysis function was conducted using our own Matlab implementation.

4 Classifiers

4.1 Definition

A multiclass classifier consists of a mapping $g : \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}$, whereby $\mathcal{X}$ is the feature space, $\mathcal{C} = \{1, \ldots, C\}$ is the set of labels and C is the number of classes. The dimension of $\mathcal{X}$ may be

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

fixed or varying with the sound, depending on the feature used. Given a feature vector (or sequence of feature vectors) $\boldsymbol{x} \in \mathcal{X}$, $g(\boldsymbol{x}; c)$ is the score of classifying $\boldsymbol{x}$ as c. A new unlabeled observation $\boldsymbol{x} \in \mathcal{X}$ is classified as: $c^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{C}} g(\boldsymbol{x}; c)$. $\mathbf{X}$ will denote the training set, i.e. a set of feature vectors $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ whose class is known, used to train the classifiers.

4.2 Four Classifiers

In this section, we present the four types of classifiers that were used in the present study. As some features are commonly used in speech and audio processing and the Signal Processing / Machine Learning communities, we present them very briefly, with links to the related literature.

k-nearest neighbors (KNN): The KNN classifier first find the subset $S_k(\boldsymbol{x}) \subset \mathbf{X}$ containing the k closest points to a given vector $\boldsymbol{x}$. $g_{k\text{NN}}(\boldsymbol{x}, c)$ is then the number of feature vectors among $S_k(\boldsymbol{x})$ that belong to the class c.

Support Vector Machines (SVMs): SVMs are a discriminative binary classification method (see [5] for a detailed description), which has already been used in sound recognition, e.g. [6, 11]. SVMs provide a discriminative function $h(\boldsymbol{x})$, learnt form a set of positive examples and a set of negative examples. The points satisfying $h(\boldsymbol{x}) = 0$ form a hyperplane in the space induced by a chosen kernel function $k(\cdot, \cdot)$. $h(\boldsymbol{x}) > 0$ means that $\boldsymbol{x}$ should be classified as positive and $h(\boldsymbol{x}) < 0$ as negative. The multi-class task uses *one-versus-rest* strategy. Also, we tested four different kernels (linear, radial basis, polynomial and sigmoid).

Gaussian Mixture Models (GMM): A GMM is a probabilistic generative model widely used in classification tasks [5]. Here, we use one GMM per sound class, which is a weighted sum of M Gaussian components. The parameter set λ_c is composed of M weights, mean vectors and covariance matrices. We thus train C sets of parameters using the well-known Expectation-Maximization (EM) algorithm. The mapping g corresponds to the likelihood of the observed data given the model parameters. GMMs can be applied directly on the mean feature vector (in such case, we simply denote this configuration with GMMs). Alternately, for a sequence of feature vectors $\boldsymbol{x} = [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^T]$, which are assumed to be independent, we calculate: $g_{\text{GMM}}(\boldsymbol{x}; c) = \text{p}(\boldsymbol{x}|\lambda_c) = \prod_{t=1}^T \text{p}(\boldsymbol{x}^t|\lambda_c)$. We denote this configuration by GMMs-T.

Hidden Markov Models (HMM): HMMs also belong to the family of generative models [5, 10]. In an HMM, the observations depend on a hidden discrete random variable called state, taking S values. The state sequence is assumed to be a first-order "left-to-right" Markovian process and the emission probability is a GMM. Thus, the model consists of the parameters of the GMMs and the parameters modeling the Markovian dynamics. All are learnt using the EM algorithm. The function g is also the likelihood of the observations given the model: $g_{\text{HMM}}(\boldsymbol{x}; c) = \text{p}(\boldsymbol{x}|\xi_c)$.

Implementations: We used the standard Matlab KNN and GMMs algorithms. The HMMs are from the PMTK3 library [12]. The SVMs are implemented using libSVM [13].

4.3 Information Fusion

In Section 3, we have seen that several kinds of features can be extracted from the baboon vocalizationsl to describe their spectro-temporal characteristics in order to be used in a supervised classification scheme. This naturally raises the question of combining those features into a multi-modal/multichannel classifier that would optimally exploit all information in an efficient manner, a problem sometimes referred to as *sensor fusion*. This makes particular sense in the present study, since we postulated in Section 3 that, as opposed to ASR, the F_0 information is expected to provide significant information about sound class, it is therefore necessary to test if this information can be used in a complementary way to the spectral envelope (for instance MFCCs) information.

The usual, and simplest approach, known as *early integration*, consists in the (row-wise) concatenation of the different features (or feature vectors) into a single vector (in which dimension is equal to the sum of the dimensions of the original feature vectors), possibly integrating some cross-modal normalization processes. This new representation can then be used directly with the different classifiers presented above. In contrast, *late integration* performs the fusion of the features at the decision level of separate classifiers [14]. Thus, a different classifier (of same or different type) can be used on each feature vector and then the outputs (crisp decision, confidence score, log-likelihood values etc.) of these classifiers are merged using a higher level process. Finally, we can consider an

intermediary common space for fusion which is neither the input space nor the output space, leading to a type of *mid-level integration*. In particular, in the field of kernel-based classifiers (such as SVMs), a new state-of-the-art fusion strategy has emerged called Multiple Kernel Learning [15]. In this approach, the fusion is made "inside" the classifier: the kernel of the classifier is computed as a combination of multiple kernels, for instance, one kernel for each feature. One advantage is the ability to choose one type of kernel and its parameters according to the features. In Section 5.3, we will test this strategy for the integration of MFCCs and F_0 features in the present task of baboon vocalization classification.

4.4 The "Sound-spotting" Task

The above techniques are all applied on elementary sounds manually extracted from vocalization sequences. In practice, it is desirable to have a system that is able to automatically perform both detection (i.e. segmentation of a series of vocalizations into elementary sounds) and classification of the detected elementary sounds from the continuous audio stream. This task can be referred to as "Sound-spotting", in reference to the "Word-spotting" task in ASR which is the detection of keywords in continuous speech signals. A naive but efficient strategy consists of applying any of the previous classifiers (that have been tuned on a training corpus of elementary sounds) on a sliding window and decide of the detection if some criterion (e.g. a likelihood function), provided by the classifier, exceeds a given threshold. Temporal integration is necessary to make this joint detection/classification robust, and this can be done at the criterion level (e.g. by averaging frame-wise likelihoods) or at the feature level (e.g. by varying the sliding window length)[1]. In the present paper, we did not conduct a deeper investigation of the "Sound-spotting" problem, but in Section 5.3, we present some elements which illustrate the feasibility of this task using the proposed classifiers.

5 Experiments

5.1 Setup

Given the database described in Section 2, different combinations of features, post-processing and classifiers have been tested. We performed 5-cross validation tests, and used the accuracy score as a metric of the performance in order to be able to statistically compare the different configurations. For each experiment reported in the next section, only the best configuration of parameters (using grid search and cross validation) has been retained due to the large number of parameters involved. For the features, MFCCs reached its best results using 20 coefficients (with the first one omited), with the derivates at the first and second order on a 10Hz-10000Hz bandwidth. As for the classifiers, SVMs have shown the best results using linear kernels and radial basis kernels with a regulation parameter equal to 0.1 and the one-versus-rest strategy. HMM have been tested with 3 to 8 states and 5 to 10 components per state. Best results with GMM-based methods needed between 5 and 10 components in the mixture.

5.2 Results with Individual Feature Sets

We first present the results obtained separately with the different feature sets, i.e. either MFCCs or ASF or F_0+harmonicity index. The accuracy scores are given in Table 1 for a selected set of configurations, and confusion matrices are given in Table 2 and Table 3 for a subset of those configurations.

The best performance are obtained with SVMs (with a radial basis kernel) applied on averaged MFCC coefficients, with an accuracy score of 94.1%. This is a very good result, even for the limited number of classes of the present problem, since there is no a priori reason to think that a vast majority of the elementary sounds of the six classes are clearly prone to discrimination: This is actually a major outcome of the present study. The confusion matrix for this configuration (Table 2b is well balanced, with no major class confusion. Best results per class are obtained for Barks (97.3% accuracy) and worst result per class are obtained for Cops with 83.8% accuracy, and 13.8% of confusion with Grunts. It is important to note that SVMs are here applied on an averaged MFCC

[1]This is reminiscent of the "early" vs "late" integration problem discussed in Section 4.3, but considering here temporal fusion and not feature fusion.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

vector (to represent the whole sound). Hence, the time structure of the spectral vector sequence does not seem to be very important, leastways not as important as in speech (even if we compare with a short word recognition task). This is confirmed by the score of the SVMs applied on time-interpolated MFCC vectors, which is a bit lower than with averaged MFCC vectors at 90.5%. And this is more severely confirmed by the scores obtained with the HMMs applied on the original MFCC vector sequences (see Table 2a): the accuracy score is here only 80.8%, which is quite deceiving. The confusion matrix exhibits notable confusions from Cops to Barks and to Grunts, and from Grunts to Cops (but not from Barks to Cops), and also from Yaks to Screams, which is surprising. This not only suggests that there is relatively poor additional information in the vector *sequence* compared to the vector mean for the task at hand, but it also suggests that the HMMs are not an appropriate tool for the modeling of such type of sounds. The latter makes sense since it is not clear so far if there exists a phonological structure in the baboon vocalizations that could be efficiently exploited by the state-space modeling of HMMs[2]. Finally, GMMs (92.7% accuracy; Table 2d) and KNN (92.4% accuracy; Table 2c), both applied on averaged vectors, are a bit below SVMs, confirming that most of the discriminative information is contained in the average vector, and that good recognition scores can be obtained with relatively basic classifiers. KNN applied on interpolated MFCC vectors are at 93.1% accuracy[3], and we did not test GMMs on interpolated MFCCs to avoid the "curse of dimensionality" problem which is typical for this model.

The scores obtained with ASF features are very deceiving. Many different combinations of ASF features were tested (with the different classifiers), and the best accuracy score is 73.2% obtained with SVMs on average ASF vectors (hence we only report this configuration in Table 1). Moreover, when using concatenation of MFCCs and ASF features (i.e. basic "early" fusion at the feature level, see Section 4.3), the scores do not improve significantly compared to using only the MFCCs, they even decrease in some configurations (that is the case for SVMs, see Table 1). Therefore, the ASF do not complement the MFCC information, which was predictable (they provide information on the global shape of the spectrum with generally less resolution than MFCCs, provided that the cepstral model order is sufficiently large). Therefore, we did not further consider those ASF features.

Generally, the results obtained with F_0 alone or F_0 concatenated with the harmonicity index are remarkable, given that it is quite rudimentary information. Here, the best results are obtained with the SVMs applied on interpolated F_0 vectors, which reach 71.0%. GMM-T comes a very close second with 70.9% accuracy. Both exploit temporal information (from interpolated or original vector sequence), but the accuracy score of the SVMs applied on the average F_0 vector is also very close at 69.6% accuracy. However, the confusion matrices for the two latter two configurations differ significantly: the matrix for GMM-T (Table 3a) is more balanced, whereas the matrix for SVMs (Table 3b) shows that the Grunts and Yaks have better results, while the Wahoos are totally confused (mainly with Barks and Cops) which is surprising. This can be explained partly by the fact that Wahoos have some prosody which is reduced by the averaging process. Note that the SVMs scores are biased by the fact that the best classification is obtained for the two classes with the higher cardinals (Grunts and Yaks), and only 3 classes out of 6 can actually be regarded as "correctly" classified. In contrast, the more well-balanced GMM-T matrix exhibits 5 classes out of 6 being fairly well classified. GMMs (68.1% accuracy; confusion matrix in Table 3d) and KNN (65.4% accuracy; confusion matrix in Table 3c), both applied on average F_0 features, are a bit below the others classifiers using F_0 as a feature, but not much. KNN applied on interpolated F_0 vectors are at 69.8% accuracy. Therefore, here also, the different classifiers for "fixed-size" features in both average and interpolated configurations are quite close to each other. Altogether, those results show that basic information about harmonicity (say F_0 range + harmonicity confidence) is enough to provide honorable classification of 6-class baboon vocalizations. Note that HMMS are, again, deceiving, with only 45.3% of correct classification.

5.3 Results with Kernel-Based Fusion of MFCCs and F_0

As announced in Section 4.3, we report the results obtained with the *mid-level integration* of MFCC and F_0 features, using fusion of SVMs kernels. As an example, Table 4 shows the results of a Multi-

[2]However, the GMMT score is also deceiving (78.5% accuracy) hence possibly pointing a problem with the use of the original MFCC sequence, and so far we cannot clearly explain this result.

[3]Hence, KNN with interpolated MFCCs is a bit better than KNN with averaged MFCCs, whereas SVMs with interpolated MFCCs is a bit lower than SVMs with averaged MFCCs. Altogether, the scores with KNN, SVMs and GMMs applied on either averaged or interpolated MFCCs are quite close to each other.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

Features	Classifier	Representation	Accuracy
MFCCs	KNN	Averaging	$92.4\% \pm 2.9\%$
MFCCs	SVMs	Averaging	$94.1\% \pm 1.2\%$
MFCCs	GMMs	Averaging	$92.7\% \pm 1.8\%$
MFCCs	KNN	Interpolation	$93.1\% \pm 3.0\%$
MFCCs	SVMs	Interpolation	$90.5\% \pm 2.9\%$
MFCCs	GMMs-T	Sequencing	$78.5\% \pm 4.8\%$
MFCCs	HMMs	Sequencing	$80.8\% \pm 3.9\%$
ASF	SVMs	Averaging	$73.2\% \pm 2.3\%$
MFCCs & ASF	SVMs	Averaging	$92.4\% \pm 2.7\%$
F_0	KNN	Averaging	$65.4\% \pm 6.9\%$
F_0	SVMs	Averaging	$69.6\% \pm 2.7\%$
F_0	GMMs	Averaging	$68.1\% \pm 7.4\%$
F_0	KNN	Interpolation	$69.8\% \pm 4.6\%$
F_0	SVMs	Interpolation	$71.0\% \pm 2.3\%$
F_0	GMMs-T	Sequencing	$70.9\% \pm 4.2\%$
F_0	HMMs	Sequencing	$45.3\% \pm 7.3\%$

Table 1: Accuracy score for different combinations of audio features, post-processing, and classifiers. "Sequencing" refers to using the original sequence of vectors.

	barks	cops	grunts	screams	wahoos	yaks
barks	**102**	0	0	1	7	0
cops	13	**89**	12	7	7	2
grunts	3	42	**312**	12	11	4
screams	1	0	0	**114**	0	4
wahoos	9	0	0	0	**55**	0
yaks	7	9	8	48	12	**252**

(a) Hidden Markov Models (HMMs)

	barks	cops	grunts	screams	wahoos	yaks
barks	**107**	0	1	0	1	1
cops	0	**109**	18	1	0	2
grunts	1	6	**369**	0	1	7
screams	0	0	0	**114**	1	4
wahoos	6	0	0	0	**58**	0
yaks	0	5	11	1	0	**319**

(b) Support Vector Machines (SVMs)

	barks	cops	grunts	screams	wahoos	yaks
barks	**106**	0	0	0	2	2
cops	3	**104**	17	1	3	2
grunts	1	9	**367**	0	0	7
screams	0	0	0	**109**	0	10
wahoos	6	0	0	0	**58**	0
yaks	0	1	2	20	1	**312**

(c) k-Nearest Neighbors (KNN)

	barks	cops	grunts	screams	wahoos	yaks
barks	**105**	0	0	3	0	2
cops	0	**115**	8	1	0	6
grunts	0	19	**350**	2	0	13
screams	0	0	1	**112**	0	6
wahoos	5	0	0	1	**57**	1
yaks	1	2	9	3	0	**321**

(d) Gaussian Mixture Models (GMMs)

Table 2: Confusion matrix for the baboon vocalization recognition systems using average Mel-frequency cepstral coefficients (MFCCs) as features for SVMs, GMMs and KNN, and using original sequence of MFCCs for HMMs.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

	barks	cops	grunts	screams	wahoos	yaks
barks	**85**	0	1	0	18	6
cops	17	**29**	44	1	36	3
grunts	9	26	**326**	0	17	6
screams	1	0	0	**90**	1	27
wahoos	14	7	0	0	**43**	0
yaks	37	7	7	32	16	**237**

(a) Gaussian Mixture Models on a sequence of vectors (GMM-T)

	barks	cops	grunts	screams	wahoos	yaks
barks	**83**	12	1	0	0	14
cops	32	**19**	72	0	0	7
grunts	13	1	**363**	0	0	7
screams	2	0	0	**67**	0	50
wahoos	29	24	8	0	**0**	3
yaks	43	3	9	18	0	**263**

(b) Support Vector Machines (SVMs)

	barks	cops	grunts	screams	wahoos	yaks
barks	**65**	13	3	0	7	22
cops	30	**30**	53	0	10	7
grunts	9	36	**328**	0	3	8
screams	1	0	1	**69**	0	48
wahoos	28	9	4	0	**14**	9
yaks	34	11	12	30	7	**242**

(c) k-Nearest Neighbors (KNN)

	barks	cops	grunts	screams	wahoos	yaks
barks	**74**	0	1	0	29	6
cops	21	**18**	50	0	35	6
grunts	6	20	**338**	0	12	8
screams	2	0	0	**94**	1	22
wahoos	9	3	3	0	**48**	1
yaks	45	1	8	56	20	**206**

(d) Gaussian Mixture Models (GMMs)

Table 3: Confusion matrix for the baboon vocalization recognition systems using average F_0 (fundamental frequency) as feature for SVMs, GMMs and KNN, and using original sequence of F_0 for GMM-T.

ple Kernel Learning experiment, in which a linear kernel has been trained on MFCC features, while another linear kernel has been trained on F_0 features, and the combination of those kernels has been computed and used in a third SVM. It can be seen that this configuration does not outperform the SVMs which uses only MFCCs as features: the accuracy scores are $88.1\% \pm 2.9\%$ for the former vs $91.2\% \pm 3.3\%$ for the latter[4]. None of the other tested configurations of kernels and hyper parameters have shown a significant improvement. One conclusion of this experiment is that, although the F_0 (and harmonicity index) feature separately carries a significant information which is exploitable for the automatic recognition of baboon vocalization, this feature was not shown in our experiments to be complementary to the MFCC features for this task. On the contrary, the combination of F_0 and MFCCs only lead so far to slightly decrease the scores obtained with MFCCs alone, which is a bit deceiving. Of course, this is also because MFCC representation initially led to impressive scores. Further investigation of the characterization of those features for the baboons vocalizations is necessary to precisely describe the redundancy between them and confirm the seeming absence of complementarity which has been observed in our experiments.

5.4 Feasibility of Sound-Spotting

In this subsection, we illustrate the feasibility of the Sound-spotting task described in Section 4.4 by applying the SVMs of Section 4.2 on an example of original (i.e. unsegmented) sequence. The SVMs were fed with MFCC vectors on a frame-by-frame basis (i.e. average of one vector at a time, corresponding to a 200ms-frame of signal, with 10ms-hop size). For each frame and class c, we retrieved $p(c|\boldsymbol{x})$ the posterior probability of the frame being part of a vocalization of class c given the input MFCC vector $\boldsymbol{x}$, which is the criterion used by the SVMs for classification [16]. Fig. 1 shows the results of this analysis. The top subfigure shows an excerpt of a vocalization waveform with the corresponding class boundaries and labels which were manually annotated. The three other subfigures plot the values of $p(c|\boldsymbol{x})$ for the Barks, Grunts, Screams and Yaks, respectively (from top to bottom; probabilities for Cops and Wahoos are not displayed for clarity). It is evident that the probability contours quite well with the actual classes, i.e. globally, the probability values are high when the corresponding class is emitted, and low when another class or background noise is

[4]This latter score is different (a bit lower) than the SVMs/MFCCs score of Table 1 because a radial basis kernel was used in the SVMs of Section 5.2.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

	barks			cops			grunts			screams			wahous			yaks		
barks	**60**	**107**	**60**	13	0	2	6	1	0	0	0	0	0	0	2	31	2	2
cops	26	1	7	**13**	**92**	**82**	81	31	36	0	1	1	0	0	1	10	5	3
grunts	11	1	2	3	7	4	**363**	**369**	**371**	0	0	1	0	1	0	7	6	6
screams	1	1	0	1	0	1	0	0	0	**94**	**109**	**96**	0	0	0	54	9	22
wahous	23	9	7	20	0	0	17	0	0	0	0	0	**0**	**55**	**55**	4	0	2
yaks	32	1	0	4	3	5	10	14	14	18	8	14	0	0	4	**272**	**310**	**299**

Table 4: Confusion matrix for one instance of Multiple-Kernel SVMs combining MFCC and F_0 features. For each cell, the three numbers from the left to the right corresponds to the result of classification for: (1) SVMs with a linear kernel on F_0, (2) SVMs with a linear kernel on MFCCs, (3) SVMs with a combination of the two precedent kernels.

emitted. For this example, a very simple detection strategy based on thresholding can be applied: Class c is detected as $p(c|\boldsymbol{x}) > 0.5$ (the probabilities for the different classes sum up to 1, hence only one class at a time can be detected). Merging the successive frames associated with the same class leads to the detected boundaries represented in the top subfigure of Fig. 4.4 with background color corresponding to the probability contours. The detection is fairly good but not perfect: for example, background noise is confused with Grunts at approx. 6s, and the boundaries between Yaks and Screams are not easy to define (nor is it easy for the human listener in this example, and manual labeling may actually be inacurate). Moreover, many sequences are not so clear. However, more refined strategies for time integration of frame-wise information, such as the ones mentioned in Section 4.4, are expected to fix these problems and be more robust in general. Part of our future work is to explore such strategies and derive an efficient and robust Sound-spotting algorithm in the present problem of baboon vocalization recognition.

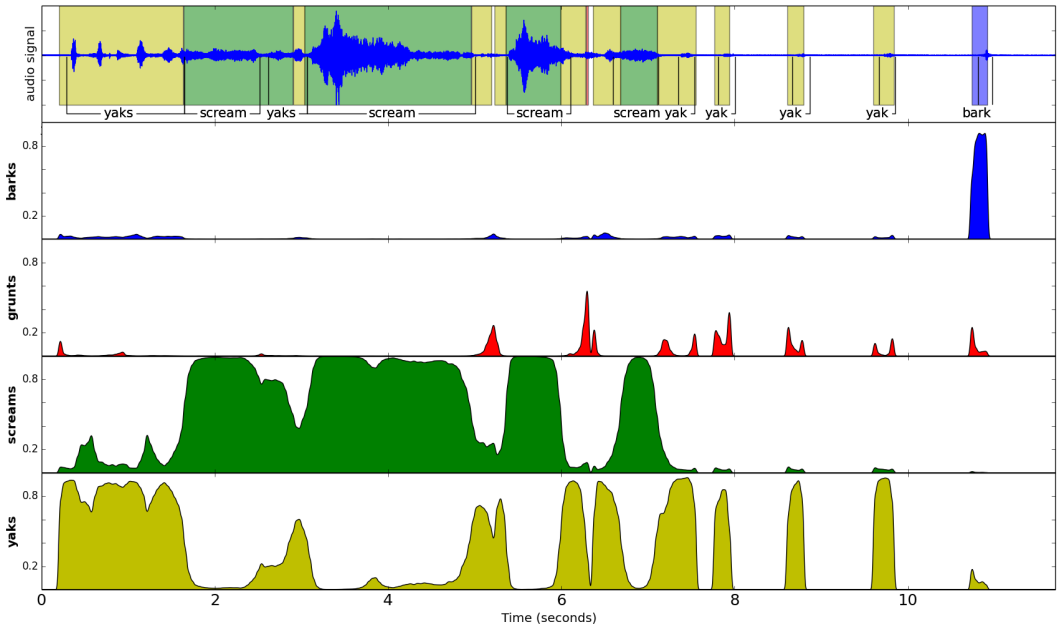


Figure 1: Example of automatic joint segmentation and classification using the SVMs of Section 4.2 (see text for details).

6 Conclusion

In this paper we have adressed the problem of automatic classification of Guinea baboon vocalizations. Six classes of sounds have been considered, and experiments have shown that several types of classifier (KNN, GMM, SVM) lead to correct classification scores higher than 90% for pre-segmented elementary vocalizations. The higher scores were obtained with SVMs applied on

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

average MFCC vectors (94.1% accuracy), and the principal remaining confusions were observed to be between grunts and copulations grunts. It is not entirely surprising that the classifiers have difficulty in distinguishing these two vocalizations; of all the sound classes, the call units of these two are the most similar from both an auditory perception and acoustic structure standpoint. This study has also shown that the fundamental frequency F_0 (alone or coupled with harmonicity index) has a significant discriminative power: several classifiers applied on these features provided approximately 70% correct classification. Indeed, analysis of the baboon vocal repertoire shows that the baboons strongly modulate their F0 between vocalizations, particularly between short- and long-distance vocal categories (Kemp et al., in prep.). However, and quite deceivingly, this information was not found to be complementary to the spectral envelope information in our study. Finally, although we did not conduct a deep investigation of the Sound-spotting problem in the present study, the observation of the good behavior of classifiers, designed on elementary sounds when applied on continuous audio streams, shows that joint segmentation and recognition is expected to be feasible with a well-grounded time integration process. This time integration can be processed at the feature level, at the classifier output level, or at some "mid-level" within the classifier, echoing the discussion of Section 4.3 on feature information fusion. Future work will concern this task, which is essential to design a real-world system. We will also consider increasing the number of classes and defining confidence measures to help the exploitation of the classification results in primatology studies.

Acknowledgments: Yannick Becker and the staff of the Rousset-sur-Arc primate center are acknowledged for technical support.

References

[1] K. Hammerschmidt and J. Fischer, "Constraints in primate vocal production," *The evolution of communicative creativity: From fixed signals to contextual flexibility*, pp. 93–119, 2008.

[2] A. Mielke and K. Zuberbühler, "A method for automated individual, species and call type recognition in free-ranging animals," *Animal Behaviour*, vol. 86, no. 2, pp. 475–482, 2013.

[3] P. Maciej, J. Fischer, and K. Hammerschmidt, "Transmission characteristics of primate vocalizations: implications for acoustic analyses," *PloS one*, vol. 6, no. 8, p. e23015, 2011.

[4] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 1060–1089, 2013.

[5] C. M. Bishop, *Pattern recognition and Machine learning*. Springer New York, 2006.

[6] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 209–215, 2003.

[7] J. Fagot and E. Bonté, "Automated testing of cognitive performance in monkeys: Use of a battery of computerized test systems by a troop of semi-free-ranging baboons (papio papio)," *Behavior Research Methods*, vol. 42, no. 2, pp. 507–516, 2010.

[8] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2004.

[9] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Int. Conf. for Music Information Retrieval (ISMIR)*, 2010.

[10] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition*, vol. 39, no. 4, pp. 682–694, 2006.

[12] K. P. Murphy, *Machine learning: A probabilistic perspective*. MIT Press Boston, 2012.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[14] L. I. Kuncheva, J. C. Bezdek, and R. P. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[15] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, pp. 2211–2268, 2011.

[16] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

6.3 UCR-USV: Software Tools for Analyzing Mice Vocalizations with Applications to Pre-Clinical Models of Human Disease

Mohammad Shokoohi-Yekta
Computer Science
UC Riverside
mshok002@ucr.edu

Jesin Zakaria
Computer Science
UC Riverside
jzaka001@ucr.edu

Sarah Rotschafer
Psychology
UC Riverside
srots001@ucr.edu

Hamid Mirebrahim
Computer Science
UC Riverside
smire002@ucr.edu

Khaleel Razak
Psychology
UC Riverside
khaleel.abdulrazak@ucr.edu

Eamonn Keogh
Computer Science
UC Riverside
eamonn@ucr.edu

Abstract

Identifying structure in mice ultrasonic vocalizations (USV) is a useful tool for investigating the role of genetics in human disorders by modifying ("knocking out") various genes in mice and examining their vocalizations for changes that may be linked to those genes, and hence the analogue genes in humans [1][2]. Thus far, it appears that all annotation and feature extraction from USV has been done manually. We believe that the lack of computational tools has been a major bottleneck in USV research. To address this problem we have previously developed an intuitive software suite that can analyze acoustic properties of USV and characterize the relationships between behavioral segments and calls [2]. Here we present a novel analytical tool that goes beyond quantifying basic acoustic properties of USVs, by characterizing the relationship between the USV syllables used during specific components of social behavior.

1 Introduction

Identifying structure in mice ultrasonic vocalizations (USV) is a useful tool for investigating the role of genetics in human disorders by modifying ("knocking out") various genes in mice and examining their vocalizations for changes that may be linked to those genes, and hence the analogue genes in humans [1][2]. In recent years this framework has emerged as an extremely promising tool for understanding human cognitive and memory disorders. Analyzing vocal behaviors of mice models in this manner has led to the discovery of the genetic cause of Autism [3], and has shown great promise for the study of Alzheimer's disease [4].

The UCR-USV tool has been implemented in MATLAB®, making it easy to extend, and essentially free for academics. The system performs five main functions: Syllable Extraction and Idealization, Analysis of Basic Acoustic Properties of USV, Syllable Classification [5][6][7][8], Visual Representation of Call Rates Annotated by Mice Behaviors and Measuring the Density of Syllables Obtained during Each Behavior Segment.

In this study we hint at the actionability of audio motif discovery by showing that motifs, once discovered, can be used to test for changes in vocal repertoire that may be attributable to genes that were deliberately deleted from the mouse genome.

We obtained six hours of vocalizations recorded during courtship/mating of various pairs of mice (only males vocalize). These sessions were annotated by the mice behaviors, from the set: {`Defensive` (D), `Ejaculate` (E), `Grooming` (G), `Intromission` (I), `Mounting` (M), `No Contact` (N), `Rooting` (R) and `Sniffing` (S)}. A basic question is does the vocal repertoire or frequency during these behaviors differ for different mice genomes. In this study we hint at the answer to this question; a fuller exposition can be found at [9].

Manipulation of particular genes has already shed light on the genetic basis of human communication disorders [15][20][21][22][23][24]. The function of ultrasonic vocalizations in adult mice is likely only to facilitate or inhibit social interaction. Understanding the types and functions of ultrasonic vocalizations emitted by laboratory rodents may enable researchers and animal car personnel to use vocalizations as an indicator of an animal's behavior and affect [25]. (Portfors, 2007) describes three types of calls emitted by rats and classifies them based on different ranges of frequencies. (Grimsley et al., 2013) [26] have applied clustering analysis on syllables emitted by a type of mouse pups and found four clusters of syllables again based on their frequency bands. Finally they introduce an Excel-Based calculator which classifies syllables by using frequency ranges and they call it an automated classification tool.

1.1 Notations

A sound spectrogram is an image of time-varying spectral representation, produced by applying the Short Fast Fourier Transform to successive overlapping frames of an audio sequence. The horizontal dimension corresponds to time and the vertical dimension corresponds to frequency. The relative spectral intensity of a sound at any specific time and frequency is indicated by the color/grayscale intensity of the image.

A 'syllable' is a unit of sound composed of one or more 'notes'. If the interval between two notes is <10 msec (user definable), they are combined into a single syllable. A 'syllable type/class' is a category of syllable, observed regularly in the animal's vocalization, distinct from other syllable types [14]. The ground truth (G) dataset is a set of annotated syllables that were manually classified by two of the authors (SR and KR). Each class in the ground truth may be represented by one or multiple exemplars. This is to allow our classification model to capture the natural variability of a class. By analogy, a handwriting recognition system must have at least two exemplars of "seven", one to match the common American style "7", and one to match the exemplars drawn with a line in the middle "7̸", the latter style being common in Europe and Latin America.

1.2 UCR-USV Tool

The tool is designed to be user-friendly. The tool performs five functions: Syllable Extraction and Idealization, Basic Analysis, Syllable Classification, Visual Representation of Call Rates Annotated by Mice Behaviors and Measuring the Density of Syllables Obtained during Each Behavior Segment.

Syllable Extraction and Idealization: Converts an audio file of USVs to a spectrogram representation. Discrete syllables from the spectrogram are then extracted and idealized.

Basic Analysis: Records the duration and range of frequencies in each syllable, and determines the gaps between the syllables which can be utilized to quantify the rate of USV calls. This step also generates separate files which include the durations, frequencies and gaps for all syllables for further analysis.

Syllable Classification: Using the GHT (Generalized Hough Transform) distance measure [5][6][7][8], all syllables are classified in separate folders. A special 'other' class is possible for syllables that our system could not confidently classify. These syllables can later be classified by a human expert, or simply discarded as they are very rarely false dismissals, but almost always simply noise/artifacts.

Visual Representation of Call Rates Annotated by Mice Behaviors: This step represents call rates for each syllable in the dictionary and maps mice behaviors to call rates.

Measuring the Density of Syllables during Each Behavior: Normalizes the number of syllables of each class obtained during a behavior segment by total number of calls and time spent in the behavior segment and characterizes the relationships between behavioral segments and calls.

2 Methodology

We discuss the modules performed by the UCR-USV tool in greater detail below:

2.1 Syllable Extraction

We use the algorithm in Table 1 to extract all the candidate syllables from the spectrogram of a mouse vocalization. The algorithm is briefly described below and additional details can be found in [2].

Table 1: Extract candidate syllables

Algorithm 1 *ExtractCandidateSyllables*(*SP*)

Require: spectrogram of a mouse vocalization
Ensure: set of candidate syllables

```
1:  I ← idealized spectrogram
2:  L ← set of connected components in I
3:  R ← row index of connected points
4:  C ← column index of connected points
5:  V← value of connected points // value ranges from 1 to |L|
6:  [A B] ← sort(V, 'ascend') // A has values of V sorted and B has the
7:  index
8:  S ← []        // set of candidate syllables in SP, initially empty
9:  c1 ← dmin, c2 ← dmax // min and max duration of a syllable
10: j ← 1, k ← 1
11: for i ← 1 to |L| do {every connected component li in L}
12:    n ← 1
13:    while A(k)=i do
14:       RWli(n) ← R(B(k)) // RWli contains row indices of li
15:       CLli (n) ← C(B(k)) // CLli contains column indices of li
16:       n ← n + 1
17:       k ← k + 1
       m←L(min(RWli):max(RWli), min(CLli):max(CLli))==i
18:                    //minimum bounding rectangle (MBR) of li
19:    [r c] ← size of m
20:    if |c| < c1 or |c| > c2
21:       continue                    // filter out noise
22:    else
23:       Sj ← m
24:       add Sj to S
25:       T1j ← min(CLli)             // start time of Sj
26:       T2j ← max(CLli)             // end time of Sj
27:       j ← j + 1
   return S, T1, T2  // candidate syllables in SP with start/end times
```

Instead of extracting candidate syllables from the original spectrogram (SP) we use an idealized version (I) of SP, as it produces fewer false negatives to be checked. In line 2, we convert the matrix I into a set of connected components, L. L has the same size as I, but it has the connected pixels marked with number 1 to |L|. The set of candidate syllables in SP is initialized with an empty set in line 7.

A syllable is a contiguous set of pixels in a spectrogram; we can thus consider it as a set of connected points in I. The `for loop` in lines 10-26 is used to search for a connected component li in I. In order to make the search time linear to the number of candidate syllables, in lines 3-5 while creating L (a set of connected components), the row and column indices and the values of all the connected points in arrays R, C and V, respectively are

saved. In line 6, the array V is sorted in ascending order and indices in B are saved. In the `while loop` in lines 12-16, the indices in B are used to find the row and column indices of a connected component li in I. The minimum and maximum values of the row and column indices are used to extract the minimum bounding rectangle (MBR) of li.

It is important to note that not all of the connected components are candidate syllables. The idealized spectrogram can still contain non-mouse vocalization sounds. In the `if block` of lines 19-20, the duration of a connected component li is checked and those li in S which are within the range of thresholds c1 and c2 are included. Since the minimum and maximum duration of syllables can vary across different mice, the values of c1 and c2 should be set after manual inspection of a fraction of the data. In our experiments, the values are set to 10 and 300, respectively. However the exact settings of these parameters are not critical to subsequent steps. In lines 24-25, the start time and end time of a syllable are saved and used for subsequent analysis. Figure 1 visually demonstrates the method. Our algorithm runs faster than real time, and thus does not warrant further optimizations for speed.

In Figure 1, a snippet spectrogram SP, matrices corresponding to the idealized version of the spectrogram I and connected components L are presented. For brevity in explanation, original matrices for I and L are resized to 10x10. Finally, the MBRs of the candidate syllables in the snippet spectrogram are marked.

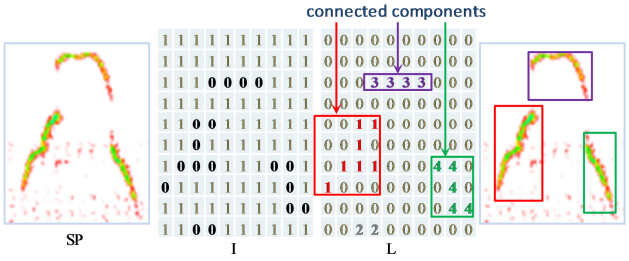


Figure 1: (*from left to right*) A snippet of a spectrogram, the resized matrix corresponding to an idealized spectrogram *I*, the resized matrix corresponding to the set of connected components *L*, and the MBRs of the candidate syllables

2.2 Basic Analysis

The tool measures basic acoustic properties of syllables such as duration, dynamic range of frequencies and gaps between syllables (Figure 2). Figure 2 shows a syllable made up of three notes (left), and one that consists of a single note (right). The maximum possible gap between single notes is a user-defined threshold (in this study we used 10 msec, which is also set as the default). Any notes which are closer than the maximum gap, are combined as one syllable. The tool reports the minimum, maximum and mean durations and produces an output file including all durations and start and end times for each syllable. Corresponding frequency dynamic range and gaps between syllables are included in separate output files. Other information such as the maximum and average gap, and the total number of syllables produced in the recording are also reported.

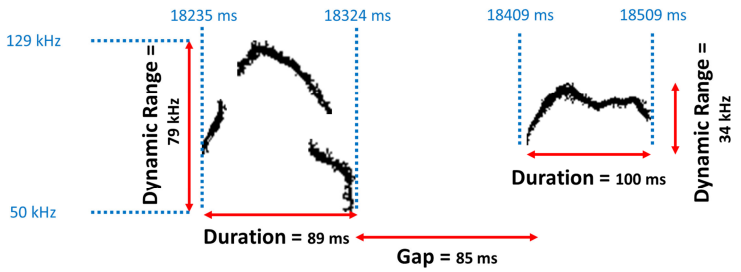


Figure 2: measurable features for syllables

2.3 Classification

Before the classification begins, the tool scans all the notes and combines the notes with a gap less than the user-defined threshold into single syllables (for example, Figure 2, left). The algorithm in Table 1 will generate a set of candidate syllables that are not classified in this step. In order to classify them, a set of annotated syllables termed Ground Truth (G), and a set of thresholds for each class of syllables are used. The candidate syllable cannot simply be assigned to the class of its nearest neighbor because a large fraction of the candidate syllables will inevitably be noise, and it is the thresholds that allow us to reject them.

A ground truth (G) dataset is a set of annotated syllables that have been classified by humans (authors SR and KR). Each class in the ground truth may be represented by one or multiple exemplars. The data set includes a small set of robust exemplars for our seven classes. The Ground Truth table consists of seven syllable classes (Figure 3) against which each candidate syllable should get compared. Our ground truth table shows consistency with other studies. For example [15] and [16] have introduced ten categories of calls which include almost all of the calls mentioned in Figure 3, except for the class of multiple notes (class 7 in this study), we have considered a single class while [15] and [16] have more than one class for multiple notes. The difference will not hurt the accuracy of our results, as most of the results concluded in this paper are based on classes with single notes. The only conclusion conducted for Class 7 could simply generalize to combining the categories of multiple notes in[15] and [16].

Furthermore a set of thresholds, one for each class, is required for the purpose of classification. Thresholds are created by simply computing the GHT distances between every annotated syllable to its nearest neighbor from the same class. Then the mean plus two standard deviations is chosen as the threshold distance for that class.

Given a set of candidate syllables S and ground truth syllables (G) with their matching thresholds (τ), the algorithm shown in Table 1 classifies syllables in S, and rejects all others as `unclassifiable`. A special 'other' folder is created for syllables that our system could not confidently classify. These syllables can later be classified by a human expert, or simply discarded as they are very rarely false dismissals, but almost always simply noise/artifacts.

Table 2: Syllable classification algorithm

Algorithm 2 *ClassifyCandidateSyllables*(S, G, T)

Require: candidate syllables, ground truth, set of thresholds
Ensure: set of labeled syllables

```
1:  // S = {S_1, S_2, … S_n} is set of candidate syllables,
2:  // G = {G_1, G_2, … G_m} is ground truth and
3:  // τ = { τ_1, τ_2, … τ_11} is set of thresholds
4:  // normalize all the syllables in S and G to equal size
5:  // initialize all syllables' class {c_S1, c_S2, …} to 0 or not classified
6:  for i ← 1 to n do        // |S| = n
7:      NNdist = inf         // initially set the NN distance to infinity
8:      for j ← 1 to m do    // |G| = m
9:          dist ← dist_GHT(S_i, G_j)  //calculate GHT between S_i and G_j
10:         if dist < NNdist
11:             NNdist ← dist   // update nearest neighbor distance
12:             NN ← j          // update nearest neighbor (NN)
13:     if NNdist ≤ τ(C_NN)     // C_NN is the class label of G_NN
14:         c_Si ← C_NN
15: return {c_s1, c_s2, … c_sn} // class labels of all candidate syllables
```

In order to classify a candidate syllable we look for its nearest neighbor in G in the `for loop` of lines 8-12. In the `if block` of lines 13-14, the class label of the nearest neighbor to a candidate syllable is assigned only if the distance between a candidate syllable and its nearest neighbor from G is less than the threshold of the nearest neighbor's class. The GHT distance measure for classifying syllables was used in this study. Although GHT is a

common and popular distance measure for this type of classification, other distance measures have also been used in this area. Hammerschmidt et al. use the log-likelihood distance measure and Schwarzsches Bayes criteria (BIC) for clustering mice calls [17].

Figure 3: a single instance for each class of syllables

This opens the question of why GHT is an appropriate distance measure to consider if a set of pixels is "sufficiently similar". GHT is fast, robust to the inevitable noise left even after idealization, and at least somewhat invariant to the significant intra-class variability observed. After careful consideration and provisional tests of dozens of possibilities, we converged on a distance measure based on the Generalized Hough Transform [5].

The Hough Transform [6] was introduced as a tool for finding well-defined geometric shapes (lines, curves, rectangles, etc.) in images [7]. Ballard et al. generalized the idea and introduced the Generalized Hough Transform to detect arbitrary shapes in images [5]. The computation time of Ballard's method is relatively expensive. It takes quadratic time, $O(n_b^2)$, to calculate the distance between a pair of windows. Here, nb is the number of black pixels in the window. However, Zhu et al. [8] augmented GHT in a way that reduces the amortized time for a single comparison significantly. Zhu et al. achieve speed-up by creating a computationally cheap tight lower bound to the GHT. Moreover, they present modifications to the classic definition that allow the measure to be symmetric and obey the triangular inequality, two properties that are highly desirable because they allow various algorithms to be used that exploit (or at least expect) these properties. We refer the interested reader to [8] for more details on GHT.

3 Experiments

We obtained six hours of vocalizations recorded during courtship/mating of various pairs of mice (only males vocalize). These sessions were annotated by the mice behaviors, from the set: {`Defensive` (D), `Ejaculate` (E), `Grooming` (G), `Intromission` (I), `Mounting` (M), `No Contact` (N), `Rooting` (R) and `Sniffing` (S)}.

We applied our motif discovery algorithm [19] to the data and found many instances of motif shown in Figure1. top. Having discovered this motif, we used a sliding window to calculate its density over time. As shown in Figure 4.middle, this particular motif occurs about 4.1 times more frequently during Sniffing than during Rooting for this particular strain of KO mice. Moreover, because we are able to automate this process (most similar research efforts resort to manual counting [4][28]) we can automatically search through a large space of motifs × behaviors × genomes, scoring the frequency differences by significant tests.

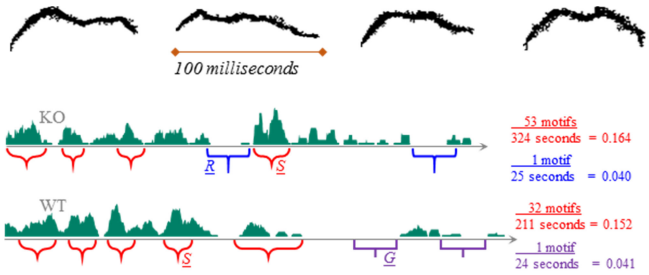


Figure 4. top) Sample instances of a motif discovered from mice vocalizations by applying our algorithm (middle) Comparing the number of motifs during S and R behaviors for a sample recording of KO mice

vocalization. bottom) Comparing the number of motifs during S and G behaviors for a sample recording of WT mice vocalization.

In Figure 4.bottom, we show another example of a similarly significant contrasting pattern, this time in WT (wild type) mice. In this case we noted a dearth of the motif during Grooming.

Moreover, because we are able to automate this process (most similar research efforts resort to manual counting [4][28]) we can automatically search through a large space of motifs × behaviors × genomes, scoring the frequency differences by significant tests.

The process of finding one or more syllable classes which distinguish WT versus KO or mice behaviors (sniffing, grooming, no contact, rooting) leds to a two group classification problem. Therefore, we can use the Fisher's Linear Discriminant metric to find the most discriminative syllable classes to distinguish WT and KO mice [27]. Fisher's Linear Discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. We calculate the class density during each behavior as a score, Figure 4, in order to calculate Fisher's Linear Discriminant values. We refer the interested reader to [9] for more detailed results.

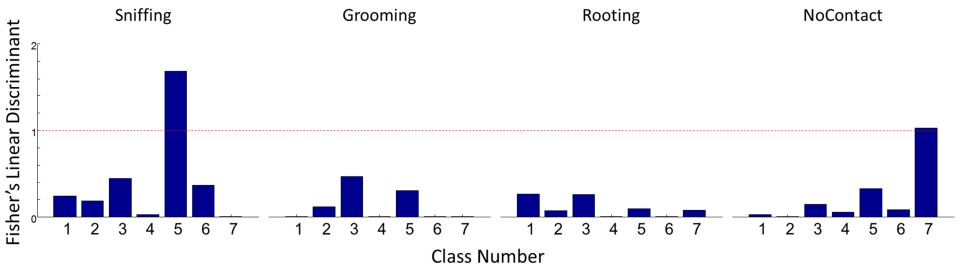


Figure 5: compares the Fisher's Linear Discriminant score for every class in each behavior

Values higher than the threshold line may happen to be significant classifiers. Figure 5 suggests Class 5 during sniffing and Class 7 during NoContact could potentially be significant discriminators for mice types, KO and WT.

4 Discussion

In contrast to many other studies, we have designed a classification algorithm for classifying syllables by considering their shape regardless of their frequencies, mice type or other basic features. Our UCR-USV tool is capable of automatically extracting syllables from mice vocalizations, idealizing the calls and classifying them to separate classes in almost real time. The tool analyzes mice vocalizations by reporting the frequencies, durations, dynamic ranges, call rates and finally characterizes correlations between the USV syllables used during specific components of social behavior.

The algorithm for classifying mice vocalization syllables described in Table 2 classifies about 90 percent of the syllables by applying the following techniques: 1- Assigning multiple instances to each group in the Ground Truth Table. 2- Idealizing the spectrogram and removing noise. 3- Using a dynamic user-defined threshold for idealizing the spectrogram.

USVs are typically analyzed in isolation from the social behaviors during which they are elicited. To the best of our knowledge this is the first time to analyze mice calls based on their social behaviors. We found out that syllables emitted by the WT mice during their sniffing behavior, overrepresented call types of class 5 (Figure 3) comparing to the KO mice. While the mice did not have any contact, KO mice produced denser calls of combined notes

(class 7 in Figure 3). We have also compared the density of calls between every pair of behaviors in a specific type of mice and the results for KO and WT mice have been shown in [9]. Higher values for Fisher's Linear Discriminant scores show a more significant discriminator among the behaviors.

Many studies have considered the rate of USV calling, [17] compares the call rates among male and female mice, they claim that during courtship in response to female intruders, females called more than males , and males called more to female than to male intruders. A comprehensive analysis has been done by Roy et al. in [29] which they have compared isolation-induced USVs generated by pups of Fmr1-KO mice with those of their wild type (WT) littermates. They claim that the total number of calls was not significantly different between genotypes, a detailed analysis of 10 different categories of calls revealed that loss of Fmr1 expression in mice causes limited and call-type specific deficits in ultrasonic vocalization: the carrier frequency of flat calls was higher, the percentage of downward calls was lower and that the frequency range of complex calls was wider in Fmr1-KO mice compared to their WT littermates.

References

[1] M. L. Scattoni, S. U. Gandhy, L. Ricceri, J. N. Crawley. *Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism*. *PLoS ONE* 3: e3067, 2008.

[2] J. Zakaria, S. Rotschafer, A. Mueen, K. Razak, E. Keogh. *Mining Massive Archives of Mice Sounds with Symbolized Representations*. *SIAM SDM*, 2012. pp 588-599.

[3] R. J. Hagerman, et.al. *Advances in the Treatment of Fragile X Syndrome*. *Pediatrics* Vol. 123 No.1, January, 2009.

[4] C. Menuet, Y. Cazals, C. Gestreau, P. Borghgraef, L. Gielis, et al. (2011) *Age-Related Impairment of Ultrasonic Vocalization in Tau.P301L Mice: Possible Implication for Progressive Language Disorders*. *PloS ONE* Jan; 6(10).

[5] D. H. Ballard, *Generalizing the Hough transform to detect arbitrary shapes*, Patt. Recognition, 13(2): 111-22 (1981).

[6] P. V. C. Hough, *Method and means for recognizing complex patterns*, U.S. Patent 3069654, (1962).

[7] R. O. Duda, P. E. Hart, *Use of the Hough transform to detect lines and curves in pictures*, Comm. ACM 15: 11–15, (1972).

[8] Q. Zhu, X. Wang, E. Keogh, S.H. Lee, *Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs*, KDD 2009, pp. 1057–1066 (2009).

[9] M. Shokoohi-Yekta, J. Zakaria, S. Rotschafer, S. H. Mirebrahim, K. Razak and E. Keogh. *Analysis of Concomitant Behaviors and Vocalizations Reveal Social Communication Deficits in a Mouse Model of Fragile X Syndrome*. *In press The Journal of Neuroscience*, 2014.

[10] D. H. Ballard, *Generalizing the Hough transform to detect arbitrary shapes*, Patt. Recognition, 13(2): 111-22 (1981).

[11] R. O. Duda, P. E. Hart, *Use of the Hough transform to detect lines and curves in pictures*, Comm. ACM 15: 11–15, (1972).

[12] P. V. C. Hough, *Method and means for recognizing complex patterns*, U.S. Patent 3069654, (1962).

[13] Q. Zhu, X. Wang, E. Keogh, S.H. Lee, *Augmenting the Generalized Hough Transform to Enable the Mining of Petroglyphs*, KDD 2009, pp. 1057–1066 (2009).

[14] T. E. Holy, Z. Guo. *Ultrasonic Songs of Male Mice*. PLoS Biol 3(12): e386. doi:10.1371/journal.pbio.0030386, 2005.

[15] Scattoni ML, Gandhy SU, Ricceri L, Crawley JN (2008) *Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism*. PLoS ONE 3(8): e3067. doi:10.1371/journal.pone.0003067

[16] E. J. Mahrt, D. J. Perkel, L. Tong, E. W. Rubel, C. V. Portfors. Engineered Deafness Reveals That Mouse Courtship Vocalizations Do Not Require Auditory Experience, The Journal of Neuroscience, 33(13):5573–5583, (2013).

[17] K. Hammerschmidt, K. Radyushkin, H. Ehrenreich, J. Fischer, *The Structure and Usage of Female and Male Mouse Ultrasonic Vocalizations Reveal only Minor Differences*. PLoS ONE 7(7): e41133. doi:10.1371/journal.pone.0041133, 2012.

[18] A. J. Doupe, P. K. Kuhl. *Birdsong and human speech: Common themes and mechanisms*. Annu Rev Neurosci 22: 567–631, 1999.

[19] Y. Hao, M. Shokoohi-Yekta, G. Papageorgiou, E. J. Keogh. *Parameter-Free Motif Discovery in Arbitrary Data Archives*. Submitted to KDD 2013.

[20] Enard W, Gehre S, Hammerschmidt K, Hölter SM, Blass T, Somel M, Brückner MK, Schreiweis C, Winter C, Sohr R, Becker L, Wiebe V, Nickel B, Giger T, Müller U, Groszer M, Adler T, Aguilar A,

Bolle I, Calzada-Wack J *(2009) A humanized version of FOXP2 affects cortico-basal ganglia circuits in mice. Cell*

[21] Wöhr M, Roullet FI, Hung AY, Sheng M, Crawley JN *(2011) Communication impairments in mice lacking Shank1: reduced levels of ultrasonic vocalizations and scent marking behavior. PLoS One 6:e20631.*

[22] Fujita E, Tanabe Y, Imhof BA, Momoi MY, Momoi T *(2012) CADM1-expressing synapses on Purkinje cell dendrites are involved in mouse ultrasonic vocalization activity. PLoS One 7:e30151.*

[23] Schmeisser MJ, Ey E, Wegener S, Bockmann J, Stempel AV, Kuebler A, Janssen AL, Udvardi PT, Shiban E, Spilker C, Balschun D, Skryabin BV, Dieck St, Smalla KH, Montag D, Leblond CS, Faure P, Torquet N, Le Sourd AM, Toro R, et al. *(2012) Autistic-like behaviours and hyperactivity in mice lacking prosap1/Shank2. Nature 486:256–260.*

[24] Srivastava DP, Jones KA, Woolfrey KM, Burgdorf J, Russell TA, Kalmbach A, Lee H, Yang C, Bradberry MM, Wokosin D, Moskal JR, Casanova MF, Waters J, Penzes P *(2012) Social, communication, and cortical structural impairments in epac2-deficient mice. J Neurosci 32:11864–11878.*

[25] Portfors CV (2007) *Types and functions of ultrasonic vocalizations in laboratory rats and mice. J Am Assoc Lab Anim Sci. 2007 Jan;46(1):28-34*

[26] Grimsley JMS, Gadziola MA and Wenstrup JJ (2013) *Automated classification of mouse pup isolation syllables: from cluster analysis to an Excel-based "mouse pup syllable classification calculator*". *Front. Behav. Neurosci*. **6**:89. doi: 10.3389/fnbeh.2012.00089

[27] R.A. Fisher. *The use of multiple measurements in taxonomic problems,* Ann. Eugenics, 7 (1936), pp. 179–188

[28] J. M. S. Grimsley, J. J. M. Monaghan, J. J. Wenstrup. *Development of Social Vocalizations in Mice*. *PLoS ONE* 6(3): e17460, 2007.

[29] S. Roy, N. Watkins, D. Heck. *Comprehensive Analysis of Ultrasonic Vocalizations in a Mouse Model of Fragile X Syndrome Reveals Limited, Call Type Specific Deficits*, PLoS ONE 7(9): e44816. doi:10.1371/journal.pone.0044816, 2012.

Chapter 7

Bird Song Classification Challenge

7.1 Multi-Instance Multi-Label Acoustic Classification of Plurality of Animals : birds, insects & amphibian

O. Dufour* H. Glotin† P. Giraudet‡ Y. Bas§
T. Artières¶

25/11/2013

1 Introduction

Nowadays, consulting firms on environment propose to evaluate impacts of transports and/or power production infrastructures on biodiversity using bioacoustic and adapted algorithms of signal processing. We present here our best algorithm (whose AUC score is 0.85%). This is our contribution to the "Neural Information Processing Scaled for Bioacoustics " (NIPS4B) workshop technical challenge [1] of NIPS 2013. Our objective was to obtain a bird-sound operational classification machine-learning model that environmental engineers (mostly ornithologists) could use to realise automatic inventories of acoustically active animals.

2 Description of the method

Our preprocessing is based on Mel-filter cepstral coefficients which have been proved useful for speech [12, 29] and bird song recognition [26]. A temporal signal is first transformed into a serie of frames (see figure 1 A and B) where each frame consists in 16 mfcc (Mel-filter cepstral coefficients), including energy (first coefficient). Each frame represents a duration of 11.6 ms (e.g 512 temporal bins of a signal sampled at 44 100 Hz). Two successive frames overlap of 33% i.e. 3.9 ms.

2.1 Detection and feature extraction

*LSIS, Université du Sud Toulon Var. olivierlouis.dufour@gmail.com
†Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France. glotin@univ-tln.fr
‡Université du Sud Toulon Var. giraudet@univ-tln.fr
§BIOTOPE. ybas@biotope.fr
¶LIP6, Université Paris 6. thierry.artieres@lip6.fr
[1] In proc. of int. symposium 'Neural Information Scaled for Bioacoustics' joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

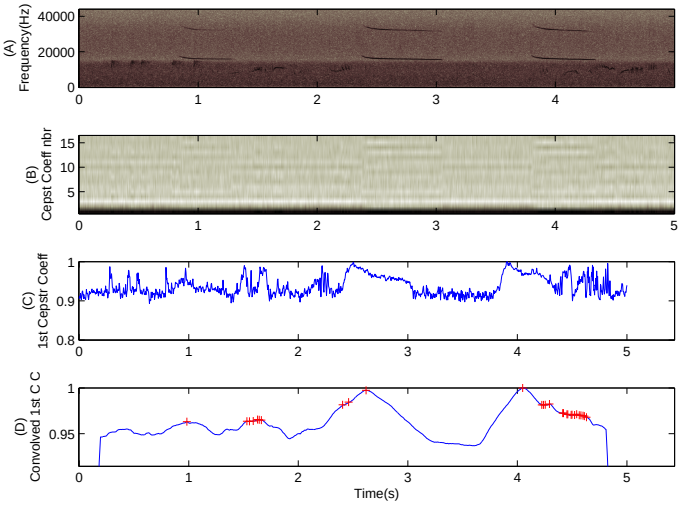


Figure 1: Main steps of the syllables detection.

Detection. To find frames of higher energy (liable to contain a bird syllable), we performed an energy-based detection step. The idea is close to the standard syllable extraction step that is used in most methods for bird identification [30, 10, 8].

1. We compute $E(t)$, $t = \{1, 2, 3, ..., N\}$ (figure 1 C). $E(t)$ is the set of values of the first MFCC from 1 to N. $E(t)$ is the value of the energy in the audio signal contained in the frame number t. N is the number of frames contained in an audio file. For a 5 seconds recording, $N \approx 860$.

2. We compute $Conv$. $Conv$ is the convolution of E by 1_{100}. 1_{100} is a 100-element vector of value "1". It corresponds to a 1.16 s duration.

$$Conv(t) = E(t) * 1_{100} \tag{1}$$

3. We note abscissas of all local maxima superior (figure 1 C) to Th such as:

$$Th = \frac{\sum_{i=50}^{N+50} Conv(i)}{N} \tag{2}$$

   or we retain abscissas of the five higher local maxima.

4. For each of D dates, we consider the values of the 16 MFCC from 16 frames before to 15 frames after the frame of the detection. Considering segments of $n = 32$ frames (i.e. about 130 ms duration) means we use windowing.

2.2 Feature extraction

The final step of the preprocessing consists in computing a reduced set of features for any segment. Recall that each segment consists in a series of n 16-dimensional feature vectors (with $n = 32$).

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 165

- **96 coefficients featuring** To get new feature vectors that are representative of longer segments, our feature extraction first consisted in computing 6 values for representing the series of n values for each of the 16 mfcc features. Let consider a particular mfcc feature v, let note $(v_i)_{i=1..n}$ the n values taken by this feature in the n frames of a window and let note $\bar{v_i}$ the mean value of v_i. Moreover let note d and D the velocity and the acceleration of v, which are approximated all along the sequences with $d_i = v_{i+1} - v_i$, and $D_i = d_{i+1} - d_i$. The 6 values we compute are defined as:

$$f_1 = \frac{\sum_{i=1}^{n}(|v_i|)}{n} \tag{3}$$

$$f_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v_i})^2} \tag{4}$$

$$f_3 = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(d_i - \bar{d_i})^2} \tag{5}$$

$$f_4 = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}(D_i - \bar{D_i})^2} \tag{6}$$

$$f_5 = \frac{\sum_{i=1}^{n-1}|d_i|}{n-1} \tag{7}$$

$$f_6 = \frac{\sum_{i=1}^{n-2}|D_i|}{n-2} \tag{8}$$

  At the end, a segment in a window is represented as the concatenation of the 6 above features for the 16 cepstral coefficients. It is then a new feature vector s_t (with t the number of the window) of dimension 96.

- **112 coefficients featuring** To get new feature vectors, we also used an algorithm implemented by Vipin Vijayan [36]. Basically this algorithm first realises a PCA on the data and then compute a LDA on the dimensionally reduced data. Contrary to 96 coefficients featuring previously mentioned, in this case the number of features (i.e 112) isn't fixed by human operator but automatically chosen.

In all cases, each audio file is finally represented as a sequence of feature vectors s_t, each representing a duration of about 130 millisecond.

2.3 Training

What makes this challenge so difficult is the fact that:

- there isn't one Multiple-Instance Single-Label training recording per class. One single-label training recording has been provided for only N classes ($K > N$; $K = 87$; $N = 51$);

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 166

- in most of test and train signals, several classes are present. Each audio file is not only represented by multiple instances but also associated with multiple class labels.

Lets consider Tsoumakas definitions from [35]. We define problem transformation methods as those methods that transform the multi-label classification problem either into one or more single-label classification problems, for which there exists a huge bibliography of learning algorithms. We define algorithm adaptation methods as those methods that extend specific learning algorithms in order to handle multi-label data directly.

Problem transformation Method The most common problem transformation method learns $|L|$ binary classifiers ($|L| = 87$), one for each different label l in L. It transforms the original data set into $|L|$ data sets Dl that contain all examples of the original data set, labelled as l, if the labels of the original example contained l and as $\neg l$ otherwise. It is the same solution used in order to deal with a single-label multi-class problem using a binary classifier. We used this approach (dubbed PT) with a Support Vector Machine classifier.

Algorithm adaptation methods One strategy can consist in separating syllables of different classes in the same training recording during preprocessing like in [10, 9]. This is an signal-processing approach. According to [30, 10, 8, 9], we chose to use a machine learning approach. We trust in learning by bag-of-instances in order to realise the tricky task.

Multi-instance multi-label learning (MIML) is a recent learning framework where each example corresponds to a bag of instances as well as a set of labels [25, 41]. To handle this MIML task, we tested different matlab toolboxes from Nanjing University [38, 39, 37]:

- MIMLRBF (MIML Radial Basis Function) is an innovative neural network style algorithm. As its name implied, MimlRbf is derived from the popular radial basis function (RBF) method [4]. Connections between instances and labels are directly exploited in the process of first layer clustering and second layer optimization. Briefly, the first layer of MIMLRBF neural network consists of medoids (i.e. bags of instances) formed by performing k-Medoids clustering on Miml examples for each possible class, where a variant of Hausdorff metric [19] is utilized to measure the distance between bags [40]. Second layer weights of MimlRbf neural network are optimized by minimizing a sum-of- squares error function and worked out through singular value decomposition (SVD) [33].

- MIML-kNN (k-Nearest Neighbor Based Multi-Instance Multi-Label Learning Algorithm) is proposed for MIML by utilizing the populark-nearest neighbor techniques. Given a test example, MIML-kNNnot only considers its neighbors, but also considers its citers whichregard it as their own neighbors. The label set of the test example is determined by exploiting the labeling information conveyed byits neighbors and citers.

- M3MIML (Maximum Margin Method for Multi-instance Multi-label Learning) assumes a linear model for each class, where the output on one class is set to be the maximum prediction of all the MIML examples instances with

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 167

respect to the corresponding linear model. Subsequently, the outputs on all possible classes are combined to define the margin of the MIML example over the classification system. Obviously, each instance is involved in determining the output on each possible class and the correlations between different classes are also addressed in the combination phase. Therefore, the connections between the instances and the labels of an MIML example are explicitly exploited by M3MIML.

Based on the feature extraction step we described above (see section *Detection*) the simplest strategy was to train a MIML classifier from feature vectors s_t which are long enough to include a syllabe or a call. We retained the idea of aggregating all vectors s_t from the same test signal to constitute a bag of present syllables and then let the classifier decide which species are present (see section *Inference*).

2.4 Inference

At test time an incoming signal is first preprocessed as explained before in section 2.2 : interesting segments are selected and feature extraction is performed. Second, a MIML learned model is used as explained before in section 2.3 to compute prediction vectors from the same audio in one K-dimension vector ($K = 87$). This yields that an input signal is represented as one bag of variable number of 96-dimension vectors.

1 000 files compose the test set. The 1 000 bags of vectors obtained after preprocessing are processed by MIML-RBF classifier to get probabilistic scores of each one of the 87 labels sets provided in the train data set.

3 RESULTS

Our detection is based on peaks of energy in time-frequency representation of animals calls and songs. Our 0.85% best score to NIPS4B challenge reveals that it is relevant to focus on higher levels of energy inside an acoustic pattern in order to counteract the intraclass variability of patterns. It is a reasonable biological hypothesis to assert that even if a given species of bird composes complex and variable strophes, it insists more (in terms of signal intensity) on some precise syllables.

Model	NIPS4B Private AUC score	short description
M1	0.7226	5 higher maxima per file + 96 features per segment + PT
M2	0.8247	5 higher maxima per file + 96 features per segment + MIMLRBF
M3	0.6837	5 higher maxima per file + 96 features per segment + MIMLkNN
M6	0.5048	5 higher maxima per file + 96 features per segment + M3MIML
M4	0.8242	all local maxima in a file + 96 features per segment + MIMLRBF
M5	0.8521	all local maxima in a file + PCA/LDA + MIMLRBF
M6	0.8290	5 higher maxima per file + PCA/LDA + MIMLRBF
Mario	0.9175	best team of NIPS4B challenge

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 168

4 Discussion

Figure 2 gives the False Negative Rate (FNR) for each class computed from model M2 predictions on data test set. One can see that the global FNR (all classes included) turns around 25%.

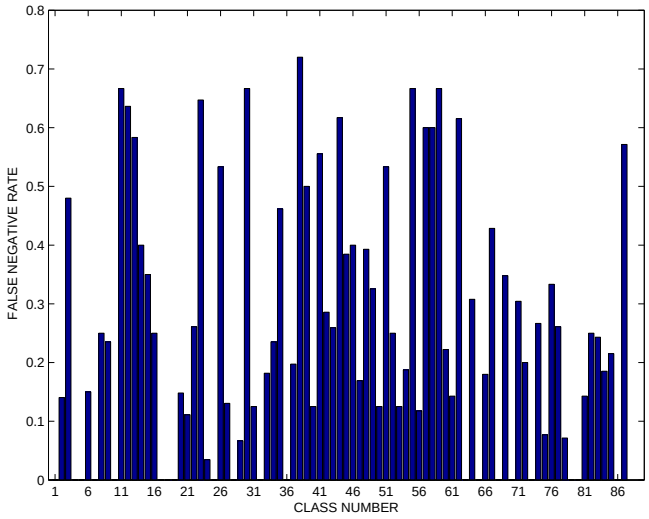


Figure 2: Class-dependant False Negative Rate for NIPS challenge

Expected comments

1. Scores are much better for classes corresponding to bird calls than for classes corresponding to bird songs. By instance, scores of classes number 36, 17, 1, 73, 18 are excellent because the concerned calls consist in strongly stereotyped signals.

2. Predictions remain generally very good for bird species whose songs stay simple and few variable (cl. 25, 65, 70).

3. A FNR of 33% for Subalpine Warbler (cl. 76) on a total of 36 test files is reasonable because it is one of the 4 most difficulty species of the challenge recognized by an ornithologist. most difficult bird species of the challenge.

4. It is well-known that European Robin produces complex and much variable songs (cl. 23). As a consequence, we reach a 65% FNR.

5. Song Thrush and European Serin (cl. 87 & 67) emit complex songs. Their respective scores are 57% et 43%. Although European Serin song is distinctive, it is also composed of a lot of syllables (50 per second). This comforts our hypothesis (see section *Improvements*) that in some cases our currently 130 ms fixed window function is well too large.

Unexpected comments

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 169

1. A 7% FNR regarding class 78 is very encouraging. Among birds species of the challenge, Sardinian Warbler is one of the 4 most difficulty recognized by an ornithologist.

2. Dartford Warbler is equally one of the 4 most complex bird species. Our flawless score must be tooked cautiously: only 3 test files contains this class.

3. Cetti's Warbler and Phylloscopus collybita (cl. 11 & 55) provide good examples of strongly stereotyped signals for which our FNRs keep too high (≈33%)).

4. The error is important concerning cl. 12 and 44 (European Greenfinch calls and Coal Tit songs) whereas their signals aren't particularly complicated because:

   - Train and test recordings providing European Greenfinch examples have a feeble signal-to-noise ratio (S.N.R);
   - In train and test recordings, Coal Tit always accompanies other species;

5. Overall, performances on insects remain disappointing

   - FNR turns around 40% for classes 82 and 14;
   - FNR regarding Common Cicada (cl. 38) is huge (72%) whereas the signal of this species is continuous and stable;

   except for

   - Pygmy Cicada (cl. 81) : 8% FNR. All train and test files concerning Pygmy Cicada come from the same location. Low FNR for this species is probably due to the fact that the model we built detect more the acoustic "signature" of the place rather than the signal of this insect;
   - and Fallow Bush-cricket (cl. 53) : 15% FNR. Its syllables keep similar to bird syllables: they are temporally-speaking punctual.

   This strengthen the idea according to which our current method isn't well compatible with uninterrupted signals. In all likelihood, a part of information concerning uninterrupted signals is lost during MFCC compression by spectral subtraction.

5 Improvements

1. According to figure 3, there is 36 classes for which we don't have any single-label recording. Plus, one can see that the volume of available training data (in seconds) varies much from one class to an other. It is very likely that this disequilibrium brakes performances of our classification algorithm. It will be interesting to watch carefully the differences of classification scores between classes and explain them: are they due to train data set disequilibrium, differences in signal complexities, variable S.N.R, acoustic properties of biotopes, etc. ?

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 170

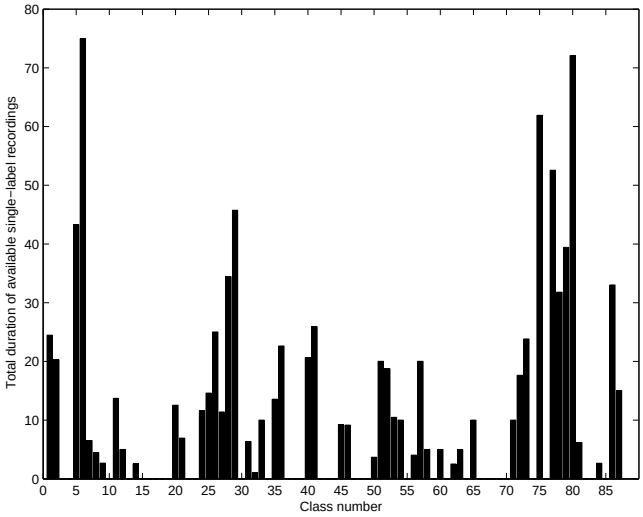


Figure 3: Available singlelabel training recordings total duration (in seconds) per class

2. Given encouraging performances of authors as Lecun, Bengio, Malikov or Abdel-Hamid models in [3, 28, 1, 23], we aim at testing in our future works MIML CNN (Convolutionnal Neural Network) algorithms.

3. One possible way of improvement consists in making variable the size of our currently fixed window function : 130 ms. Some species of birds emits more syllabes per second than others (between 1 up to 60 [6]). Moreover, we could improve learning vectors by adding the information: "Is there others detected syllables close to the considered syllable?".

4. Organizers of the challenge made the effort to label and to provide (Dr Yves Bas) 100 audio files containing only parasite sounds. Parasites constitute the most diversified class because then can be created by an infinity of different ways: car, bike or plane passage, wind, rain, walking sounds, etc. Parasites sounds designates the same type of frequency-and-temporal continuous signals than animals syllables. This is the reason why they complicate the classification task. An other way to improve our model consists in gathering all s_t vectors of all training files and separating them. On the one hand, we have the set S_a containing s_t vectors belonging to animals classes. On the other hand, we have the set S_p containing s_t vectors belonging to parasite class. It is easy to realise separately an optimized clustering of S_a (in $K1$ classes) and S_p vectors (in $K2$ classes). Thus, one can create a $K1 + K2$ multiclassification model by one-vs-all learning approach (binary relevance). This way, after the extraction of s_t from train and test files, we can identify and exclude s_t vectors similar to parasites s_t vectors. This should facilitates afterwards MIML classification task.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 171

6 Acknowledgments

PhD funds of 1st author are provided by Agence De l'Environnement et de la Maîtrise de l'Energie (mila.galiano@ademe.fr) and by BIOTOPE company (Dr Lagrange, hlagrange@biotope.fr, R&D Manager). We thank Y. Bas and S. Vigant (from BIOTOPE company) who provided and labeled the challenge data.

References

[1] O. Abdel-Hamid, L. Deng, and D. Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *INTERSPEECH*, 2013.

[2] M. Acevedo, C. Corrada-Bravo, H. Corrada-Bravo, L. Villanueva-Rivera, and T. Aide. Automated classification of bird and amphibian calls using machine learning: A comparaison of methods. *Ecological Informatics 4 206214*, 2009.

[3] Y. Bengio and Y. Lecun. Convolutional networks for images, speech, and time-series, 1995.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[5] B. Bogert, M. Healy, and J. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking. In E. M. Rosenblatt, editor, *Symposium on Time Series Analysis, Chapter 15, p 209-243*, 1963.

[6] A. Bossus and F. Charron. Guide des chants d'oiseaux d'europe occidentale : Description et comparaison des chants et des cris, 2010.

[7] F. Briggs et al. The 9th Annual MLSP Competition: New Methods for Acoustic Classification of Multiple Simultaneous Bird Species in a Noisy Environment. In *IEEE Workshop on Machine Learning for Signal Processing, MLSP 2013*, 2013.

[8] F. Briggs, X. Fern, and R. Raich. Acoustic classification of bird species from syllables: an empirical study. Technical report, 2009.

[9] F. Briggs, X. Z. Fern, and J. Irvine. Multi-label classifier chains for bird sound. *CoRR*, abs/1304.5862, 2013.

[10] F. Briggs, B. Lakshminarayanan, L. Neal, X. Fern, R. Raich, M. Betts, S. Frey, and A. Hadley. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *Journal of the Acoustical Society of America*, 2012.

[11] C.-C. Chang. Libsvm. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2008.

[12] L. Chang-Hsing, L. Yeuan-Kuen, and H. Ren-Zhuang. Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications Vol. 1, pp.17-23*, 2006.

[13] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *Big Learning 2011 : NIPS 2011 Workshop on Algorithms, Systems, and Tools for Learning at Scale*, 2011.

[14] H. G. E. Deng, L. and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *International Conference on Acoustic Speech and Signal Processing (ICASSP)*, 2013.

[15] O. Dufour, T. Artières, H. Glotin, and P. Giraudet. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. *International Machine Learning Conference*, 2013.

[16] O. Dufour, P. Giraudet, T. Artières, and H. Glotin. Automatic bird classification based on mfcc clusters, ranked 4th @ icml4b kaggle 2013 competition. In *Listening in the Wild*, page 11, 2013.

[17] O. Dufour, H. Glotin, T. Artières, and P. Giraudet. Classification de signaux acoustiques : Classification de matrices cepstre par support vector machine. Technical report, Laboratoire Sciences de l'Information et des Systèmes, Université du Sud Toulon Var, 2012.

[18] O. Dufour, H. Glotin, T. Artières, and P. Giraudet. Classification de signaux acoustiques : Recherche des valeurs optimales des 17 paramètres dentrée de la fonction melfcc. Technical report, Laboratoire Sciences de l'Information et des Systèmes, Université du Sud Toulon Var, 2012.

[19] G. A. Edgar. *Measure, topology, and fractal geometry*. Undergraduate texts in mathematics. Springer-Verlag, New York, Berlin, Paris, 1990. Réimpression en 1992, 1995.

[20] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. online web resource.

[21] H. Glotin and O. Dufour. *Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification*. INTECH, 2013.

[22] H. Glotin and J. Sueur. Overview of the first international challenge on bird classification, 2013. online web resource.

[23] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. C. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R.-T. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, C. Zhang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. In *ICONIP (3)*, pages 117–124, 2013.

[24] A. Graves, A. rahman Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013.

[25] Z. hua Zhou and M. ling Zhang. Multi-instance multilabel learning with application to scene classification. In *In Advances in Neural Information Processing Systems 19*, 2007.

[26] joint to Int. Conf. on Machine Learning. *The 1st International Workshop onf Machine Learning for Bioacoustics (ICML 2013)*, Atlanta, USA, june 2013. Glotin H. et al. http://sabiod.univ-tln.fr/ICML4B2013$_p$*roceedings.pdf*.

[27] E. Kasten, M. Philip, and G. Stuart. Ensemble extraction for classification and detection of bird species. *Ecological Informatics 5 153166*, 2010.

[28] H. Lee, Y. Largman, P. Pham, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems 22*, pages 1096–1104. 2009.

[29] A. Michael Noll. Short-time spectrum and cepstrum techniques for vocal-pitch detection. *Journal of the Acoustical Society of America, Vol. 36, No. 2, pp. 296-302*, 1964.

[30] L. Neal, F. Briggs, R. Raich, and F. X. Time-frequency segmentation of bird song in noisy acoustic environments. In *International Conference on Acoustics, Speech and Signal Processing*, 2011.

[31] A.-V. Oppenheim and R.-W. Schafer. From frequency to quefrency: a history of the cepstrum. *Signal Processing Magazine, Vol 21, Issue 5, pp 95 - 1015*, 2004.

[32] J. Placer and C. Slobodchikoff. A method for identifying sounds used in the classification of alarm calls. *Behavioural Processes 67: 8798*, 2004.

[33] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1988.

[34] L. Ranjard, H. Ross, and H. Ross. Unsupervised bird song syllable classification using evolving neural networks. *Journal of the Acoustical Society of America, Volume 123, Issue 6, pp. 4358-4368*, 2008.

[35] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.

[36] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.

[37] M.-L. Zhang. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *ICTAI (2)*, pages 207–212. IEEE Computer Society, 2010.

[38] M.-L. Zhang and Z.-J. Wang. Mimlrbf: Rbf neural networks for multi-instance multi-label learning. *Neurocomputing*, 72(16-18):3951–3956, 2009.

[39] M.-L. Zhang and Z.-H. Zhou. M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning. In *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 688–697, Washington, DC, USA, Dec. 2008. IEEE Computer Society.

[40] M.-L. Zhang and Z.-H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31(1):47–68, Aug. 2009.

[41] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Miml: A framework for learning with ambiguous objects. *CoRR*, abs/0808.3231, 2008.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al. 175

7.2 Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition*

Mario Lasseck
Animal Sound Archive
Museum für Naturkunde Berlin
Mario.Lasseck@mfn-berlin.de

Abstract

The challenge of the NIPS4B competition is to identify 87 sound classes of birds and other animals present in 1000 audio recordings, collected in the field. The difficulty of this task lies in the large number of species and sounds that have to be identified in various contexts dealing with different levels of background noise and simultaneously vocalizing animals. The solution presented here ranks first place on the kaggle private leaderboard and achieves an Area Under the Curve of 91.7% (AUC).

1 Introduction

The audio data was recorded at different places in Provence France and is provided by the BIOTOPE society, having one of the largest collections of wildlife recordings of birds in Europe. The nearly 2 hours of recordings are split into smaller clips ranging from 0.25 to 5.75 seconds. The recordings were done with Wildlife Acoustics SM2 and are presented in uncompressed WAV format with a sample rate of 44.1 kHz. The 87 individual sound classes within these recordings represent different bird species and their songs, calls and drumming. Other animal species living in the same environment like insects and one amphibian are also included. The training set consists of 687 audio files. Each file is paired with the subset of sound classes present in that recording. Some recordings are empty, containing only background noise, others contain up to 6 different simultaneously vocalizing birds or insects. Each species is represented by nearly 10 training files within various contexts, different background noises and an arbitrary number of other species. The goal of the competition is to identify which of the 87 sound classes of birds and amphibians are present in 1000 continuous wildlife recordings, using only the provided audio files and machine learning algorithms for automatic pattern recognition.

2 Preprocessing and Segmentation

The method of segmentation has a big influence on classification results. Several different approaches were tested. The one that works best regarding leaderboard score is surprisingly simple. Audio files are first resampled to 22050 Hz. After applying the STFT using a hanning window with a size of 512 samples and 75% overlap the resulting spectrogram is normalized to a maximum of 1.0. The 4 lowest and 24 highest frequency bins are removed,

* In proc. of 'Neural Information Scaled for Bioacoustics' joint to NIPS, http://sabiod.org/nips4b, Nevada, dec. 2013, Ed. Glotin H. et al.

leaving 228 frequency bins or spectrogram rows representing the relevant frequency range of approximately 170 to 10000 Hz. The narrowed spectrogram of each audio file is treated as grayscale image and further processed for noise reduction and segmentation.

To reduce background noise each pixel value is set to 1 if it is above 3 times the median of its corresponding row (frequency band) AND 3 times the median of its corresponding column (time frame), otherwise it is set to 0. This *Median Clipping* per frequency band and time frame removes already most of the background noise. Variable noise levels in different frequency regions are compensated and short, broadband distortions coming from rain, wind or microphone handling are attenuated.

The resulting binary image is further processed using standard image processing techniques (e.g. closing, dilation, median filter). Finally, all connected pixels exceeding a certain spatial extension are labeled as a segment and a rectangle with a small area added to each direction is used to define its size and position. Figure 1 gives an example of the preprocessing steps involved and Figure 2 shows the outcome of a complete segmentation process.

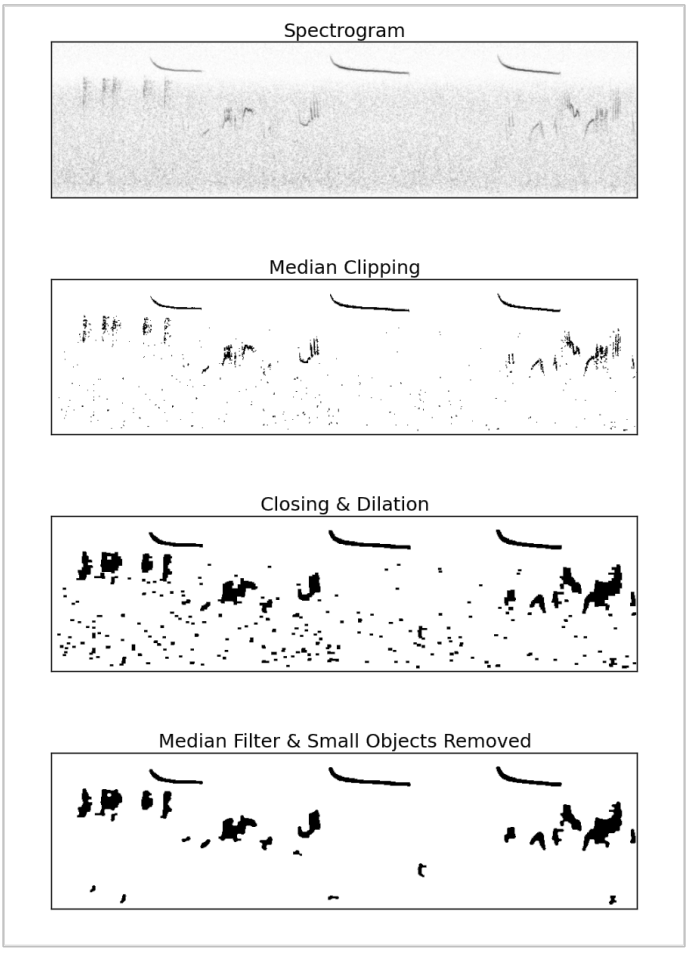


Figure 1: Preprocessing of Spectrogram Image

Preprocessing the spectrograms extracts 9198 segments from the training data and 16726 segments from the test recordings.

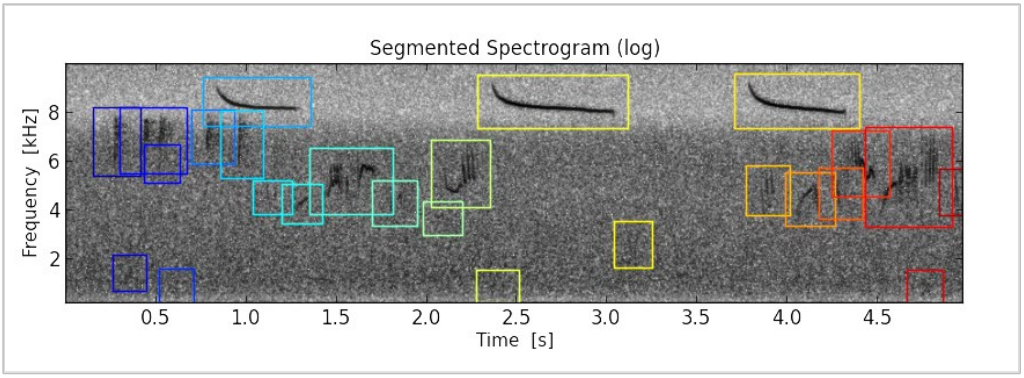


Figure 2: Spectrogram Image with marked Segments

3 Feature Extraction

Features are calculated for both, training and test files, coming from three different sources: File-Statistics, Segment-Statistics and Segment-Probabilities.

File-Statistics include minimum, maximum, mean and standard deviation taken from all values of the unprocessed spectrogram. Additionally the spectrogram is divided into 16 equally sized and distributed frequency bands and their minima, maxima, means and standard deviations are also included.

For Segment-Statistics the number of segments per file plus minimum, maximum, mean and standard deviation for width, height and frequency position of all segments per file are calculated.

In order to find Segment-Probabilities a variation of Fodor's method [1] is used which was already successfully applied in the MLSP 2013 Competition. The highest matching probability of all segments extracted from training files associated with one or more sound class is determined in all files by template matching using normalized cross-correlation [2]. A Gaussian blur with a sigma of 1.5 is applied to segment and target spectrogram before matching. Best matches are only searched for within the frequency range of the segment (± a small tolerance of 4 pixels). Unlike Fodor, the template matching uses only absolute-intensity spectrograms and for better performance the OpenCV library [4] is used.

File- and Segment-Statistics produce 81 features per file scaled to the range [0 1]. Segment-Probabilities create, corresponding with the number of extracted segments from the training set, 9198 features per file.

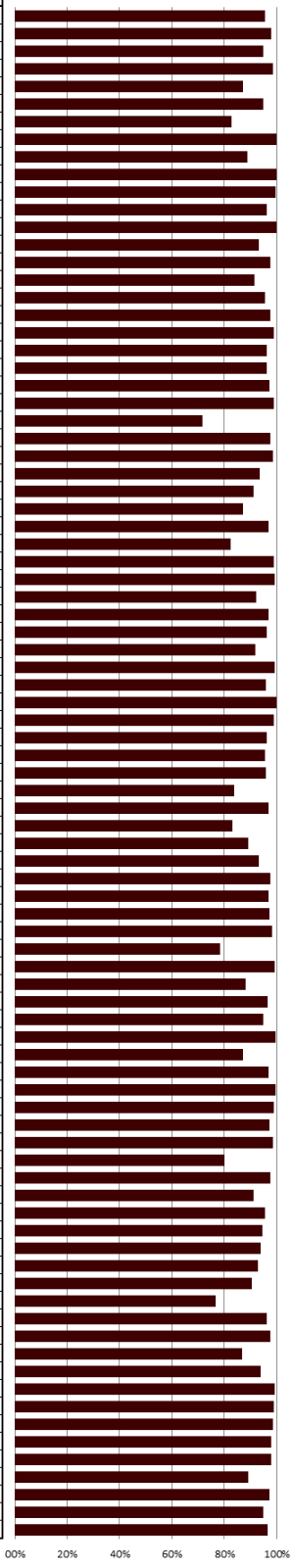
4 Feature Selection

As in [1] already suggested the multi-instance multi-label classification problem is turned into 87 individual classification problems. In that way the probability for each target sound class is calculated separately for all files. Each of the 87 classifiers uses all File- and Segment-Statistics. But as for the Segment-Probabilities, only those belonging to segments extracted from training files associated with the corresponding target sound class are included. This way the number of features to be taken into account for learning and predicting a particular sound class can be reduced significantly which produces much better classification results.

To give an example: Sound class 86 appears in 8 training files (107, 172, 264, 353, 387, 504, 510 and 596). For learning and predicting this particular sound class, only matching probabilities belonging to segments extracted from these 8 files are included as features. The number of relevant features selected from Segment-Probabilities for each sound class is listed in Table 1.

Table 1: Number of selected features and estimators per sound class plus AUC scores

Class			Features				
No.	Name	Scientific Name	File Statistics	Segment Statistics	Segment Probilities	Estimators	AUC
1	Aegcau_call	Aegithalos caudatus	68	13	148	458	95.6%
2	Alaarv_song	Alauda arvensis	68	13	213	500	97.8%
3	Anttri_song	Anthus trivialis	68	13	271	500	95.0%
4	Butbut_call	Buteo buteo	68	13	131	424	98.5%
5	Carcan_call	Linaria cannabina	68	13	244	500	87.1%
6	Carcan_song	Linaria cannabina	68	13	137	436	94.9%
7	Carcar_call	Carduelis carduelis	68	13	192	500	82.6%
8	Carcar_song	Carduelis carduelis	68	13	301	500	99.8%
9	Cerbra_call	Certhia brachydactyla	68	13	183	500	88.7%
10	Cerbra_song	Certhia brachydactyla	68	13	298	500	99.8%
11	Cetcet_song	Cettia cetti	68	13	226	500	99.6%
12	Chlchl_call	Chloris chloris	68	13	203	500	96.4%
13	Cicatr_song	Cicadatra atra	68	13	240	500	100.0%
14	Cicorn_song	Cicada orni	68	13	98	358	93.3%
15	Cisjun_song	Cisticola juncidis	68	13	118	398	97.5%
16	Colpal_song	Columba palumbus	68	13	103	368	91.6%
17	Corcor_call	Corvus corone	68	13	158	478	95.7%
18	Denmaj_call	Dendrocopos major	68	13	157	476	97.7%
19	Denmaj_drum	Dendrocopos major	68	13	392	500	98.8%
20	Embcir_call	Emberiza cirlus	68	13	208	500	96.1%
21	Embcir_song	Emberiza cirlus	68	13	283	500	96.2%
22	Erirub_call	Erithacus rubecula	68	13	325	500	97.3%
23	Erirub_song	Erithacus rubecula	68	13	275	500	98.8%
24	Fricoe_call	Fringilla coelebs	68	13	361	500	71.6%
25	Fricoe_song	Fringilla coelebs	68	13	230	500	97.6%
26	Galcri_call	Galerida cristata	68	13	331	500	98.5%
27	Galcri_song	Galerida cristata	68	13	152	466	93.4%
28	Galthe_call	Galerida theklae	68	13	117	396	91.1%
29	Galthe_song	Galerida theklae	68	13	155	472	87.0%
30	Gargla_call	Garrulus glandarius	68	13	235	500	97.1%
31	Hirrus_call	Hirundo rustica	68	13	76	314	82.5%
32	Jyntor_song	Jynx torquilla	68	13	166	494	98.9%
33	Lopcri_call	Lophophanes cristatus	68	13	252	500	99.3%
34	Loxcur_call	Loxia curvirostra	68	13	377	500	92.1%
35	Lularb_song	lullula arborea	68	13	323	500	97.1%
36	Lusmeg_call	Luscinia megarhynchos	68	13	211	500	96.3%
37	Lusmeg_song	Luscinia megarhynchos	68	13	307	500	91.7%
38	Lyrple_song	Lyristes plebejus	68	13	235	500	99.3%
39	Motcin_call	Motacilla cinerea	68	13	176	500	95.8%
40	Musstr_call	Muscicapa striata	68	13	154	470	99.9%
41	Oriori_call	Oriolus oriolus	68	13	96	354	99.0%
42	Oriori_song	Oriolus oriolus	68	13	255	500	96.4%
43	Parate_call	Periparus ater	68	13	446	500	95.7%
44	Parate_song	Periparus ater	68	13	609	500	96.0%
45	Parcae_call	Cyanistes caeruleus	68	13	253	500	83.8%
46	Parcae_song	Cyanistes caeruleus	68	13	351	500	96.9%
47	Parmaj_call	Parus major	68	13	312	500	83.3%
48	Parmaj_song	Parus major	68	13	532	500	89.3%
49	Pasdom_call	Passer domesticus	68	13	308	500	93.1%
50	Pelgra_call	Pelophylax kl. grafi	68	13	101	364	97.7%
51	Petpet_call	Petronia petronia	68	13	202	500	96.8%
52	Petpet_song	Petronia petronia	68	13	176	500	97.3%
53	Phofem_song	Pholidoptera femorata	68	13	279	500	98.2%
54	Phycol_call	Phylloscopus collybita	68	13	168	498	78.4%
55	Phycol_song	Phylloscopus collybita	68	13	500	500	99.4%
56	Picpic_call	Pica pica	68	13	185	500	88.1%
57	Plaaff_song	Platycleis affinis	68	13	287	500	96.7%
58	Plasab_song	Platycleis sabulosa	68	13	324	500	94.9%
59	Poepal_call	Poecile palustris	68	13	335	500	99.7%
60	Poepal_song	Poecile palustris	68	13	253	500	87.2%
61	Prumod_song	Prunella modularis	68	13	304	500	96.8%
62	Ptehey_song	Pteronemobius heydenii	68	13	132	426	99.7%
63	Pyrpyr_call	Pyrrhula pyrrhula	68	13	120	402	99.0%
64	Regign_call	Regulus ignicapillus	68	13	264	500	97.2%
65	Regign_song	Regulus ignicapillus	68	13	375	500	98.7%
66	Serser_call	Serinus serinus	68	13	178	500	80.1%
67	Serser_song	Serinus serinus	68	13	465	500	97.5%
68	Siteur_call	Sitta europaea	68	13	147	456	91.3%
69	Siteur_song	Sitta europaea	68	13	466	500	95.6%
70	Strdec_song	Streptopelia decaocto	68	13	66	294	94.7%
71	Strtur_song	Streptopelia turtur	68	13	118	398	94.0%
72	Stuvul_call	Sturnus vulgaris	68	13	155	472	92.9%
73	Sylatr_call	Sylvia atricapilla	68	13	178	500	90.7%
74	Sylatr_song	Sylvia atricapilla	68	13	152	466	76.7%
75	Sylcan_call	Sylvia cantillans	68	13	227	500	96.2%
76	Sylcan_song	Sylvia cantillans	68	13	344	500	97.4%
77	Sylmel_call	Sylvia melanocephala	68	13	287	500	86.8%
78	Sylmel_song	Sylvia melanocephala	68	13	224	500	93.9%
79	Sylund_call	Sylvia undata	68	13	48	258	99.4%
80	Sylund_song	Sylvia undata	68	13	189	500	98.8%
81	Tetpyg_song	Tettigettula pygmea	68	13	222	500	98.7%
82	Tibtom_song	Tibicina tomentosa	68	13	159	480	98.1%
83	Trotro_song	Troglodytes troglodytes	68	13	400	500	97.9%
84	Turmer_call	Turdus merula	68	13	358	500	89.0%
85	Turmer_song	Turdus merula	68	13	441	500	97.1%
86	Turphi_call	Turdus philomelos	68	13	22	206	94.8%
87	Turphi_song	Turdus philomelos	68	13	386	500	96.6%



5 Classification

The scikit-learn library is used for classification [3]. For each sound class an ensemble of randomized decision trees (sklearn.ensemble.ExtraTreesRegressor) is applied. The number of estimators is chosen to be twice the number of selected features per class but not greater than 500. The winning solution considers 4 features when looking for the best split and requires a minimum of 3 samples to split an internal node. During 12-fold cross validation the probability of each sound class in all test files is predicted and at the end, after removing the lowest and highest value, averaged.

Good classification results are possible even without calculating File- and Segment-Statistics and therefor without the need to segment the test recordings. Just with Segment-Probabilities, using the same parameter settings as mentioned above, a score of 91.6% AUC on the private leaderboard can be achieved. A score around 84% is achievable using File- and Segment-Statistics exclusively.

By ranking feature importance returned from the decision trees during training one can find important segments to identify each sound class. Figure 3 and 4 show the ten most important segments to identify the songs of Cetti's Warbler (sound class 11) and Common Chiffchaff (sound class 55). Both sound classes achieve very good classification results with a score close to 100%. Figure 5 gives an example of a sound class with poor classification results. The feature ranking returned from decision trees to identify the call of the European Serin (sound class 66) is partly incorrect and segments are not properly assigned.

To give an idea how well individual species can be identified, a score per sound class is calculated on one third of the training data during 3-fold cross validation. The average of this score is listed and visualized in Table 1.

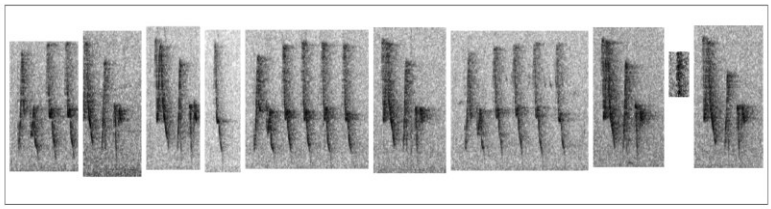


Figure 3: Important segments to identify the song of Cettia cetti (Cetti's Warbler)

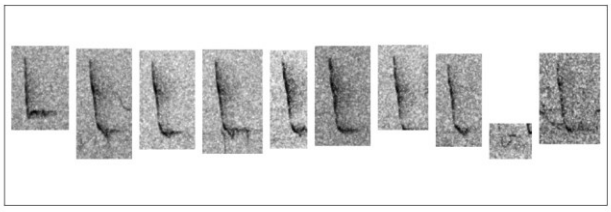


Figure 4: Important segments to identify the song of Phylloscopus collybita (Common Chiffchaff)

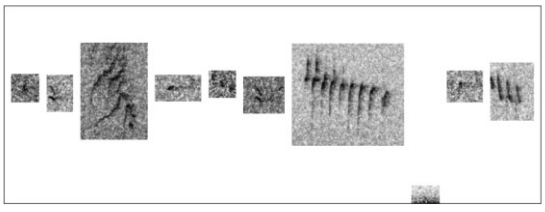


Figure 5: Important segments to identify the call of Serinus serinus (European Serin)

6 Conclusion

This working note describes the winning solution of the NIPS4B 2013 multi-label Bird Species Classification Challenge. The solution of the MLSP 2013 Competition, implemented and described by Fodor, was used as a starting point for further development. The here proposed method includes an efficient way of extracting single sound events and connected sequences of bird calls and syllables in complex acoustic scenes and noisy environments. An ensemble of randomized decision trees is used to learn and predict the binary relevance of each sound class separately with individually selected features per class. The complete source code to reproduce the classification results and additional figures are available at www.animalsoundarchive.org/RefSys/Nips4b2013.php.

Acknowledgments

I would like to thank Prof. Hervé Glotin for organizing this competition, BIOTOPE and ADEME for financing the corpus constitution and kaggle for providing the competition platform. I especially thank Gabor Fodor for documenting his approach and publishing his code for the 2013 MLSP Challenge. I also want to thank Dr. Karl-Heinz Frommolt for supporting my work, sharing his knowledge and providing me with the access to the resources of the Animal Sound Archive [5] at the Museum für Naturkunde Berlin.

References

[1] Fodor G. (2013) The Ninth Annual MLSP Competition: First place. Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on, Digital Object Identifier: 10.1109/MLSP.2013.6661932 Publication Year: 2013, Page(s): 1 - 2

[2] Lewis J.P. (1995) Fast Normalized Cross-Correlation, Industrial Light and Magic

[3] Pedregosa F. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: 2825-2830

[4] Bradski G. (2000) The OpenCV Library. Dr. Dobb's Journal of Software Tools, http://docs.opencv.org/modules/imgproc/doc/object_detection.html

[5] Animal Sound Archive URL: http://www.animalsoundarchive.org/ [25. 11. 2013]

7.3 Feature design for multilabel bird song classification in noise (NIPS4B challenge)

Dan Stowell and Mark D. Plumbley
Centre for Digital Music, Queen Mary University of London, London, UK
dan.stowell@qmul.ac.uk

Bird vocalisations are highly varied, containing natural variation across a range of timescales. In recent work we have modelled the transitions between syllables [5], and combined this with representations which capture fine FM variations within syllables [3]. Following on from such work we are exploring feature design for the representation of temporal structure in sounds such as birdsong.

The 2013 NIPS4B bird song challenge is concerned with automatically recognising the presence of a number of species, from sound alone. For training, 687 audio clips are provided (each annotated as containing 0–6 species); for testing, 1000. The clips are around 5 seconds long, recorded by automated monitoring units, and often noisy and with the target sounds both distant and quiet.

Our submission to the challenge therefore focusses on feature design for two goals: noise robustness, and the representation of temporal structure. We first analyse each sound file into basic features, either MFCCs (13 MFCCs plus delta features) or our peak-chirplet representation [4]. Importantly, both of these feature algorithms are modified to apply noise reduction in their spectral analysis step, simply by median-filtering: taking a spectral profile across time (the energy 75-percentile), subtracting this profile from the values, and keeping only the positive values.

We then reduce the time-series data for each file down to an atemporal summary vector. We summarise our noise-reduced MFCCs by their mean and standard deviations, a simple and common baseline approach. We summarise our chirplets by a histogram of all the bigrams found in the file: in other words, for every transition from one packet of energy to another, we record the time separation as well as the frequency and chirp-rate values, and these parameters form the axes of the high-dimensional histogram we create (related to the method in [4]). The time-separation between bigram pairs is not constant: we examine all possible transitions shorter than 1 second. Note that we avoid any need to perform segmentation on the input audio files. The histogram represents the set of all transitions observed in the audio data, and is used directly for multilabel classification.

For multilabel classification we use Random Forests [1] implemented in scikit-learn [2]. In variations of our submission, we use either MFCC statistics (52 dimensions), chirplet histograms (up to 20,000 dimensions), or both. We experimented with dimension reduction but found this unnecessary. We also experimented with other multilabel classifiers, but found they generally reduced performance relative to Random Forests.

Figure 1 illustrates the two types of feature, as well as the effect of noise reduction. For both feature types, even the noise-reduced plots are visually noisy, but a strong difference between the features is visible: the chirplet representation captures many of the fine-grained pitch trajectories in the segments containing bird vocalisations. We observed informally that the chirplet features performed well for tonal songbird sounds, for which the features were originally designed, but did not completely generalise to the less tonal vocalisation types in this challenge. In practice, this led to the MFCCs outperforming the chirplet features (by only a small margin) when considered in isolation. As future work we intend to consider schemes to combine aspects of these features, which go beyond the simple stacking or PCA tested during this challenge.

By the Area Under the Curve (AUC) score, we attain 89.5% on the public leaderboard and 88.5% on the held-out leaderboard.

In Proc. of 'Neural Information Scaled for Bioacoustics' joint to NIPS, http://sabiod.org/nips4b, Nevada, Dec. 2013, Ed. Glotin H. et al.

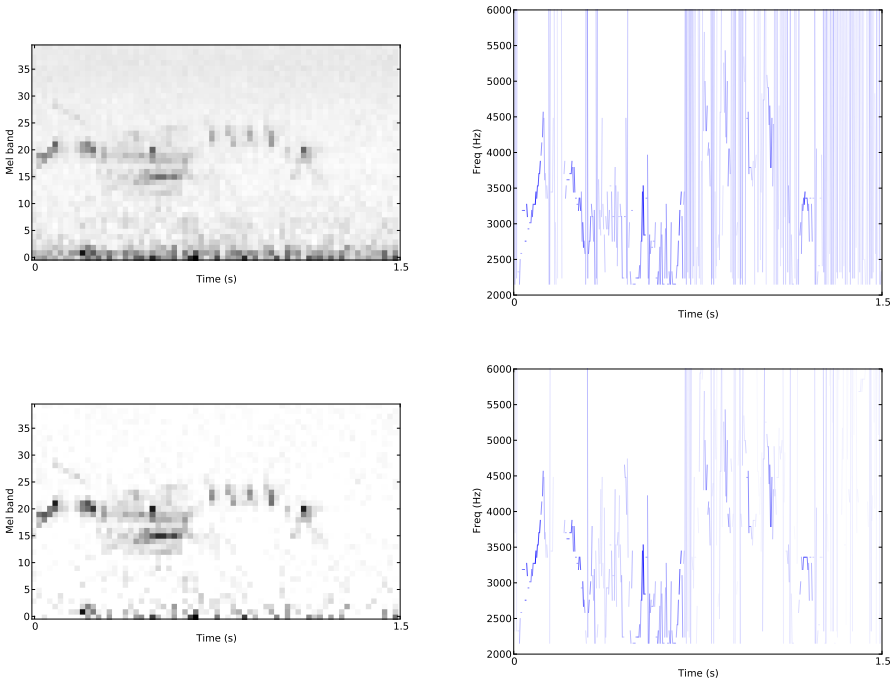


Figure 1: Illustration of features for an excerpt of training file 007, comparing Mel spectra (left) against peak chirplet data (right). The lower plots show the same features after noise reduction. Note that the left plots show Mel spectra which we further process to MFCCs, and the right plots show chirplets which which we further process to bigram histograms.

Acknowledgments

DS & MP are supported by an EPSRC Leadership Fellowship EP/G007144/1.

The challenge was organised by Prof Hervé Glotin and the SABIOD project team, with data provided by the BIOTOPE society and ADEME.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] D. Stowell, S. Muševič, J. Bonada, and M. D. Plumbley. Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering. In *Proceedings of the International Conference on Audio and Acoustic Signal Processing (ICASSP)*, 2013. preprint arXiv:1302.3642.

[4] D. Stowell and M. D. Plumbley. Framewise heterodyne chirp analysis of birdsong. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 2694–2698, 2012.

[5] D. Stowell and M. D. Plumbley. Segregating event streams and noise with a Markov renewal process model. *Journal of Machine Learning Research*, 14:1891–1916, 2013.

7.4 Learning multi-labeled bioacoustic samples with an unsupervised feature learning approach *

Eneldo Loza Mencía
Knowledge Engineering Group
Technische Universität Darmstadt
eneldo@ke.tu-darmstadt.de

Jinseok Nam
Knowledge Engineering Group
Technische Universität Darmstadt
nam@kdsl.informatik.tu-darmstadt.de

Dong-Hyun Lee
sayit78@gmail.com

Abstract

Multi-label Bird Species Classification competition provides an excellent opportunity to analyze the effectiveness of acoustic processing and mutlilabel learning. We propose an unsupervised feature extraction and generation approach based on latest advances in deep neural network learning, which can be applied generically to acoustic data. With state-of-the-art approaches from multilabel learning, we achieved top positions in the competition, only surpassed by teams with profound expertise in acoustic data processing.

1 Introduction

Acoustic data is a common and natural representative of multilabel data, i.e. data in which an example, in this case an acoustic sample, can be mapped to several, non-exclusive classes or categories. Examples are the popular *emotions* benchmark with the objective of assigning emotions to music or the *hifind* dataset where the tasks is to identify used instruments, genres, moods, languages, styles etc. in songs [12]. In the Multi-label Bird Species Classification competition (NIPS4B) the task was to identify 87 birds, insects and amphibians in short audio recordings. Of course, these could appear in the same sample. More specifically, the objective was to maximize the area under the ROC curve (AUC) on 1000 unlabeled recordings, a common measure for the quality of a label ranking.

The challenge was thus two fold: On the one hand it was necessary to process the data in a way appropriate for machine learning approaches, since the data was only available in a raw format or in very basic preprocessed format. This article presents a combination of recent and state-of-the-art approaches from neural network and deep learning which allows an unsupervised generation of an aleatory number of features, appropriate for being processed by standard machine learning algorithms. It basically consists of random patching, a Denoising Autoencoder unit and subsequent convolution and represents a general approach for processing acoustic data.

On the other hand, it was essential to learn the data accurately in order to produce high quality predictions and to get the most out the provided information in form of input feature and binary (relevant/irrelevant) label information. We tried out three approaches: firstly, a pairwise ensemble of SVMs which actually is geared towards the base of AUC, the correct order of pairs of labels. The popular and effective LibSVM library was specifically adapted to allow pairwise learning and the modifications are made available. Secondly and thirdly, random decision trees and a single layer neural network were applied. The diversity of the classifiers ensured that the combination of the

*In proc. of int. symposium *Neural Information Scaled for Bioacoustics* joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

predictions was effective. Our final ranking in the very competitive contest show that our acoustic prepocessing provides a good base for the following machine learning step and that the multilabel learner exhaust this.

2 A Multilabel Bird Species Classification Dataset

The dataset for the Multi-label Bird Species Classification competition contains 687 labeled training examples, including 100 noise samples which are not labeled with any bird species, and 1,000 unlabeled test examples for measuring the generalization performance. The recordings belong to 87 categories (bird species like the subalpine warbler), each of which is associated with approx. 13 training instances. The average labelset size is 2.00 excluding noise samples, maximally 6, and there are 265 distinct labelsets in the training data. The dataset comprises two format: one is in raw wav format in which bird songs and calls are recorded with distant insects, and the other is Mel-Frequency Cepstral Coefficients (MFCCs) of the wav files following preprocessing steps in [4] where each time frame is represented with 17 coefficients. Each audio clip in both train and test data varies in length.

3 Unsupervised Feature Generation and Extraction for Acoustic Clips

Let $\mathbf{s} \in \mathbb{R}^{m \times T}$ be the input vector for an audio clip where T is the total number of frames in time and each time frame t consists of an m dimensional feature vector. In our first attempts, we just padded (repeated) smaller samples so that in the end all samples had the same length $\max_i T_i = 1288$, resulting in 21,896 total number of features (referred to as *raw* dataset). However, the results were not satisfactory, thus we applied the following operations and methods from unsupervised feature learning. These were already successfully applied e.g. on image data, hence the question was whether they would work for acoustic data.

Firstly, we extract M_{tr} and M_{ts} random patches, totally $M = M_{tr} + M_{ts}$, whose size is $psz = m \times wnd$ from training and test data, respectively, where wnd denotes the size of window. An extracted patch is then normalized along the time frame axis which makes each coefficient has zero mean and unit variance. Secondly, the randomly sampled patches are concatenated to form training examples $\Phi \in \mathbb{R}^{psz \times M}$ for Denoising Autoencoder or DAE [14], which learns hidden representations from inputs in an unsupervised way. A DAE is a neural network architecture consisting of *encoder* f_{enc} and *decoder* g_{dec} with parameters $\theta = \{W, b, c\}$ to minimize the squared error loss function $\|\varphi - g_{dec}(W^T f_{enc}(W\tilde{\varphi} + b) + c))\|_2^2$ where $W \in \mathbb{R}^{F \times psz}$ is the weights matrix connecting visible units and hidden units, b and c are biases for hidden units and visible units, and $\tilde{\varphi}$ is the corrupt input by adding Gaussian noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2)$ to an input φ . Once training DAE is done, each column of the weights W^T acts as a feature detector. F feature detectors can be considered in total, and each feature detector has the same size as the randomly extracted patches, that is, the k^{th} feature detector is $W^T_{\cdot\cdot k} \in \mathbb{R}^{m \times wnd}$. Finally, we can obtain a fixed feature representation for an input signal $\mathbf{s}$ in terms of T while convolving it with learned feature detectors.

$$\mathbf{a}_k = f_{conv}\left(s * W^T_{\cdot\cdot k} + b_k\right) \qquad \mathbf{a}_{k+F} = f_{conv}\left(s * (-W^T_{\cdot\cdot k}) + b_k\right) \tag{1}$$

$$x_k = \sum_{j=1}^{T-wnd+1} a_{k,j} \qquad x_{k+F} = \sum_{j=1}^{T-wnd+1} a_{(k+F),j} \tag{2}$$

where $*$ stands for a 2D discrete convolution operator[1] and f_{conv} is to provide non-linearity to the convolved feature representations. We use ReLUs $f(x) = \max(0, x)$ for nonlinear function f_{conv}. In order to make use of the negative part as well as the positive part of inputs to ReLUs, we apply polarity splitting [2] in Eq. 1. Then, we sum up the convolved feature values $\mathbf{a}_k$ over time which is analogous to accumulated activations of $\mathbf{s}$ with respect to the feture detector $W^T_{\cdot\cdot k}$ (Eq. 2). For instance, x_k will be higher if a feature k defined by $W^T_{\cdot\cdot k}$ is detected many times in $\mathbf{s}$.

For training DAE, we extracted 100,000 17×80 patches randomly from training data and 100,000 from test data in the MFCC format. We then trained two DAE models with Gaussian noise $\sigma = 0.2$

[1] The convolution operator yields a matrix of size $1 \times (T - wnd + 1)$.

on the patches; the models have 400 and 800 hidden units resulting in 800 (*small*) and 2000 (*big*), respectively. We used ReLU for the *encoder* and sigmoid $f(x) = 1/(1+\exp(-x))$ for the *decoder*.

4 Multilabel Learning

Multilabel classification refers to the task of learning a function that maps instances $\mathbf{x}_i \in \mathcal{X}$ to label subsets or label vectors $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n}) \in \{0,1\}^n$, where $\mathcal{L} = \{\lambda_1, \ldots, \lambda_n\}$, $n = |\mathcal{L}|$ is a finite set of predefined labels and where each label attribute y_i corresponds to the absence (0) or presence (1) of label λ_i. In the following, we will present the different learning algorithms we applied on the NIPS4b dataset.

Pairwise Support Vector Machines The most common approach for multilabel classification is to use an ensemble of binary classifiers, where each classifier predicts if an instance belongs to one specific class or not (*binary relevance* or BR). An alternative is to do *pairwise decomposition*. Here, one classifier is trained for each pair of classes, i.e., a problem with n different classes is decomposed into $\frac{n(n-1)}{2}$ smaller subproblems [6]. More precisely, for each pair of classes (λ_u, λ_v), $u < v$, we learn a binary base classifier $h_{u,v}$, whose training set is composed of examples for which λ_u is a relevant class and λ_v is an irrelevant class, or vice versa. All other examples are ignored for this particular subproblem [10]. During classification, all of the $\frac{n(n-1)}{2}$ base classifiers make a prediction for one of the both corresponding classes, which is interpreted as a full vote (0 or 1), hence resulting in a full ranking over the labels.[2]

Pairwise learning method is often regarded as superior to BR because it profits from simpler decision boundaries in the subproblems [6, 8]. The reason is that each of the pairwise classifiers contains fewer examples. In fact, it has also been shown that the complexity for training an ensemble of pairwise classifiers is comparable to the complexity of training a BR ensemble [6, 10]. During prediction, however, we have a quadratic number of classifiers we have to evaluate. But particularly for support vector machines this problem is alleviated by the fact that easier (sub-)problems lead to less support vectors and that support vectors can be shared among the pairwise SVMs.

Multilabel LibSVM Because of this and because SVMs trained in a pairwise fashion already obtained state-of-the-art results on standard benchmark datasets [12] in previous works [11], we decided to use the very popular and effective SVM software library LibSVM [1] for training our pairwise SVMs. However, during preliminary experiments, we found out that just plugging in LibSVM was not feasible since a simple experiment on the dataset apparently required more than 20 GB of memory. The reason is that the used Java interface copies every training instance each time for every base learner. Additionally, each of the 3741 LibSVM instantiation could request up to 40 MB of cache. We thus extended LibSVM directly in order to support the pairwise learning of multilabel data. Our extension does not copy a training instance more than once and also shares a common cache for Kernel computations, so that we managed to perform an experiment in less than 250 seconds for training 618 instances and 9 seconds for testing 68 instances and with less than 100 MB of memory (worst cases, respectively) despite the quadratic number of models to be trained, stored and evaluated. The LibSVM modifications and interfaces for the multilabel learning toolkit MULAN [13] are available from http://www.ke.tu-darmstadt.de/resources/multilabellibsvm.

Random Decision Trees Zhang et al. [15] recently proposed to use ensembles of random decision trees (RDTs) for learning multilabel data and also provide a software library.[3] The main idea is to generate k_1 RDTs with random attribute tests at the inner nodes and maximal depth k_2. Comparably small values of k_1 and k_2, around 10 or 20 and maximally 100, are sufficient in their experiments. During the extremely fast training, the leafs incrementally collect statistics about the label distributions $\mathbf{y}$ which passed all tests to the leafs. Hence, each RDT predicts an average distribution, which is subsequently averaged over all trees. RDTs are very suitable for data with a high number of examples and labels, since the costs are bounded by the selection of k_1 and k_2. However, they may have problems with high number of features and particularly sparse features, which is not the case for NIPS4B.

[2] Ties in the final votes counting are broken by using the prior probabilities of the labels.

[3] http://www.dice4dm.com/

Neural Networks with a Single Hidden Layer Neural networks (NNs) have attracted increasing interest in recent years thanks to success of NNs with multiple levels of trainable feature extractors, namely *deep learning* in various domains such as object recognition and speech recognition. In order to achieve state of the arts performance, one usually trains deep neural networks on a large amount of training examples or initialize parameters by pretraining the networks on *unlabeled* data in an unsupervised manner, followed by learning whole parameters including a classification layer using labeled instances. As the bird species classification dataset has only 587 labeled examples and only 11 positive training instances are available per label, on average, we decided to train NNs with a only single hidden layer rather than ones with multiple hidden layers.

The single hidden layer NNs perform surprisingly well when we combine them with AdaGrad [3], which makes it possible to adapt the learning rate per parameter, and Dropout [7] to prevent overfitting and hence improving generalization performance. The output $\hat{\mathbf{y}}$ of NNs for a given training example $\mathbf{x}$ is computed by using the following composition of non-linear functions $\hat{\mathbf{y}} = f_o(W^{(2)} f_h(W^{(1)}\mathbf{x} + b^{(1)}) + b^{(2)})$ where $f_o(x) = 1/(1+\exp(-x))$ and $f_h(x) = \max(0, x)$ are activation functions for the output layer and the hidden layer, respectively. At the output layer, we compute the cross entropy error $CE(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_l)\log(1-\hat{y}_i)$ where $\hat{y}_i$ is the predicted score for label λ_i. We run Stochastic Gradient Descent (SGD) to train NNs with 1,000 hidden units for 50,000 epochs, which corresponds to 300,000 parameter updates, and use mini-batches of size 100 for computing gradients.

5 Experimentation

In order to estimate the performance on the public and private test set, we performed 10 fold cross validation on the available labeled training data.

Evaluation Measures The competition submissions were evaluated by computing the area under the ROC curve of the label rankings and then averaging over the instances. This measure can be defined as

$$AUC(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|P||N|} \sum_{\lambda_i \in P} \sum_{\lambda_i \in N} [[\hat{y}_i > \hat{y}_j]] + \frac{1}{2}[[\hat{y}_i = \hat{y}_j]] \tag{3}$$

where $[[x]]$ denotes the indicator function and $\hat{y}_i$ the predicted score for λ_i, e.g. the inverted ranking position. It is obvious that $1-AUC$ corresponds to the popular *ranking loss* used for evaluating multilabel classification [5]. There are several discrepancies in computing this measure, e.g. sometimes the second term is skipped and tied pairs are arbitrarily counted as wrong or correct, or sometimes test instances with an empty labelset are skipped. We compute the score for each instance, i.e. we additionally set $AUC(\emptyset, \hat{\mathbf{y}}) = 1$, but note that for the cross validation results it is easy to obtain the other less optimistic version with $AUC' = (687 \cdot AUC - 100)/587 = 1.17 \cdot AUC - 0.17$. However, this does not explain the discrepancies between the estimated AUC values and the values on the test set, since our best 0.94 would be only reduced to 0.93.

Results Table 1 shows our estimated results and the AUC values on the public and private test set. The first observation is that our preprocessing approach substantially improved the ranking quality over using the provided raw MFCC features. The achieved improvement is greater than the possible improvement by any other tried approach or combination of approaches. This demonstrates the applicability and effectiveness for acoustic data of our neural network based unsupervised feature generation process. Before heading to the comparison between the used approaches, we also note that there is an important discrepancy between the CV estimations on the training set and the test set results which cannot be explained by overfitting or differences in computing AUC (cf. Sec. 5).

We see that the pairwise LibSVM approach (SVM), the random decision trees (RDT) and the neural network (NN) with a single hidden layer obtain similar results on the test sets, with a small advantage for the NN approach. This submission obtained the 5$^{\text{th}}$ rank on the public test set and the 8$^{\text{th}}$ rank on the private test set.[4] With the arrival of the *big* dataset the last day of the competition, used for training the RDTs, and some struggling in merging teams and results, so that only the predictions of the SVMs with $\gamma = 0.5$ and the RDTs could be merged, we managed to reach the 4$^{\text{th}}$ and 6$^{\text{th}}$

[4] http://www.kaggle.com/c/multilabel-bird-species-classification-nips2013/leaderboard/public & ../private.

Table 1: Results for the different multilabel approaches and settings in terms of AUC, estimated on the training data via 10 fold cross validation (with standard deviation) or a train/test split of 600/86 and computed on the public and private test set. Training and prediction times are given in seconds. The features column indicates which feature set was used. The second block shows post-competition results.

Approach	features	CV	public	private	training	predicting
SVM $C = 2000\ \gamma = 10^{-3}$	raw	0.8994	–	–	681.8	49.76
SVM $C = 10^6\ \gamma = 0.5$	small	0.93883 ± 0.0180	0.89202	0.88967	60.18	19.39
SVM $C = 10^6\ \gamma = 1$	small	0.93915 ± 0.0179	0.89130	0.88996	60.96	19.44
RDT 50000 trees	big	0.93718 ± 0.0189	0.89129	0.88195	13444.44	290.83
NN	big	0.92595 ± 0.0200	0.89650	0.89374	5585.15 (GPU)	0.01 (GPU)
SVM & RDT	–	–	**0.90104**	**0.89525**	–	–
SVM $C = 5 \cdot 10^4\ \gamma = 0.1$	big	0.94022 ± 0.0181	0.88699	0.88710	119.71	45.98
SVM $C = 2 \cdot 10^5\ \gamma = 0.1$	big	0.93976 ± 0.0178	0.88752	0.88696	109.26	45.96
SVM & RDT & NN	–	–	**0.90331**	**0.89824**	–	–
RDT & NN	–	–	0.89903	0.89279	–	–
SVM & NN	–	–	0.89807	0.89556	–	–

positions on the public and private leaderboard, respectively. The ranking merging effect had a small impact on absolute numbers, but a considerable effect on the test set ranking due to the high competitiveness in the contest.

It seems clear that merging rankings exploits the diversity of the underlying classifiers by reinforcing predictions if the individual classifiers agree and by (tendentially) correcting rankings if for some instances some of the rankers fail. For binary decision ensembles it can be shown that the accuracy approximates 1 with increasing number of voters, though assuming a certain diversity (in the sense of probabilistic independence) [9, Sec. 4.2.1]. We could confirm this when joining predictions of classifiers of the same family, which did not lead to any improvement. But as the post-competition results in the second half of Table 1 show, combining different approaches almost always improved the AUC. Indeed, if we had submitted a joined prediction of all three approaches, we would have been ranked 4$^{\text{th}}$ on both test sets. This is just below the three competitors using advanced and sophisticated acoustic signal processing techniques relying on expert knowledge and on own processing of the raw acoustic data, as reflected by the relatively big gap between them (with AUC greater than 0.91) and the rest of the competitors.

However, please note that our best approaches with almost 0.93 AUC had roughly 10 wrongly paired labels per instance (cf. Eq. 3). It holds that this number e equals $\sum_{i=1}^{|P|} r_i - |P|(|P|+1)/2$ with r_i being the ranks of the positive labels in P, thus for examples with $|P| = 1$ the label is on average ranked at the 11$^{\text{th}}$ position, for two labels e.g. on positions 5 and 6. On the other hand, additional evaluations show that for approx. 64.6% of the instances a relevant label was predicted on the first position (one-error loss), and that we could obtain an F1-score of 75% using perfect thresholding. Remind however, that our CV results are overestimations.

6 Conclusions

We have presented a general and unsupervised approach for processing acoustic data, particularly for short recordings of birds', insects' and amphibian sounds. It is based on recent findings and state-of-the-art approaches from the field of neural networks and deep learning. The generated features, which basically are activation signals by using learned feature detectors, achieved an important improvement over using the unprocessed MFCCs in terms of AUC.

The three applied multilabel approaches, which we applied on the data, were highly suited for the particular task and carefully optimized so that we were able to obtain top results in the competition. By combining the individual approaches we could exploit the diversity among them and obtain the 4$^{\text{th}}$ rank on the public test set and 6$^{\text{th}}$ position in the final ranking. Unfortunately, we did not manage to combine all three classifiers on time, since this would have allowed us to obtain 0.90 AUC and hence the overall 4$^{\text{th}}$ rank, right after the three solutions based on expert knowledge and therefore unreachable with our means.

We see some space for improvement in the pairwise learning approach, which currently ignores examples in the label overlaps, and the neural network approach, which still has a lot of unexplored degrees of freedoms for optimizing. However, the narrow range of AUC results in the top 10 (excluding the top 3) indicates that we already got nearly the most out of the provided acoustic representation and multilabel learning. Next steps hence include to find new representations directly from the raw data, and we believe from our work with the birds sounds dataset that supervised and unsupervised techniques from neural networks and deep learning can make important contributions to this.

References

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Manual, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[2] Adam Coates and Andrew Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning*, pages 921–928, 2011.

[3] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[4] O. Dufour, T. Artières, H. Glotin, and P. Giraudet. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. In *The 1st International Workshop on Machine Learning for Bioacoustics (ICML 2013)*, pages 89–93, Atlanta, USA, june 2013. Glotin H. et al.

[5] Şeyda Ertekin and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, November 2011. ISSN 1532-4435.

[6] Johannes Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

[7] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[8] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[9] Ludmila I. Kuncheva. *Combining Pattern Classifiers : Methods and Algorithms*. Wiley-Interscience, 2004. ISBN 0471210781.

[10] Eneldo Loza Mencía and Johannes Fürnkranz. Pairwise learning of multilabel classifications with perceptrons. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN-08)*, pages 2900–2907, Hong Kong, 2008. IEEE. ISBN 978-1-4244-1821-3. doi: 10.1109/IJCNN.2008.4634206.

[11] Eneldo Loza Mencía, Sang-Hyeun Park, and Johannes Fürnkranz. Efficient voting prediction for pairwise multilabel classification. *Neurocomputing*, 73(7-9):1164 – 1176, March 2010. ISSN 0925-2312.

[12] Grigorios Tsoumakas. Mulan: A java library for multi-label learning, dataset repository. Website, January 2012. URL http://mulan.sourceforge.net/datasets.html. last accessed at 2013-12-01.

[13] Grigorios Tsoumakas, Eleftherios Spyromitros Xioufis, Jozef Vilcek, and Ioannis P. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.

[14] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[15] Xiatian Zhang, Quan Yuan, Shiwan Zhao, Wei Fan, Wentao Zheng, and Zhong Wang. Multi-label classification without the multi-label cost. In *Proceedings of the Tenth SIAM International Conference on Data Mining*, April 2010.

7.5 Ensemble logistic regression and gradient boosting classifiers for multilabel bird song classification in noise (NIPS4B challenge)

Luca Massaron
Independent marketing research director & data scientist
Verona, Italy
lucamassaron@gmail.com

Abstract

This technical report details the author's approach in the NIPS4B competition which led to a final result of an area under the ROC curve of 0.89575 in the public leaderboard and 0.89041 in the private one. The described approach involved building an ensemble of generalized linear models, such as a logistic regression and a classification model by hinge loss as provided by the Vowpal Wabbit, an open source learning system library and program based on stochastic gradient descent optimization, and boosted trees ensembles provided by Scikit-learn library in Python.

1. Description of the competition

The contest, held on the big data predictive analytics Kaggle web site (www.kaggle.com), required participants to identify which of 87 sound classes of birds (for some species the contest required to discriminate the song from the call) and their ecosystem are present in 1000 continuous wild recordings provided by the BIOTOPE society from different locations in Provence, France.

The training set contained 687 .wav files, each one featuring one or more species. Each species was overall represented by nearly 10 training files (within various context / other species). The files were recorded at a frequency sample of 44.1 kHz on an SM2 system.

The test set, matching the training set conditions, was composed of 1000 files. All species in the test set were also in the training set, posing quite an interesting discrimination challenge in distinguishing signals proper to each species.

The organizers of the competition have also provided some baseline features on the train and test .wav files. These were the optimized MFCC features, as described in the ICML4B 2013 bird challenge [1]. The format is a matrix 17xN: 17 cepstral coefficients x N frames (frame size 11.6 ms, frame shift 3.9 ms, one line per frame).

2. Data preparation

First of all the presented approach is entirely based on the original MFCC data, without the creation of further new features. The original MFCC data has just been manipulated in order to fit the training schedule of the different machine learning algorithms involved.

The MFCC matrices have been transposed in the matrix format Nx17, so that cepstral

coefficients become variable columns of the matrix and each row represented a time unit. Then, for each one of the matrices, it has been creating a "sliding window" of various sizes, from 3 to 20, horizontally stacking contiguous rows from the matrices, thus opening an "observation window" for the learning algorithm to evaluate at the same time more instants of the analyzed sounds.

Empirically the author found that the best windows to feed a linear model with were in the range of 15 to 20 horizontally stacked rows. The same observation proved true for different learning algorithms, such as gradient boosting classifiers.

Basically, with a sliding window of 20 rows, the learning algorithm had 340 variables to learn from each example (time unit).

3. Training, hyper-parameters choice

The choice was to learn a single model for each of the 87 species involved in the study, though the model choice and the parameters were generally chosen.

Therefore, the author first trained 87 first logistic (Logistic loss: L(p,y) = log(1+exp(-y*p))) and then hinge (Hinge loss: L(p,y) = max(0,1-y*p)) regression models relying on the computational speed of the open source software *Vowpal Wabbit* [2].

In order to let the learning process discriminate at best the different species, for every target species in each model the author over-weighted its instances in order that the sum of the weight of the target species was equal to the sum of the weight of the other species under analysis (a one against all approach).

As for as *Vowpal Wabbit* hyper-parameters, the best results were obtained by an 24bits hashing with 5 passes over the data. No regularization (L1/L2) has been used.

4. Predictions from single models and ensemble

The author, after estimating the probabilities of species being present in a time unit in a sound file (as for as logistic regression models using its link function, as for as hinge regression models by rescaling and clipping the results), simply averaged logistic and hinge probability results and therefore obtained a first ensemble forecast of the presence of every singular species in every row of every target transposed test MFCC matrix.

In order to turn the results relative to single time units into overall probabilities of species presence in each sound file, the author empirically experimented that using for each test matrix a moving average of 200 rows and retaining for each species the maximum probability result allowed to obtain a prediction whose public AUC was 0.87791 and its private one was 0.87120.

Noticing, by direct inspection of the fitted results on the train set and on the test results, that the estimations had surely an high recall of the species (systematically a large number of species had high scores for each test MFCC matrix, pointing out the likelihood of many false positives) but were likely lacking the necessary precision to reach higher scores on the Kaggle's leaderboard, the author decided to integrate the linear models by a different approach based on gradient boosting classifiers [3], as implemented in the Scikit-learn library [4] in Python (using the function *GradientBoostingClassifier*).

The underlying idea was that gradient boosting classifier (GBC), allowing interactions, has surely less bias than the linear models (thus an increased precision) but were suffering from an higher variance in estimates.

The ensemble approach required to create a new training dataset, resampling the initial one, in order to obtain, for each target species, all the examples of the target species itself and a 5% of examples available in each training MFCC matrix.

As for as hyper-parameters, it has been used a GBC with 30 trees, learning rate 0.1, max depth of 10 interactions and minimum sample split of 30 cases.

By itself alone, this sole model when submitted to Kaggle obtained a public AUC of 0.87779 and a private one of 0.87143, results analogous to the ensemble of logistic and hinge regression.

By examining some randomly chosen predictions from the test set, it can be observed that, as depicted in figure 1 for test sound file no. 500, an ensemble of logistic and hinge regressions tends to polarize the results in high and low probability ends and to mark many species as possibly present in the sound file.

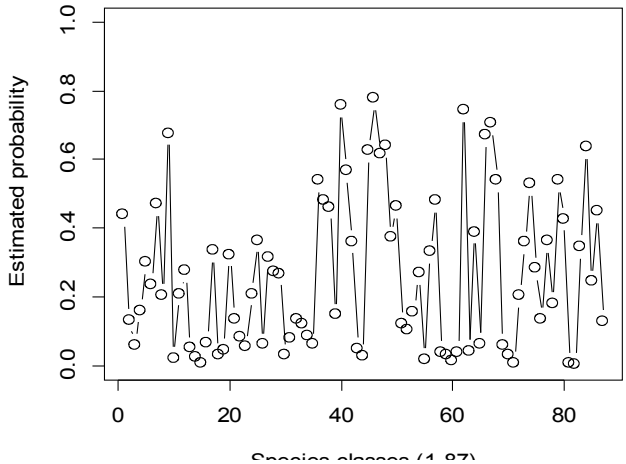


Figure 1. Estimated probabilities of test file no. 500
by an ensemble of logistic and hinge regressions

The same graphical inspection for the GBC model reveals a completely different patter, being the estimated probabilities limited in value, with spikes relative to only the most likely ones. Such a pattern, repeated all over the test file, confirms the author's expectation of the gradient boosting approach to point out only the species certainly present with an high probability and confidence, thus penalizing the recall of other species whose presence is suspected, but with equivalent certainty, confirmed.

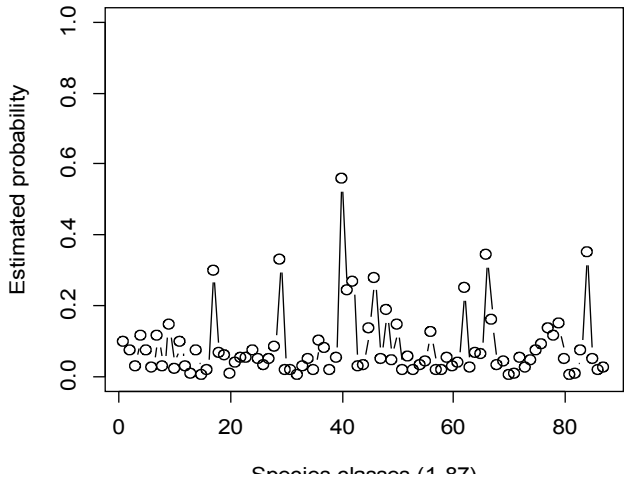


Figure 2. Estimated probabilities of test file no. 500
by a single gradient boosting classifier constituted by 30 trees

Finally, on the basis of such an insight, it has therefore been created an ensemble bringing together the results from both the averaged linear models and the gradient boosting classifier, by means of an harmonic mean, brought the final result of a public AUC of 0.89575 and a private one of 0.89041.

It is observed in figure 3 how the previously polarized predictions have naturally arranged themselves into probability tiers, allowing a better probability estimation, as for as the AUC measure.

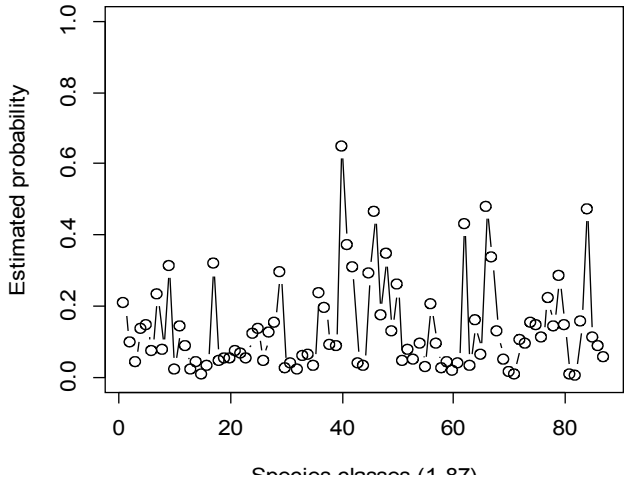


Figure 3. Estimated probabilities of test file no. 500
by an ensemble (by harmonic mean) of the previous models (logistic/hinge, GBC)

5. Reflections and open opportunities for improvement

The proposed approach highlights how an ensemble mixing high bias / low variance models and low bias / high variance ones may prove an effective strategy in bioacoustics problems. Moreover, the gradient boosting classifiers are a tree based machine learning methodology that should, in the author's opinion, better explored. The author recognizes that there are furthermore open opportunities in further tuning of models' hyper-parameters and in simplifying the ensemble strategy.

Acknowledgments

The Neural Information Processing Scaled for Bioacoustics (NIPS4B) bird song competition has been organized by BIOTOPE, ADEME, SABIOD.ORG and Prof. Hervé Glotin.

References

[1] O. Dufour, T. Artières, H. Glotin, P. Giraudet "Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification" , The 1st International Workshop on Machine Learning for Bioacoustics, LSIS, pp. 89-93, ICML 2013, Atlanta, USA, 2013

[2] Vowpal Wabbit by John Langford, Lihong Li, Alex Strehl, 2007

[3] "Boosting and Additive Trees" in T. Hastie, R. Tibshirami, J. Friedman. "Elements of Statistical Learning", 2nd ed., 2009

[4] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". Journal of Machine Learning Research, 12:2825–2830, 2011

7.6 A Novel Approach Based on Ensemble Learning to NIPS4B Challenge

Wei Chen
Institute for Infocomm Research,
Agency for Science, Technology and Research (A*STAR), Singapore
chenwei@i2r.a-star.edu.sg

Gang Zhao
School of Computing, National University of Singapore,
Singapore
zhaogang@comp.nus.edu.sg

Xiaohui Li
Institute for Infocomm Research,
Agency for Science, Technology and Research (A*STAR), Singapore
lixh@i2r.a-star.edu.sg

Abstract

Bioacoustic data science aims at analyzing and modeling animal sounds for neuroethology biodiversity assessment. The goal of competition is to automatically identify which species of bird is present in an audio recording using supervised learning. Devising effective algorithms for bird species classication is a preliminary step toward extracting useful ecological data from recordings collected in the field.

In the competition, we analyze a real-world data which contains 1000 continuous wild recordings from different places in Provence, France We identify prominent features from windowing mfccs with overlap and leverage them to build a a ensemble classifier which is a blend of different classifiers (Gradient Boosting Tree models, Random Forest models and Lasso and elastic-net regularized generalized linear model etc). Our evaluation and final private leaderboard shows that our Team DB2 method was capable of classifying large number of the bird songs, which put us on the 4th place in the final ranking.

1 Preprocessing and Feature Extraction

Our preprocessing is based on mfcc cepstral coefficients which have been proved useful for bird song recognition[1,2]. A signal is first transformed into a series of frames where each frame consists in 17 mfcc (mel cepstra feature coefficients) feature vectors, including energy. Each frame represents a short duration (e.g. 512 samples of a signal sampled at 44kHz).

Besides, we also following [1] to do the Windowing, silence removal and feature extraction step. In Final step, a reduced set of features for any remaining segment / window can be computed. In final

step, we will have each segment consists in a series of n 17-dimensional feature vectors (with n in the order of hundreds).

2 Training

Our solution consists of a blend of many single predictors. The standard way of training a single predictor consists of two steps. In the first step, the validation set is created from the training dataset and the model is trained on the remaining dataset. Then predictions for the validation set are stored. In the second step training is done on all available data with the same meta parameters as in the first step, such as number of tree in gradient boosting machine and so on. Last, the predictions for the test set are stored. We use the following algorithm for the training model:

2.1 Gradient Boosting Machine

Gradient boosting is a machine learning technique used for classification problems with a suitable loss function, which produces a final prediction model in the form of an ensemble of weak prediction decision trees[3].

2.2 RandomForest

Random forests[4] are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. We set number of tree as 500 for the configuration.

2.3 Lasso and elastic-net regularized generalized linear model

In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that combines the L1 and L2 penalties of the lasso and ridge methods.

We used the implementation of Generalized Boosted regression Models (GBM) RandomForest and Lasso and elastic-net regularized generalized linear model in R's package.

3 Blending

As blending algorithm we use a simple basic blending - linear blending. A linear blender is easy to implement. The most basic blending method is to compute the final prediction simply as the mean over all the predictions in the ensemble. Better results can be obtained, if the final prediction is given by a linear combination of the ensemble predictions. In this case, the combination coefficients have to be determined by optimization procedure, in general by regularized linear regression. In our case, All inputs (the predictions) are normalized to [0...+1]. We use linear blending approach and the coefficients are determined by cross-validation performance on average precision. For validation we use 10-fold cross-validation. The performance of each individual model and blending is shown in table 1.

Table 1: Performance for different algorithm

Method	ROC (public)	ROC (private)
GBM	88.762%	88.732%
RandomForest	88.850%	88.239%
GLMNET	86.079%	86.332%
Blending	89.740%	89.624%

Besides the overall performance, we also try to investigate the performance of individual classification performance, we found that class 38/57/80 is easy to predict with ACU more than 98 % and the class 20/66/78/85 is the difficult class to predict. We suspect that it is because the data sparsity problem which hurt the performance of classifier.

4 Conclusion

We described our approach to NIPS4B challenge. In summary, we try extract the feature using MFCC with time overlap from various statistics such as velocity acceleration etc. With the feature extraction, we building our effective classifiers using Gradient boosting machine, RandomForest and Lasso and elastic-net regularized generalized linear model. Each single predictor is trained individually. A linear blending of single predictor is used for final prediction. The experiments presented in this paper, and the ranking on the Privateleaderboard[1] , suggests that our methods are effective in multi label and multi instance classification tasks.

Acknowledgments

I would like to thank the BIOTOPE society for collecting, labeling and preparing the dataset, and SABIOD-LSIS with ADEME for having organized the great competition under the supervision of Pr. Glotin. We would also thanks other competitors who make this competition intresting.

References

[1] Dufour, O. and Artières, T. and Glotin, H. and Giraudet, P. Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification *The 1st International Workshop onf Machine Learning for Bioacoustics (ICML 2013)*

[2] Briggs, F.; Raich, R.; Fern, X.Z., "Audio Classification of Bird Species: A Statistical Manifold Approach," *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference*

[3] Jerome H. Friedman. 2002. Stochastic gradient boosting *Comput. Stat. Data Anal*

[4] Leo Breiman. 2001. Random Forests *Mach. Learn.*

[1] http://www.kaggle.com/c/multilabel-bird-species-classification-nips2013

Chapter 8

Whale Song Clustering

8.1 Analyzing the Temporal Structure of Sound Production Modes within Humpback Whale Sound Sequences[1]

Eduardo Mercado III
Department of Psychology
University at Buffalo, SUNY
Buffalo, NY 14031
emiii@buffalo.edu

Abstract

Past analyses of humpback whale song have often focused on classifying the sequential structure of discrete sound types. An alternative approach is to analyze the temporal dynamics of sound production by singing whales. Temporal variations in sound production modes within and across individual sounds can by quantified in terms of duty-cycle measures. Low duty-cycle values correspond to pulse trains, medium duty-cycle values correspond to tonal sounds, and high duty-cycles correspond to higher-pitched sounds. The temporal structure within song sessions can then be analyzed by performing time-frequency analyses of sequences of duty cycle measures. Advantages of this approach include that it readily accommodates graded variations within and across individual sounds, it does not involve arbitrary/subjective criteria for sorting sounds, and it reveals modulation in song structure that would be lost in analyses of symbolic sequences.

1 Introduction

The first step in most past structural analyses of humpback whale songs has been to sort individual sounds into discrete types (typically based on subjective visual or aural criteria). Even when automated sorting techniques such as self-organizing maps have been used to sort individual sounds [1], the success of these quantitative methods has been judged relative to how human observers sort the same sounds. This approach is highly problematic because many sounds that humans judge to be different might be perceptually equivalent for humpback whales and sounds that humans judge to be highly

similar might be completely different from a whale's perspective. Consequently, analyses of song structure based on such discrete classification approaches likely reveal more about human perception than they do about the actual structural features that whales are producing and receiving.

If the individual sounds within humpback whale songs were stereotypically produced, such that the acoustic structure of each sound could be matched to a specific acoustic template, then mismatches between human and humpback perception would be largely irrelevant to identifying structure within songs. For instance, birds most likely do not perceive their songs in the same way as humans, but because the notes they use within their songs are often quite distinctively different, one can be confident that structural organization is being accurately described by symbolic sequences representing the notes within songs. In contrast, humpback whales are continuously morphing the features of individual sounds that they produce within songs along multiple acoustic dimensions [2], to the extent that the sound repertoire that a single whale uses in one year can differ considerably from the sounds it was using five years earlier.

The kinds of acoustic transformations present within humpback whale songs are similar in certain respects to modulations present within human music. Analytical approaches commonly used to analyze musical structure suggest one way of handling the graded acoustical transitions present within humpback whale songs [3-4]. Specifically, analyses of the temporal structure within a sequence can reveal sequential periodicities, even when the individual sounds giving rise to regularities are constantly changing.

The individual sounds produced by singing whales can all be described as the result of variations in the vibrating modes of paired membranes [2]. A major factor that determines the aural qualities of any particular sound is the rate of vibration. Because the vibrating membranes act as a relaxation oscillator, low-rate vibrations generate more pulsatile sounds, whereas higher-rate vibrations tend to produce more sinusoidal sounds. Consequently, although the sounds fall along a graded continuum of different vibration rates, sounds at the endpoints of this continuum sound dramatically different and also will appear qualitatively different in time-frequency representations. Using measures of waveform features that directly reflect production mechanisms (e.g., vibration rate) provides a way to avoid the subjective biases that can be introduced by the nature of this continuum. Duty-cycle measures provide a convenient way of quantifying how sinusoidal versus pulsive a signal is [2], providing a way to convert produced oscillations into points along a continuum. Temporal structure analyses such as those applied to musical recordings can then be used to analyze the structural properties within sequences of measures of whale-generated vibration modes.

2 Methods

The analysis presented here serves to illustrate how temporal structural analysis can be applied to a recording of a singing whale. To get precise measures of production-related features, it is important that the recording be made from relatively close to the singer. The 26-min recording analyzed in this study was collected from a short distance, with minimal background noise (media/NIPS4B_Humpback_Darewin_LaReunion_Jul_03_2013-

001_26min.wav).

The time-domain waveform was analyzed in its raw form, with no pre-processing. The recording was segmented into sequential 100 ms duration frames and each frame was converted into a single duty-cycle measure. Duty-cycle was calculated regardless of whether the frame corresponded to a sound produced by the singer or to a silent interval between sounds (for details of duty-cycle calculation, see [2,5]). The duty-cycle measure provides a ratio scale measurement in which zero corresponds to no deflections. Trains of pulses produce a lower duty-cyle value (near 0) and more sinusoidal signals produce a higher value (1 for a perfect sinusoid). This measure has previously been used to analyze the vocalizations of false killer whales [5] and singing humpback whales [2]. The sequence of duty-cycle (DC) measures is referred to hereafter as a DC-gram. Conversion of the 26 min recording into a DC-gram (15410 elements, 32 kB file) took ~10 s using Matlab R2010b on a 3.1 GHZ Apple iMac.

3 Results

The DC-gram is similar to the envelope of the waveform, but provides additional information, such as evidence of dynamic transitions between sound production modes within individual sound units, even when signal amplitude remains constant (Fig. 1). Although the DC-gram consists only of positive values between 0 and 1, it can still be analyzed using standard time-frequency representations. Here, the DC-gram was analyzed by transforming it into a spectrogram (Fig 2). A spectrographic analysis of a DC-gram produces a representation comparable to a rhythm spectrogram (e.g., see [2,6]). Unlike the rhythm spectrogram, however, the spectrogram of the DC-gram reveals temporal patterns in production modes rather than (or in addition to) temporal patterns in amplitude or pitch. Consequently, it is more sensitive to modulations that a singer is controlling.

The Figure 2 shows transitions in the temporal structure within the 26 min recording of a singing whale. The spectrogram shows that overlaying a relatively constant .5 Hz rate of sound generation, the singer is modulating the timing/modes of its sound production in stereotypical ways.

4 Discussion

Temporal structure is aurally evident in all recordings of singing humpback whales, yet few attempts have been made to measure this structure (however, see [7-8]). Instead researchers have focused on analyzing subjectively salient, repeating sound patterns. Analyses of apparent hierarchical structure within sequences of subjectively categorized sound types may bear little relation to the informational structures that are produced and used by humpback whales [9]. The current analysis suggests that more objective methods of analyzing song structure are feasible and may reveal previously unsuspected temporal dynamics within humpback whale sound sequences.

Ultimately, the criteria for a successful structural analysis depend on the purpose of the analysis. If the goal is to systematically describe the vocal behavior of humpback whales in a particular year and locale, then

classification strategies that focus on clustering individual units (or subunits, or phrases) into discrete types clearly provide a useful way of doing this [10]. If, however, the goal is to identify functionally relevant structure within the sound sequences produced by humpback whales, then it is important to keep in mind that the information bearing properties of elements within these sequences remain unknown.

Figure 1.(top) DC-gram of the first minute of the 26 min recording. Each frame corresponds to a 100 ms segment of the recording. Here, duty-cycles above ~0.2 reflect energy within units. The DC-gram shows that production modes vary more in some sounds than others (unimodal peaks reflect continuous gradations in production), that the overall rate of sound production is quite stable, and that the duration of units and intervening intervals are also stable (at least during this 1 min segment).

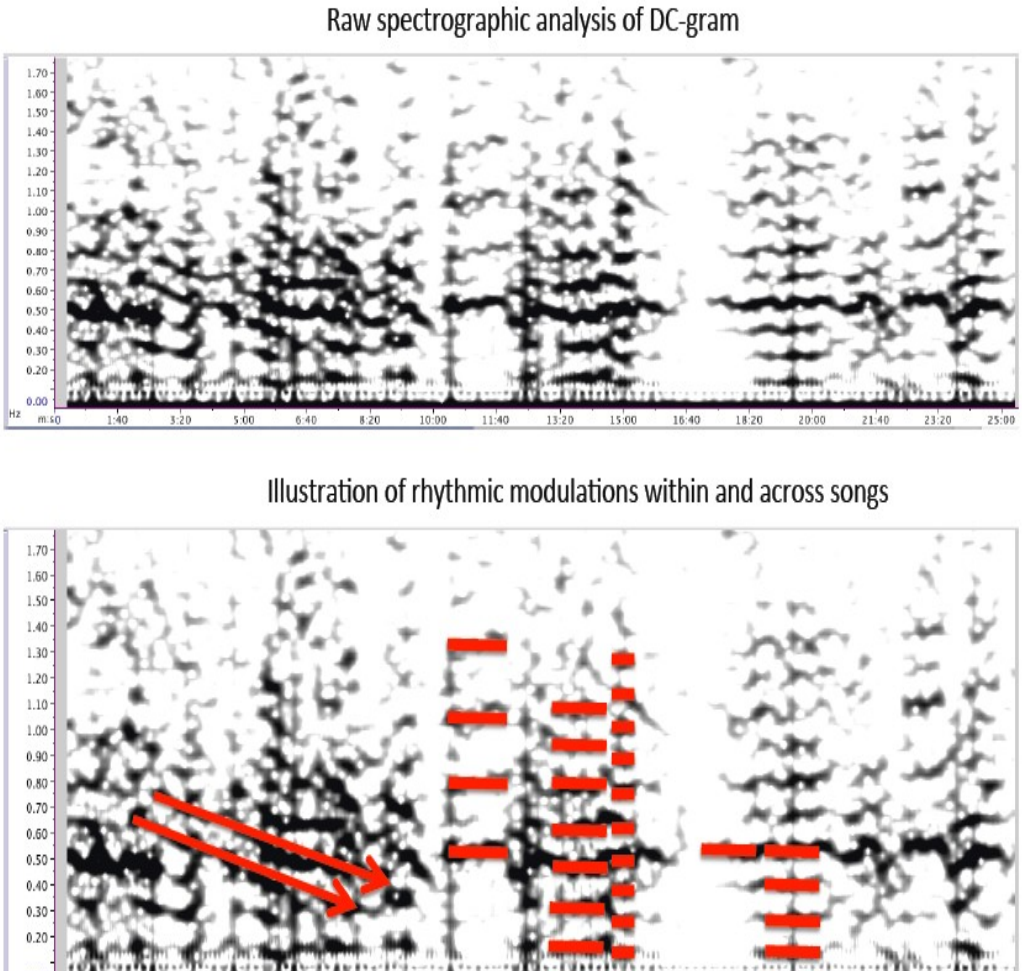


Figure 2. Spectrographic DC-gram of a 26 min song bout. In the first 9 min of singing, the whale slows down the rhythm of sound production, while maintaining production rate (~.5 Hz). After 10 min, the whale shifts to a rhythm that closely matches the rate of sound production. Near 12 min, the whale transitions through several rhythms before settling into a new mode near the 13 min mark. Interestingly, the time spent in this new mode matches that spent in the rate-synchronized mode; later, near the 17 min mark, the whale repeats this temporal pattern, again producing each mode for ~1 min.

Acknowledgments

I thank S. Handel and H. Glotin for providing useful feedback on an earlier version of this manuscript.

References

[1] R. Suzuki, J. R. Buck, and P. L. Tyack, "Information entropy of humpback whale songs," J. Acoust. Soc. Am. 119, 1849-1866 (2006).

[2] E. Mercado, III, J. N. Schneider, A. A. Pack, and L. M. Herman, "Sound production by singing humpback whales," J. Acoust. Soc. Am. 127, 2678-2691 (2010).

[3] J. Paulus, M. Müller, and A. Klapur, "Audio-based music structure analysis," Proceedings of the International Society for Music Information Retrieval, 625-636 (2010).

[4] R. J. Weiss, and J. P. Bello, "Unsupervised discovery of temporal structure in music," IEEE Journal of Selected Topics in Signal Processing 5, 1240-1251 (2011).

[5] S. O. Murray, E. Mercado, III, and H. L. Roitblat, "Characterizing the graded structure of false killer whale (*Pseudoorca crassidens*) vocalizations," J. Acoust. Soc. Am. 104, 1680-1688 (1998).

[6] S. Saar, and P. P. Mitra, "A technique for characterizing the development of rhythms in bird song," PLoS 3, e1461 (2008).

[7] S. Handel, S. K. Todd, and A. M. Zoidis, "Rhythmic structure in humpback whale (*Megaptera novaeangliae*) songs: preliminary implications for song production and perception," J. Acoust. Soc. Am. 125, EL225-230 (2009).

[8] S. Handel, S. K. Todd, and A. M. Zoidis, "Hierarchical and rhythmic organization in the songs of humpback whales (*Megaptera novaeangliae*)," Bioacoustics 21, 141-156 (2012).

[9] E. Mercado, III, and S. Handel, "Understanding the structure of humpback whale songs," J. Acoust. Soc. Am. 132, 2947-2950 (2012).

[10] F. Pace, F. Benard, H. Glotin, O. Adam, and P. White, "Subunit definition and analysis for humpback whale call classification," Appl. Acoust. 71, 1107-1112 (2010).

8.2 Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture

Marius Bartcus[1], Faicel Chamroukhi[1,2], Joseph Razik[1,2], and Hervé Glotin[1,2,3]

[1]Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
[2]Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
[3]Institut Universitaire de France, France

Abstract

In this work we propose to extend the finite parsimonious Gaussian mixture to the infinite case so that the classification of our data could be performed in one stage. We implemented the eigenvalue decomposition of the covariance matrix of each cluster to the Infinite Gaussian mixture model and made it parsimonious. We developed an MCMC algorithm (Gibbs sampling) to learn the various models and we named this approach the bayesian non-parametric parsimonious approach for cluster analysis. The new approach will be more flexible in terms of modeling and will automatically provide the partition of the data and the number of clusters. This approach will be applied into the challenging problem of Whale song decomposition NIPS4B challenge. These algorithms would also give efficient clustering on complex sequence of pulses, and then may allow muti-source/multi-animals labelling.

1 Introduction

Clustering is one of the essential tasks in machine learning and statistics. One of the main problem in data analysis is to estimate the number of clusters that fits best the data. For that we find different approaches in the literature, where one of the most popular is the model-based clustering [1, 2]. These finite parsimonious Gaussian mixtures rely on the eigenvalue decomposition of the covariance matrix, allowing the models to change between the simplest spherical one to the more general [3]. The model parameters can be estimated in a Maximum Likelihood (ML) framework by the Expectation Maximization (EM) algorithm [4] or in a Maximum A Posteriori estimation (MAP) [5] framework or by using MCMC sampling techniques[6, 7]. In this approach, as well as in standard model-based clustering techniques, the selection of the number of clusters is performed by using penalized likelihood criteria such as the Bayesian Information Criteria (BIC) [8], Akaike Information Criterion [9], Integrated Classification Likelihood (ICL)[10], etc. So we need to perform a two stages for classification, first estimate the number of clusters and then run th EM algorithm for the classification of the data.

An alternative well-principled approach for the difficult problem of model selection is to use the Bayesian Non-Parametric (BNP) [11] methods for clustering, one of them being the infinite Gaussian mixture model (IGMM) [12]. Indeed, the principle of IGMM is based on the one of the Chinese Restaurant Process (CRP) [13, 14, 15, 16, 17] which is well suited to the problem of non-parametric clustering. This alternative gives us the possibility to obtain the number of clusters in the same stage of clustering so that as the new data will be observed the number of model parameters can be changed. The general (full GMM) model used in IGMM is not so flexible as in the case of the model-based clustering [3, 18] where the covariance matrix can take different forms, depending on the volume shape and orientation. Therefore we proposed to develop a new approach that will rest being an infinite Gaussian mixture model approach that will give us the possibility to automatically

provide the number of classes but know with an eigenvalue decomposition of the covariance matrix giving more flexibility for the model.

The paper is organized as follows. Section 2 briefly discusses previous work on finite Gaussian mixture clustering, in particular we show the model-based clustering approach. Then, Section 3, presents the proposed approach and Section 4 shows experiment results after application to the Whale song decomposition NIPS4B challenge of the EM algorithm with ML and MAP frameworks and the proposed bayesian non-parametric parsimonious approach.

2 Parametric parsimonious Gaussian clustering

It is supposed that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a sample of n i.i.d observations in $\mathbb{R}^d$, and $\mathbf{z} = (z_1, \ldots, z_n)$ is the corresponding unknown cluster labels where $z_i \in \{1, \ldots, K\}$ represents the cluster label of the ith data point $\mathbf{x}_i$, K being the possibly unknown number of clusters.

In the model-based clustering [1, 2, 5] the data X is proposed to be generated from a mixture model with the density:

$$p(\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \theta_k) \tag{1}$$

having f_k a distribution with parameters θ_k and the non-negative mixing proportions π_k that sum to one.

We will suppose in particular the multivariate Gaussian Mixture Model (GMM) [1] to cluster the data X so that in this case we have f_k being a multivariate Gaussian distribution (equation 2) with the parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ which are respectively the mean vector and the covariance matrix for the kth Gaussian component density.

$$f_k(\mathbf{x}_i|\theta_k) = \mathcal{N}_k(\mathbf{x}_i|\mu_k, \Sigma_k) \equiv 2\pi|\Sigma_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right\} \tag{2}$$

The finite parsimonious GMM by the eigenvalue decomposition of the covariance matrix makes the model more flexible, giving a possibility to variate each cluster density by volume, orientation and shape. The parametrization of the covariance matrix is given in equation 3.

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{3}$$

where λ_k is a scalar that defines the volume, $\mathbf{D}_k$ a orthogonal matrix that defines the orientation and $\mathbf{A}_k$ is a diagonal matrix with determinant 1 witch defines the shape. This decomposition leads to fourteen flexible models [3] going from simplest spherical models to the complex general one.

One of the most used algorithm for learning the model is the Expectation Maximization(EM) algorithm that maximizes the likelihood [19, 20] is an iterative algorithm consisting of two stages, the expectation of the complete data log-likelihood named the E-step and the maximization of the expected complete data log-likelihood named the M-step. Maximizing the likelihood (ML framework) will maximize the mixture likelihood $p(\mathbf{X}|\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$.

$$p(\mathbf{X}|\pi_k, \boldsymbol{\mu}_k, \Sigma_k) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

The maximizing of the posteriori (MAP framework) can be also performed by the EM algorithm [5]. It leads by adding a prior to the mixtures parameters so that it maximizes the following posterior parameter distribution $p(\boldsymbol{\theta}|\mathbf{X})$

$$p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$$

where $p(\boldsymbol{\theta}) = p(\Sigma)p(\boldsymbol{\mu})$ is the prior for the parameters of the mixture. Also we find in the literature different extension of the EM algorithm like CEM, GEM, etc. that could also be used to learn the model. Another alternative to learn the models are the Markov Chain Monte Carlo (MCMC) algorithms (like Gibbs sampling) [7, 21, 22].

However before learning the model with one of these finite gaussian mixture model we must have the answer to what is the number of mixtures in our model. For that we pose K_{max} that is a maximum number of cluster possible and we compute the penalized log-likelihood criteria (BIC, AIC, ICL, etc.) After choosing the optimal number of clusters that fit best the data we can run one of the learning algorithms.

3 Bayesian non-parametric parsimonious clustering

First off all we make attention that the term of non-parametric learning does't mean at all that the model doesn't have parameters, indeed it means that it could have an infinite number of them as the data grows, in other words it is assumed that the observed data are governed by an infinite number of clusters, but only a finite number of them does actually generates the data. Bayesian non-parametric (BNP) mixtures for clustering offers a good alternative to infer the number of clusters form data within one stage, rather then in two stages like in the case of the parametric modeling [11, 23, 24, 12]. BNP approach proposes to pose a prior on an infinite partitions in such a way that a finite number of clusters will be active. We could use the Chinese Restaurant Process (CRP) prior [25, 26, 23] or a Dirichlet Process Mixture (DPM) [27, 23, 28].

In this work we proposed to develop the previous work called the infinite Gaussian mixture model [12], based on the full GMM, by extending it to a more flexible mixture model where the covariance matrix has an eigenvalue decomposition [3, 18]. We call the new approach the bayesian non-parametric parsimonious approach. We assumed the Chinese Restaurant Process (CRP) prior for the cluster assignments.

Indeed CRP provides a distribution on the infinite partitions of the data, that is a distribution over the positive integers $1, \ldots, n$. Considering the following joint distribution of the unknown cluster assignments: $p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2) \ldots p(z_n|z_1, z_2, \ldots, z_{n-1})$ we can compute each term by using the CRP distribution. The problem of the Chinese Restaurant Process can be expressed by a real human situation if supposing a restaurant that could be extended in a real time by having the possibility to add an infinite number of tables if the number of customers grows. So the CRP is explained as follows: supposing we have this kind of restaurant where one customer is visiting it. This customer enters and sits at the first table. When the second customers enters the restaurant he will sit with a probability $\frac{1}{1+\alpha}$ to the first table and with probability $\frac{\alpha}{1+\alpha}$ to the second table where α will be a dispersion parameter. Going future we say that the n-th customer will be sitting at a new table with a probability equal to $\frac{\alpha}{n-1+\alpha}$ or at the table k with the probability $\frac{n_k}{n-1+\alpha}$ where n_k is the number of customers sitting at table k. The idea of this model is that humans adaptively learn the number of categories of their observations. In the clustering problem the customers are the the observations, so that the new observations can enter the clustering method and choices the table meaning the cluster. This can be explicitly formulated as follows

$$p(z_i = k|z_1, ..., z_{i-1}) = \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_+ \end{cases} \tag{4}$$

where K_+ is the number of tables that have customers sitting on that table $n_k > 0$ or it is also known as active classes. We note $k \leq K_+$ when the k-th table is occupied or in clustering problem the new data observed will be associated to the k-th cluster and $k > K_+$ when a new table will be occupied or the new observation will form a new cluster.

It is also used a prior for the mixture parameters as in MAP approach or the MCMC Gibbs sampling. This priors are used to be conjugate priors so that for example we have the normal inverse-Wishart prior distribution for the mean and the covariance matrix if we use a full GMM. We note this prior distribution as G_0 so that we can show the following generative process.

$$\boldsymbol{\theta}_i \sim G_0 \tag{5}$$

$$z_i \sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \tag{6}$$

$$\mathbf{x}_i \sim p(.|\boldsymbol{\theta}_{z_i}). \tag{7}$$

According to this generative process we see that $\boldsymbol{\theta}_i$ exhibit a clustering property so that the unique values of the parameters are the number of mixtures that fits the data. G_0 is called the base distribution [27, 23]. The distribution over the partition z_i as it was talked before is a CRP distribution. We proposed to develop the infinite parsimonious Gaussian mixture, where the covariance matrix is parameterized in term of eigenvalue decomposition to provide more flexibility of this model. So the priors on the parameters depends on the type of the parsimonious model. Having chosen the MCMC Gibbs sampling [12, 29, 16, 23] for learning the model we will have different sampling depending on the covariance matrix decomposition.

Indeed, yet we investigated seven parsimonious models, covering the three families of the mixture models which are the general, the diagonal and the spherical family. The parsimonious models

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

therefore go from the simplest spherical one to the more general full model. In table 1, we summarize the considered models and the corresponding prior for each model used in Gibbs sampling.

Nr.	Decomposition	Model-Type	Prior	Applied to
1	$\lambda\mathbf{I}$	Spherical	$\mathcal{IG}$	λ
2	$\lambda_k\mathbf{I}$	Spherical	$\mathcal{IG}$	λ_k
3	$\lambda\mathbf{B}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{B}$
4	$\lambda_k\mathbf{B}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{B}$
5	$\lambda\mathbf{DAD}^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
6	$\lambda_k\mathbf{DAD}^T$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$
7	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Table 1: Considered Parsimonious GMMs via eigenvalue decomposition and the associated prior distribution for the covariance. Note that $\mathcal{I}$ means that it is an inverse distribution, $\mathcal{G}$ means that it is a Gamma distribution and $\mathcal{W}$ means that it is a Wishart distribution.

4 Experiments

We compared our Bayesian non-parametric parsimonious mixture with model-based clustering (ML-based and MAP-based) approaches. For the ML and MAP approaches, we used the EM algorithm to estimate the model parameters. The model selection is performed by ICL for values of K between 1 and 60. For each value of K, we considered 10 runs of EM, with different initializations, to estimate the mixture model parameters and the one providing the best solution (corresponding to the maximum value of the log-likelihood is selected). Then, the value of K corresponding to the highest ICL value is considered as the best solution with the optimal number of clusters.

For the Bayesian non-parametric approach (IGMM), we used the Gibbs sampler by running it ten times and selecting the best solution in the sense of the posterior.

We illustrate the estimations of the number of classes for Gibbs samplings for 2 spherical models $\lambda\mathbf{I}$ and $\lambda_k\mathbf{I}$, 2 diagonal models $\lambda\mathbf{B}$ and $\lambda_k\mathbf{B}$ and two general models $\lambda_k\mathbf{DAD}^T$ and $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ in the histograms of figure 1. Note that we dont take in consideration the first 50 iterations of the Gibbs sampling. For this whale song data we can conclude that for these models we have been estimated a different number of clusters, that could be compared when estimating the number of clusters by using the information criteria.

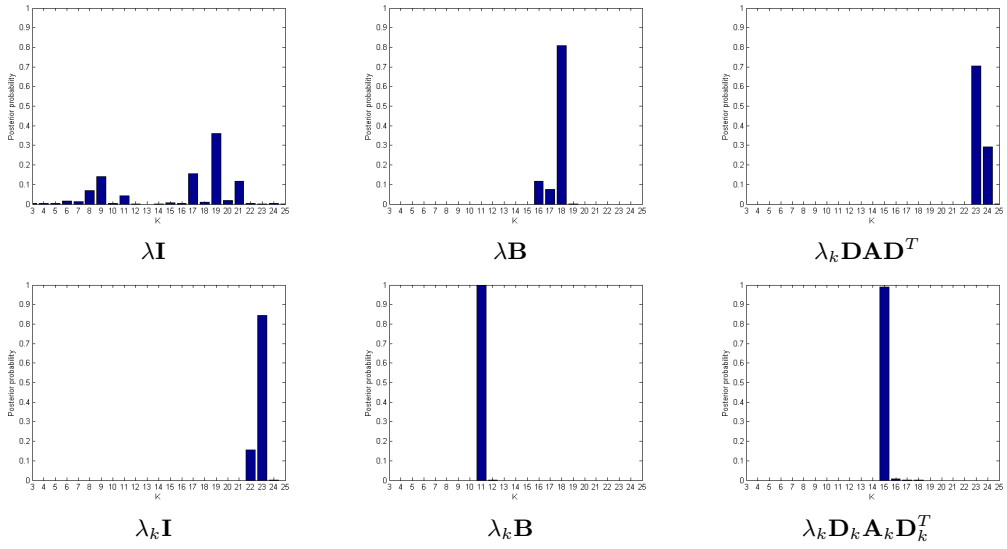


Figure 1: Posterior distribution of the number of clusters obtained by the proposed bayesian non-parametric approach.

The table 2 shows the log-likelihood values that are divided by 10^6 and the number of estimated classes obtained by using the Expectation Maximization (EM) algorithm with one of the information criteria and the proposed bayesian non-parametric parsimonious method for clustering the data. By analysing the results we can conclude that the best solution is by using the more general model with the eigenvalues decomposition of the covariance matrix $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$, meaning that the volume, the orientation and the shape can vary for each cluster. The best likelihood obtained here is by using the EM with maximum a posteriori framework algorithm, that estimates 18 classes. On the other hand, the bayesian non-parametric model estimates 15 classes. By using the spherical models, the one with the equal volumes $\lambda \mathbf{I}$ and the other one with different volumes $\lambda_k \mathbf{I}$, we notice that the estimation of classes are taken to be the maximum, equal to 60, when using the finite Gaussian Mixture Models (GMM), while for the infinite case we have estimated 9 classes for the $\lambda \mathbf{I}$ model and 23 classes for the $\lambda_k \mathbf{I}$ model. Also, for the diagonal models, we have the model with equal volumes $\lambda \mathbf{B}$ that estimates 22 classes for the finite mixture models when using the EM ML approach or EM MAP approach with the Integrated Classification Likelihood (ICL) criteria, and 18 classes when using the proposed non-parametric bayesian clustering. [1]

Table 2: Log-likelihood values (divided by 10^6) and the number of estimated classes obtained for the whale song data set by using the Expectation Maximization approach with maximization of the likelihood (ML) approach and with the maximization a posteriori (MAP) approach and the proposed bayesian parsimonious approach (IPGMM).

	EM ML		EM MAP		IPGMM	
Model	$\hat{K}$	log-lik	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda \mathbf{I}$	60	-2.2198	60	-2.1924	9	-2.3413
$\lambda_k \mathbf{I}$	60	-2.1129	60	-2.0858	23	-2.2133
$\lambda \mathbf{B}$	22	-2.1435	22	-2.1339	18	-2.1958
$\lambda_k \mathbf{B}$	59	-2.0059	53	-1.9595	11	-2.1900
$\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T$	-	-	34	-2.0815	33	-2.1695
$\lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T$	51	-1.9811	-	-	24	-2.1589
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	19	-1.9418	18	$\mathbf{-1.9381}$	15	-2.1234

In the figure 2 we show the spectrograms of the whale songs obtained with the proposed bayesian non-parametric approach with the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$. We chose to show these spectrograms of the whale songs because we obtained the best log-likelihood solution when using the new method. On the vertical axes the frequency is showed and on the horizontal axes we have the frames, each frame being represented by 10 ms. As we observe in the table 2 we have 15 clusters for the $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ model when using the infinite Gaussian mixture model, so in figure 2 we show the 6 spectrograms of the whale songs that the time repass 10 ms.

By classification the whale song data with the infinite gaussian mixture model using the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ we see in the figure 3 the song that where observed for each observation. The songs (classes) 8, 12 and 15 are uniformly activated in time, therefore we may figure out that they are representing the sea noise. Whereas the songs (classes) 10,13 and 14 are clearly conveying information (low entropy).

5 Conclusion

This work presents a new Bayesian non-parametric approach for clustering. It is based on an infinite Gaussian mixture with an eigenvalue decomposition of the cluster covariance matrix and a Chinese Restaurant Process prior. It allows deriving several flexible models and avoids the problem of model selection in maximum likelihood-based and Bayesian parametric Gaussian mixture. We applied this method on the Whale song decomposition NIPS4B challenge. The obtaining results highlight the interest of using parsimonious Bayesian clustering as a good alternative namely to finite parsimonious GMM clustering. We saw that the infinite parsimonious Gaussian mixture model (IPGMM) is

[1] The missing values for the two state of art models ($\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T$ model for EM ML and the $\lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T$ model for EM MAP) are due to some trobles when executing the em algorithm for this data and are currently being fixed.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

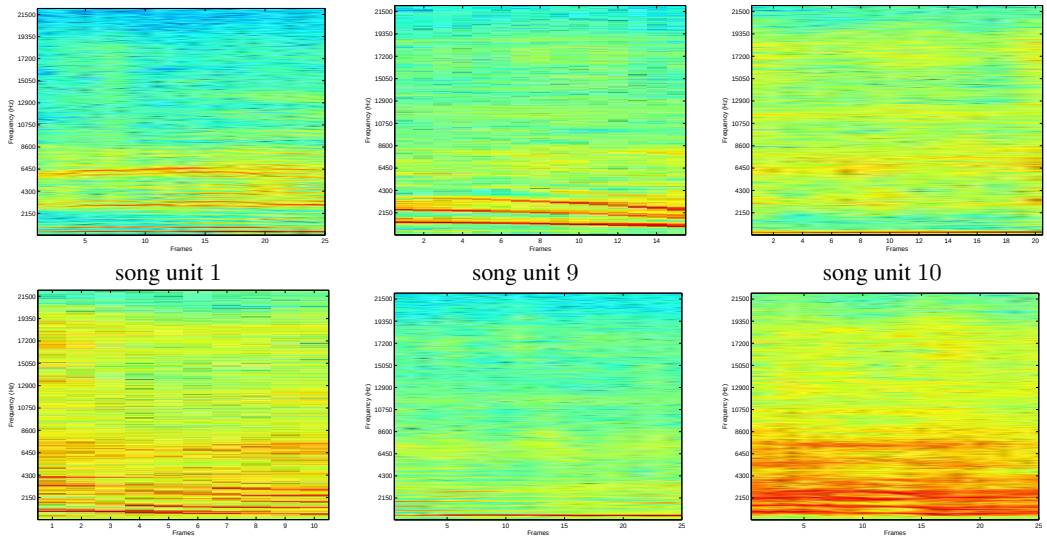


Figure 2: Spectrograms for the whale songs obtained with the proposed bayesian non-parametric approach with the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$.

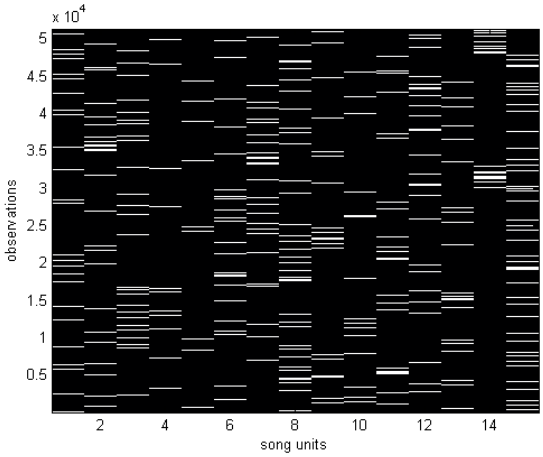


Figure 3: Clusters activities versus time sea noise obtained by IPGMM with $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ model

more flexible in terms of modeling and automatically provides a partition of the data and the number of clusters for the data needed to be clustered.

References

[1] G. J. McLachlan and D. Peel. *Finite mixture models*. New York: Wiley, 2000.

[2] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[3] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[4] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3):315–332, October 1992.

[5] C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, september 2007.

In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

[6] Halima Bensmail. *Regularized models in discrimination and Bayesian classification*. PhD thesis, University Paris 6, 1995.

[7] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

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

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

[10] 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.

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

[12] C. Rasmussen. The infinite gaussian mixture model. *Advances in neuronal Information Processing Systems*, 10:554 – 560, 2000.

[13] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Phd. thesis, MIT, Cambridge, MA, 2009.

[14] D. Görür. *Nonparametric Bayesian Discrete Latent Variable Methods for Unsupervised Learning*. Phd. thesis, Berlin Institute of Technology, Berlin, D83, 2007.

[15] Xiaodong Yu. Gibbs sampling methods for dirichlet process mixture model: Technical details. Technical report, University of Maryland, College Park, University of Maryland, College Park, September 2009.

[16] F. Wood, Thomas 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.

[17] F. Wood and M. J. Black. A nonparametric bayesian alternative to spike sorting. *Journal of neuroscience methods*, 173(1):1–12, 2008.

[18] Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[19] 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.

[20] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. New York: Wiley, 1997.

[21] H. Bensmail and J. J. Meulman. Model-based clustering with noise: Bayesian inference and estimation. *J. Classification*, 20(1):049–076, 2003.

[22] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9(4):639–650, 1998.

[23] J. Gershman Samuel and David M. Blei. A tutorial on bayesian non-parametric model. *Journal of Mathematical Psychology*, 56:1–12, 2012.

[24] Erik B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[25] D.J. Aldous. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Math. 1117.

[26] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. of Statistics. UC, Berkeley, 2002.

[27] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[28] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[29] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

8.3 Classifying Humpback whale sound units by their vocal physiology, including chaotic features

Dorian Cazau
Institut Jean Le Rond d'Alembert
University UPMC Paris 6, CNRS UMR 7190
Equipe Lutheries Acoustique Musicale (LAM)
cazaudorian@aol.com

Olivier Adam*
Institut Jean Le Rond d'Alembert
University UPMC Paris 6, CNRS UMR 7190
Equipe Lutheries Acoustique Musicale (LAM)
olivieradam@upmc.fr

Abstract

Following a production-based approach, this paper proposes a new kind of representations of humpback whale songs. Simple acoustic descriptors are used to characterize specific features of vocal sounds (e.g. fundamental frequency, formants, chaos), which can be traced back to their vocal mechanisms. Such representation of songs allow to interpret acoustic features of vocalizations in terms the characteristics of vocal organs. It may be a very useful tool for researchers dealing with the acoustic behavior of humpback whales.

1 Introduction

The vocal repertoire of humpback whales (*Megaptera novaeangliae*) ranges widely, with a great variety of bandwidths, durations and intensities of the emitted sounds. More specifically, the vocal diversity of humpback whales includes acoustic features such as harmonic sounds with a huge fundamental frequency range, noise-like sounds, formant structures [1], pulse-like sound units and various non-linearities (e.g. frequency-jumps). Scientists have particularly been interested in the complex stereotyped songs that male individuals of humpback whales, one of the most studied species of mysticetes, emit during the winter-spring breeding season. One major topic of research when analysing these songs is the characterization and classification of their constitutive sounds. Two main approaches can be distinguished in this task. [2] proposed a famous hierarchical framework (songs - themes - phrases - sub-phrases - sound units), in which the temporal structure of these songs has been longly studied. The method used is to perform spectrogram analysis to determine salient acoustic features characterizing discrete sound patterns, which are further used to build an organized structure in regards to their relative presence/prevalence within a song. Numerous studies (see review in [3]) have followed this approach, including [4] who studied temporal song evolution and [5] who proposed a manual categorization of humpback whale sounds. A second approach is the use of machine learning methods, with the obvious advantages of building objective, automatic, time-saving tools of analysis for humpback whale songs. Baseline surveys include [6] who used

*Also at Centre de Neurosciences Paris Sud - CNRS UMR 8195 - The Bioacoustics Team - 91400 Orsay France

information theory techniques to study song structure, [7] who developed a cluster-based learning to classify sound units.

Recently, [8] have called for the rejection of [2]'s framework, prompted by the involvement of human subjectivity in the characterization of information-carrying vocal sounds whose semantic and syntax are unknown to humans. Especially, how can you tell if this hierarchical descriptive structure is relevant for the whales when producing their sounds? A similar critic could be formulated for the machine learning technique, whose representations of humpback whale songs do not directly provide an understandable insight in their acoustic communication.

To deal with this issue, some authors [9, 10, 8, 11] have proposed to adopt a production-based approach by analysing the factors involved in the overall sound production system of humpback whales. This approach consists in the study of the whale sound producing anatomy in order to evaluate the acoustic characteristics of potential vocal organs (e.g., the fundamental frequency range of a sound generator, or the maximal airflow [1] amplitude), but also of other constraint-like factors that are not directly "controlled" by the whale, being either physical (e.g., composition of internal respiratory gases) or environmental (e.g., depth-related ambient pressure). Humpback whale vocal production should reflect the physical interaction between all these factors, through for example the temporal sequencing or the different acoustic contrasting types of sound patterns. Therefore, by studying the vocal material of humpback whales in direct relation with their sound producing mechanisms, and with any other factors which may influence them, this work should bring us more relevant and objective descriptive information to understand humpback whale vocal behavior. The general production-based approach has been widely applied to different mammal species [12, 13]. This approach most often combines predictions based on functional morphology, supported by computational model of sound production systems (e.g., see [14] for doves, or [15] for dolphins). One of the reasons why similar research on mysticete cetacean in general has fallen largely behind is that their vocal production mechanisms are still under investigation. Although agreements have been found on the fact that vocal production results from air movements and is included in the process of internal air recirculation, the precise acoustic origins of the different vocal features observed in humpback whale songs have remained elusive. Recent advances have been made with the work of [16, 17], focusing respectively on two laryngeal components : the U-fold, so called for its particular U-like shape and for its similarities with vocal folds, and the laryngeal sac, which is an air sac located at the ventral aspect of the larynx. While [1] and [8] mostly based their studies on computational analyses of humpback whale sound sequences, the authors have used anatomical data to develop a biomechanical modeling of the vocal production mechanisms, and quantify the acoustic features of synthesized sound units [10, 11].

2 Production-based classification of sound units

2.1 Physiological division of the respiratory tractus into three different configurations

The anatomical scheme of figure 1 has been derived from the work of [16, 17]. The two symmetric lungs are connected to a short broad cone-shaped canal called trachea. In the laryngeal region, three respiratory valves are present. The U-fold (see photo 1) possesses some strong similarities with common mammal vocal folds [16], particularly in regards to its geometrical dimensions (length to thickness ratio) and slit-like glottal shape, tissue composition (presence of a thin mucosa in the outermost layer) and associated laryngeal muscles and cartilages (e.g. presence of homologous thyroid and arytenoid cartilages). The two sides of the U-fold do not attach directly the thyroid cartilage, but fuse together caudally into one midline connective tissue ridge within the laryngeal sac lumen. In front of the epiglottis, whose classical valve function is to protect the lower respiratory system from foreign bodies (mainly water), a third pair of lips-like organ called corniculate cartilage flaps (see photo 1) is present and is characterized by its long shape, the high elasticity of its tissues and the proximity of its two symmetric lips. The laryngeal sac can be seen as a soft extensible oval balloon with an extensive surrounding musculature. On the opposite side of the laryngeal sac, the nasal region is composed of the nasopharynx, a short flexible region of muscles and soft tissues, and of two tube-like parallel nasal cavities with rigid walls. On the top side, these tubes both end on thick nose plugs.

[1] Although the term of air will be mostly used through this text, a reference is actually implicitly made to any kind of internal gases flowing into the respiratory tractus of the whale.

It is noteworthy that the strong muscular structures surrounding both the laryngeal sac and the lungs (i.e. respectively abdominal muscles, the diaphragm and the intercostal muscles) should have three different roles: 1) pressurize all the system to a specific pressure different from the depth-related ambient pressure, 2) manage the different airflows in the different configurations and 3) allow them to withstand large variations in hydrostatic pressure and avoid “chest squeeze” while diving and surfacing. The hypothesis that airspaces inside the whale are undisturbed by the depth-dependent ambient pressure is then anatomically reasonable.

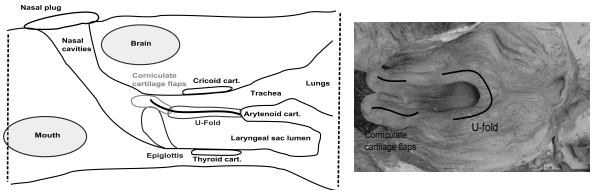


Figure 1: On the left, simplified anatomical scheme of the respiratory tractus. On the right, photo of a dissected whale larynx, hilighting the locations of the two sound generators with black lines. These pictures come from [18].

Considering this overall system through successive respiratory phases, two processes may induce different configurations of this system: the laryngeal valves (i.e. the epiglottis, the U-fold and the corniculate cartilage flaps) movements and the air recirculation. The physiological properties of the lungs and the laryngeal sac allow them to ensure alternatively two opposite functions, i.e. storing the incident airflow or emitting it back in the respiratory tractus for further uses. By combining valve states (open when an airflow can pass through it, or closed otherwise) and air source locations, we highlight the formation of three mutually exclusive configurations of the respiratory tractus, further subdivided by the direction of the airflow, which we will now describe.

Configuration 1 , where only the U-fold and the epiglottis are open. Based on the left graph of first row in figure 2, the air flow in this configuration can either come from the lungs or the laryngeal sac, and then pass through the U-fold. Here, the folds are parted and the epiglottis is pressed against the wall ;

Configuration 2 , where only the corniculate cartilage flaps and the epiglottis are open. Referring to the second row of graphs in figure 2, the U-fold is now replaced by the corniculate cartilage flaps and the pulmonary air going through is directly guided into the nasal cavities, before being stored in the laryngeal sac. The longitudinal shape of these flaps do not allow them to let pass an airflow in the opposite direction, which eliminates the laryngeal sac as an air source. In this configuration the glottal airflow will be more distant from the laryngeal sac than in the first configuration ;

Configuration 3 , where only the U-fold is open. We consider in this configuration 3, illustrated with the two bottom schemes in figure 2, that both the lungs and the laryngeal sac may act alternatively as an acoustic source. The epiglottis and the corniculate cartilage are tightened together, making the nasal cavities inaccessible to the airflow. The lungs, the trachea, the U-fold and the laryngeal sac are the sole anatomical components to be taken into account here, providing a certain symmetry to this configuration.

2.2 Acoustic characterization of each sound unit

Previous acoustic models developed by the authors [11, 18] have allowed the association of quantitative spectral features to the previously described physiological configurations. Briefly, considering the configurations 1 and 2, acoustic resonances inside the respiratory tractus result mainly from the acoustic parallel coupling between the nasal cavities and the laryngeal sac. The tube-like nasal cavities with its rigid walls generate formants through acute acoustic resonances, characterized by harmonically related frequencies (whose dispersion depends on the length tube) spanning a large spectrum. The large laryngeal sac with its softer elastic walls is expected to play a role of low-pass filter with a poor sustaining of higher frequency resonances. Also, the U-fold is more likely to

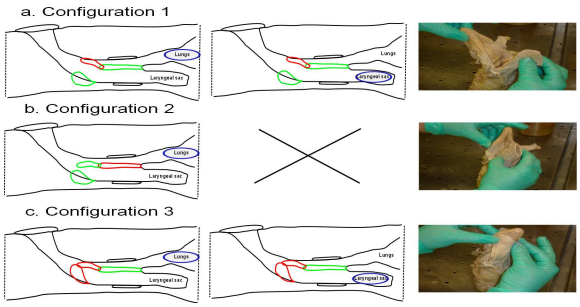


Figure 2: Schemes of the three physiological configurations. The blue circles stand for the acoustic source. A closed valve is drawn in red, while an opened one is drawn in green. On the right, photos illustrating the different configurations of the pair epiglottis / corniculate cartilage flaps, from top to bottom : epiglottis lowered and flaps closed (config. 1), epiglottis lowered and flaps open (config. 2) and epiglottis lifted and flaps open (config. 3)

generate higher fundamental frequencies trough a thickness-to-length ratio and a laryngeal muscle structure quite similar to common mammal vocal folds. This preliminary analysis allows us to also discriminate the three physiological configurations in regards to the active sound generator, producing either a low-pulse F_0 range (below 50 Hz) with the corniculate cartilage flaps (configuration 2), or a medium (50 - 800 Hz) to high (above 800 Hz) frequency range with the U-fold (configuration 1 & 3, with a reversed U-fold for opposite airflow). Figure 3 illustrates the different types of vocalizations formed through the three physiological configurations, with real vocalizations taken from recorded whale songs. The three types of calls proposed allow to explain anatomically the acoustic differences in the fundamental frequency range (low / medium-high) and the presence of formants. Also, the movable precarious structure of formant patterns observed across humpback whale vocal displays fits well the idea of an active temporal dynamic, based on a temporally-shaped laryngeal sac impacting the overall acoustic behavior of the respiratory system. Further studies will be needed to explain source-related acoustic features, such as the noisy characteristic of a sound and the modifications due to a reverse glottal flow. Basic LPC and F_0 tracking [19] have been used to automatically detect values of these acoustic features within each sound unit.

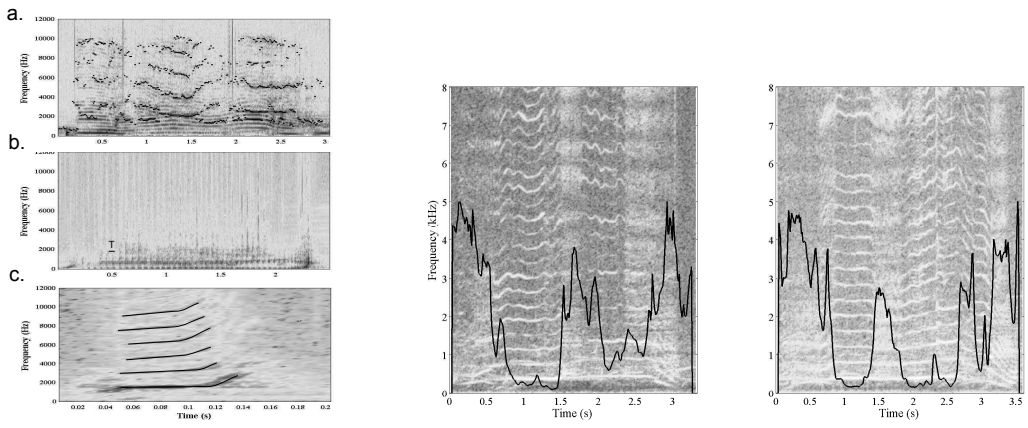


Figure 3: On the left, spectrograms of three types of real sounds units extracted from recordings of humpback whales, corresponding to the three configurations (labeled from a to c in reference to figure 2). The first type presents an harmonic structure with a medium fundamental frequency and modulated by formants, the second type presents a pulse sound with a period T $\approx 60ms$ (i.e. $F_0 = 17Hz$), and the third type presents an harmonic structure with a high fundamental frequency without formants. From [18]. On the right, illustrations of detection of chaotic segments.

To these three basic sound unit types, only discriminated based on fundamental frequency range and presence/absence of formants, we can add an acoustic feature traducing the occurrence of vocal non-linearities. Mostly two types of vocal non-linearities are present in humpback whale songs :

frequency-jumps and chaos. These two features are very interesting in a production-based approach, as easily tractable to specific vocal mechanisms. Indeed, frequency jumps result either from a cross-over between F_0 and a formant [20], or from a register transition during a F_0 modulation [21]. Chaos result exclusively from chaotic oscillations of the vocal folds. We propose to define a chaotic descriptor as the sum of three classical tools from non-linear dynamic methods [22] : the **Entropy**, the **Lyapunov Exponents** and the **Correlation Dimension**. The **Entropy E** quantifies the rate of loss of information about the state of a dynamic system as it evolves over time [22]. For regular behaviors (i.e. static states, periodic and quasi-periodic oscillations), the entropy is equal to zero. For chaotic systems with finite degrees of freedom, the entropy is finite. Chaotic systems display a sensitive dependence on initial conditions. Such a property deeply affects the time evolution of trajectories starting from infinitesimally close initial conditions, and **Lyapunov exponents** are a measure of this dependence. These characteristic exponents give a coordinate-independent measure of the local stability properties of a trajectory. A trajectory is chaotic if there is at least one positive exponent. The **correlation dimension D**, proposed by [23], quantifies the complexity or irregularity of a trajectory in phase space, describing the geometrical scaling property of a vocal sound in a state space. Figure 3 on the right illustrates the use of such descriptors in the discrimination between harmonic / chaotic segments in two different sound units.

3 Discussion and Conclusion

At the final step of analysis, we can use the production-based call types previously described and identify their occurrences over different songs of humpback whales. Figure 4 illustrates two representations of humpback whale songs based on this production-based approach. All sound units are preliminary isolated, characterized individually and then concatenated over time.

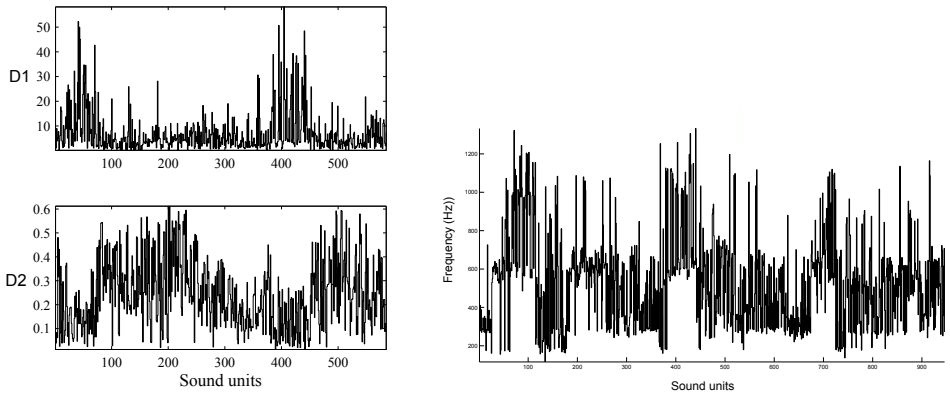


Figure 4: Curves of two descriptors D_1 and D_2, detecting respectively chaotic vocalizations and formant occurrences. On the right, a fundamental frequency tracking over a song is represented.

A more "high-level" representation could be extracted from figure 4, with the identification of the different configurations developed in section 2.2, based on the continuous temporal distribution of formants and fundamental frequencies. This would potentially provide an insight into internal respiratory mechanisms of the whale, and its management of sound production. Also, the temporal distributions of vocal non linearities give us an interesting support to speculate on what they could traduce within humpback whale communication framework (e.g. exceptional vocal features, lack of vocal skills, pathological signs). The main goal of the proposed representation is then to provide a useful tool for biologists. This kind of representation offers a meaningful support for direct interpretation and discussion of humpback whale vocal strategies within their communication framework.

References

[1] E. Mercado, J. Schneider, A. A. Pack, and L. M. Herman, "Sound production by singing humpback whales," *J. Acoust. Soc. Am.*, vol. 127, pp. 2678–2691, 2010.

[2] R. Payne and S. McVay, "Songs of humpback whales," *Science*, vol. 173, pp. 585–597, 1971.

[3] D. M. Cholewiak, R. S. Sousa-Lima, and S. Cerchio, "Humpback whale song hierarchical structure: Historical context and discussion of current classification issues," *Marine Mammal Science*, vol. 29, pp. 312–332, 2013.

[4] M. J. Noad, D. H. Cato, M. Bryden, M. N. Jenner, and K. C. S. Jenner, "Cultural revolution in whale songs," *Nature*, pp. 408–537, 2000.

[5] W. W. L. Au, A. A. Pack, M. O. Lammers, L. M. Herman, M. H. Deakos, and K. Andrews, "Acoustic properties of humpback whale songs," *J. Acoust. Soc. Am.*, vol. 120, pp. 1103–1110, 2006.

[6] R. Suzuki, J. R. Buck, and P. L. Tyack, "Information entropy of humpback whale songs," *J. Acoust. Soc. Am.*, vol. 119, pp. 1849–1866, 2006.

[7] H. Ou, W. W. L. Au, L. M. Zurk, and M. O. Lammers, "Automated extraction and classification of time-frequency contours in humpback vocalizations," *J. Acoust. Soc. Am.*, vol. 133, p. 301310, 2013.

[8] E. Mercado and S. Handel, "Understanding the structure of humpback whale songs," *J. Acoust. Soc. Am.*, vol. 132, pp. 2947–2950, 2012.

[9] E. Mercado, "Computational models of sound production and reception in the humpback whale," Master's thesis, University of Hawai, 1998.

[10] D. Cazau, "Acoustics of the mysticete cetacean (baleen whale) vocal production system," Master's thesis, Université Pierre and Marie Curie, 2012.

[11] O. Adam, D. Cazau, N. Gandilhon, B. Fabre, J. T. Laitman, and J. S. Reidenberg, "New acoustic model for humpback whale sound production," *Applied Acoustics*, vol. 74, pp. 1182–1190, 2013. Applied Acoustics, Elsevier.

[12] W. T. Fitch and M. D. Hauser, "Voice production in non-human primates : acoustics, physiology, and functional constraints on honest advertisement," *American Journal of Primatology*, vol. 37, pp. 191–219, 1995.

[13] W. T. Fitch, J. Neubauer, and H. Herzel, "Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production," *Animal Behaviour*, vol. 63, pp. 407–418, 2002.

[14] T. Riede, M. J. Owren, and A. C. Arcadi, "Nonlinear acoustics in pant hoots of common chimpanzees (pan troglodytes): Frequency jumps, subharmonics, biphonation, and deterministic chaos," *American Journal of Primatology*, vol. 64, pp. 277–291, 2004.

[15] T. W. Cranford, P. Krysl, and J. A. Hildebrand, "Acoustic pathways revealed: simulated sound transmission and reception in cuvier's beaked whale (ziphius cavirostris)," *Bioinsp. Biomim.*, vol. 3, pp. 1–10, 2008.

[16] J. S. Reidenberg and J. T. Laitman, "Discovery of a low frequency sound source in mysticeti (baleen whales): anatomical establishment of a vocal fold homolog," *Anatomical Record*, vol. 290, pp. 745–760, 2007.

[17] J. S. Reidenberg and J. T. Laitman, "Sisters of the sinuses: Cetacean air sacs," *Anatomical Record*, vol. 291, pp. 1389–1396, 2008.

[18] D. Cazau, O. Adam, J. T. Laitman, and J. S. Reidenberg, "Understanding the intentional acoustic behavior of humpback whales: a production-based approach," *J. Acoust. Soc. Am.*, vol. 134, pp. 2268–2273, 2013.

[19] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.

[20] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, pp. 2733–2749, 2008.

[21] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal vibratory mechanisms: The notion of vocal register revisited," *Journal of Voice*, vol. 23, pp. 425–438, 2009.

[22] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement," *Journalof Voice*, vol. 20, pp. 2–17, 2006.

[23] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D.*, vol. 9, pp. 189–208, 1983.

8.4 Gabor Scalogram for Robust Whale Song Representation

Randall Balestriero
Université de Toulon
av de l'Université
La Garde, France
randallbalestriero@gmail.com

Hervé Glotin
Inst. Universitaire de France
Bt St Michel Paris
& Université de Toulon
glotin@univ-tln.fr (correspond. author)

1 Introduction

It has been well documented that Humpack whales produce songs with a specific structure [Payne]. The NIPS4B challenge provides 26 minutes of a remarkable Humpback whale song recording produced at few meters distance from the whale in La Reunion - Indian Ocean, by "Darewin" research group in 2013 at a frequency sampling of 44.1kHz, 32 bits, mono, wav format (Fig 1).

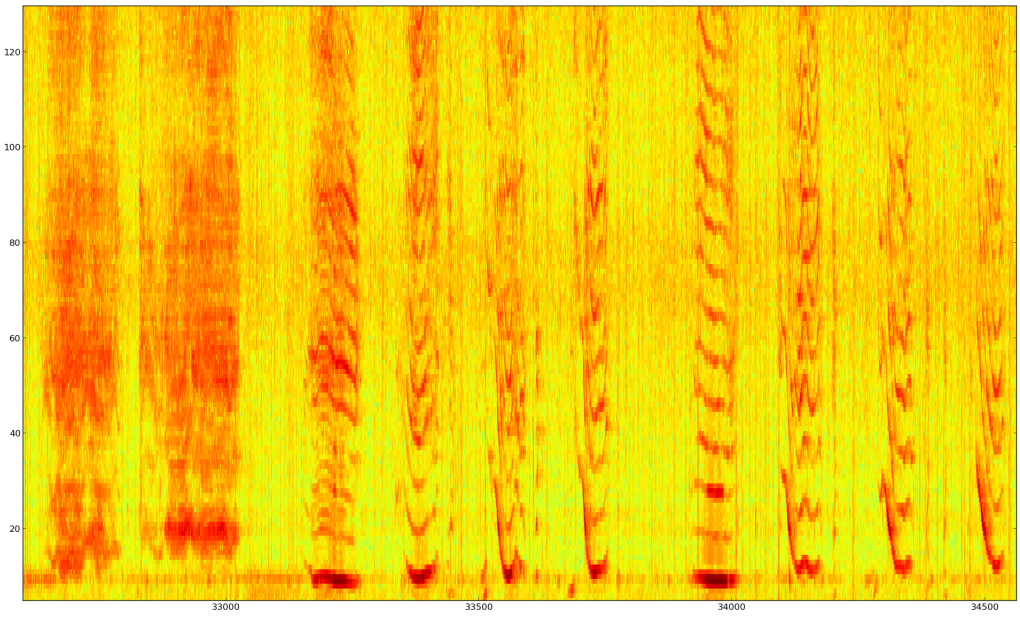


Figure 1: Spectrum of around 20 seconds of the given song of Humpback Whale (start from about 5'40 to 6'. 0 to 22.05 kHz - frameshift of 10 ms)

Usually, the Mel Filter Cepstrum Coefficients are used as parameters to describe these songs [Pace and al.] We propose here another efficient representation, the scalogram, and we demonstrate that the sea noise is efficiently removed, even in the case of lower SNR recordings, allowing robust song representations.

2 Scalogram for robust whale song unit extraction

We compute the first layer of the scattering transform of the ScatNet Toolbox to perform the Gabor wavelet transform. We then generate different scalograms on the challenge 2 wav file, but also on some of others whale songs recorded in 2013 in New Caledonia with low cost material in order to emphasize the potential of this representation for bio-acoustic analysis even at low SNR.
According to our experiments, the parameters that were performing the best were the Gabor mother wavelet, with opt.Q=8 , opt.J=62 , opt.T=948.1 = Q*2^(J/(Q+1) . T is then a the minimal physiological scale (<2ms).

Then the first layer appears to loose few units which are also missing in the other scattering layers, however it has a strong energy coefficient. However some specific patterns appear and could possibly be used to describe and identify the singer. For example, the chirps have a specific length and slope as shown with some examples extracted from 4 different samples recordings in the next sections (the original figure are at : http://sabiod.univ-tln.fr/pimc/rapport/).

We give the scalogram and spectrogram of around 2 minutes on each signals. For all the scalogram none additionnal non-linear transformation has been applied. This comparison emphasizes the strength of the scattering decomposition conpared to the spectrogram containing the sea noises.

We illustrate this with different occurrences of some specific patterns, computed on window lasting 2^16 samples which is the maximum window length we can use in ScatNet toolkit.

3 Challenge results

Results on the NIPS4B_humpback.wav challenge data are in Fig 2. (http://sabiod.univ-tln.fr/nips4b/challenge2.html) (http://sabiod.univ-tln.fr/pimc/RAPPORT_NIPS4B_humpback_J62_Q8_T948.0957/).We give in figure 3 some extracted examples of a recurrent particular shape (spectrogram window = 128, overlap = 64)

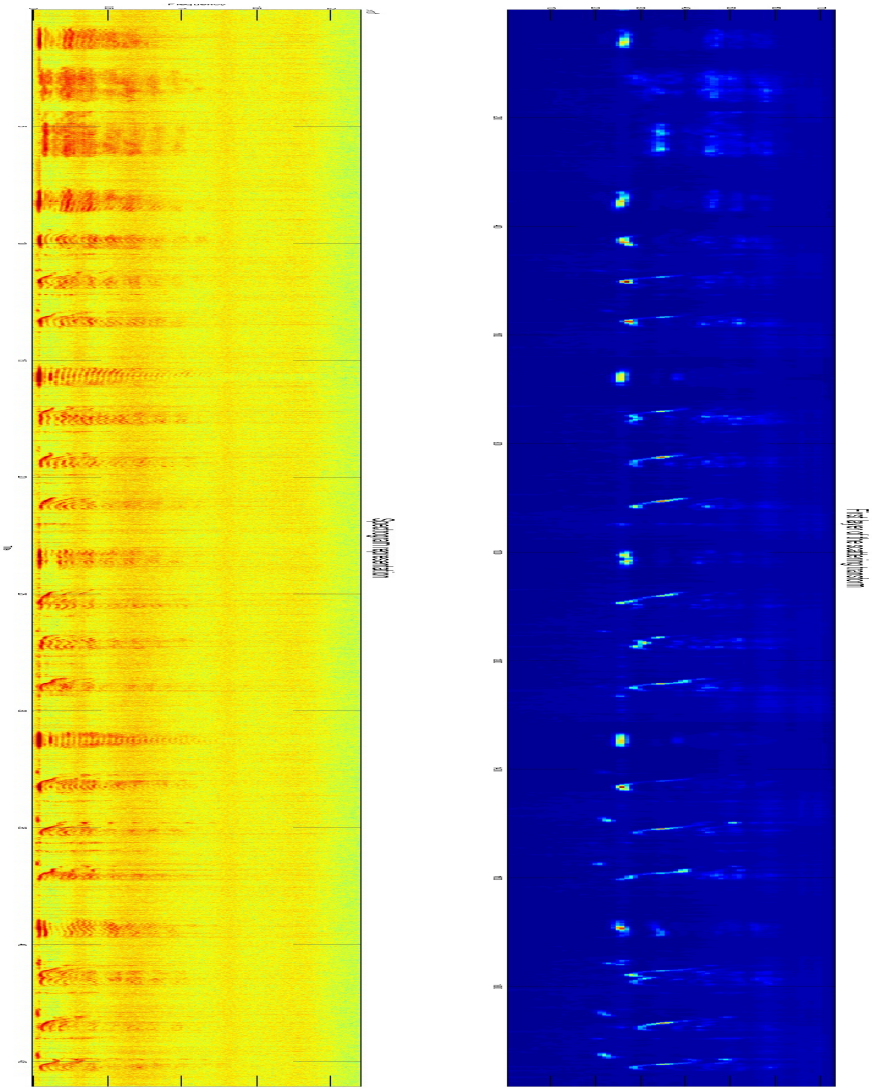


Figure 2 : scalogram and spectrogram of the challenge data including the 20 seconds of the challenge part 8,J=62, Q=8, T=948.0957

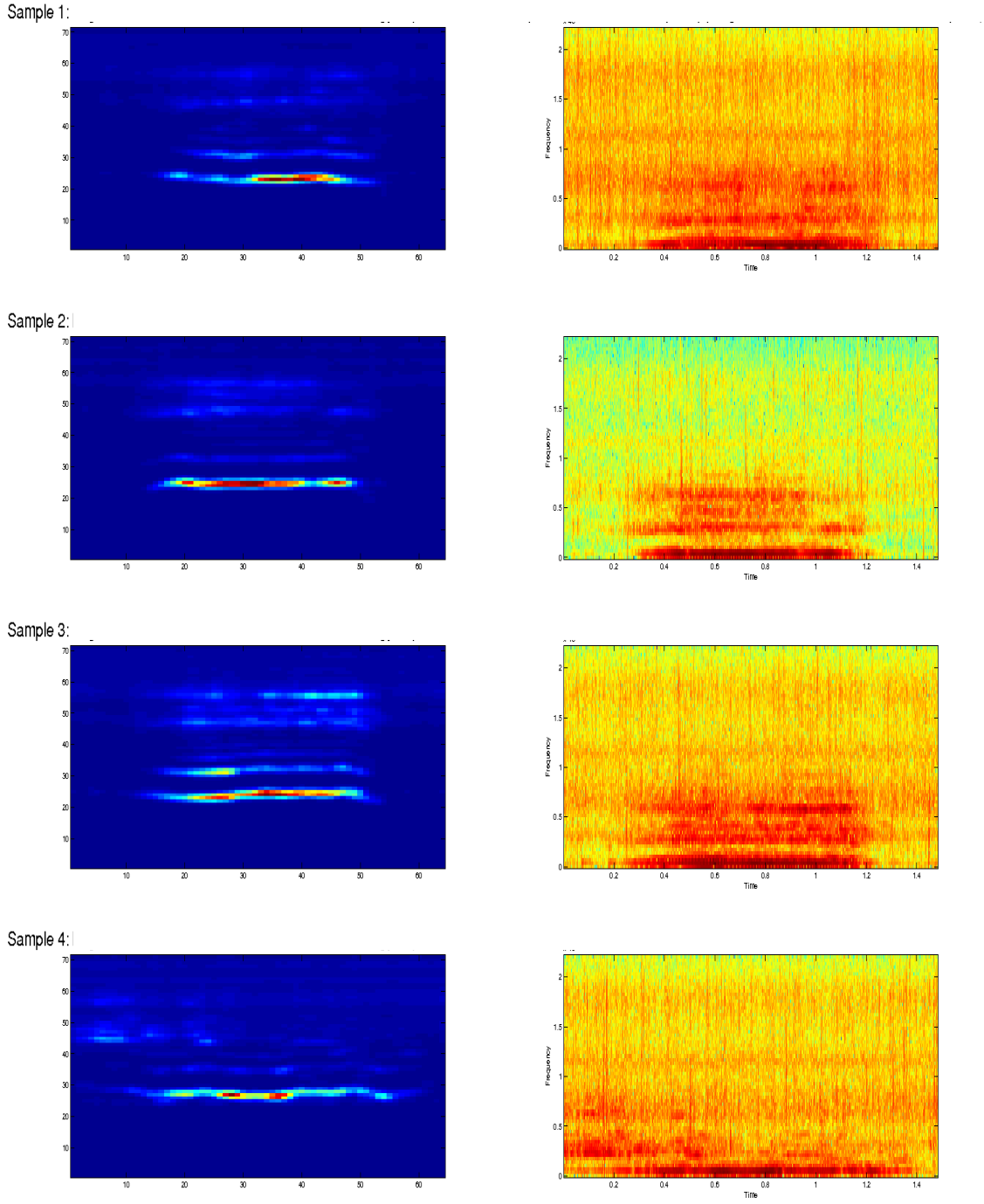


Figure 3: Chirp extracted from the same challenge data and corresponding times. Duration of each window : 1.49 sec. Begin time, Sample 1 : 0.11 sec., Sample 2 : 38.19sec., Sample 3 : 44.52sec., Sample 4 : 59.04sec.

3. Results at low SNR of various songs on same area and different days

In this section we compute with the same parameters the scalogram on a noisy recording taken in the New Caledonian Lagoon.
/NAS3/PIMC/SITE/FGAB_WAV_all/20130720_BB_en_plusieurs_points/DECAV_20130720_113312.wav
The full results are at

http://sabiod.univ-tln.fr/pimc/RAPPORT_DECAV_20130720_113312_J62_Q8_T948.0957/
A sample is given in figure 4 below, showing again clear chirps.

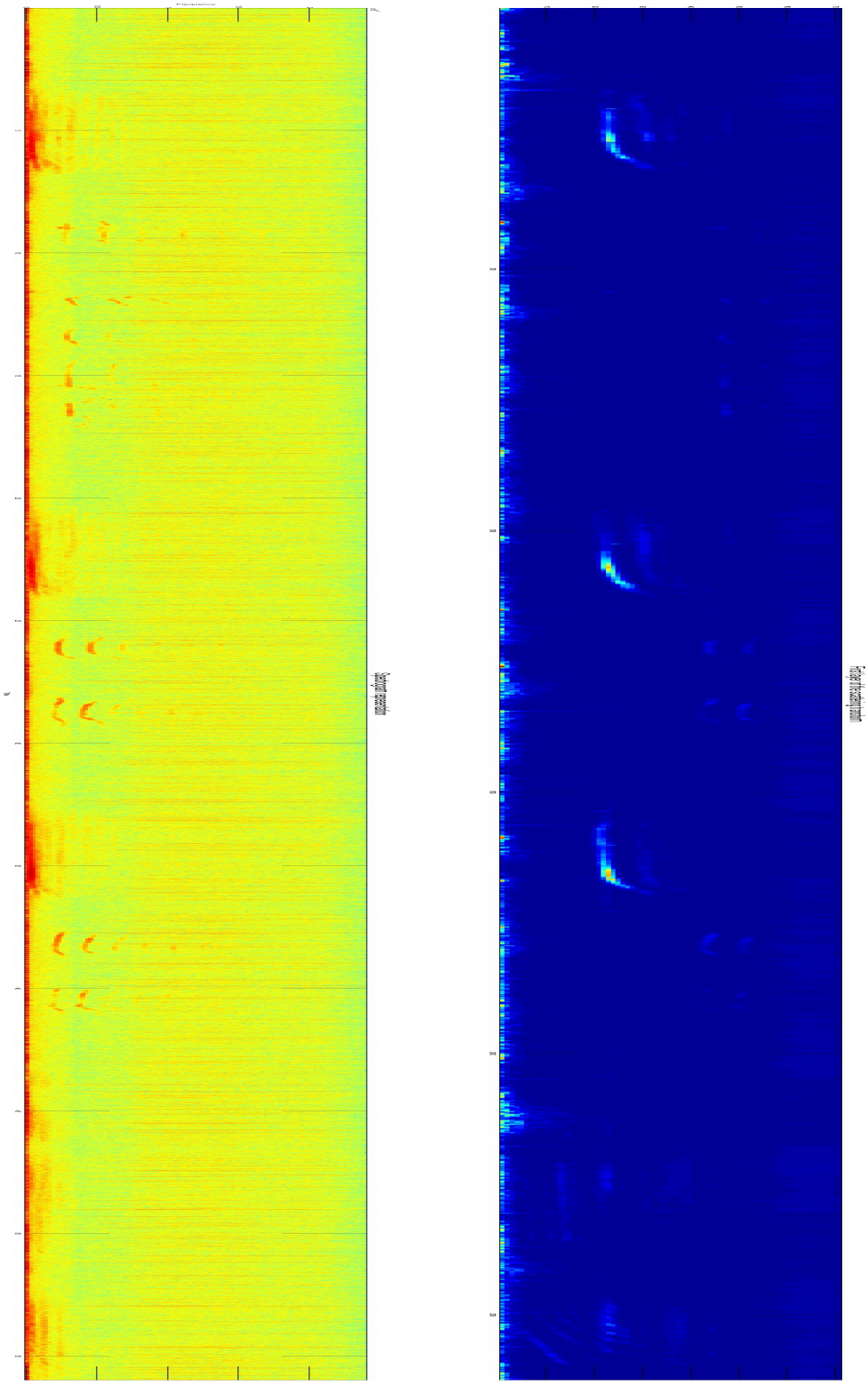


Figure 4: 2-minute scalogram and spectrogram of this file part 2,J=62, Q=8, T=948.0957

The figure 5 shows for the same file a recurrent particular chirp :

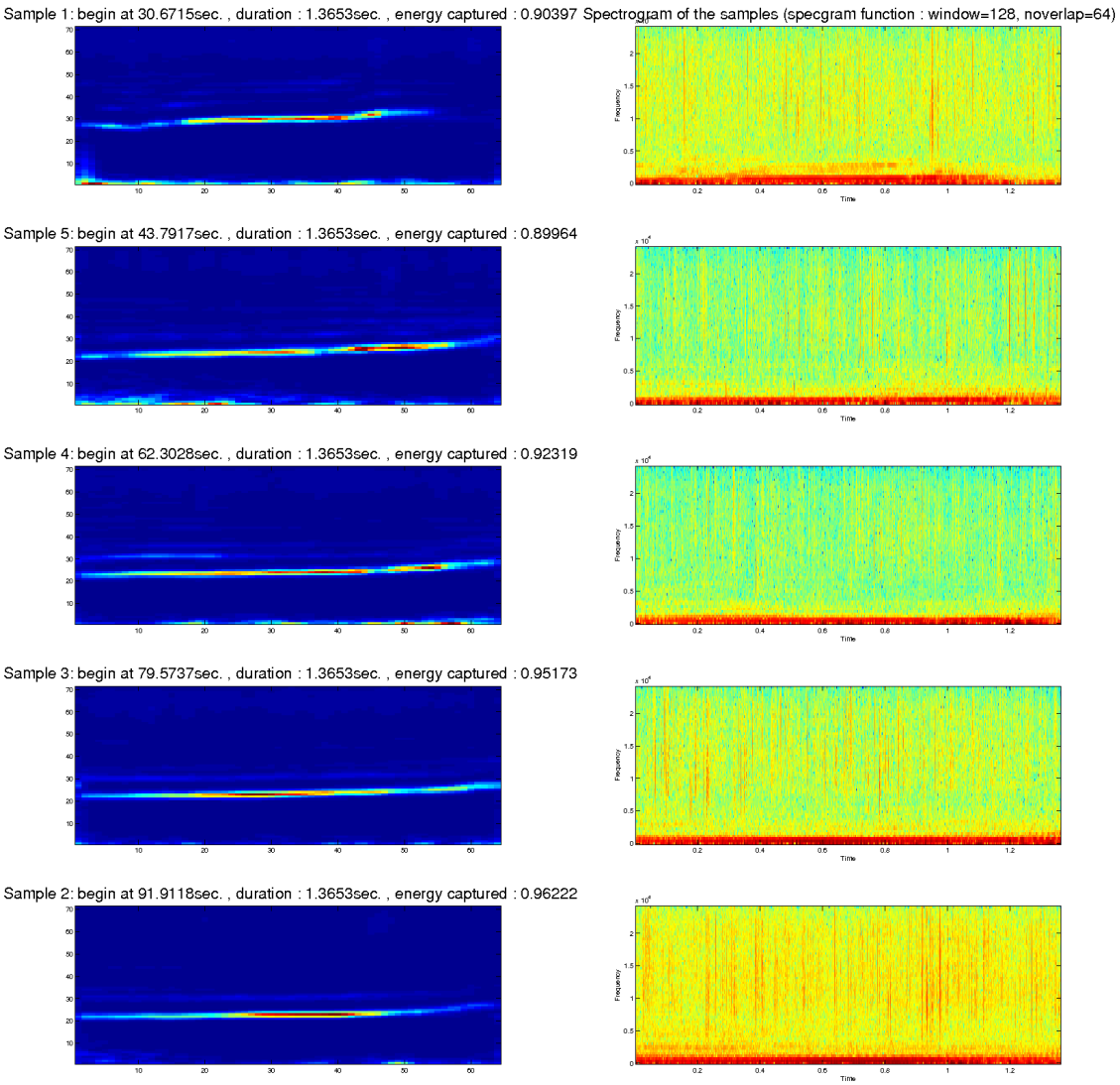


Figure 5: Chirp extracted from the same recording.

Another similar analysis is conducted on a whale recorded two days laters at the same place, on SABIOD data:
/NAS3/PIMC/SITE/FGAB_WAV_all/20130722_triangulation_avec_GOPRO/DECAV_20130722_103948.wav

The full representation of the wav using the scattering decomposition and the FFT is at
http://sabiod.univ-tln.fr/pimc/RAPPORT_DECAV_20130722_103948_J62_Q8_T948.0957/
We give one sample below (figure 6), showing other kind of pattern, from another kind of song units.

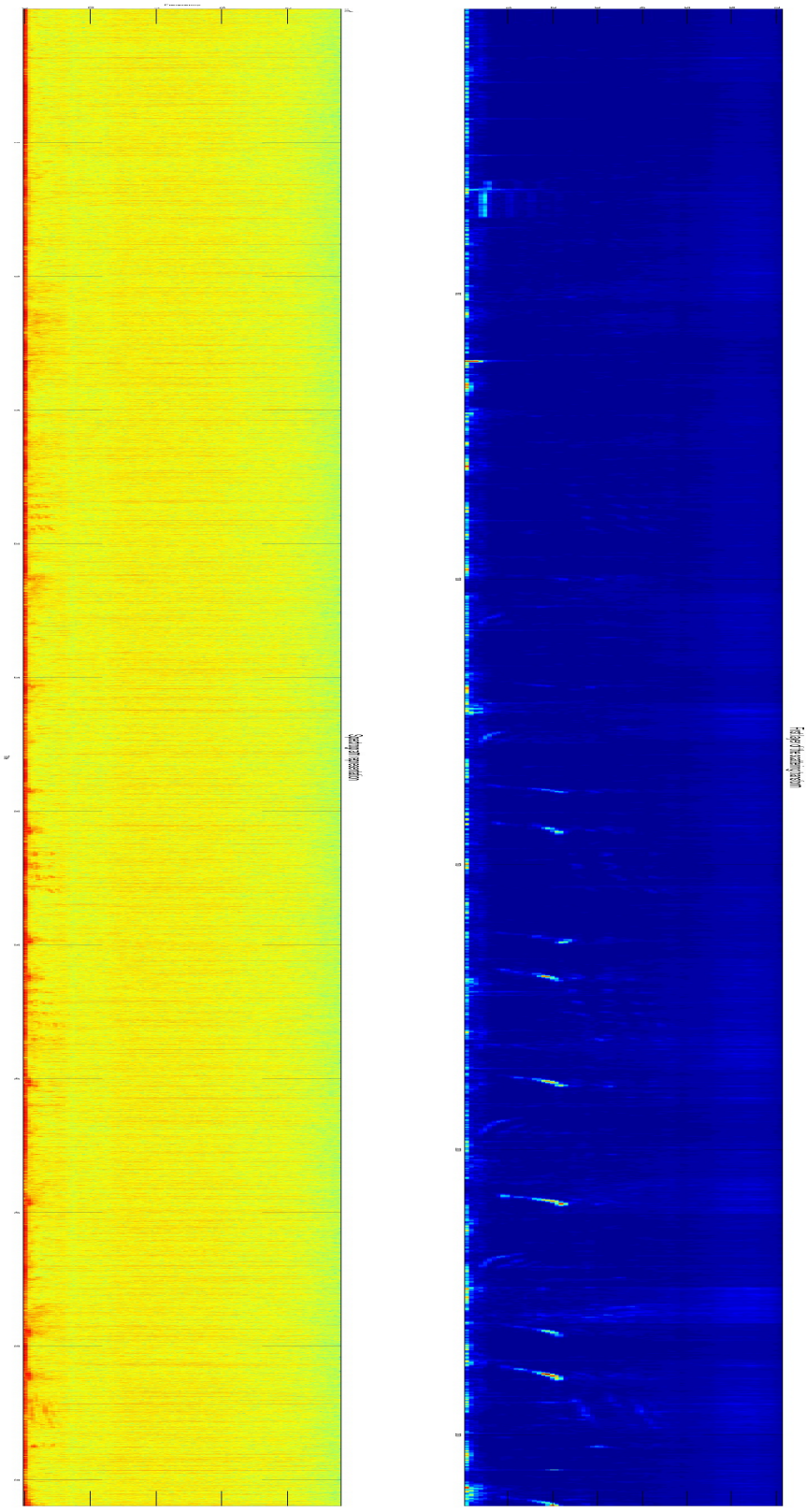


Figure 6: 2-minute scalogram and spectrogram of this file part 2,J=62, Q=8, T=948.0957

Here is the zoom on other units found into these file (figure 7):

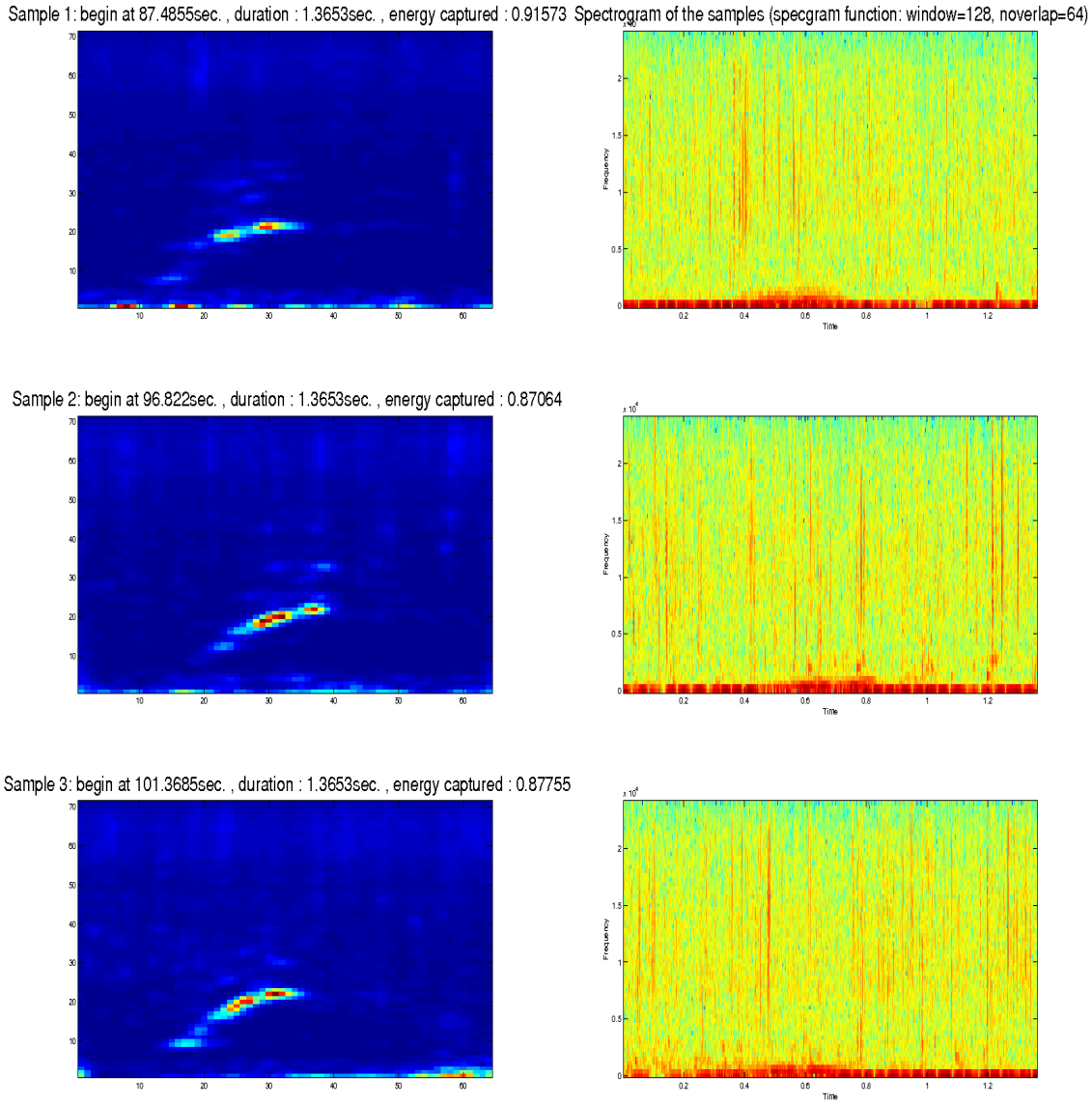


Figure 7: Chirp extracted from the same recording.

We conduct the same analysis 3 days later, showing again different kind of units:
/NAS3/PIMC/SITE/FGAB_WAV_all/20130725_triangulation_et_TASCAM/DECAV_20130725_093238.wav

The full representation is available at :
http://sabiod.univ-tln.fr/pimc/RAPPORT_DECAV_20130725_093238_J62_Q8_T948.0957/

A 2-minute sample already shows different patterns (figure 8) :

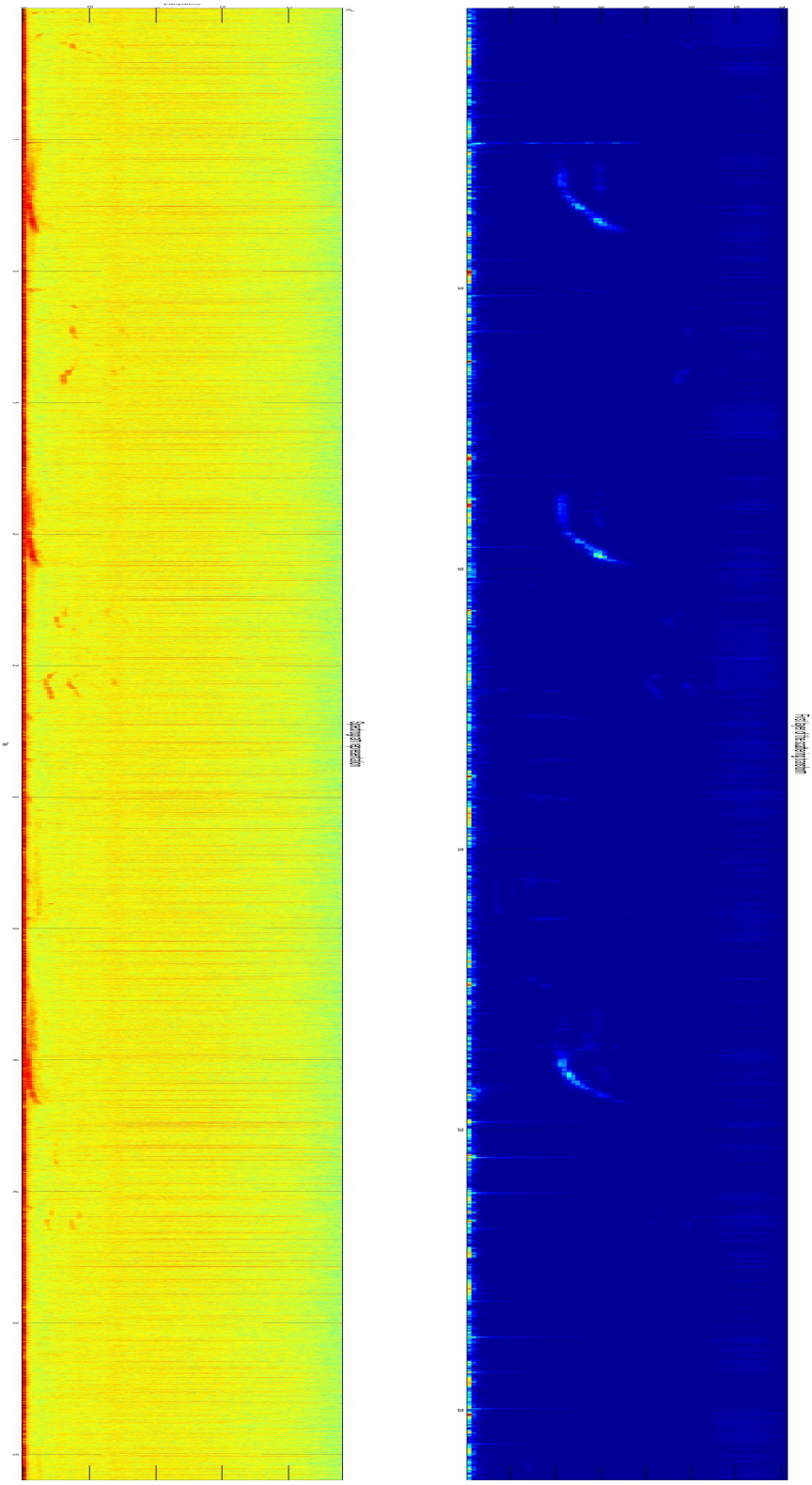


Figure 8: 2-minute scalogram and spectrogram of this file part 3,J=62, Q=8, T=948.0957

And here (figure 9) a recurrent chirp appearing multiple times on this file :

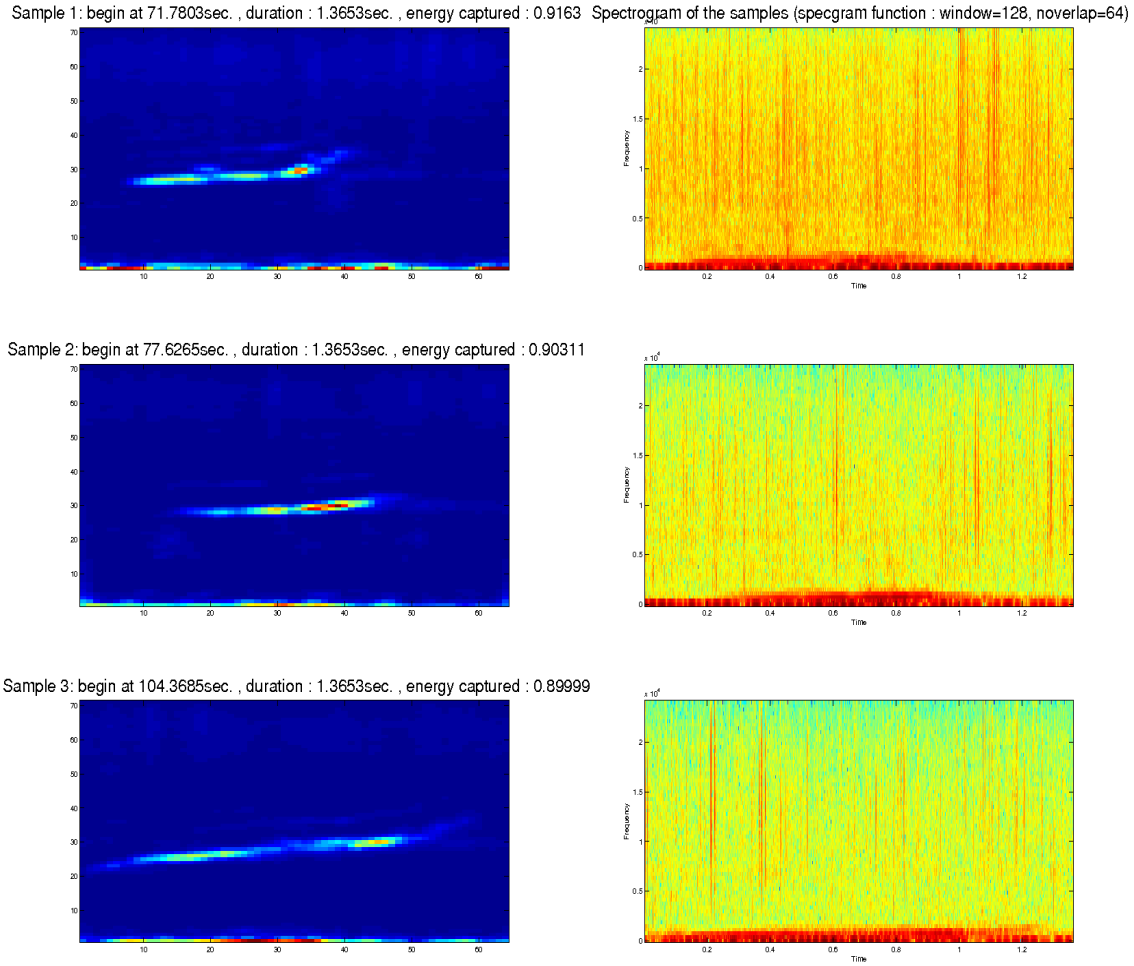


Figure 9: Chirp extracted from the same recording.

4 Conclusion

We demonstrate the advantage of Gabor scalogram to reveal humpback whale songs analysis : it distinguishes fine details that are possibly linked to individual signature. This representation may be usefull for research on whale identification [Cazeau 2013, in this workshop].

Looking at the recurrent units found in each file, we can see that the NIPS4B_humpback.wav has some really flat and mid-sized units of approximately 0.7 to 1 second, the DECAV_20130720_113312.wav file (figures 4 and 5) has longer chirps (lasting for the whole time window taken, about 1.35 second) and also caracterized by a small positive slope.

For the DECAV_20130722_103948.wav file (figures 6 and 7), the chirps are smaller (the length is about three times smaller than the previous example) and the slope is greater, also note the concave shape. For another whale, in the record DECAV_20130725_093238.wav (figures 8 and 9), we see chirps with mid-sized length, a small positive slope, and a convex pattern. These, are the kind of signature we are looking for individual indexing

Even if a log spectrogram may have also revealed some interesting patterns, we demonstrate the advantage of the scalogram representation compared to spectrogram according to the sea noise level that has been removed into the scalogram.

References

Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. (2010) Subunit definition for humpback whale call classification int. journal Applied Acoustics, Elsevier, 11(71)

ScatNet http://www.di.ens.fr/data/software/scatnet/documentation/

8.5 Automatic analysis of a whale song

Ilyas Potamitis
Dep. of Music Technology & Acoustics
Technological Educational Institute of Crete
Heraklion, Crete, Greece
potamitis@staff.teicrete.gr

Stavros Ntalampiras
Joint Research Center, EC
Ispra
Varese, Italy
stavros.ntalampiras@jrc.ec.europa.eu

Abstract

Male whale vocalizations have the characteristics of a song. Male whales form temporal sequences of different syllabic types making repeated phrases. We provide a method for the automatic quantitative analysis of a single humpback whale song by decomposing these songs in their constituent syllabic types and studying their temporal sequencing. This work describes our approach to the humpback whale song processing challenge organized and hosted as part of the 2013 Neural Information Processing Scaled for Bioacoustics: NIPS4B.

1 Introduction

Animals use vocalization for reasons that are vital for their existence. Mate selection, courtship rituals, coordination, alarming and marking of territory are the most important ones. Cetaceans vocalize for these reasons although the complete mapping of vocalization to behavioral modes is not yet fully clarified. The growing concern of signal processing and pattern recognition applications with marine mammals is associated with assessing the impact of anthropogenic noise on cetaceans [1], providing means to avoid collisions with ships [2] and detecting species that are in endangered status [3]. In this work we focus on a humpback whale song. Both female and male humpback whales produce sounds that are referred to as 'social sounds', while 'songs' are produced exclusively by male humpbacks. In order to be clear with the terminology we also adopt the widely accepted definition of a single vocalization of a humpback whale constituting a 'syllable' which is a distinct compact segment of continuous sound separated from other syllables by silence. A 'phrase' is a sequence of heterogeneous syllables in close succession [4] whereas a song is a structured stereotyped repetition of phrases. For more elaborate representation of the types of structural components typically present in sequences of sounds produced by singing humpback whales please refer to [5]. It has been well documented that humpback whales produce songs with a specific structure [6]. A song typically lasts from 10-20 minutes and is repeated continuously, possibly over hours with small variations in its phrase composition. The song changes gradually from year to year and the songs of populations around the globe are distinctively different. There is a long list of studies concerning whale songs (see [6] and the references therein).

The 2013 Neural Information Processing Scaled for Bioacoustics: NIPS4B featured a signal processing and pattern recognition challenge on the domain of song whale processing [7]. The NIPS4B event provides 26 minutes of a single, high quality humpback whale song recording produced at a few meters distance from the whale in La Reunion - Indian Ocean. The purpose of this study is to propose an efficient representation of a given humpback whale song that helps to study its structure, as well as discover and index its units [8]. Our contribution is an automatic segmentation approach of a whale song and clustering of the resulting syllables into an approximate alphabet of syllables. Moreover, we discuss a sequence modelling of the whale song based on modelling their succession with N-grams.

2 Signal Analysis

2.1 Elementary Vocalizations and Distinct Syllable Types

We manually examined the spectrogram and listened to the 26-minute song and ended with 9 distinct syllables. This is in accordance with [6] that also reports 9 distinct syllables of humpback whales. The syllables consist of a series of 200-19 kHz spectral chunks. Syllables are of varying duration uttered at rates of about 30 per minute and are both amplitude and frequency modulated.
A description of the units of this particular humpback whale song follows:
S1: has a flat tonal. It is broadband with very strong harmonics reaching up to 16 kHz. Its duration is of 1 sec mean.
S2: has a small low-frequency flat tonal subunit followed by a downsweep of frequencies. Its mean duration is of 1.5 sec.
S3: has weak harmonics and an initial downsweep followed by a vibrating flat tonal. Duration of 1 sec.
S4: is relatively broadband, with strong frequency modulated harmonics reaching up to 12 kHz ending with an upsweep. Its mean duration is of 2 sec.
S5: is broadband with harmonics plus noise reaching up to 15 kHz. Its mean duration is of 2 sec.
S6: is relatively narrowband having a mostly noisy structure. Its duration is of 2 sec.
S7: has bird like chirps, a strong tonal and is relatively narrowband. Its mean duration is of 1 sec.
S8: is a small low-frequency flat tonal subunit followed by a short-time chirp segment. It possesses strong harmonics. S8 appears alone or with one or two following characteristic subunits. These subunits never appear in isolation but when present they come always after S8. Its mean duration is of .75 sec for the main syllable and about .5 sec for two following subunits.
S9: is strongly tonal with higher frequency harmonics, demonstrating an almost chirp-like character. Its mean duration is 1 sec.

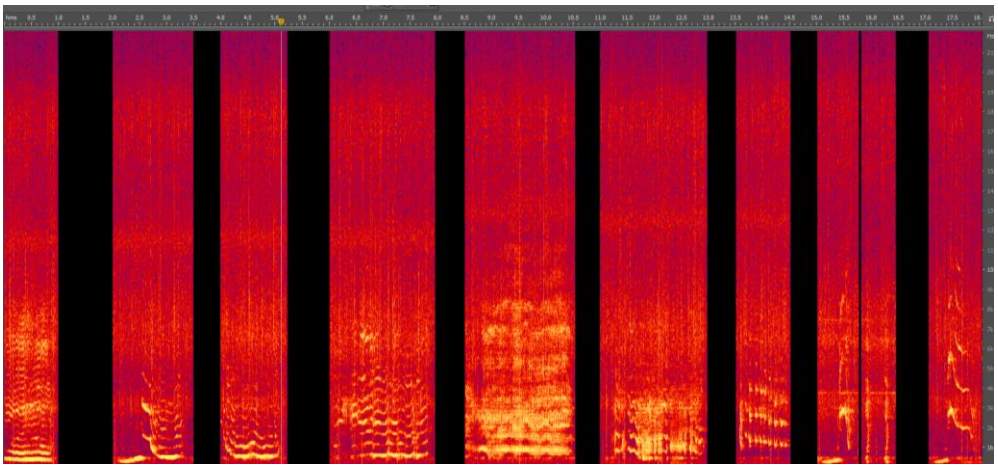


Fig. 1. The 9 syllabic units of the humpback whale song

Most energy is concentrated between 200-2500 kHz except syllable S1 that has high energy harmonics up to 16 kHz. Manual examination shows silent intervals ranging in duration from .1-3 sec. In Fig. 2 we give a distribution of durations as found by our automatic segmentation system (see par 3.1 and Fig. 3).
Once we had a clue about what the syllables should look like we proceeded into examining which features might serve as a basis for classification and employing machine learning techniques to examine the possible number of clusters.

3 Signal Processing & Pattern recognition

3.1 Signal Segmentation

Let *x(n)* denote the discrete time-domain signal holding the original recording, where *n* is the discrete-time index. The recording is sampled at 44.1 kHz, 16 bit and downsampled to 16 kHz as we are interested in deriving cluster indices of segments and the signal energy beyond 8 kHz is small. The SNR is quite good as the hydrophones were close to the whale [7]. In order to extract the useful audio event from its background we applied the Hilbert follower, also known as envelope follower. The Hilbert follower follows the characteristic shape of the time-domain audio envelope of vocalizations. We briefly describe its derivation and function:
Let, $x_h(n)$ = Hilbert($x(n)$) return a complex sequence called the analytic signal of $x(n)$. The analytic signal $x_h(n) = x(n) + jx_i(n)$ has a real part $x(n)$ which is the original data, and an imaginary part, xi(n), which contains the Hilbert transform of $x(n)$. The envelope $y(n)$ of the sampled time-domain recording is calculated as:

$$y(n) = \left(X_h(n) \otimes \hat{X}_h(n)\right)^{1/2} \quad (1)$$

where $\hat{X}_h(n)$ stands for the conjugate of $X_h(n)$ and X for component wise multiplication.
The envelope in Eq. 1 is compared against a threshold θ. When $y(n) > \theta$ the sample $x(n)$ is classified as belonging to the activity class otherwise to the non-activity class (see Fig. 2). The threshold is calculated from the whole recording. The envelope $y(n)$ is sorted by value and a conservative threshold is calculated as $\theta = 3*\theta_1$ where θ_1 is the mean of the 90% of the lowest values of the envelope. Let $x_e(n)$ hold the recordings for which $y(n) > \theta$.

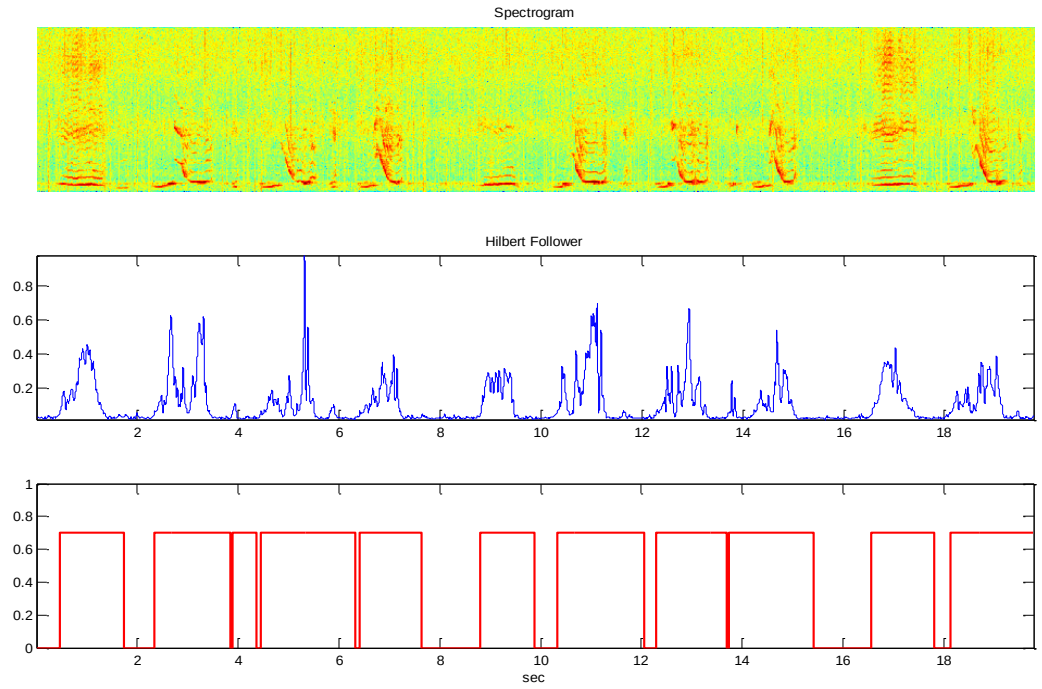


Figure 2: Top: Spectrogram of the first 10 syllables. 2. Middle: envelope of the Hilbert detector. Last: Detection result and segmentation

In Fig. 3 we give a distribution of silence and syllable durations as found by our automatic segmentation system. As regards inter-syllable silence durations, the mean is 0.66 sec with a standard deviation of 0.53 sec whereas the syllables have a mean of 1.25 sec and standard deviation of 0.46 sec.

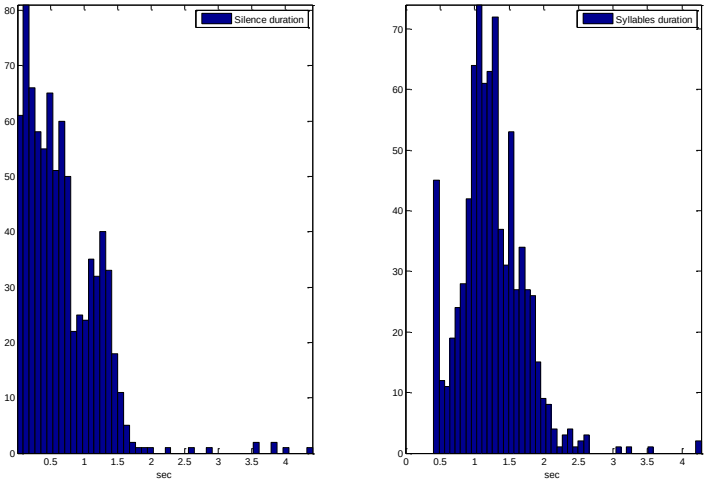


Figure 3: Histogram of inter-syllable silence durations. Right: Histogram of syllable durations

3.2 Features

In the next steps we outline the manner in which *x(n)* is transformed into a set of low dimensional descriptors that subsequently are fed to the clustering stage.

Once the song is segmented into syllables, each syllable undergoes a transformation in order to derive the characteristics that are distinct for each syllable and will allow its clustering to a syllable class. The following features are derived on a per frame basis for each segment provided by the segmentation procedure. Subsequently, the features are averaged on a per-segment basis and provide one feature vector per segment. We included features that can capture the general shape of the syllable as well as its tonal or possible noise character. We included the following audio descriptors used in the MPEG-7 standard:

1. Mel Frequency log spectrum (12 coefficients). The spectrum of the signal is passed through 12 bandpass filters. The Mel frequency filter bank is used to reduce the dimensionality of the spectrum and to reduce the within band spectrum variability. The application of log energy decreases the dynamic range of the spectrum.

2. Delta MFCC (12 coefficients): Deltas are calculated as MFCC difference between frames and serves to measure the dynamic variation of the signal.

$$\Delta C_m(t) = \frac{\sum_{\tau=-M}^{M} \tau C_m(t+\tau)}{\sum_{\tau=-M}^{M} \tau^2}$$

where, M is the number of frames before and after the frame t, and in our case M=2.

3. Spectral Bandwidth (1 coefficient): the frequency extent of each segment.

4. Spectral Entropy (8 feature): the Shannon entropy, calculated as:

$$H(x) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i)$$

where *p(x)* is the spectrum amplitude normalized so that it can be considered a probability distribution. Spectral entropy is sensitive to predominant peaks or spectral flatness. We calculated spectral entropy every 1 kHz (8 bands in total) and averaged the results. *N* is the total number of spectral amplitudes.

5. Spectral Crest (1 feature): a measure of how noisy/tonal is the signal. It is by definition the ratio of the maximum value of the spectral amplitude to the arithmetic mean of the energy spectrum. We calculate spectral crest independently for 8 bands and sum.

$$SCr(t) = \frac{\max_n A_t(n)}{\sum_{n=1}^{N} A_t(n) \ / \ N}$$

6. Spectral Centroid (1 feature): the center of mass of the distribution of spectral amplitude. It has a robust connection with the impression of "brightness" of a sound.

$$SC(t) = \frac{\sum_{n=1}^{N} nA_t^2(n)}{\sum_{n=1}^{N} A_t^2(n)}$$

where, $A_t(n)$ is the magnitude of the Fourier transform at frame t and frequency bin n. A higher centroid correspond to a spectrum with dominant high frequencies.

7. Spectral Flatness (1 feature): a measure of how noisy/tonal is the signal. It is by definition the ratio of the geometric mean to the arithmetic mean of the energy spectrum.

$$A(t)_L = \log(A(t))$$

$$SF(t) = \frac{\exp\left(\sum_{n=1}^{N} A_{L,t}(n) \ / \ N\right)}{\sum_{n=1}^{N} A_t(n) \ / \ N}$$

We measured spectral flatness across the whole spectral band.

8. Spectral Slope (1 feature): represents the decreasing slope of the spectral amplitude and is calculated as a linear regression of the spectral amplitude.

9. Spectral Roll-off (1 feature): the cutting frequency c below which the signal energy is 85% of the total signal energy. It is an indicator of the general shape of the spectrum.

$$\sum_{i=1}^{C} A_t(i) = 0.85 * \sum_{i=1}^{N} A_t(i)$$

10. Skewness of Spectral Flux (1 feature): skewness is a measure of asymmetry of a distribution around its mean value and spectral flux represents the variation of the spectrum along time. The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions

$$F(t) = \sum_{i=1}^{N} \left(A_t(i) - A_{t-1}(i)\right)^2$$

where A_t and A_{t-1} are the normalized magnitude of the Fourier transform at the current time frame t, and the previous time frame $t-1$, respectively. The spectral flux is a measure of the amount of local spectral change.

The feature extraction procedure results into a 39 dimensional array and therefore the segmentation and feature extraction produce a real matrix S of dimension 612x39.

11. Segment duration (1 feature): the duration of a segment in seconds. It is derived from the signal segmentation procedure described in par. 3.1.

3.3 Estimating the number of Clusters

We manually selected the distinct syllables of the song by carefully observing the spectrogram of each syllable and listening to every syllable. However, this is not possible for recordings of many whales, or for recordings scaling up to months or years. Since we wanted a generic approach to automatically label a song we tried some methods that do not require the number of clusters to be set *a-priori* but try to discover it themselves. We analyzed these syllables using an approach that embeds a set of high dimensional data points, estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbors to represent high dimensional data in lower dimensions. What we would need is a visualization technique that would map the high dimensional space of features to 2 or 3 dimensions so that we have an idea of the presence of

several clusters indicating distinct syllable types. We would then expect the number of clusters to match (in principle) the number of classes. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a powerful technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets [10]. In Fig. 4 we see the results of applying this technique on humpback syllable segments. The method indeed is capable of bringing this real-life data set into clusters and discovering the intrinsic dimension of the number of syllables. Although the correct number of clusters is unknown to us and the manual labelling of Fig. 4 can be done in many ways, we can see that we have a number of clusters between 9 and 15. This estimation now allows to set the number of classes in several clustering methods and examine the outcome.

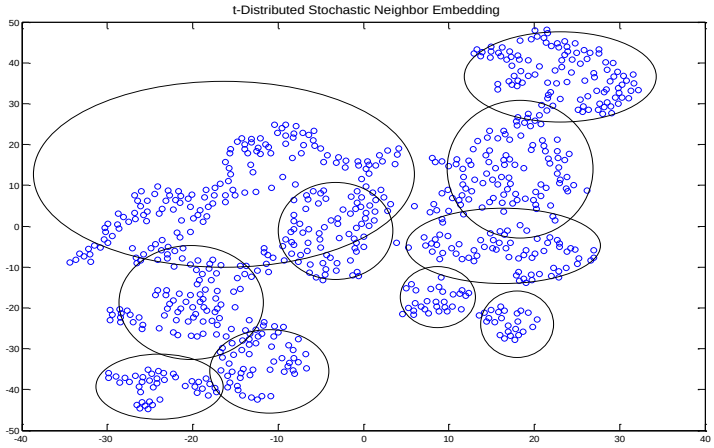


Fig. 4: t-DSNE approach on whale syllables. Ellipses inserted manually

4 Clustering Models and Results

The k-means is a well-known method of vector quantization and quite popular as a means to apply clustering to a dataset. k-means aims to partition the observations in clusters where the mean of the cluster serves as prototype. As soon as the observations are clustered a new mean is calculated for each cluster. The process is repeated until no significant change is detected. The k-means requires that the number of clusters is set *a -priori*. The hand-labelled syllables and the t-DSNE lead us to set the number of clusters in K-means from 8-11 and to perform a series of experiments.

Affinity Propagation is based around the idea of examining the suitability of having one observation to be the exemplar of the other. Therefore messages between observations are sent until convergence. The dataset is then described by the small number of exemplars [11]. Affinity Propagation is quite suitable to our problem because rather than requiring the number of clusters to be known *a-priori*, it discovers it itself. A threshold of preference=-8 is set in this algorithm which discovers 10 clusters [12].

For visualization clarity we show the clustering results of the 34 first members of S1, S2 and S5 as classified by K-means in Fig. 5 and using Affinity Propagation in Fig. 6 we show S8,S5 and S1 syllable clusters.

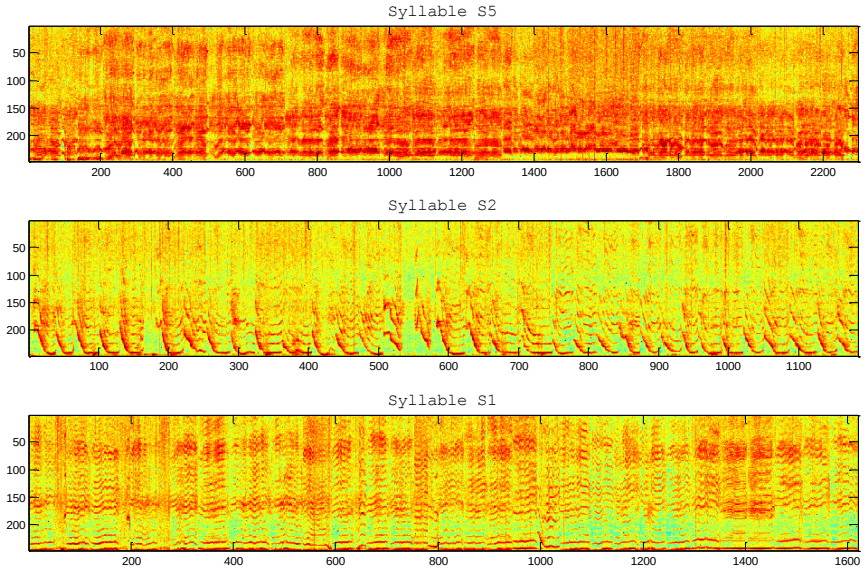


Fig. 5: k-means as applied to syllables dataset. The first 34 syllables of each cluster.

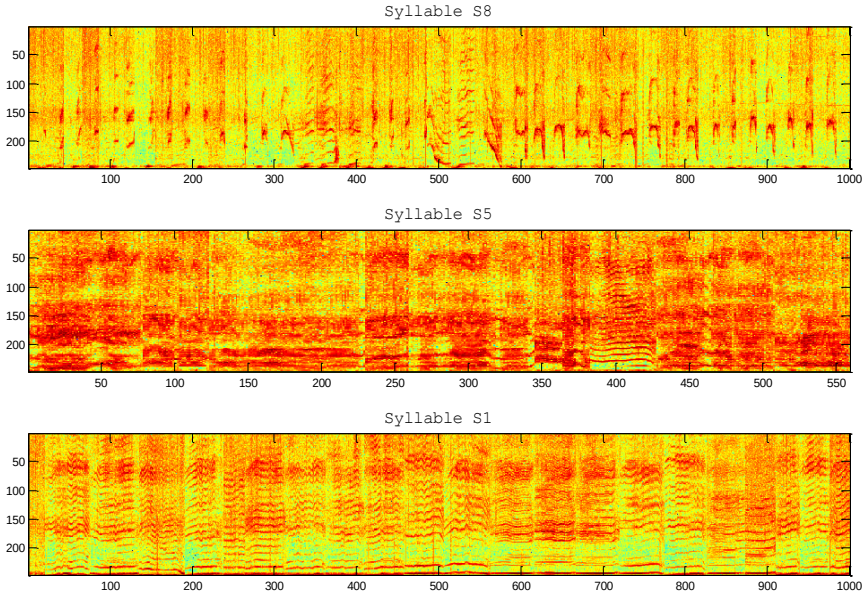


Fig. 6: Affinity Propagation. Spectrograms of the first syllables in each cluster.

5 Temporal Evolution

The humpback whale song demonstrates strong temporal regularities. That is, the syllables do not appear in random order but seem to construct phrases that are uttered as regular and repeated temporal sequences. In order to model this effect we used the cluster labels of k-means as an alphabet and tagged the whole recording. We then proceeded in calculating the transition probabilities from syllable to syllable (bigrams). Silence duration is modelled as Gaussian distributed with mean and variance as derived from Fig. 3. Although the duration is clearly bimodal as a first approach we sampled silence duration from a Gaussian having as mean 0.66 and standard deviation 0.534.

In Fig. 7 (left) we have the number of unigrams, that is, the frequency of each syllable in the alphabet and in Fig. 7 (right) the bigram frequencies in a Hinton diagram. A Hinton Diagram gives an immediate view of the probability of moving from one syllable to another. Bigram frequencies are calculated by counting the number of occurrences of every transition from syllable to syllable and normalizing by the number of all transitions. The larger the box the larger the transition probability. One should note that modelling of temporal evolution is based on the derived alphabet tagged by the k-means and Affinity projection methods and cannot correct their mistakes. The following results as shown in Fig. 7 are based straight on the raw classification and only shown as a proof of concept. A detailed n-gram of manual labels as well as a more accurate clustering of the syllable in order to have a realistic model of syllable transitions will be shown elsewhere. We propose that bigrams are an indispensable tool for studying phrase composition of whale phrases and subsequently of songs based on phrase transitions.

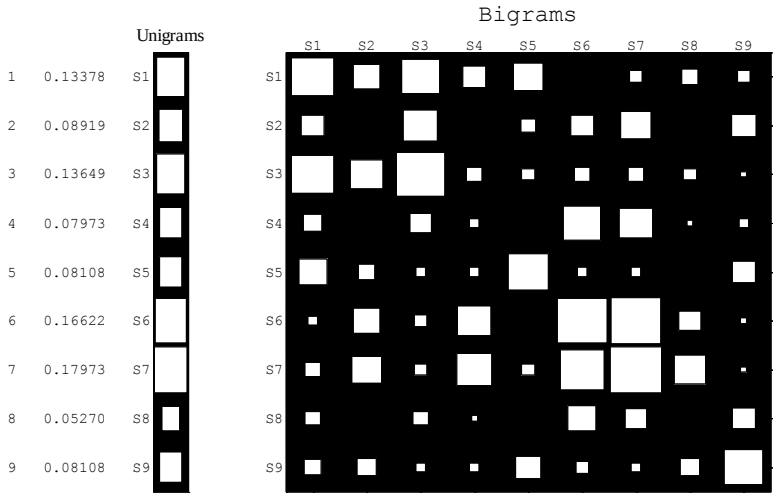


Fig. 7 Left: Frequency of humpback whale alphabet (unigrams). Right: the alphabet with transition probabilities from syllable to syllable (bigrams).

Discussion

We have shown that it is possible to have an accurate decomposition of a raw recording of a whale song into its constructing syllables and a way to model the transition between syllables in order to study the sequencing of the song. The whole process is quite practical as it takes under half minute from the raw 26 minute recording to the final clustered segments on an i7, 16 GB RAM computer. One practical significance of temporal modelling beyond studying the transition pattern and phrases of whales is that it can be used in games either as standalone devices or programs or in audio-books. These usually use pre-recorded signal segments that are very small due to memory constraints that the device imposes and therefore completely predictable to the point of being annoying after a while. With our approach the device only needs to store the syllables and an endless song can be derived on the spot by sampling first a syllable from its cluster with uniform probability, then a silence duration from the Gaussian distribution that we fitted on real data and then moving to the next syllable with probability given by the bigram calculations. This process can be repeated as long as is needed and produces a rich repertoire that is quite realistic and actually costs less in memory than a single recording of a song.

References

[1] Cox T. et al., Understanding the impacts of anthropogenic sound on beaked whales, Journal of Cetacean Resources Management 7(3):177–187, 2006

[2] Laist D., Knowlton A., Mead J., Collet A., Podesta M., Collisions between ships and whales, Marine Mammal Science 17(1): 35–75, 2001.

[3] Adam O. & Samaran F., Detection, Classification and Localization of Marine Mammals using passive acoustics. 2003-2013: 10 years of international research, 2013.

[4] Payne RS, McVay S. Songs of humpback whales. Science 173: 585–597, 1971.

[5] Mercado, E., III, Herman, L. M., & Pack, A. A, Stereotypical sound patterns in humpback whale songs: Usage and function. Aquatic Mammals 29: 37-52, 2003.

[6] Au W, Pack A, Lammers M/, Herman L, Deakos M, Andrews K. Acoustic properties of humpback whale songs. Journal of the Acoustical Society of America 120 (2):1103-10, 2006.

[7] http://sabiod.univ-tln.fr/nips4b/challenge2.html (date last viewed 21/11/2013)

[8] Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. Subunit definition for humpback whale call classification. Applied Acoustics, Elsevier: 11(71), 2010.

[9] Peeters G., A large set of audio features for sound description (similarity and classification) in the CUIDADO project, Report 2004.

[10] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9 (Nov): 2579-2605, 2008.

[11] Brendan F., Delbert D., Clustering by Passing Messages between data points. Science Feb. 2007.

[12] Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12: 2825-2830, 2011.

Chapter 9

Big Bio-Acoustic Data

9.1 NIPS4B: Neural ĩnformation ĻJrocessing ăcaled for ōioacoustics

Lake Tahoe, Nevada, November 10, 2013

Cabled Observatory Acoustic Data: Challenges and Opportunities

Maia Hoeberechts
Ocean Networks Canada
University of Victoria
TEF160 - 2300 McKenzie Ave
Victoria, BC V8P 5C2
Canada
maiah@uvic.ca

Ocean Networks Canada operates several world-leading ocean observatories including the NEPTUNE observatory off the West coast of Vancouver Island, British Columbia, the VENUS coastal observatory in the Salish Sea, British Columbia and the Cambridge Bay mini-observatory in the Canadian Arctic. These observatories connect instruments to shore stations via submarine cables permitting the collection of data on physical, chemical, biological and geological aspects of the ocean over long time periods. Analysis of data from co-located sensors and integration with complementary data sets enables interdisciplinary research on complex Earth-Ocean processes. Among the wide variety of instruments deployed are hydrophones in diverse settings: in the Strait of Georgia, a busy marine shipping channel (100 m – 300 m), in Barkley Canyon, a submarine canyon on the continental slope (400 m – 1000 m), at Cascadia Basin, in the middle of an abyssal plain (2660 m), at Folger Passage, a near-shore site at the mouth of an inlet channel (100 m) and at Cambridge Bay in the Arctic (7 m). Researchers have used Ocean Networks Canada hydrophone data in a multitude of ways, from the study of seismicity to the detection of fish sounds, from the analysis of shipping traffic to the study of marine mammals.

Cabled observatory deployments permit data to be acquired continuously, over long time periods. This capability presents a “big data” challenge to the scientist using and accessing the data, and likewise to the designers of the observatory data archive. Automated analysis, including the classification of acoustic signals, event detection, data mining and machine learning to discover relationships among data streams are techniques which promise to aid scientists in making discoveries in an otherwise overwhelming quantity of acoustic data. Increasing numbers of deployed instruments and the ever-growing data archive necessitate scalable, efficient solutions for data analysis and delivery. This talk will survey research in bioacoustics and automated analysis currently underway with observatory hydrophone data, and outline challenges and open problems which present opportunities for interdisciplinary collaboration and new innovations in bioacoustics.

DEEP SEA SCIENCE.

OCEAN NETWORKS CANADA SCIENCE

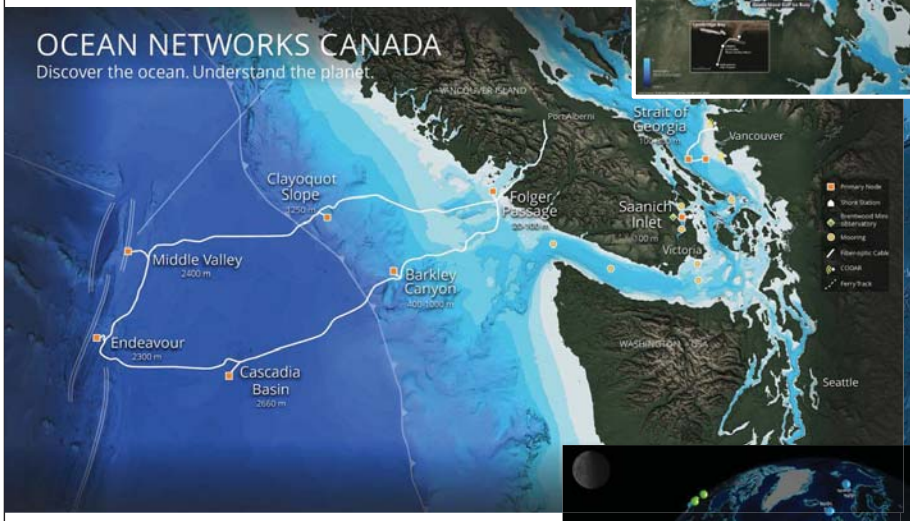
Cabled Observatory Acoustic Data: Challenges and Opportunities

Dr. Maia Hoeberechts
NIPS4B Workshop, Lake Tahoe, NV
December 10, 2013

AN INITIATIVE OF University of Victoria

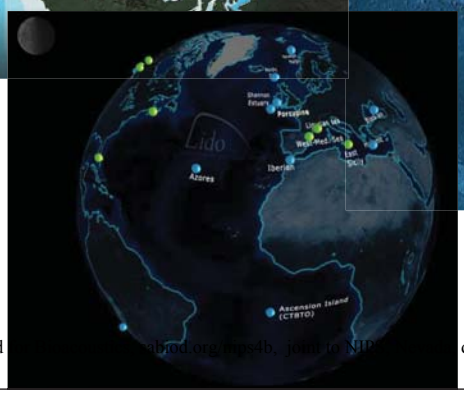
OCEAN NETWORKS CANADA SCIENCE

World-wide expansion



Ocean Networks Canada (ONC)

Ocean Observatories Initiative (OOI)



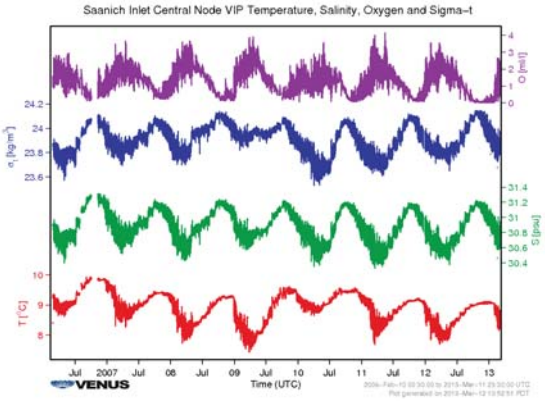
In: proc. of int. symp. Neural Information Scaled ... ...ebiod.org/nips4b, joint to NIPS ... dec. 2013, Ed. Glotin H. et al. ...

University of Victoria

OCEAN NETWORKS CANADA SCIENCE

Characteristics of observatory instrumentation

» Continuous presence
» High sampling frequency
» Co-located sensors



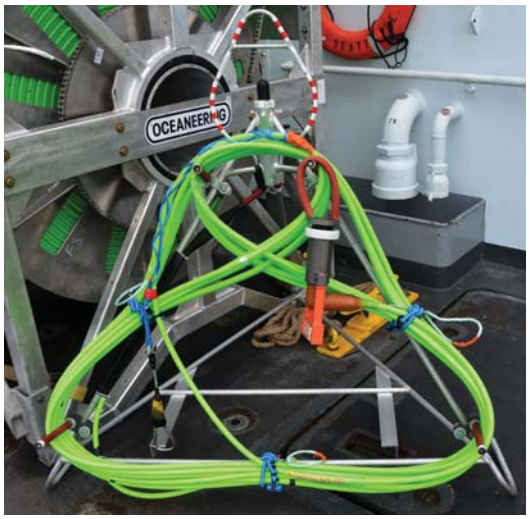
» Interactivity
» Abundant power
» Event detection

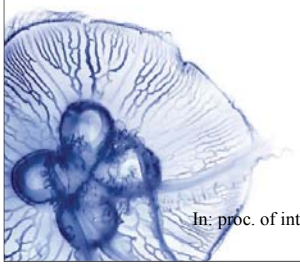
CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

OCEAN NETWORKS CANADA SCIENCE

Hydrophones (example)

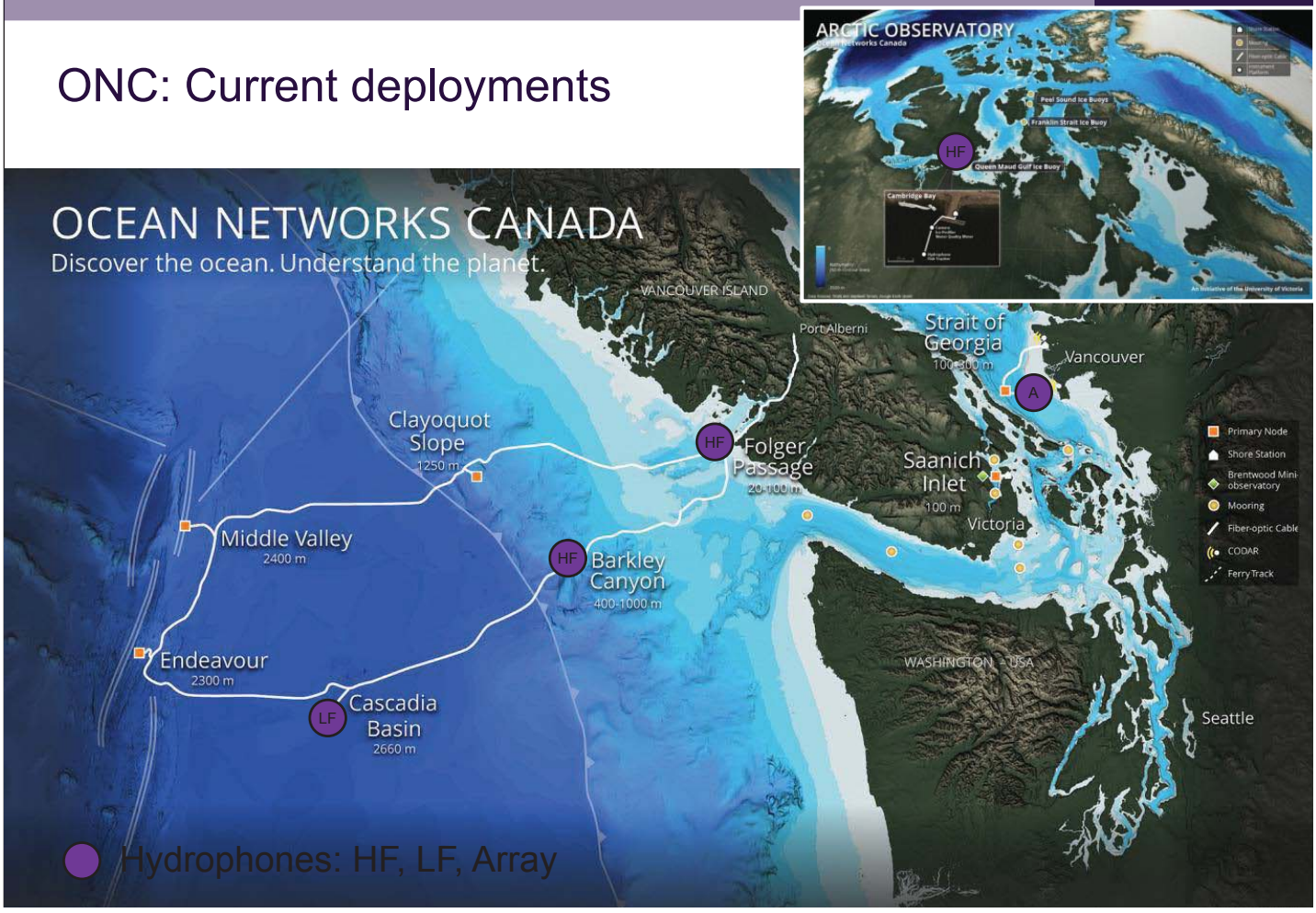
» OceanSonics icListen Smart Hydrophones
» Ethernet-ready
» 24-bit HF (10 Hz – 200 kHz)
» 24-bit LF (1 Hz – 1600 Hz)





University of Victoria

ONC: Current deployments



CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

OCEAN NETWORKS CANADA SCIENCE

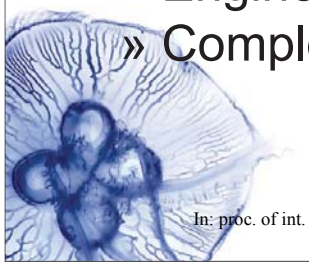
Big Data Challenge

On Ocean Networks Canada's observatories:

> 250 science instruments in the water
> 8.5 million point measurements / day

All parsed, calibrated, QC checked and archived upon arrival: *~100 million data manipulations / day !*

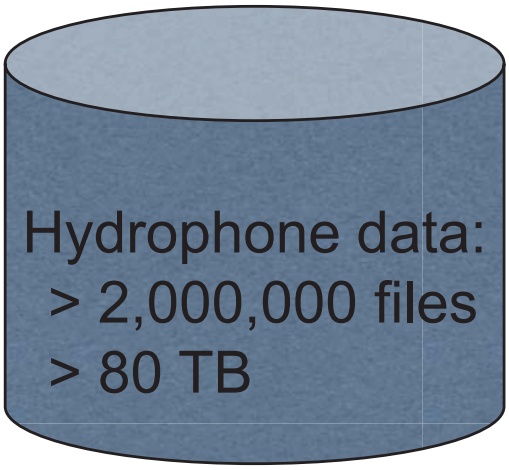
Plus:
» Engineering data = 5x this amount!
» Complex data, video, hydrophone etc. > 30 TB / year



OCEAN NETWORKS CANADA SCIENCE

ONC archived data

HF hydrophone:
24 bits, 96 kHz

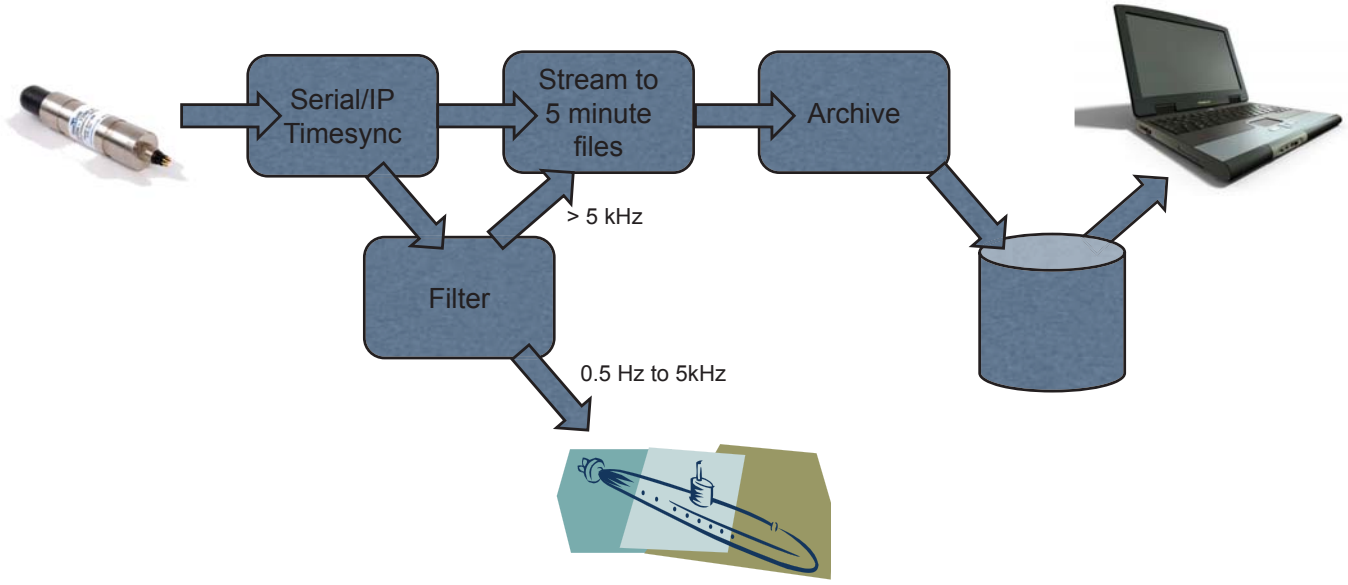


+ active acoustic instruments…

CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

OCEAN NETWORKS CANADA SCIENCE

Data acquisition process



In: proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

University of Victoria

OCEAN NETWORKS CANADA SCIENCE

Selected research highlights (1/3)

Averaging underwater noise levels for environmental assessment of shipping.
Nathan D Merchant, Philippe Blondel, D Tom Dakin, John Dorocicz,
The Journal of the Acoustical Society of America, Volume: 132, Issue: 4 (2012)

Addressed need for method of averaging local shipping noise to assess effects on marine mammals.

Analysed 110 days of continuous data from Strait of Georgia (VENUS observatory).

© Mark Malleson

CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

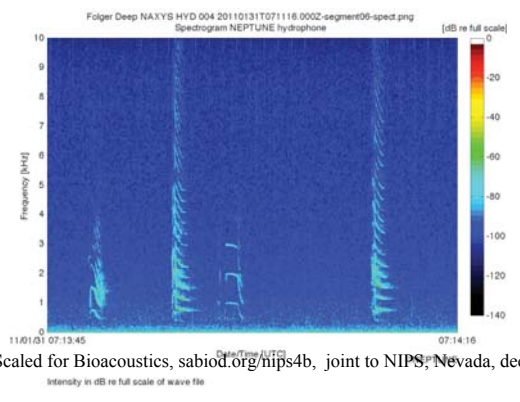
OCEAN NETWORKS CANADA SCIENCE

Selected research highlights (2/3)

Automatic Event Detection for Long-term Monitoring of Hydrophone Data.
F. Sattar, P.F. Driessen, G. Tzanetakis, S.R. Ness, W.H. Page
In Communications, Computers and Signal Processing (PacRim), 2011 IEEE Pacific Rim Conference on (pp. 668-674), IEEE.

Automated event detection using two-stage denoising process followed by event detection function which estimates temporal predictability.

Analysed data from NAXYS hydrophones (NEPTUNE observatory).



Transient killer whale calls recorded on NEPTUNE observatory

University of Victoria

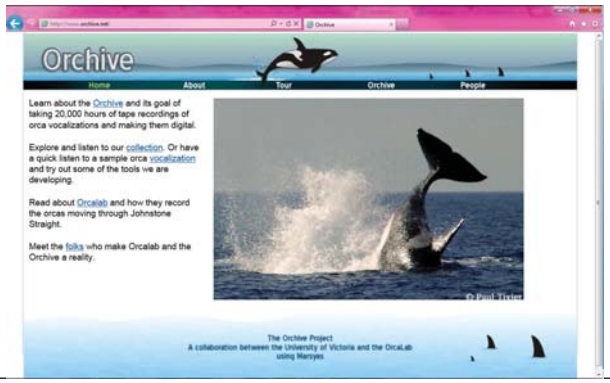
OCEAN NETWORKS CANADA SCIENCE

Selected research highlights (3/3)

Automatic Event Detection for Long-term Monitoring of Hydrophone Data.
F. Sattar, P.F. Driessen, G. Tzanetakis, S.R. Ness, W.H. Page
In Communications, Computers and Signal Processing (PacRim), 2011 IEEE Pacific Rim Conference on (pp. 668-674), IEEE.

Collaborative web-based interface for annotation enhanced with automatic retrieval and classification. Mobile client for crowdsourcing introduced.

Based on data from OrcaLab (Hanson Island, BC) but method also applicable to observatory data.



Visit http://www.orchive.net

AN INITIATIVE OF University of Victoria

CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

OCEAN NETWORKS CANADA SCIENCE

Data access & tools

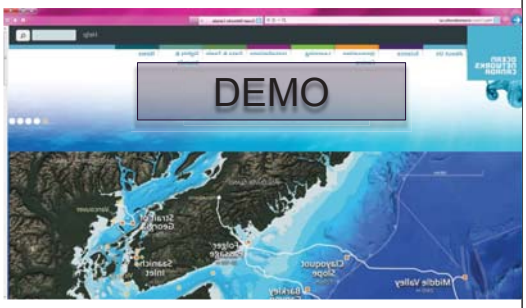
» Web-based data search tools

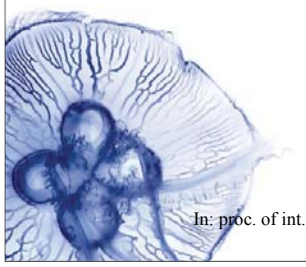
» Code Runner (cloud processing tool)

» Web services

» Audio highlights

» Online visualization (prototype)





University of Victoria

OCEAN NETWORKS CANADA SCIENCE

Feedback: What tools are useful in your research?

- » Visualization ideas?
- » On-line annotation?
- » Event detection?
- » Collaborative tools?
- » Data formats?

Neural Information Processing Scaled for Bioacoustics… we have a long way to go!

AN INITIATIVE OF University of Victoria

CABLED OBSERVATORY ACOUSTIC DATA: CHALLENGES AND OPPORTUNITIES

OCEAN NETWORKS CANADA SCIENCE

Connect With Us!

OCEANNETWORKS.CA

Explore Ocean Networks Canada's online resources
Plotting Utility | Data Search | Sea Tube | Digital Fishers

Download free publications, ibooks and apps
Invitation to Science | Deep-Sea Marine Field Guide | Coastbusters

Like us on Facebook. OceanNetworksCanada

Follow us on Twitter @Ocean_Networks

Watch the Ocean Networks Canada Youtube Channel

Contact
Dr. Maia Hoeberechts | Associate Director, User Services |
maiah@uvic.ca

AN INITIATIVE OF University of Victoria

In: proc. of Int. Symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013, Ed. Glotin H. et al.

9.2 A challenge for computational bioacoustics

Lars Kindermann
Alfred Wegener Institute
for Polar and Marine Research
Bremerhaven, Germany
lars.kindermann@awi.de

Abstract

Antarctic Minke whales are the most abundant baleen whale species on earth. As the main target of today's controversial "scientific whaling" and possibly of a re-established commercial whaling enterprise as proposed by some countries, they are in the focus of interest for many NGOs and the public. Until few month ago nothing was known about their vocal behavior, so they had no "own voice" and no bioacoustic methods could be used to investigate the many open questions about them. On the other hand, for several decades a strange sound of unknown origin has been recorded repeatedly in the Southern Ocean – but only during polar winter when the sea is covered almost completely by a dense layer of ice. Long term recordings from our acoustic observatory at the ice shelf show it is in fact the dominant acoustic emission around Antarctica during that time. Tenth of thousands of hours of this sound have been recorded during the last 8 years and are published under an open access policy. And recently, during a winter expedition to Antarctica we could finally assign this sound to the Minkes. We invite everybody to look into that data using advanced methods to extract definitely new knowledge about this important species.

1 Antarctic Minke Whales Acoustics

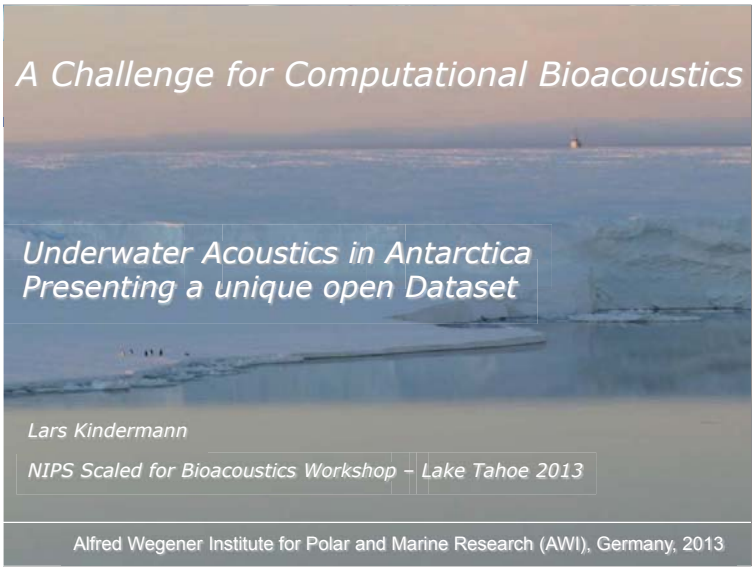
Throughout the southern ocean a unique rhythmic underwater sound with a frequency range of 100 Hz to 20 kHz has been recorded repeatedly by many researchers and navy sonar officers [1-6]. The first published evidence of its existence dates back to 1964 where it appeared in an audio recording as an "unidentified signal in the background" [1]. The crew of an Australian submarine designated the sound "the bioduck" because of is auditory impression [3,4]. The PALAOA observatory, located north of Neumayer Station on the Antarctic Eckström ice shelf [8] and several moored audio recorders throughout the Weddell sea pick up this sound regularly - but strictly during austral winter only, which explains much of the difficulty in its investigation. From end of April to begin of November it is continuously audible and most of that time it even constitutes the most intense sound source in the southern ocean. However, the source of this signal remained a mystery until recently.

The largest inhabitant of thee winterly pack ice is the Antarctic Minke whale, *Balaenoptera bonaerensis*. Up to 10 meters long and weighing 10 tons it is a rather small member of the baleen whale family. While its larger relatives like blue, fin, and humpback whales mostly leave Antarctica during winter for their subtropical mating grounds, this species has adapted for a permanent life in the ice. Little is known about this most frequent of all great whale species, population estimates differ between 360.000 to 1.000.000 individuals and there are contradicting opinions whether the stock is growing or shrinking. To the public it became famous as the main target of the controversially discussed contemporary whaling. Especially during polar night the study of this animals is extremely difficult.

While it had been suggested, that the unidentified rhythmical sound might be produced by Antarctic Minke whales [2] along with some known irregular downsweeps [7] this was proven by several parties only recently in 2013 [e.g. 9].

So far, no detailed study of the acoustic behavior of this species has been performed yet. The 8 year continuous acoustic recordings from the PALAOA observatory thus provide a unique opportunity to investigate this species for the very first time. As the minke sounds are present continuously from April to November more than 20.000 hours are available in total, making it a good candidate for modern computational methods.

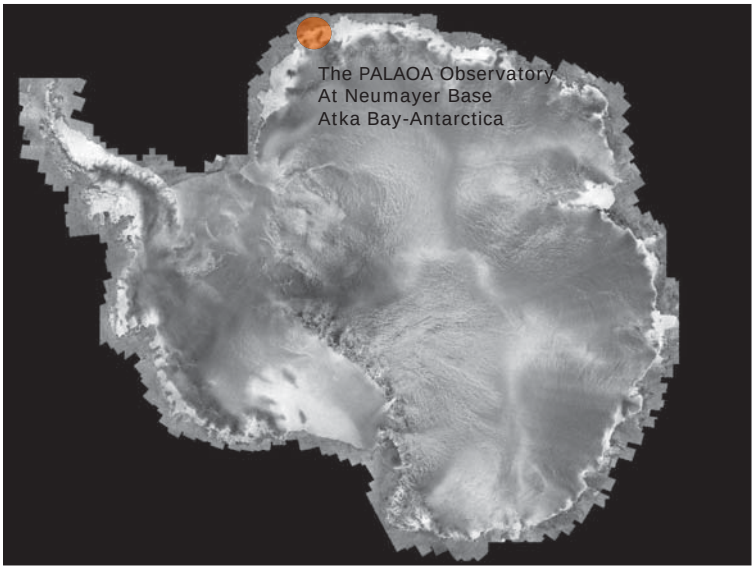
A livestream of the under-ice hydrophone is available under http://www.awi.de/PALAOA and the complete data set (2005-2013) is published in the PANGAEA database of the World Data Center for Marine Environmental Sciences [10]: http://doi.pangaea.de/10.1594/PANGAEA.773610


A Challenge for Computational Bioacoustics
Underwater Acoustics in Antarctica
Presenting a unique open Dataset
Lars Kindermann
NIPS Scaled for Bioacoustics Workshop – Lake Tahoe 2013
Alfred Wegener Institute for Polar and Marine Research (AWI), Germany, 2013


PALAOA-S
Neumayer

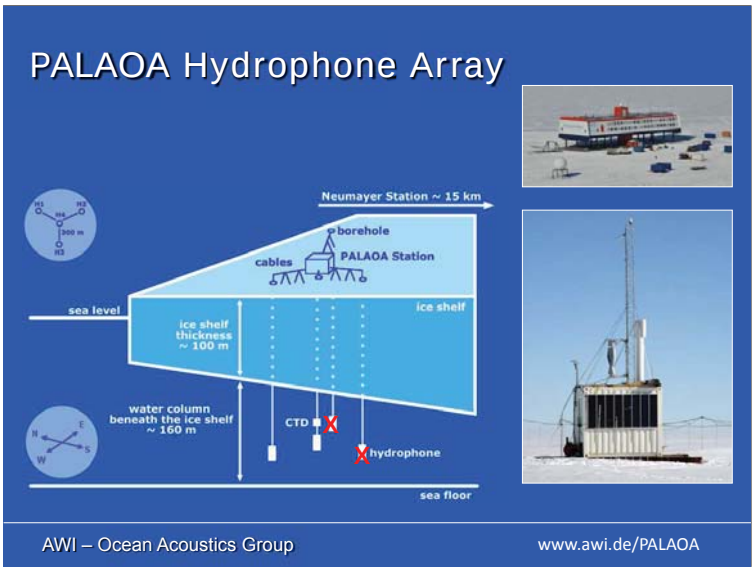

The Southern Ocean - One of the most productive ecosystems on earth during austral summer

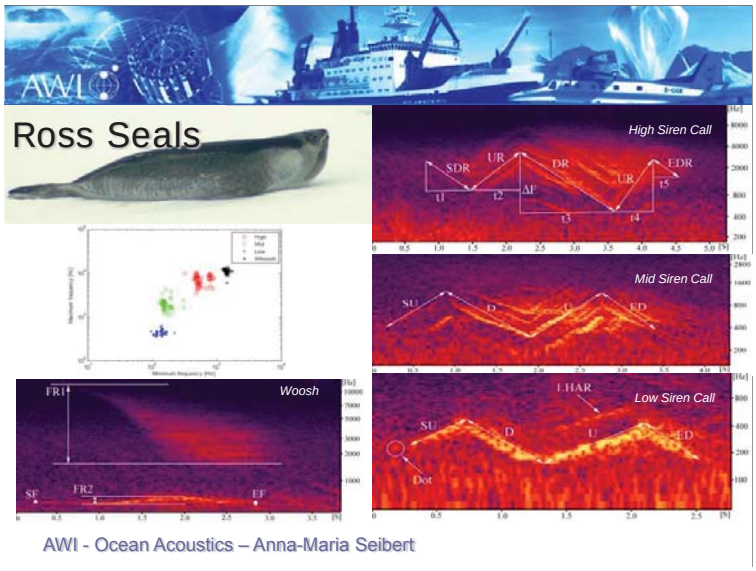

“Ice finger” at the tip of the Eckström Ice shelf
PALAOA site


The PALAOA Observatory
At Neumayer Base
Atka Bay-Antarctica

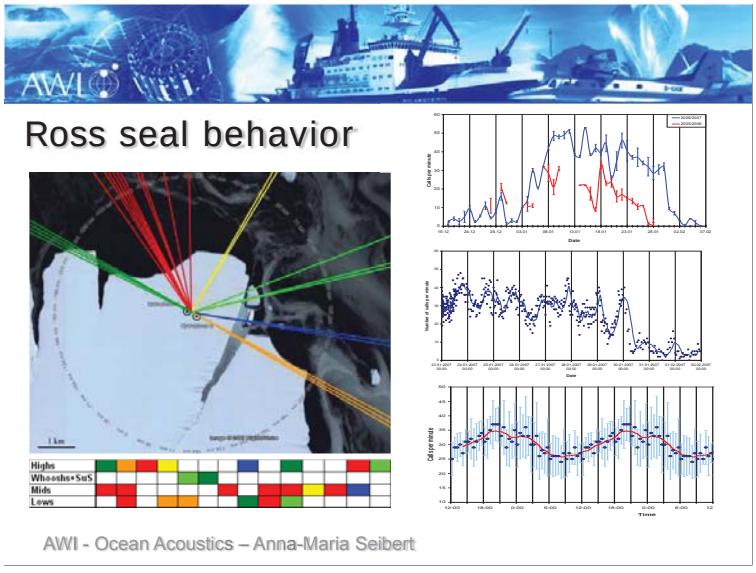

Drilling camp


AWI hot water drilling facility

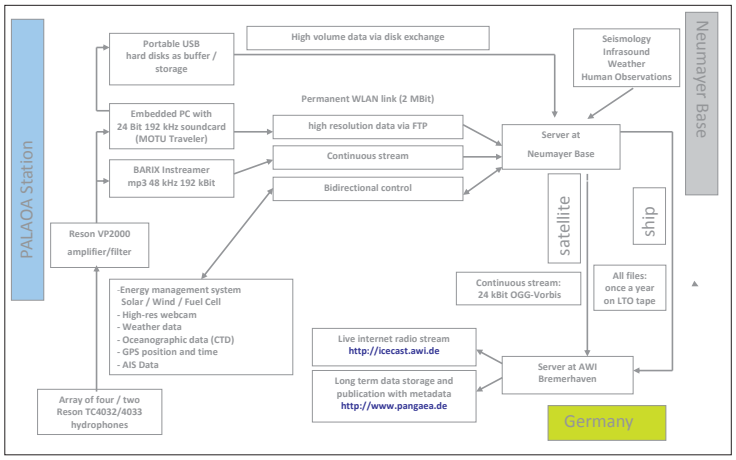

PALAOA Hydrophone Array
Neumayer Station ~ 15 km
borehole
cables
PALAOA Station
sea level
ice shelf
ice shelf thickness ~ 100 m
water column beneath the ice shelf ~ 160 m
CTD
hydrophone
sea floor
AWI – Ocean Acoustics Group
www.awi.de/PALAOA

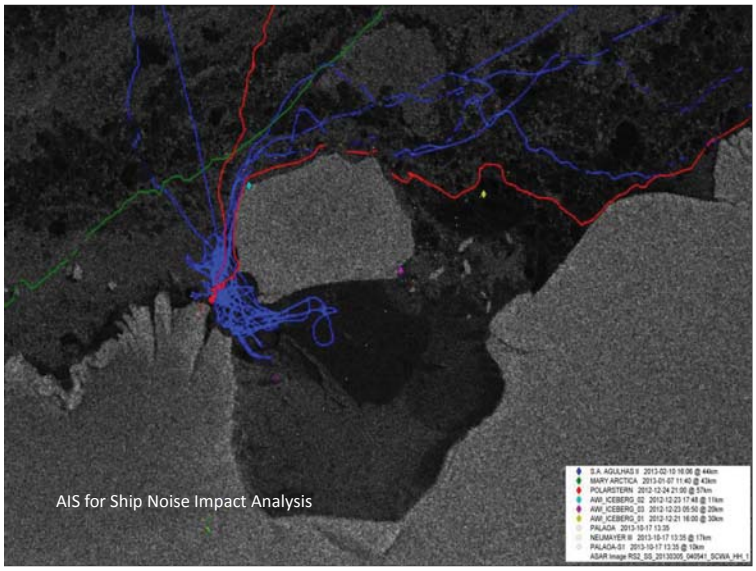

Ross Seals
High Siren Call
Mid Siren Call
Low Siren Call
Woosh
AWI - Ocean Acoustics – Anna-Maria Seibert


PALAOA

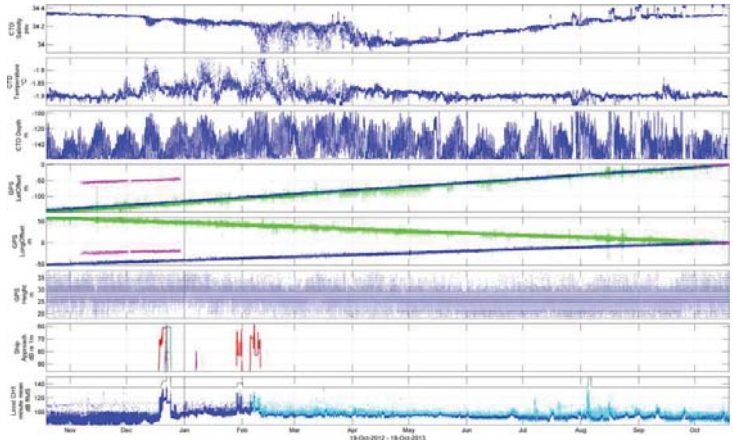

Ross seal behavior
Highs
Whoosh+SnS
Mids
Lows
AWI - Ocean Acoustics – Anna-Maria Seibert

Data and control flow of the PALAOA observatory

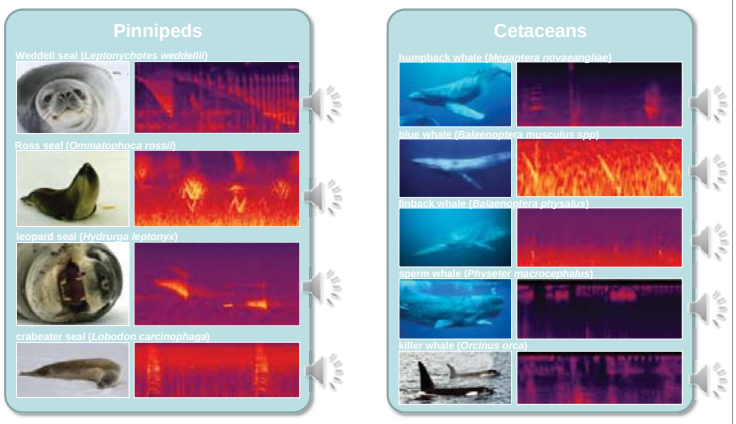




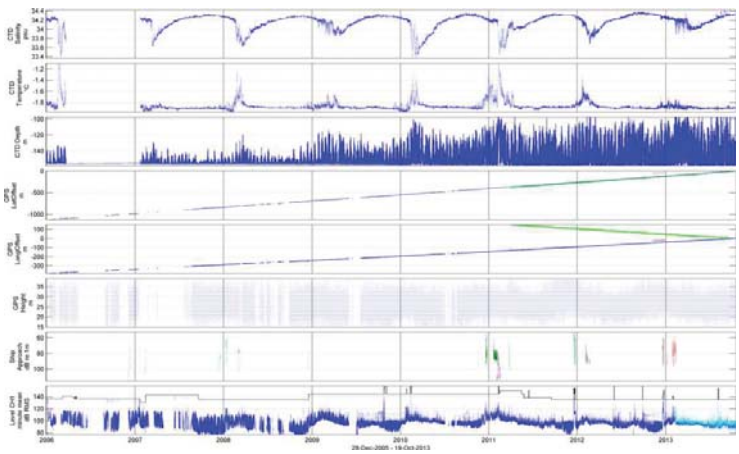
CTD - GPS - AIS – Noise levels: 1 year data

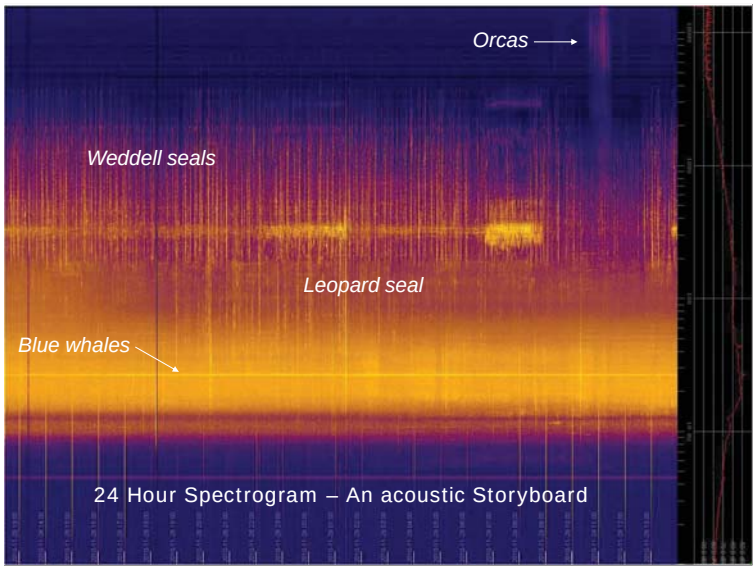


Identified vocalizations in the PALAOA recordings

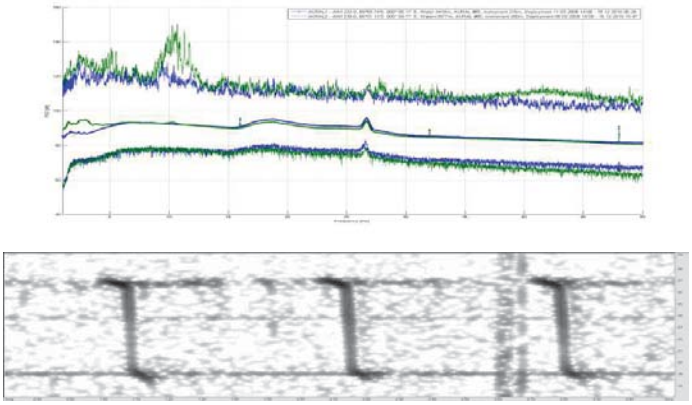


CTD - GPS - AIS – Noise levels: 8 Year data



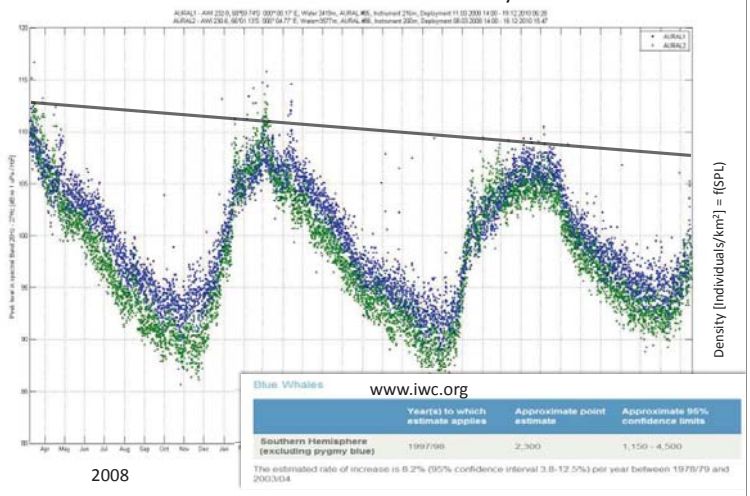


Blue Whale Z-Calls

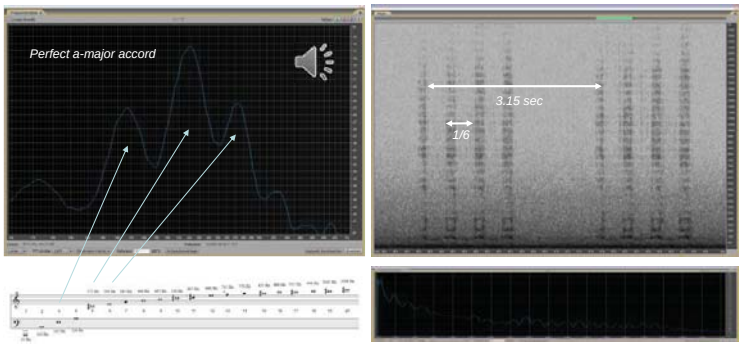


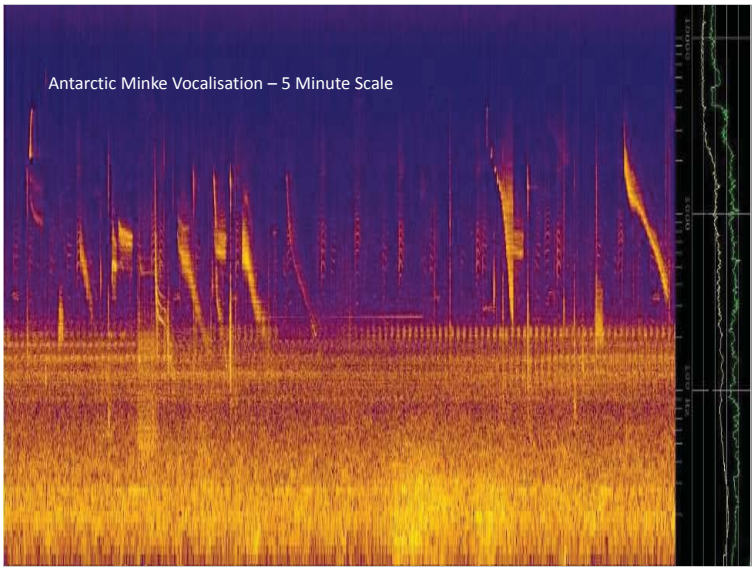
Spectrogram of blue whale "Z"-calls, recorded by MARU#1 on 28.12.2008 17:55

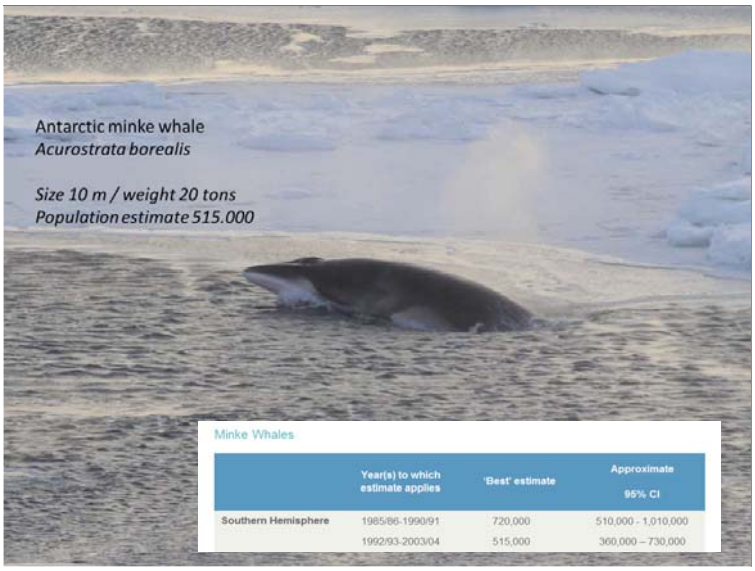
Blue Whale Chorus Intensity

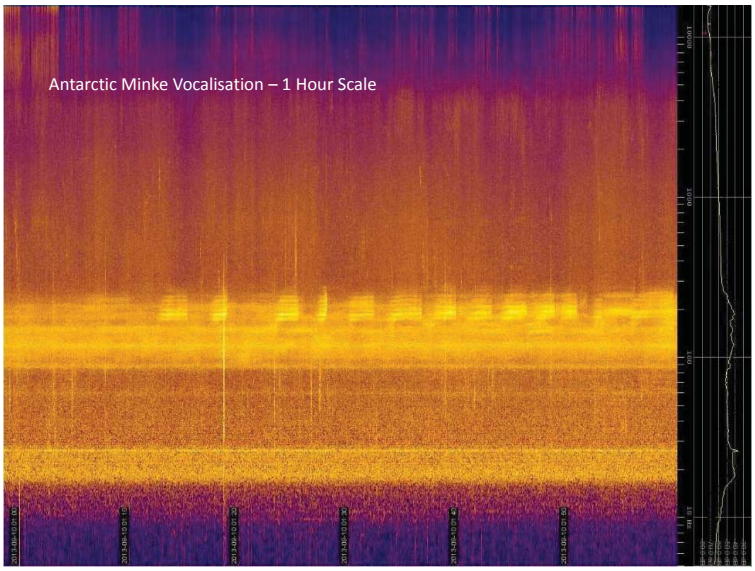


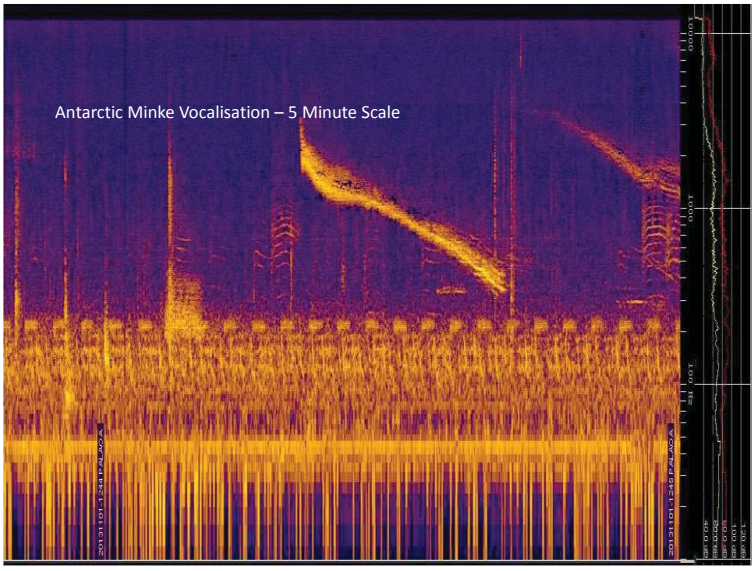
Just identified: A unique vocalisation

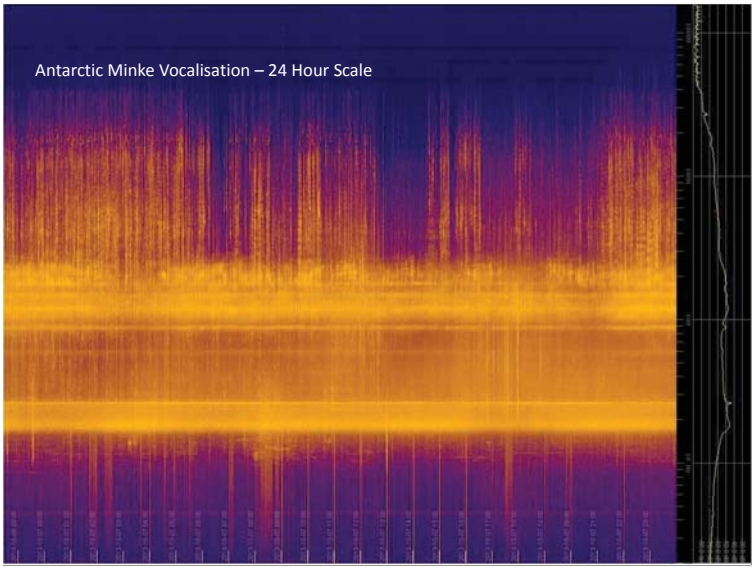


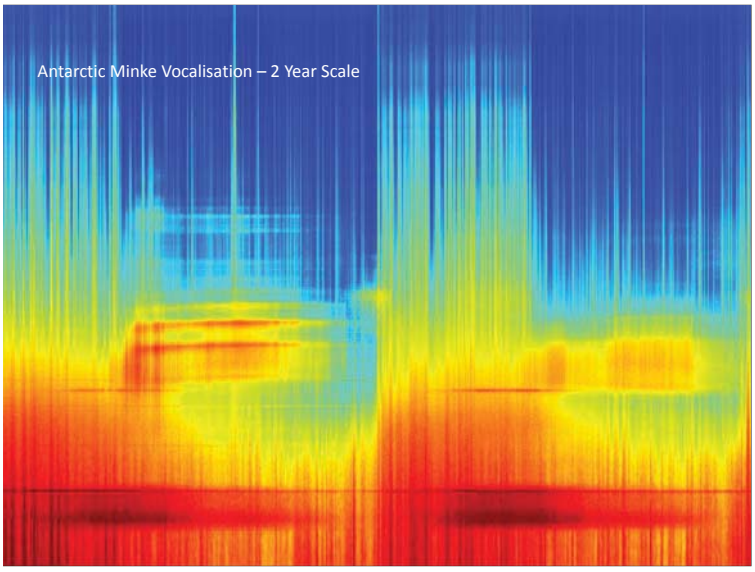











Antarctic Minke Vocalisation – 2 Year Scale


Open Access Live Data: www.awi.de/PALAOA
2005-2013: 50.000 Hours of Audio and Metadata
CTD - Weather – Webcam - AIS – Satellite Images - Spectrograms
Published In the World Data Center for Marine Environmental Data (WDC-MARE)
www.PANGAEA.de

References

1. Poulter, T. Recording # 120342 (December 26th, 1964). Macaulay Library (1964).
2. Matthews, D., Macleod, R. & McCauley, R. D. Bio-Duck Activity in the Perth Canyon. An Automatic Detection Algorithm. Proceedings of Acoustics 1–4 (2004).
3. McCauley, R. Western Australian Exercise Area Blue Whale Project: Final Summary Report. CMST Report R2004-29, Project - 350 1–73 (2004).
4. Dolman, S. J., Swift, R. J., Asmus, K. & Thiele, D. Preliminary analysis of passive acoustic recordings made in the Ross Sea during ANSLOPE III in 2004. Paper SC/57/E10 presented to the Scientific Committee of the International Whaling Comission 1–8 (2005).
5. Klinck, H. & Burkhardt, E. Marine mammal automated perimeter surveillance - MAPS. Reports on Polar and Marine Research 580, 114–121 (2008).
6. van Opzeeland, I., Rettig, S., Thomisch, K., Preis, L., Lefering, I., Menze, S., Zitterbart, D., Monsees, M., Kindermann, L. & Boebel, O. in Boebel: Cruise Report of the Expedition of the Research Vessel "Polarstern" to the Antarctic in 2012/13 (ANT-XXIX/2). To appear in: Reports on Polar and Marine Research.
7. Schevill, W. E. & Watkins, W. A. Intense low-frequency sounds from an antarctic minke whale, Balaenoptera acutorostrata. Breviora 388, 1–8 (1972).
8. Kindermann, L; Boebel, O; Bornemann, H et al. (2007): A perennial acoustic observatory in the Antarctic Ocean. Computational bioacoustics for assessing biodiversity: proceedings of the international expert meeting on IT-based detection of bioacoustical patterns, December 7th until December 10th, 2007 at the International Academy for Nature Conservation (INA), Isle of Vilm
9. Kindermann L. & Cabreira, A. Acoustic Ecology of Antarctic Minke Whales. In: Lemke, P., Cruise Report of the Expedition of the Research Vessel "Polarstern" to the Antarctic in 2013 (ANT-XXIX/6). To appear in: Reports on Polar and Marine Research.
10. Kindermann, L. (2013): Acoustic records of the underwater soundscape at PALAOA with links to audio stream files, 2005-2011. Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Bremerhaven, doi:10.1594/PANGAEA.773610

Schedule of the Workshop

07:30 Introduction
Glotin - A Bioacoustic Turing test ?

07:40 * **Natural Neural Bioacoustic Learning**
07:40 Tchernichovski - Physiological brain processes that underlie song learning
08:10 Pollack - Neuroethology of hearing in crickets: embedded neural process to avoid bat
08:45 Stathopoulos - Bat call classification

09:00 Coffee Break

09:25 * **Representation for bioacoustics**
09:25 Glotin & Razik - Sparse coding & whale tracking and song evolution
09:45 Halkias - SAE & DBN for whale classif.

10:05 * **Advanced ANN**
10:05 LeCun - ConvNets & DNN for Bioacoustics
10:40 Kindermann - ANN for sequences interpolation

10:55 * **Learning to Track by Passive Acoustics**
10:55 Doh - Inter Spectral Attenuation ANN: Range & Bearing Physeter est.
11:03 Paris - Physeter Localization: Sparse coding & Fisher vectors
11:20 Mathias - Physeter Multiple Range estim.
11:30 Mishchenko - Bat tracking with LM acceleration

11:35 * **Non Human speech processing**
11:35 Trone - Speech of Dolphin : transient formant ?
11:45 Janvier - Speech of Monkey ?
11:50 Shokoohi - Mouse Genome & Biocoustics

12:10 Lunch break
13:30 Posters (1/2) & Discussion

15:30 * **Bird song multilabel multi-instance Challenge Kaggle NIPS4B**
15:30 Dufour - Challenge overview / Baseline
15:40 Lasseck - Winner of Kaggle Bird NIPS4B
15:48 Stowell - http://vimeo.com/81440385
15:55 Potamitis - Bird syllabic classif.

16:00 * **Whale song Challenge**
16:00 Mercado - www.youtube.com/watch?v=YHM18JmC9Eo
16:05 Potamitis - Eff. syllabic clustering
16:10 Bartcus - Infinite Parcim. GMM
16:15 Cazeau - Chaos and whale song
16:20 Randall - Gabor Scatnet filtering
16:20 Posters (2/2) & Discussion
17:00 Coffee break

17:30 * **Feature learning**
17:30 Elie - Data driven bearing features

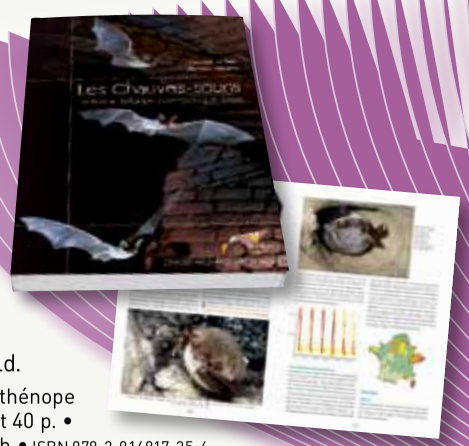
17:50 * **BIG Bioacoutic DATA**
17:50 Hoeberechts - Canadian submarine bioacoustic Big Data
18:05 Kinderman - Antartic submarine bioacoustic Big Data
18:20 Glotin - Mediterranean submarine Big Data
18:25 Panel Discussion : AI & Bioacoustics

18:45 Closing

Publishing house

Les Chauves-souris de France, Belgique, Luxembourg et Suisse

Bats of France, Belgium, Luxembourg and Switzerland



This book presents the 34 bats species living in Western Europe. The book opens on general chapters on bats, their ecology, relationship with man... The reader can find detailed descriptions of all species with distribution maps, sonograms... Generously illustrated, this book also contains a removable identification booklet to be taken in the field.

Authors: Laurent ARTHUR and Michèle LEMAIRE • Parthénope series – Coproduction Biotope/MNHN • 544 p. + booklet 40 p. • 16.5 × 24 cm • Price: 42.65 € excl. VAT • Book in French • ISBN 978-2-914817-35-6

Écologie acoustique des chiroptères d'Europe

Acoustic ecology of European chiropters

DVD included

This book provides an overview of the knowledge gathered during 20 years on ultrasound detection. It brings a new dimension to the first purely hearing approaches: the identification of ultrasounds by computer analyses. Thanks to high technology devices and computer analysis, the author is demonstrating that it is now possible to identify 80% of ultrasounds and to lead inventories and deep studies without disturbing bats. This new indispensable handbook includes a DVD containing sound samples illustrating the methods as well as graphical files required for species identification.

Author: Michel BARATAUD • Inventaires et biodiversité series – Coproduction Biotope/MNHN
340 pages • 16.5 × 24 cm • Price: 42.65 € excl. VAT • Book in French • ISBN 978-2-914817-82-0

Online shop: www.leclub-biotope.com

Nature picture library

This flyer has been designed by Biotope and illustrated with photographs from our picture library. We can produce all kind of communication documents and provide nature photographs.
Simply ask us!

www.biotope.fr/phototheque/piwigo
Contact: phototheque@biotope.fr

BIOTOPE – 22 boulevard Maréchal Foch – BP 58 – 34140 MÈZE (FRANCE)
Phone: (+33) 04 67 18 46 20 • Fax: (+33) 04 67 18 46 29 • **www.biotope.fr**

Production: Biotope (communication@biotope.fr); graphic design: Joanna Raynal; photos: T. Disca/Biotope

French leader in environmental engineering

biotope

Chirotech

Scientific device

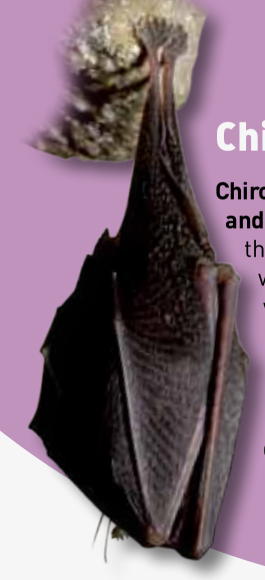
Engineering office

SonoChiro® NEW

Books

Picture library

biotope

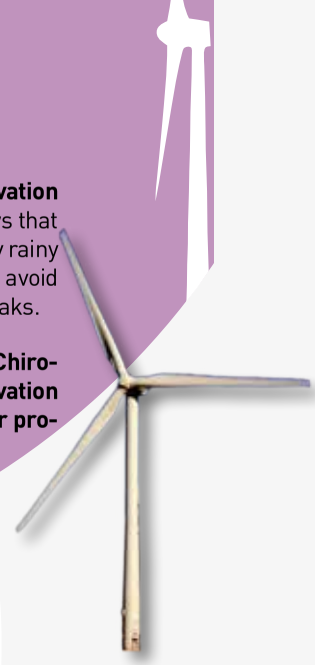


Chirotech

Chirotech© is a tool that makes possible to combine bats' preservation and wind power development. The analysis of bats' behaviour shows that they only fly under given conditions: for instance they rarely hunt by rainy weather, they slow down their activity during colder periods and avoid windy conditions which correspond to wind power production peaks.

By stopping wind turbines during bats' activity periods, the Chirotech© system makes possible the combination of the preservation of these protected species and the development of wind power production.

Contact:
Jim Buzon (jbuzon@biotope.fr)



Scientific device

The SM2BAT+ Stereo box is designed to record every sound, from the audible to the ultrasound, with a good restitution quality.

A very efficient and polyvalent tool!

Studies led by many chiropterologists have demonstrated that the SM2BAT+ Stereo is perfectly adapted to:

- identifying the European bats.
- ensuring two remote clew with cables for SM2BAT+ microphones.
- studying the moving direction of bats.
- quantifying the bats' populations going out of a given place.

This weather-resistant device can continuously monitor and record during long periods of time echolocation sounds of bats but also birds, insects and amphibians sounds.

Bat expert essential kit: (1 545 € excl. VAT)

- 1 SM2BAT+ Stereo
- 2 SMX-US Ultrasonic Microphone
- 2 memory cards SDHC 16 GB
- 4 rechargeable batteries 10000 mAh
- 2 microphone cables (10 and 50 meters)
- 1 security box

Sales: leclub@biotope.fr / www.leclub-biotope.com

The engineering and consulting office: expertise in France and abroad

Leader of ecological engineering consultancies in France, Biotope carries out ecological surveys and Environmental Impact Assessments (EIA) for all kind of planning projects (roads, motorways, railways, etc.) and energy development projects (wind farms, solar farms, hydroelectric systems, thermal power plants, etc.). **Biotope aims at meeting regulatory requirements at local and international levels and helping its clients to build a project that is both economically efficient and environmentally aware**. This approach offers innovation and optimal legal security. Biotope is also a major stakeholder in the framework of nature conservation with an **important participation to the implementation of the Natura 2000** network in France and in its neighboring countries.

Biotope represents 250 employees, 22 agencies in France and abroad and 15 chiropterologists.

Contact: Jean-Yves Kernel (international@biotope.fr)

SonoChiro® software NEW

Automated analysis of European bats ultrasound recordings from SM2BAT+ or other full-spectrum recorders

Biotope has set up an efficient software for automatic processing of ultrasonic bat calls: SonoChiro®.

This software provides a valuable help for all users of ultrasonic bat detectors, particularly in the case of automatic recorders, accumulating large volumes of data.

Its various features make it useful in different contexts of analysis: species identification, bat pass counting, feeding buzz quantification, social activity highlighting, etc.

SonoChiro® is provided as a 1 year subscription including utilization of the software for 1 year, updates and support (by sonochiro@biotope.fr and through an internet forum).

Key features:

- swift and efficient processing of your bats recordings
- identification at 2 levels: species and groups
- automatic detection of feeding buzzes and social calls
- > 90% time-saving for most analyses
- the result of a long beta-test: 15 bat specialists during 2 years!

1 year subscription 705 € (excl. VAT)

Sales: leclub@biotope.fr / www.leclub-biotope.com
Technical contact: sonochiro@biotope.fr

Neural Information Processing Scaled for Bioacoustics - from Neurons to Big Data -

Glotin H., LeCun Y., Artieres T, Mallat S., Tchernichovski O., Halkias X.
Toulon, New-York, Paris - http://sabiod.org

This book is the content of the 1st Big Bioacoustics Data [NIPS4B] that took place at Tahoe lake, Nevada, in december 2013, during the NIPS international conference. The 40 attendies provided further insights into the analysis of large scale bioacoustic data and modeling of animal sounds, not only from a neuro- perspective, but also by highly reinforcing the need to approach these unique signals within the machine learning community.

As a result both the bioacoustics community and the mainstream NIPS community met, leading to new collaborations: the communications ranged from the complexity of bioacoustics to scaled analyses, from understanding and monitoring bird song ontogeny, to cricket auditory neural functions, from use of sparse architectures for whale sound classification, to range estimation and bat tracking...

Although, in recent years, the majority of the existing applications lend themselves to advanced acoustic signal processing methodologies, our efforts are successfully integrating robust processing and machine learning algorithms for scaled analysis of these abundant recordings. Major issues such as data repositories and the need for standardizations within the bioacoustics field discussed and addressed.

We exchanged ideas on how to proceed in understanding bioacoustics to provide methods for biodiversity indexing, and to open a novel paradigm toward a Bioacoustic Turing Test: one might model animal communication before tackling the original Turing test for human being.

The scaled bioacoustic data science is a novel challenge for artificial intelligence that require new methods. For example Minke whales, observed all around the planet have been recorded by Kindermann's acoustic observatory at the ice shelf around Antarctica during 8 years. Big data scientists are today invited to look into that data using advanced methods to definitely new knowledge about this important species.

Similarly, large cabled submarine acoustic observatory deployments permit data to be acquired continuously, over long time periods. For examples, Neptune observatory in Canada, Antares or Nemo neutrino workshop on Neural Information Processing Scaled for observatories in Mediterranean sea are 'big data' challenges to the scientists. Automated analysis, including the classification of acoustic signals, event detection, data mining and machine to discover relationships among data streams are techniques which promise to aid scientists in discoveries in an otherwise overwhelming quantity of acoustic data as it is presented in this book.

We acknowledge the SABIOD.ORG MASTODONS project for its support. ISBN : 979-10-90821-04-0