The 1st International Workshop on Machine Learning for Bioacoustics

joint to

The 30th International Conference on Machine Learning (ICML 2013)

Atlanta, USA, on June 20 – June 21, 2013

Volume 1: Proceedings

Edited by Hervé Glotin, Christopher Clark , Yann LeCun, Peter Dugan, Xanadu Halkias and Jérôme Sueur

ISBN : 979-10-90821-02-6

"This proceedings are available on line, and shall be cited as : Glotin H., Clark C., LeCun Y., Dugan P., Halkias X. and Sueur J., (2013), Proc. of the first workshop on Machine Learning for Bioacoustics, Vol.1, 104 pages, joint to Int. Conf. on Machine Learning, ICML 2013, Atlanta, ISBN 979-10-90821-02-6, http://sabiod.univ-tln.fr/ICML4B2013_proceedings.pdf

The Slides of the talks are available on YouTube (urls are in the section 1.2 below) and in the Vol2: http://sabiod.org/ICML4B2013_proceedings_slides.pdf

Email list of participants

Arik Kershenbaum arik@nimbios.org
Chris Clark: cwc2@cornell.edu
Christian Mason cmason@g.hmc.edu
Dan Stowell dan.stowell@eecs.qmul.ac.uk
Daniel Sheldon sheldon@cs.umass.edu
David Ledbetter ledbetdr@gmail.com
Emily Hockman ehockman@utk.edu
Faicel Chamroukhi chamroukhi@univ-tln.fr
Florencia Noriega flo@nld.ds.mpg.de
Forrest Briggs fbriggs@gmail.com
Gianni Pavan gianni.pavan@unipv.it
Guangzhi Qu gqu@oakland.edu
Harold Mills harold.mills@gmail.com
Herve Glotin: glotin@univ-tln.fr
Ilyas Potamitis potamitis@staff.teicrete.gr
Ishanu Chattopadhyay ishanu.chattopadhyay@cornell.edu
Jean-Marc Prévot: jean-marc.prevot@univ-tln.fr
Jeff Knewstubb jeffk759@gmail.com
Jerome Sueur: sueur@mnhn.fr
Marian Popescu cp478@cornell.edu
Marie Trone mtrone@valenciacollege.edu

Mike Izbicki mike@izbicki.me
Mohammad Pourhomayoun mpourhoma@gmail.com
Nicole Bender Nicole.Bender1@marist.edu
Peter Dugan peterdugan68@gmail.com
Peter Dugan: pjd78@cornell.edu
Sarah Hallerberg shallerberg@nld.ds.mpg.de
Scott Dobson dobsons13@hotmail.com
Sebastian Huebner info@sejona.de
Sebastien Paris sebastien.paris@lsis.org
Steven Ness sness@sness.net
Sunil Shahi sunil.shahi@selu.edu
Wiffried Segretier Wiffried.segretier@gmail.com
Xanadu Halkias xanadu.halkias@univ-tln.fr
Yann Lecun yann@cs.nyu.edu
Yu Shiu ys587@cornell.edu
Yuan Hao yhao@cs.ucr.edu
Ziang Xie zxie@berkeley.edu

Contents

1.Workshop at a glance

1.1. Contributions

Invited talks

Prof. Christopher Clark, W - Cornell University, NY, USA
Prof. D. Sheldon and T. G. Dietterich - Oregon State University, USA
Prof. Hervé Glotin - USTV, Inst.Univ.de France, CNRS LSIS, FR
Dr. Xanadu Halkias - CNRS LSIS and USTV, FR
Prof. Y. Bengio - Department of Computer Science and Operations Research Canada Research Chair in Statistical Learning Algorithms
Prof. Diana Reiss - Hunter College - CUNY, NY USA
Prof. Gianni Pavan Pavia - Italy
Prof. Ofer Tchernichovski - Hunter College - CUNY, NY, USA
Dr. Peter J. Dugan - Cornell University, NY, USA

Full papers

Rami Abousleiman - Oakland University, Department of Electrical and Computer Engineering, Rochester, MI, USA
Guangzhi Qu - Oakland University, Department of Computer Science and Engineering, Rochester, MI, USA
Osamah Rawashdeh - Oakland University, Department of Electrical and Computer Engineering, Rochester, MI, USA
Steven Ness - Department of Computer Science, University of Victoria, Canada
Helena Symonds - OrcaLab, P.O. Box 510 Alert Bay, BC, Canada
Paul Spong - OrcaLab, P.O. Box 510 Alert Bay, BC, Canada
George Tzanetakis - Department of Computer Science, University of Victoria, Canada
Sebastien PARIS - DYNI team, LSIS CNRS UMR 7296, Aix-Marseille University
Yann DOH - DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Herve GLOTIN - DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Xanadu HALKIAS - DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Joseph RAZIK - DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Marian Popescu - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Peter J. Dugan - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Mohammad Pourhomayoun - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Denise Risch - Northeast Fisheries Science Center, Woods Hole, MA, USA, 02543
Harold W. Lewis III - Department of Systems Science and Industrial Engineering, Binghamton University, NY, USA, 13850
Christopher W. Clark - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Mohammad Pourhomayoun - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Peter J. Dugan - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Marian Popescu - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Denise Risch - Northeast Fisheries Science Center, Woods Hole, MA, USA, 02543
Harold W. Lewis III - Department of Systems Science and Industrial Engineering, Binghamton University, NY, USA, 13850
Christopher W. Clark - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Mohammad Pourhomayoun - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Peter J. Dugan - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Marian Popescu - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Christopher W. Clark - Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850
Erick Stattner - LAMIA Lab. University of the French West Indies and Guiana, France
Wilfried Segretier - LAMIA Lab. University of the French West Indies and Guiana, France
Martine Collard - LAMIA Lab. University of the French West Indies and Guiana, France
Philippe Hunel - LAMIA Lab. University of the French West Indies and Guiana, France
Nicolas Vidot - LAMIA Lab. University of the French West Indies and Guiana, France

Short papers

Ales MISHCHENKO - SATT sud-est / DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Herve GLOTIN - DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var
Evgeny Smirnov - Saint-Petersburg State University, Universitetskii prospekt 35, Petergof, Saint-Petersburg, Russia 198504

Bird Challenge worknotes

Emmanouil Benetos - Department of Computer Science, City University London, London, UK.
Forrest Briggs - Oregon State University, Corvallis, OR, 97333, USA
Olivier Dufour - LSIS, Universite du Sud Toulon Var
Thierry Artieres - LIP6, Universite Paris 6
Herve GLOTIN - Universite de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
Pascale Giraudet - Universie du Sud Toulon Var
Dan Stowell and Mark D. Plumbley - Centre for Digital Music, Queen Mary, University of London
Rafael Hernandez Murcia - Carlos III University of Madrid, Spain
Victor Suarez Paniagua - Carlos III University of Madrid, Spain
Jennifer G. Turner, Charles J. Turner - Academic Technology Services, UC Davis, Davis CA 95616 USA

1.2. Videos of presentations

06/20/2013 morning:

1 http://youtu.be/TDnmXgJdbj0 : C. Clark 'Advanced High-performance-computing for Mapping Marine Mammals overs ecologically meaningful scales'
2 http://youtu.be/Ubce2i1MmZY : O. Tchernichovski "Physiological brain processes that underlie song learning"
3 http://youtu.be/u4wU0bUbh-w : Y. Bengio 'Deep Learning : Looking Forward'
4 http://youtu.be/9Tf4uJu-Jnw : X. Halkias 'Classification of Mysticete : Extracting spectrotemporal structures using Sparse Architectures'

06/20/2013 afternoon:

1 http://youtu.be/VB_DBWCsfSs : H. Glotin 'Sparse coding for deformed marine or terrestrian events/ Bird or Whale Cocktail Party 3D Tracking'
2 http://youtu.be/JPXExcR634Q : Segretier et al. 'Song-based Classification techniques for Endangered Bird Conservation'
3 http://youtu.be/pJmuCNQMMWQ : D. Sheldon et al. 'Machine Learning and Ecology'
4 http://youtu.be/PbivO_7aVng : Briggs et al. 'Multi Label Classif Chains for Bird Sound'
5 http://youtu.be/_YELQRwHYmo : G. Pavan 'Monitoring bioacoustic diversity for research, conservation and education'

06/21/2013 morning :

1 http://youtu.be/2-rwaKR3BwQ : P. Dugan 'Pratical considerations for high performance on continuous passive acoustic data'
2 http://youtu.be/OBUv1Ba0MfU : Pourhomayoun et al. 'Human Scoring joint to ANN for classif.'
3 http://youtu.be/XG3Bdxg-nYw : G. Pavan 'Challenges in monitoring / indexing bioacoustic diversity'
4 http://youtu.be/GWeR-zBWOzw : Pourhomayoun et al. 'Classification on continuous region'
5 http://youtu.be/M51T0Iv3o3Q : Popescu et al. 'Pulse Classification'

06/21/2013 afternoon:

1 http://youtu.be/RtPNEFi9nIE : Paris et al. 'Sparse Coding for whale localization'
2 http://youtu.be/_CzUJC7VpQ0 : Abousleiman et al. 'Whale Classification'
3 http://youtu.be/1zTsohmGm-M : Ness et al. 'Orca Big Data'
4 http://youtu.be/qO5NEgdf0DI : Challenge Results, Herve Glotin

Slides of presentations are available at http://sabiod.univ-tln.fr

1.2. Workshop Abstract

Biodiversity assessment remains one of the most difficult challenges encountered by ecologists and conservation biologists. There is a critical need to describe and quantify the spatio-temporal dynamics of biodiversity over ecologically meaningful scales and to provide timely syntheses and interpretations so as to enable responsible decisions that reduce risks to endangered species, populations and habitats from anthropogenic activities.

This task has become even more urgent with the current increase of habitat loss and global environmental changes as a result of global commercial and industrial activities. The field of animal bioacoustics has received increasing attention due to its diverse potential benefits to science and society, and is increasingly 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. The bottleneck at this point is not access to raw data. It is the inability to efficiently process, visualize and interpret large volumes of data within an advanced, data management system.

This workshop brings together a cohort of world class scientists with expertise in animal bioacoustics, digital signal processing and machine learning to specifically address the emerging field of bioacoustic machine learning, from basic to applied research.

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. Most recently, researchers in the field have been exploring and identifying possible links and correlations between the dynamics of animal sound development and the evolution of human speech. These observations have led to both quantitative and qualitative advancements such as using MRIs for monitoring bird song ontogeny and human brain activity associated with linguistic metaphors, or the use of genetic algorithms to identify a possible common framework in the evolution of human and non-human cultural relationships. From an applied perspective, very basic, semi-automated systems for near-real-time acoustic detection of species of concern are being used by regulatory agencies to dynamically monitor and mitigate human activities, and there is increasing demand for such near-real-time capabilities.

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 due to multiple and diverse challenges. Specifically, the dynamic and variable factors in the collection and analysis of raw data from both wild and captive environments often require the use of real-time or near-real-time systems that minimize manual interaction/supervision. This requirement can be strongly coupled with the creation and employment of on-line algorithms and stochastic optimization techniques allowing field researchers to assess the computational and accuracy trade-offs without compromising the data collection process. Eventually, results from intelligent, open-access systems could offer significant societal benefits by raising public awareness of natural phenomena and exposing possible hazardous interactions between wildlife and humans allowing for swift mitigation procedures.

An additional, yet critical issue in present bioacoustic analysis strategies is the inability to provide comprehensive, accurate species validation across the full suite of signals available in very large sets of raw data. The process of extracting ground-truth, typically involves manual interaction by experts, which is an intractable task. This inherent bottleneck significantly limits our ability to identify a species' complete signal variability across the multiple dimensions of its acoustic signals, which thereby constrains our ability to process data at scales commensurate with the spatial-temporal-spectral biodiversity needs. The application of advanced, unsupervised learning algorithms offers a possible solution to this problem because it would enable rapid computational access into the unique, underlying characteristics of the species-specific features, which would accelerate the recognition task. Successful completion of this stage could then be combined with supervised methodologies to yield a robust, iterative system for automatically processing very large amounts of data and visualizing those data products over appropriate ecological scales.

Moreover, automatic and accurate species recognition remains a top priority in the field. This is a highly complex and challenging task. To be effective it needs to mirror the complexities of the hierarchical acoustic structures so often found within animal acoustic signaling behaviors, which would involve the application of both discriminative and generative approaches. Depending on the type of species under study, shallow or deep architectures might be favored. However, the diversities of the vocalization repertoires of the different species combined with their underlying biological structures indicate that any analysis and modeling would greatly benefit by integrating sparse constraints in order to increase the discriminative power of the models.

Finally, the lack of standardization and unified comparative framework, combined with the different environments and contexts of large scale data collection creates a unique domain adaptation and transfer learning framework whereby the proposed machine learning methodologies need to provide an adequate intra- and inter-species generalization.

In conclusion, the application of machine learning processes to bioacoustic signal recognition analysis and modeling of large data sets promises to yield significant theoretical and applied advances in present understandings of complex, learned animal vocal behaviors and in the quantitative description of biodiversity over ecologically meaningful spatio-temporal-spectral scales.

1.3. Workshop objectives

The main objectives of this workshop are two-fold:

1. Firstly, the workshop aims at bringing together experts from the machine learning and computational auditory scene analysis fields with experts in the field of animal acoustic communication systems to promote, discuss and explore the use of machine learning techniques in bioacoustics.
2. Secondly, by presenting current approaches, their limitations and open problems in bioacoustics to the ICML community, this workshop will encourage interdisciplinary, scientific exchanges and foster collaborations among the workshop participants.

The proposed workshop is organized jointly by experts in the field of animal bioacoustics, digital signal processing and machine learning and depending on participation rates, it will take place over two days. The target audience covers researchers working in the fields of bioacoustics signal analysis and detection-classification, as well as researchers from the whole ICML community sharing an interest in real-world applications ranging from natural to cultural sounds. Given the combined participation of computer scientists and bioacousticians, the invited speakers will be asked to give talks with a tutorial character and make the covered material accessible for both communities.

A special technical challenge on automated computer recognition of bird and marine mammal sounds will be organized in order to foster a common, quantitative framework bridging the two communities, while creating an initial, open-access and standardized data library for the communities.

The proposal and all future additional information can be found on line at

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

1.4. Invited talks overview

- "Application of advanced analytics and high-performance-computing technologies for mapping occurrences of acoustically active marine mammals over ecologically meaningful scales"

C. W. Clark[1]; P. J. Dugan[1]; Y. A. LeCun[2]; S. M. Van Parijs[3]; D. W. Ponirakis[1]; A. N. Rice[1]

[1]*Bioacoustics Research Program, Cornell University, 159 Sapsucker Woods Road, Ithaca, New York 148504, USA*

[2]*The Courant Institute of Mathematical Sciences, New York University, 715 Broadway, New York, New York 10003, USA*

[3]*Northeast Fisheries Science Center, Woods Hole Oceanographic Institute, 166 Water Street, Woods Hole, Massachusetts 02543, USA*

Abstract

Marine mammals are adapted to produce and perceive a great variety of sounds that collectively span 4-6 orders of magnitude along the dimensions of frequency, time and space. Thus, for example, blue and fin whales produce intense, long, very-low-frequency songs that can be acoustically detected and tracked at ranges of 1500 miles over periods of many weeks. In contrast, sperm whales hunting for squid at half-mile depths produce intense, very short, broadband echolocation pulses that can be acoustically detected and tracked at ranges of a few miles over periods of hours. This perspective leads to two important concepts referred to here as acoustic ecology and acoustic habitat; where acoustic ecology is the study of the acoustics involved in the interactions of living organisms, while acoustic habitat is the ecological space that is acoustically utilized by a particular species. Marine mammals are dependent on access to their normal acoustic habitats for basic life functions, including communication, food finding, navigation and predator detection. Acoustic masking from anthropogenic sounds (vessel noise, energy exploration, commercial activities) can result in measurable losses of marine mammal acoustic habitats. Masking leads to a reduction in the space within which an animal effectively operates, which is ecologically a reduction in an animal's acoustic habitat. Traditional mechanisms for detecting, classifying and analyzing acoustically active marine mammals are insufficient for mapping the ecological scales over which animals normally operate and anthropogenics influence their acoustic habitats. Here we process a relatively large acoustic data set (40 months, 6-10 channels) using advanced detection-classification analytics combined with a high-performance-computing system to explore the spatio-temporal dynamics for a suite of acoustically active marine mammals (fin, humpback, minke, and right whales) and a fish species (haddock) whose sounds can be confused with whales. The results yield insights into mechanisms for optimizing the analytical system as well as dynamic maps and metrics that describe the species-specific, spatio-temporal variability for these acoustically active animals as well as the spatio-temporal variability of their background noise environments. When considered from the large-scale, ecological perspective, these results point to an entirely novel approach for analyzing, visualizing and understanding ocean acoustics at scale.

- "Machine learning and Ecology"

Prof. D. Sheldon and T. G. Dietterich - Oregon State University, USA

Abstract

This talk will discuss current work and open problems in applying machine learning to conservation ecology. It will begin with a broad overview of challenges and opportunities for machine learning in

ecology. It will then discuss two example problems: approximate Bayesian inference to infer the velocities of migrating birds from weather radar data, and species distribution modeling. Finally, it will highlight the important role of latent process models in ecology and discuss some of the algorithmic challenges related to these models.
The work discussed in the talk is joint work between University of Massachusetts Amherst, Oregon State University, and the Cornell Lab of Ornithology.
Short Bio:Daniel Sheldon is an assistant professor in the School of Computer Science at the University of Massachusetts Amherst. The primary goal of his research is to develop new algorithms to understand and make decisions about the environment using large data sets. He leads the UMass portion of the NSF-funded BirdCast project for developing novel machine learning algorithms to model and forecast bird migration, in collaboration with Oregon State University and the Cornell Lab of Ornithology.

- "Sparse operators for deformed marine or terrestrian bioacoustic event classification / challenges in bird and whale cocktail party labeling"

Hervé Glotin; Joseph Razik; Sébastien Paris - USTV, Inst.Univ.de France, CNRS LSIS, FR

Abstract

We first recall the machine learning baseline developped for automatic speech classification. We discuss on efficient approaches for classification of animal sound units : sparse coding. We illustrate their advantages with various cases of species, from birds to whales.
For example, since Humpback whale calls present several similarities to speech, including voiced and unvoiced type vocalizations, a great variety of methods have been used to analyze them. Most of the studies of these songs are based on the classification of sound units, however detailed analysis of the vocalizations showed that the features of an unit can change abruptly throughout its duration making it difficult to characterize and cluster them systematically. We then show how sparse coding can help to determine in a song the stable components versus the evolving ones. This results in a separation of the song components, and then highlights song copying between males.
We finaly discuss how such combined models are relevant for the derivation of statistical algorithms for solving ill-posed inverse problems like the source localisation, applied to bird or to whales. We'll present a challenge on 3D whale localization using passive acoustics to illustrate this perspective.

- "Classification of Mysticete Sounds: Extracting spectro-temporal structures of calls using spare architectures"

Dr. Xanadu Halkias - CNRS LSIS and USTV, FR

Abstract

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.

- "Deep Learning : Looking forward"

Prof. Y. Bengio - Department of Computer Science and Operations Research Canada Research Chair in Statistical Learning Algorithms

Abstract

Deep learning research aims at discovering learning algorithms that discover multiple levels of distributed representations, with higher levels representing more abstract concepts. Although the study of deep learning has already led to impressive theoretical results, learning algorithms and breakthrough experiments, several challenges lie ahead.

This talk proposes to examine some of these challenges, centering on the questions of scaling deep learning algorithms to much larger models and datasets, reducing optimization difficulties due to ill-conditioning or local minima, designing more efficient and powerful inference and sampling procedures, and learning to disentangle the factors of variation underlying the observed data. It also proposes a few forward-looking research directions aimed at overcoming these challenges for AI applications such as those involving images, text or acoustics.

Accompanying paper: http://arxiv.org/abs/1305.0445

- "Gaining insights into the structure and use of dolphin whistle repertoires"

Prof. Diana Reiss - Hunter College - CUNY, NY USA

Abstract

In sharp contrast with descriptions of contact calls in all other species, the contact or cohesion calls used by bottlenose dolphins, Tursiops truncatus, in contexts of social isolation have been historically described as individually distinctive and categorically different whistle types, termed "signature whistles". These whistle types have been proposed to function as labels or names of conspecifics. Other studies have reported an absence of signature whistles and have demonstrated that dolphins, like other species, produce a predominant shared whistle type that probably contains individual variability in the acoustic parameters of this shared whistle type. To further understand the discrepancies between different studies on dolphin whistle communication and the vast differences reported between the isolation calls of dolphins and other species, we conducted a study replicating the approach and methodologies used in the studies that originally and subsequently characterized signature whistles. In contrast to these studies, we present clear evidence that, in contexts of isolation, dolphins use a predominant and shared whistle type rather than individually distinctive signature whistles. This general class of shared whistles was the predominant call of 10 of the 12 individuals, the same shared whistle type previously reported as predominant for individuals within both socially interactive and separation contexts. Results on the further classification of this predominant shared whistle type indicated that 14 subtle variations within this one whistle type could be partially attributed to individual identity.

Short Bio: Prof. Reiss earned her Ph.D. in Speech and Communication Science from Temple University and is an internationally recognized researcher in animal cognition and communication. In 1982, she developed a laboratory at Marine World in California, where she investigated the nature of dolphin communication and cognitive abilities.

Her research focuses on marine mammal cognition and communication, comparative animal cognition, and the evolution of intelligence. Her past work includes cognitive studies with interactive keyboards with dolphins to investigate their learning and communicative abilities, research in mirror self-recognition in marine mammals, marine mammal vocal repertoires and vocal and behavioral development in dolphins. Her work also involves the rescue and rehabilitation of stranded marine mammals. She was one of the scientists instrumental in the campaign to protect dolphins from being killed in tuna nets that resulted in the labeling of "dolphin safe" tuna.

Prof. Reiss's work has been published in numerous international scientific journals and book chapters and has been featured in many television science programs, included Nature, National Geographic, Wild Kingdom, the Today Show and several BBC nature shows.
Prof. Reiss au lephant, The fallacy of "signature whistles" in bottlenose dolphins: a comparative perspective of "signature information" in animal vocalizations, Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence, and others.

- Monitoring bioacoustic diversity for research, conservation and education

Prof. Gianni Pavan Pavia - Italy
Centro Interdisciplinare di Bioacustica e Ricerche Ambientali, Department of Earth and Environment Sciences, University of Pavia, Italy, Gianni.pavan@unipv.it

Abstract

Bioacoustics is an emerging technology in biodiversity science and conservation: from the recognition and monitoring of individual species through to soundscape description in terrestrial and aquatic environments, it provides new insights and approaches.
However, the complexity of the acoustic world is difficult to manage and requires new dedicated smart algorithms to process the data and extract useful and easy to handle information.
Soundscape analysis, or sonic environment analysis, also provides insights into the noise pollution problem. Natural soundscapes can be contaminated by the noise produced by human activities; this may produce behavioural and physiological changes and interfere with the communicative sounds used by animals (masking). Noise may have a severe impact on their life and an impact on natural habitats; this is particularly true in the underwater environment where sound propagates well and animals use sound as a primary system to communicate, navigate and find food.
Examples of sound monitoring and sonic environment analysis will be presented in the framework of wildlife conservation and acoustic ecology issues.

- "Physiological brain processes that underlie song learning"

Prof. Ofer Tchernichovski - Hunter College - CUNY, NY, USA

Abstract

Sleep affects learning and development in humans and other animals, but the role of sleep in developmental learning has never been examined. Here we show the effects of night-sleep on song development in the zebra finch by recording and analysing the entire song ontogeny. During periods of rapid learning we observed a pronounced deterioration in song structure after night- sleep. The song regained structure after intense morning singing. Daily improvement in similarity to the tutored song occurred during the late phase of this morning recovery; little further improvement occurred thereafter. Furthermore, birds that showed stronger post-sleep deterioration during development achieved a better final imitation. The effect diminished with age. Our experiments showed that these oscillations were not a result of sleep inertia or lack of practice, indicating the possible involvement of an active process, perhaps neural song-replay during sleep. We suggest that these oscillations correspond to competing demands of plasticity and consolidation during learning, creating repeated opportunities to reshape previously learned motor skills.

Short Bio: Ofer Tchernichovski is a professor at Hunter College - CUNY. His research uses the songbird to study mechanisms of vocal learning. Like early speech development in the human infant, the songbird learns to imitate complex sounds during a critical period of development. The adult bird cannot imitate any more - we do not know why. His lab studies the animal behavior and dynamics of vocal learning and sound production across different brain levels. The lab aims to uncover the specific physiological and molecular (gene expression) brain processes that underlie song learning. He has extensive publications in Nature and Science as Nature Letter Vol 459, 28 May 2009, "De novo establishment of wild-type song culture in the zebra finch"

- "Practical considerations for using high performance computing for applied detection classification on continuous-passive-acoustic data"

Dr. Peter J. Dugan - Cornell University, NY, USA

P. J. Dugan (1), C. W. Clark (1), Y. A. LeCun (2), S. M. Van Parijs (3), D. W. Ponirakis (1), M. Popescu (1), M. Pourhomayoun (1), Y. Shiu1, A. N. Rice (1)

(1) Bioacoustics Research Program, Cornell University, NY USA
(2) The Courant Institute of Mathematical Sciences, New York University, USA
(3) Northeast Fisheries Science Center, Woods Hole Oceanographic Institute, MA USA

Abstract

From biology to technology, the rate of data collection often far exceeds the ability to process the information. Processing large data sets is becoming a major point of interest for every field of science. The ease of digital data collection allows for the capture of many terabytes of data, yet this often creates major computational bottlenecks when trying to analyze such datasets. This talk focuses on a new system developed by Cornell University that uses high performance computing (HPC), and combines it with parallel and distributed processing approaches to process large amounts of bioacoustic data.

This work will discuss how the HPC system was developed using commercial off the shelf (COTS) tools creating a flexible client-server model that is expandable, flexible and portable. The presentation will demonstrate a strategy for providing a flexible software interface for running a plurality of data mining algorithms using a dense computer cluster called the Acoustic Data Accelerator, or HPC-ADA. In addition, a variety of tools have been developed to complement the system, providing efficient methods for data processing.

The authors will also summarize a specific example for processing multiple months of multi-channel, continuous data recorded in the Stellwagen Bank National Marine Sanctuary, MA, USA. Results show distinct seasonal distribution patterns of species-specific vocalization for right whales (Eubalaena glacialis) and minke whales (Balaenoptera acutorostrata).

These examples will also show other related acoustic activity from a variety of other marine animals. Results from these data products illustrate daily and seasonal patterns as shown across multiple sensors. As the scale of data collection continues to expand (the bioacoustics community will soon be faced with the challenges of processing pedabytes of data), such high-throughput computational approaches will be essential in bringing passive acoustic monitoring and analysis into the realm of big data science.

2. Organisation committee

Workshop Chairs:

- Pr. H. Glotin - Institut Universitaire de France, CNRS LSIS and USTV, FR
  Email: glotin@univ-tln.fr
- Pr. Y. LeCun - Computational and Biological Learning Lab at New York University, USA
  Email: yann@cs.nyu.edu
- Pr. C. Clark - Director of Bioacoustics Research Program at Cornell University, NY, USA
  Email: cwc2@cornell.edu

Co-Organizers:

- Dr. X. Halkias - CNRS LSIS and USTV, FR
  Email: xanadu.halkias@univ-tln.fr
- J-M. Prévot - USTV, FR

Technical Session Chairs:

- Dr. Peter Dugan - Cornell University, NY, USA
  Email: pjd78@cornell.edu
- Associate. Pr. Jérôme Sueur, Habilitated - Muséum National d'Histoire Naturelle, Paris, FR
  Email: sueur@mnhn.fr

Preparation of the Proceedings

- Ales Mishchenko SATT sud-est, CNRS LSIS and USTV, FR

Email: alesmichtchenko@mail.ru

2.1. Organizers short CV

Prof. Hervé Glotin **- Institut Universitaire de France, CNRS LSIS and USTV, FR**

Hervé Glotin is a Professor at the Insitut Universitaire de France and Univ. of Toulon, in the Systems & Information Sciences CNRS lab. He is leading the DYNI team on stochastic multimodal information retrieval. He received a diploma in computer science from University Pierre et Marie Curie-Paris. During his master thesis he proposed the first modelisation of vocalic system evolution, addressing the emergence of a common phonetic code in a society of communicating speech agents using evolutionary learning, which has been extended in many other works. He carried out his PhD at the Inst. of Perceptual Artificial Intelligence (IDIAP), CH and Inst. of Spoken Communication - Perception Team Grenoble on "Robust adaptive multi-stream automatic speech recognition using voicing and localization cues". In 2000 he was involved as an expert at the Johns Hopkins CSLP lab with the IBM human language team in audiovisual Large Vocabulary Speech Recognition. After two years as a research engineer at CNRS lab on phonology and Semantic analysis, he became an assistant professor at the University of Toulon in 2003. His research focuses on multimodal pattern analysis and retrieval systems, audiovisual indexing, cognitive models and bioacoustics. He is the co-author of one hundred of international refereed articles, and of an international (US, CANADA...) patent on a real-time bio-acoustic indexing algorithm. Herve Glotin is leading the CNRS interdisciplinary project 2012-2016, Scaled Acoustic Biodiversity with LIP6 Paris 6, the CNPS, MNHN and LIG. He is invited as a keynote speaker at the American Society of Acoustics workshop in June 2013 - Montreal for the special session on "Conditioning, Segmentation and Feature Extraction in Bioacoustics".

Prof. Yann LeCun - New York University, USA

Yann received a Diplôme d'Ingénieur from the Ecole Superieure d'Ingénieur en Electrotechnique et Electronique (ESIEE), Paris in 1983, a Diplôme d'Etudes Approfondies (DEA) from Université Pierre et Marie Curie, Paris in 1984, and a PhD in Computer Science from the same university in 1987. His PhD thesis was entitled "Modèles connexionnistes de l'apprentissage" (connexionist learning models) and introduced an early version of the back-propagation algorithm for gradient-based machine learning. In 1987, he joined Geoff Hinton's group at the University of Toronto as a research associate. He then joined the Adaptive Systems Research Department at AT&T Bell Laboratories in Holmdel, NJ in 1988. In 1991, he spend six months with the Laboratoire Central de Recherche of Thomson-CSF in Orsay, France, after which he returned to Bell Labs. Shortly after AT&T's second breakup in 1996, he became head of the Image Processing Research Department, part of Larry Rabiner's Speech and Image Processing Research Lab at AT&T Labs-Research in Red Bank, NJ. In 2002, he became a Fellow of the NEC Research Institute (now NEC Labs America) in Princeton, NJ. He joined the Courant Institute of Mathematical Sciences at New York University as a Professor of Computer Science in 2003. He was named Silver Professor in 2008. Yann LeCun has been associate editor of PLoS ONE (2008-present), IJCV (2003-present), IEEE Trans. PAMI (2003-2005), Pattern Recognition and Applications, Machine Learning Journal (1996-1998), IEEE Transactions on Neural Networks (1990-1991). Yann LeCun has published over 130 technical papers and book chapters on machine learning. He is leading the Computational and and Biological Learning Lab at NYU.

Dr. Christopher Clark - Cornell University, NY, USA

Christopher Clark is currently the Imogene P. Johnson Director for the Bioacoustics Research Program at the Cornell Lab of Ornithology, and a senior scientist at the Department of Neurobiology and Behavior at Cornell University, NY. He oversees and directs a vigorous, multidisciplinary program that is actively engaged in both basic and applied research. Dr. Clark is an expert in engineering design and implementation of automatic acoustic detection, classification, localization and tracking systems as applied to animal acoustic communication, behavioral ecology and quantifying potential risks to wildlife from anthropogenic activities. Projects include migratory bird monitoring on DOD installations, nicaloise effects on endangered bird species, rare bird monitoring, miniaturized radio tracking transmitters and advanced radio tracking receiver networks. His scientific conservation research on a variety of large whale species continues throughout the world's oceans.

***Dr. Xanadu Halkias* - CNRS LSIS and USTV, FR**

Xanadu Halkias received her PhD from the Electrical Engineering Department of Columbia University, NY. Her research focused on advanced signal processing and machine learning as it applies to bioacoustics. She is currently a post-doctorate fellow at the Université du Sud - Toulon working on machine learning and specifically deep architectures and their applications.

***Dr. Peter Dugan* - Cornell University, NY, USA**

Peter Dugan is currently the PI on the National Oceanic Partnership (NOPP) Grant focusing on detection, classification and localization of marine mammals. He received his PhD in Electrical Engineering and Combined behavioral biology from Binghamton University in NY. Prior to Cornell University he held positions in the industry in companies such as Hughes Link Flight Simulation and Lockheed Martin. He also has an extensive publication and patent portfolio showcasing advanced methodologies in machine learning for marine mammal vocalizations. His interests and motivations include the research and development of computationally intelligent systems, by combining traditional "shallow systems" with "deep learning systems" for object detection and classification in order to enhance system accuracy. The NOPP grant has been awarded 1M$ for the years 2012-2015. As the PI, his goal is to investigate new approaches and deliever comparative studies working on integrated teams representing Science, Technology, Engineering and Mathematics (STEM).

***Associate Prof. Jérôme Sueur* - Muséum National d'Histoire Naturelle, Paris, FR**

Jérôme Sueur is currently an habilitated Associate Professor at the museum of natural history in Paris, France. With a strong international academic background in biological sciences, his interests and expertise can be found in: acoustics ecology i.e biodiversity assessment through acoustics; Animal audition, i.e. the nano-mechanics of tympanal audition in insects; Animal behavior and animal systematics. He has an extensive list of international publications in journals such as Ecological Indicators, Journal of Experimental Biology and Plos ONE.

Dr. Ales Mishchenko SATT sud-est, CNRS LSIS and USTV, FR

Ales Mishchenko is currently a post-doctorate fellow at the Université du Sud - Toulon and senior researcher of SATT sud-est. His current research focuses on advanced signal processing, machine learning and Neural networks with application to bioacoustic data and EEG/MEG/fMRI data processing.

3. Workshop challenges

3.1. Challenges overview

Data and submission of results through Kaggle.com
Avian Data
Whale Data

The field of bioacoustics faces similar challenges to those defined in the general machine learning and advanced signal processing communities. However, in the former case, the unknown extent of signal feature variability, unknown informational significance of features and limited knowledge of the signal's semantics may hinder the robustness of existing machine learning algorithms as a result of highly tuning and adapting these methodologies for the different collections of data.

As previously mentioned, bioacoustics aims at analyzing and modeling animal sounds for biodiversity assessment. However, given the large amount of the collected data along with the different taxonomies of the different species and their environmental contexts, it would be intractable to attempt to analyze each of the different sounds or even to offer general solutions that would be applicable across different species.

The goal of the technical session is to provide possible solutions for targeted applications that have been popular in the field, but remain either partially solved or require significant manual interaction. Moreover, we hope that by building a representative, standardized collection of validated acoustic data we will provide a much needed comparative framework within the bioacoustics community.

In the hopes that the proposed workshop could become a repeated event, the desire would be for any subsequent technical challenges to focus on increasingly complex acoustic phenomena representative of a particular set of signal types which generally occupy a high-dimensional, acoustic feature space: for example, short duration, broadband transients; or hierarchically organized combinations of stereotypic, frequency-modulated sounds, syllables, and phrases repeated in long bouts; or hierarchically organized combinations of highly variable, two-voiced, frequency-amplitude-modulated sounds, repeated in long bouts. Further complexity could include acoustic scenes with multiple, higher variable sources or data from coherent, multi-sensor systems. These would provide a diverse set of algorithms to allow parallel or cross-comparison classification schemes and would aid in elucidating how different algorithms perform on different levels of acoustic complexity. The expected outcome would be the significant advancement of our abilities to explore and understand different classes of bioacoustics data within natural context, and across the broad spectrum of biodiversity.

Technical Challenge Data

The automatic analysis of marine mammal sounds has long been an interest in the bioacoustics community given the sounds' intrinsic complexities, the underlying concerns for these over-exploited species, and the fact that in the U.S. there are explicit statutes protecting marine mammals. The same level of interest easily applies to avian songs and calls, which are easily experienced, aesthetically pleasing, and have served as the basis for a rich and productive history in fundamental research on vocal ontogeny, memory, and computational neuroscience.

For that reason we propose that the first Technical Challenge is focused on sounds of marine mammals and birds.

A training, development and test set will be provided for each of the data sets. In both the avian and marine mammal sets. The development sets will be comprised of artificial mixtures offering a controlled environment for system creation. On the other hand, the test sets will include real life mixtures, thus exposing the created systems to the variability and diversity of bioacoustic recordings.

Sample examples of the avian data set are shown in the images below.

Figure : Avian data set examples; Left: Aegcau species, Right: Alaarv species

Data	Avian (from MNHN Paris)	Marine mammals (from Cornell Univ)
Description	TRAIN: Clean recordings from the MNHN soud library[1] TEST: Recordings form the Regional Park of the Upper Chevreuse Valley in France, daily for 30min prior to sunrise.	TRAIN: Recordings of single species (one call per hour), Right Whale (3 types of calls) TEST: Recordings in the wild from the Stellwagen Bank National Marine Sanctuary
Duration	TRAIN: 30 sec x 35 bird recordings = 18min TEST: 150sec x 3 mics x 90 recordings = 11.2 hours	TRAIN: approximately 48 hours x 1 species TEST: approximately 96 hours
Equipment	TRAIN: 1 microphone, 16bit, 44.1kHz TEST: 3 microphones on trees in the same area, into different forest state (mature, young, open), 16bit, 44.1kHz[2] Audio format: .wav	Hydrophones, 16 bit, 2-10kHz MARU, bit depth = 11, Sensitivity -167.5 dB re: 1uPa/V Audio format: .aif
SNR	TRAIN: 20-60 dB TEST: 5-9 dB	TRAIN: 1-20 dB TEST: 1-20 dB
Ground truth	Sample level: Call/no call Call level: Species class	Sample level: Call/no call, frequency start/end Call level: Species class (by spectrogram expert analysis and/or by on the field visual identification)
Features	Raw audio Mel Cepstral coefficients are given	Raw audio
Meta-data	TRAIN / TEST : The phylogenetic	Location (longitude, latitude), Time,

tree (usual format, including R package to read/manipulate) of the target species	Depth, ambient sound level
Location, Time, Temperature, locationMeteo	

[1] The data for this challenge are copyright of Fernand Deroussen Jerome Sueur of the Musee National d Histoire Naturelle, their usage is restricted to this challenge. The competition test data was graciously provided by Jerome Sueur. More details on the train data are given in : Deroussen, F., 2001. Oiseaux des jardins de France. Nashvert Production, Charenton, France ; Deroussen, F., Jiguet, F., 2006. La sonotheque du Museum: Oiseaux de France, les passereaux. Nashvert production, Charenton, France ; http://naturophonia.fr. naturophonia.fr

[2] Depraetere M, Pavoine S, Jiguet F, Gasc A, Duvail S, Sueur, J - Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. Ecological Indicators, 13: 46-54

The ground truth of each test set will be used to score each system and distributed after the deadline for your working notes. * The participants must not try to handlabel the test set for tuning their models. Guidelines on submitting results are simple. Please follow the guidelines on KAGGLE as shown in the links below:

Avian Data
Whale Data

Tasks

As part of the challenge, different thematic approaches can be proposed that will reflect on the most commonly encountered problems in the field. In both cases, avian and marine mammal sounds, the major thematic approaches include:

- (i) Species detection
- (ii) Species classification
- (iii) Call extraction
- (iv) Tracking/Localization

In this proposal we recommend a technical challenge in the area of automatic classification as it appears to be a task that will immensely benefit the bioacoustics community. Our proposed challenge is outlined below:

Task 1: Species Recognition/Clustering

The goal of the task is to recognize the avian and marine mammal species in their respective recordings. This is a multi-class problem and allows participants to explore both supervised and unsupervised methodologies. Emphasis will be given on the recognition rate rather than the computational cost of the methodologies.

Task 2: Free challenge

Using the bird data set we can offer participants the chance to provide any meaningful environmental/ecological information using machine learning tools. The goal is to extract possible meaningful ecological correlations between the avian recordings and the provided meta-data (phylogenetic data which could match some acoustic cues, meteo (wind, sun...) for each test set.

Challenge descriptions are given below:

3.2. ICML Bird challenge (web link)

The data for this challenge are copyright of Fernand Deroussen Jerome Sueur of the Musee National d Histoire Naturelle, their usage is restricted to this challenge. The competition test data was graciously provided by Jerome Sueur.

Train Data extracted from :

Fernand Deroussen naturophonia.fr

Deroussen, F., 2001. Oiseaux des jardins de France. Nashvert Production, Charenton, France
Deroussen, F., Jiguet, F., 2006. La sonotheque du Museum: Oiseaux de France, les passereaux. Nashvert production, Charenton, France

Here is the link to KAGGLE WEB SITE with description and RUN SUBMISSION (+SCORING).
We are pleased to announce that the challenge on bird classification at ICML workshop is now running on Kaggle web site [https://www.kaggle.com/c/the-icml-2013-bird-challenge].
=> thus each of your run (~not limited number) are automatically scored : this leaderboard is calculated on approximately 33% of the test data.

The final rank will be based on the other 67%, so the final standings may be different. Kaggle competitors are used to selecting 5 models at the end of the competition (16th of june).

You find below the data sets, also available at Kaggle on other format (CVS). If you participate to this challenge, please Email to icml4b@gmail.com for free inscription, and to get updated news if any on these data during the challenge.

Task description : for each test file, you have to index the 35 bird species given in the official training set (below).

Runs submission : each run will be submitted into the Kaggle web site according the details given on the Kaggle web site. Each run gives the Pij (90x35) probabilities : Pij = P('The species j sings in the test file i') , with j and i in alphabetical order.

You can submit up to 5 runs in the official contest, choosen from the ones you submitted to Kaggle until the 16th of june. Your run may include if possible at least one using the given MFCC features.

Evaluations will be computed on ROC.

* You are free to use any additional external data or recordings (as wikipedia wav samples linked in the list below, or taxonomia for hierarchical classification, ...), but in that case you must specify it in you run description (.txt) and this run with modified train set will not count for the prize of the challenge.

Sorted list of the 35 bird species (= classes) with wikipedia links

link to the TRAIN SET : 35 .WAV files and suggested FEATURES (see also Kaggle for various format)

link to the TEST SET : 90 .WAV files and suggested FEATURES (see also Kaggle for various format)

** This TEST SET includes the results for three test files to be used as a small development set (please include them at their right place into your run).

The suggested MFCC features were computed according to a minimum error (in average on all the species) reconstruction signal of the signal. The scripts are given here :

MFCC SCRIPTS

METADATA : you may use some metadata or other training set to enhance your model, such as :

PHYLOGENETIC tree to reveal some acoustic cues between species. This tree, with distances, is given there. The phylogenetic tree, with distances, of the target species. WEATHER : The weather, wind speed, humidity, sun conditions... of each test set files.

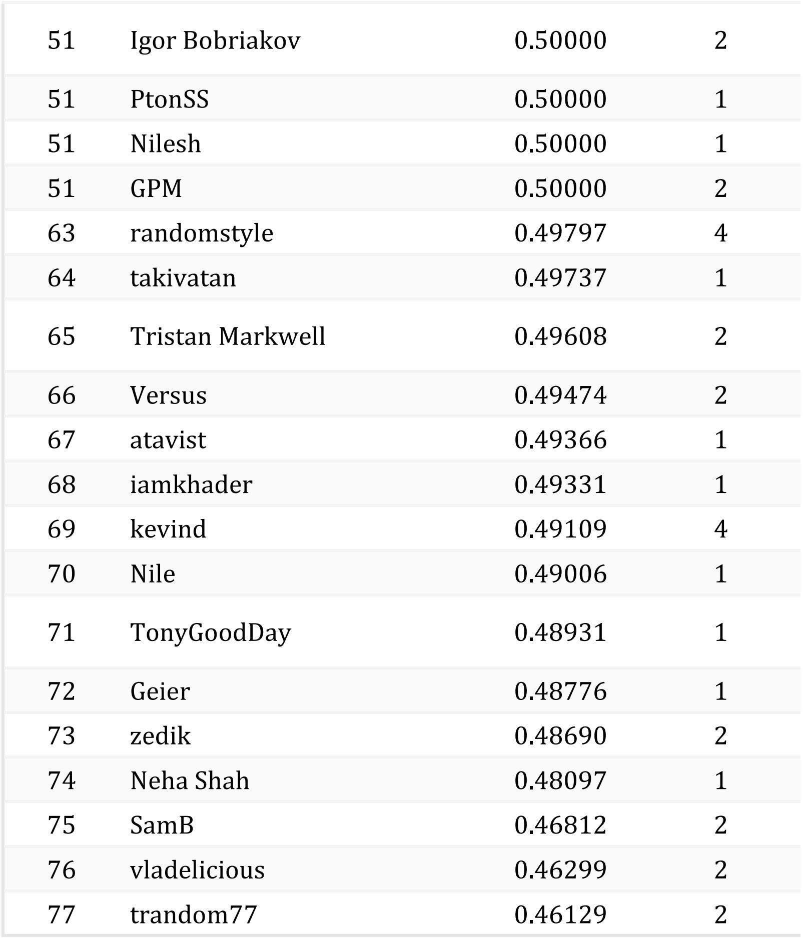
SUBMISSION : The ground truth of each test set will be used to score ROC. * The participants must not try to hand label the test set for tuning their model: every parameter will be automatically set.

Guidelines on submitting results are simple as explained in Kaggle. You are invited to send a working note with your partial results by the 30th of may following the ICML template: it will be published into this ICML workshop Proceedings.

CHALLENGE RESULTS

Team #	Team Name	Score	Entries
1	ATD_Madbirds (winner)	0.69454	24
2	AMPires	0.69015	26
3	ikretus	0.65246	5
4	Dufour_Org_Baseline	0.64639	41
5	Phoenix	0.64166	20
6	OSUBioacoustics	0.64029	6
7	Bent Wing	0.62843	33
8	Damian Mingle	0.61266	15
9	oo__oo	0.60406	10
10	ATD-GPM	0.60314	2
11	Benoit Plante	0.60160	4
12	Aspartame	0.60026	12
13	emmanouilb	0.59168	33
14	Ben Horsburgh	0.58764	10
15	Rafael	0.58177	51
16	ZhaoHFUT	0.57874	8
17	casia	0.57725	13
18	SophomoreOlinHackers	0.56604	11

19	Kirill Makukhin	0.56548	14
20	SonoChiro	0.56278	4
21	Aupathletic	0.56269	3
22	lamkelf	0.55978	4
23	zjfu	0.55332	1
24	Alphia	0.54693	20
25	adm4749	0.54521	7
26	hejk	0.54478	4
27	Egor Lakomkin	0.54293	1
28	Slam Johnson	0.54235	7
29	ws45	0.54197	5
30	Anil Thomas	0.54136	3
31	toto	0.54035	2
32	danstowell	0.53874	13
33	MasterDean	0.53541	1
34	IDCMember	0.53462	4
35	David Ledbetter	0.53180	6
36	atd_mps	0.53120	1
37	UWB	0.52857	20
38	VISAL	0.52796	3
39	Olivier	0.52707	6
40	Muppet	0.52334	5
41	Hog	0.52269	3
42	Clickthisway	0.51639	4
43	Jonathan Dursi	0.51379	3
44	zty263263	0.51377	1
45	AR 2	0.50675	4
46	DataMonkeys	0.50629	2
47	kichong	0.50466	2
48	ATD_Fail	0.50354	10
49	BigNate	0.50146	12
50	James Lyons	0.50043	4
51	InvincibleGuy	0.50000	1
51	Data Geek	0.50000	2
51	BigFish	0.50000	1
51	paper plates	0.50000	2
51	David Völgyes	0.50000	1
51	kenny	0.50000	1
51	Abhishek	0.50000	1



51	Igor Bobriakov	0.50000	2
51	PtonSS	0.50000	1
51	Nilesh	0.50000	1
51	GPM	0.50000	2
63	randomstyle	0.49797	4
64	takivatan	0.49737	1
65	Tristan Markwell	0.49608	2
66	Versus	0.49474	2
67	atavist	0.49366	1
68	iamkhader	0.49331	1
69	kevind	0.49109	4
70	Nile	0.49006	1
71	TonyGoodDay	0.48931	1
72	Geier	0.48776	1
73	zedik	0.48690	2
74	Neha Shah	0.48097	1
75	SamB	0.46812	2
76	vladelicious	0.46299	2
77	trandom77	0.46129	2

3.3. The ICML 2013 Whale Challenge (web link)

Challenge to develop recognition solutions to detect and classify right whales for BIG data mining and exploration studies

(right whale illustration courtesy of Pieter Folkens, ©2011)

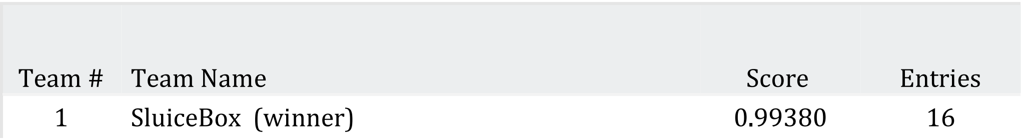
This competition complements the previously held Marinexplore Whale Detection Challenge, in which Cornell University provided data from a ship monitoring application termed "Auto Buoy", or AB Monitoring System. In the Marinexplore challenge we received solutions that exceeded 98% accuracy and will ultimately advance the process of automatically classifying North Atlantic Right Whales using the AB Monitoring Platform.

Since the results from the previous challenge proved so successful, we decided to extend the goals and consider applications that involve running algorithms on archival data recorded using portable hydrophone assemblies, otherwise referred to as Marine Autonomous Recording Unit (or MARU's). Since Cornell and its partners have been using the MARU for over a decade, a sizable collection of data has been accumulated. This data spans several ocean basins and covers a variety of marine mammal species.

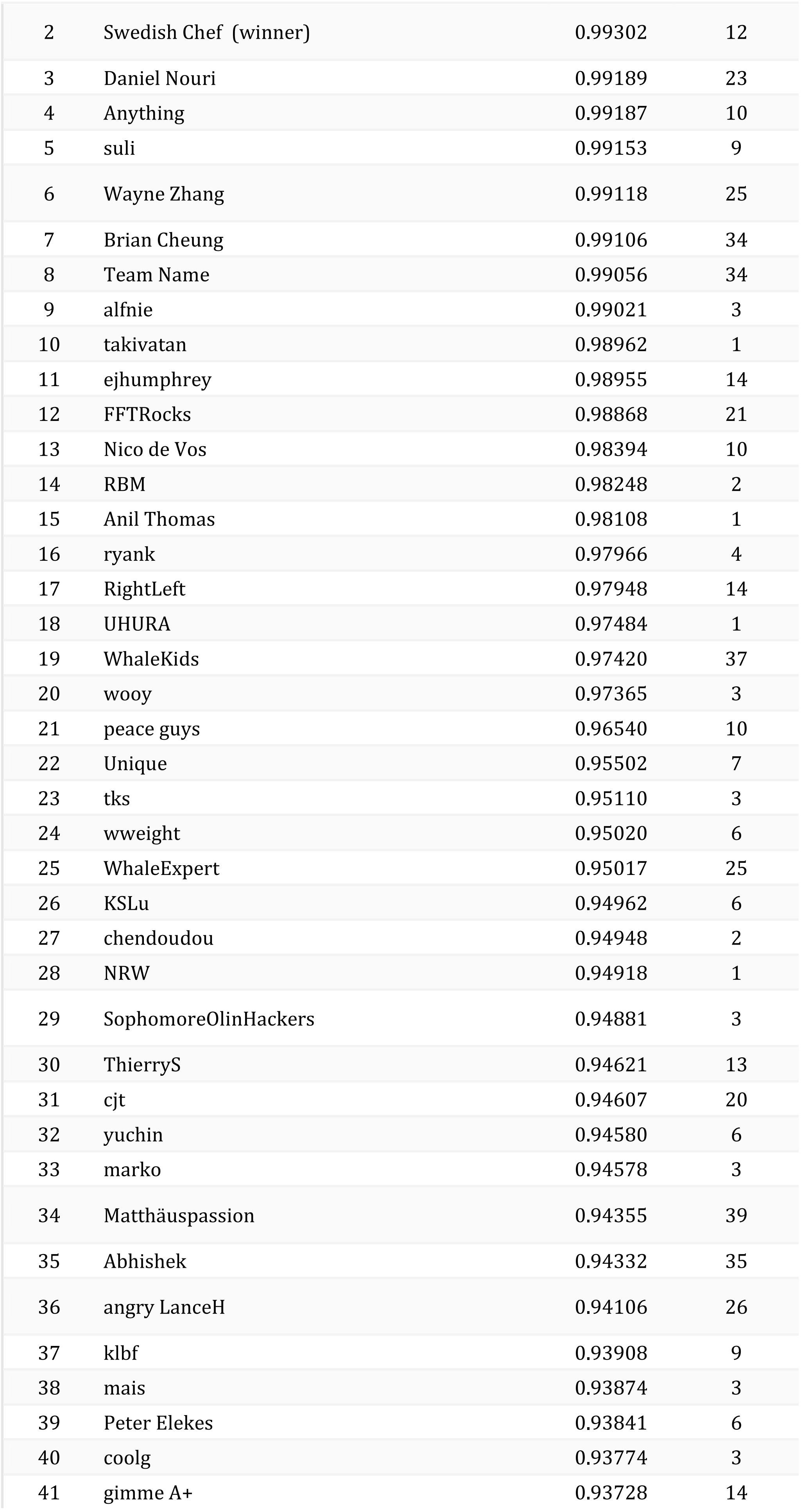
Solutions to this challenge will be ported to a High Performance Computing (HPC) platform, being developed in part through funding provided by the Office of Naval Research (ONR grant N000141210585, Dugan, Clark, LeCun and Van Parijs). Together, Cornell will combine algorithms, HPC technologies and its data archives to explore data using highly accurate measuring tools. We encourage participants who developed prior solutions (through the collaboration with Marinexplore) to test them on this data.

The results will be presented at the Workshop on Machine Learning for Bioacoustics at ICML 2013.

CHALLENGE RESULTS



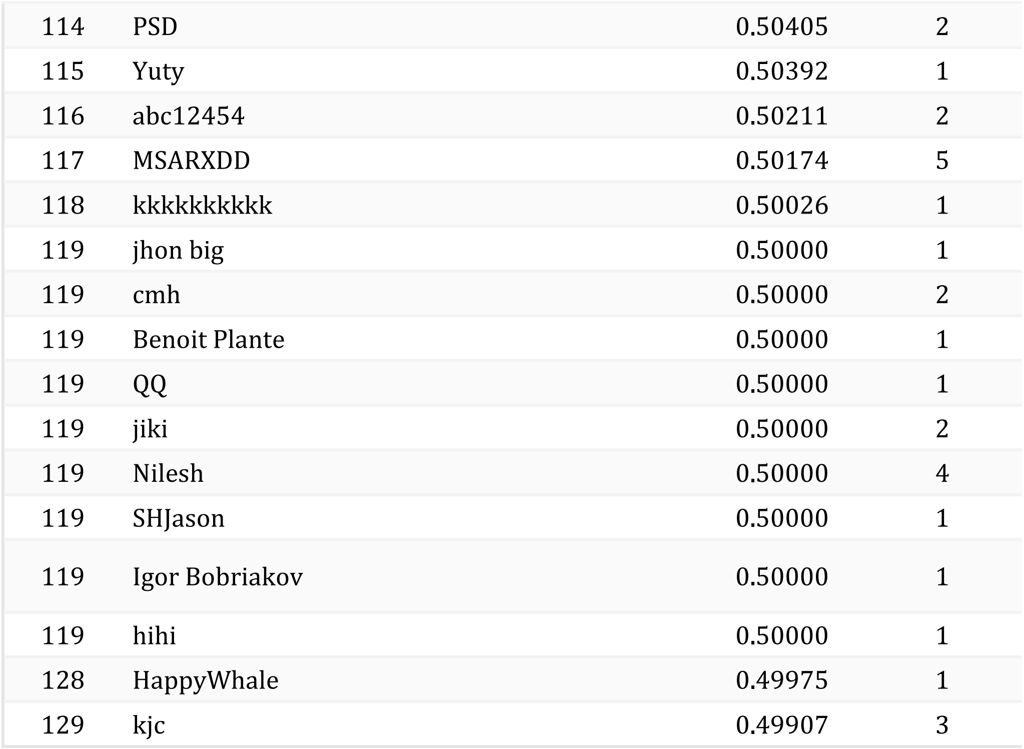
Team #	Team Name	Score	Entries
1	SluiceBox (winner)	0.99380	16



2	Swedish Chef (winner)	0.99302	12
3	Daniel Nouri	0.99189	23
4	Anything	0.99187	10
5	suli	0.99153	9
6	Wayne Zhang	0.99118	25
7	Brian Cheung	0.99106	34
8	Team Name	0.99056	34
9	alfnie	0.99021	3
10	takivatan	0.98962	1
11	ejhumphrey	0.98955	14
12	FFTRocks	0.98868	21
13	Nico de Vos	0.98394	10
14	RBM	0.98248	2
15	Anil Thomas	0.98108	1
16	ryank	0.97966	4
17	RightLeft	0.97948	14
18	UHURA	0.97484	1
19	WhaleKids	0.97420	37
20	wooy	0.97365	3
21	peace guys	0.96540	10
22	Unique	0.95502	7
23	tks	0.95110	3
24	wweight	0.95020	6
25	WhaleExpert	0.95017	25
26	KSLu	0.94962	6
27	chendoudou	0.94948	2
28	NRW	0.94918	1
29	SophomoreOlinHackers	0.94881	3
30	ThierryS	0.94621	13
31	cjt	0.94607	20
32	yuchin	0.94580	6
33	marko	0.94578	3
34	Matthäuspassion	0.94355	39
35	Abhishek	0.94332	35
36	angry LanceH	0.94106	26
37	klbf	0.93908	9
38	mais	0.93874	3
39	Peter Elekes	0.93841	6
40	coolg	0.93774	3
41	gimme A+	0.93728	14

42	Ankush Shah	0.93532	2
43	changyu	0.93524	2
44	IfElseForWhale	0.92579	25
45	TeamSMRT	0.92042	3
46	bitspersecond	0.91969	10
47	eatgod	0.91832	5
48	V	0.91743	7
49	NTU_msar_r01922021	0.91660	17
50	ikretus	0.91546	3
51	msar.r01942108	0.90521	3
52	SyuPei	0.90515	8
53	MSAR_r01942130	0.89964	2
54	msar_whale	0.89913	5
55	sidney78	0.89013	6
56	Laboon	0.88301	8
57	kaggel	0.87727	1
58	a3785lex	0.87289	1
59	Alekos	0.87144	3
60	Rafael	0.87097	4
61	fb	0.86994	3
62	LOL	0.86723	5
63	Zephyr	0.86710	6
64	libmsar	0.86677	18
65	ApesTeam	0.86521	5
66	gg3be0	0.85952	12
67	MSAR_b98901024	0.85930	5
68	QoQ	0.85529	47
69	ManOfSteel	0.85311	8
70	whitehsu	0.84731	3
71	bench	0.84428	10
72	zzzz	0.83973	4
73	._______.	0.83741	4
74	MNOP	0.83596	5
75	3pi	0.83221	4
76	Aalto University	0.82419	1
77	offside	0.82377	3
78	MSAR_r01944023	0.80851	5
79	Renniw111	0.80697	5

80	Apo_	0.80375	5
81	w01ph	0.79824	2
82	Shaun Lippy	0.79689	7
83	Belethia	0.79060	5
84	TTNP OC!	0.78793	1
85	tsubasa_wing	0.78555	5
85	kc_karenchou	0.78555	3
87	WannaCry	0.77289	2
88	KennyChuang	0.76753	5
89	WhaleWhaleWhale	0.76723	6
90	oops	0.75911	4
91	chuchu	0.75486	6
92	BOOOOOO	0.74259	3
93	dism	0.73261	3
94	BrianKuo	0.71083	7
95	Doyapiya	0.70296	12
96	WhenPigsFly	0.69339	4
97	Mélanie	0.68519	16
98	Darkbtf	0.68194	2
98	Tony LO.	0.68194	1
100	tig_007	0.65520	1
101	Fronage	0.62359	4
102	Darth Dave Diode	0.62100	4
103	Aff	0.61631	3
104	dencbc	0.61194	23
105	panda orca penguin dalmatian zebra skunk walk into a bar.	0.60980	20
106	Chetan Nichkawde	0.56579	2
107	Jomican	0.55056	5
108	Zealous	0.54621	1
109	David Völgyes	0.54349	4
110	CC	0.51717	14
111	Anand Subhash	0.51658	3
112	anketer	0.51103	8
113	Montag	0.50408	3



114	PSD	0.50405	2
115	Yuty	0.50392	1
116	abc12454	0.50211	2
117	MSARXDD	0.50174	5
118	kkkkkkkkkk	0.50026	1
119	jhon big	0.50000	1
119	cmh	0.50000	2
119	Benoit Plante	0.50000	1
119	QQ	0.50000	1
119	jiki	0.50000	2
119	Nilesh	0.50000	4
119	SHJason	0.50000	1
119	Igor Bobriakov	0.50000	1
119	hihi	0.50000	1
128	HappyWhale	0.49975	1
129	kjc	0.49907	3

4. Workshop Schedule

The workshop follows a "mini-conference" format, and explicitly offer opportunities for free discussion between researchers.

- Each talk may be followed by a brainstorming session whereby participants and speakers will explore possible solutions. Their conclusions and outcomes will be presented at the end of the workshop.
- The second day includes sessions on the Technical Challenge for the avian and marine mammal data sets.

Program (due to last changes, be sure to refresh the upload of this page)

Day 1 : Thursday the 20th of june

08:30	Opening Session/Welcome	Why ICML4B? H. Glotin and C. Clark
08:45	Keynote Speaker	C. Clark 'Advanced High-performance-computing for Mapping Marine Mammals overs ecologically meaningful scales'
09:20	Keynote Speaker	O. Tchernichovski "Physiological brain processes that underlie song learning"
10:00	Coffee	
10:30	Keynote Speaker	Y. Bengio 'Deep Learning : Looking Forward'
11:15	Keynote Speaker	X. Halkias 'Classification of Mysticete : Extracting spectro-temporal structures using Sparse Architectures' (slides .pdf)
11:55	Memory photography	
12:00	Lunch	
14:00	Keynote Speaker	H. Glotin 'Sparse coding for deformed marine or terrestrian events / Bird or Whale Cocktail Party 3D Tracking'
14:30	Accepted paper	Segretier et al. 'Song-based Classification techniques for Endangered Bird Conservation'
14:55	Keynote Speaker	D. Sheldon et al. 'Machine Learning and Ecology'
15:30	Coffee	
16:00	Accepted paper	Briggs et al. 'Multi Label Classif Chains for Bird Sound'
16:30	Keynote Speaker	G. Pavan 'Monitoring bioacoustic diversity for research, conservation and education' 1/2
17:10	Poster and short papers	Popescu et al. 'Large Scale Classification' Dugan et al. 'High Perf Computing for Bioacoustics' Mishchenko et al. 'Target Optimization funct for bird songs' Smirnov, 'CNN for whale classification'

Day 2 : Friday the 21th of june

08:30	Keynote Speaker	P. Dugan 'Pratical considerations for high performance on

		continuous passive acoustic data'
09:05	Accepted paper	Pourhomayoun et al. 'Human Scoring joint to ANN for classif.'
09:30	Accepted paper	Abousleiman et al. 'Whale Classification'
10:00	Coffee	
10:30	Keynote Speaker	G. Pavan 'Challenges in monitoring / indexing bioacoustic diversity' Demo (2/2)
11:30	Accepted paper	Pourhomayoun et al. 'Classification on continuous region'
12:00	Lunch	
14:00	Accepted paper	Paris et al. 'Sparse Coding for whale localization'
14:30	Accepted paper	Popescu et al. 'Pulse Classification'
15:00	Accepted paper	Ness et al. 'Orca Big Data'
15:30	Coffee	
16:00	Challenge Results 1/2	Bird song classification : Methods and results of the 78 participants, Prize, and next..? Benetos et al. working note Briggs et al. working note Turner et al. working note (5th) Stowell al. working note
16:30	Challenge Results 2/2	Whale Challenge : Methods, Results (136 participants), Prize, and next..?
17:00	General discussion and closing	Organizers and all Participants

5.1. Workshop Full papers

North Atlantic Right Whale Contact Call Detection

Rami Abousleiman RDABOUSL@OAKLAND.EDU
Oakland University, Department of Electrical and Computer Engineering, Rochester, MI, USA

Guangzhi Qu GQU@OAKLAND.EDU
Oakland University, Department of Computer Science and Engineering, Rochester, MI, USA

Osamah Rawashdeh RAWASHD2@OAKLAND.EDU
Oakland University, Department of Electrical and Computer Engineering, Rochester, MI, USA

Abstract

The North Atlantic right whale (Eubalaena glacialis) is an endangered species. These whales continuously suffer from deadly vessel impacts alongside the eastern coast of North America. There have been countless efforts to save the remaining 350 - 400 of them. One of the most prominent works is done by Marinexplore and Cornell University. A system of hydrophones linked to satellite connected-buoys has been deployed in the whales' habitat. These hydrophones record and transmit live sounds to a base station. These recording might contain the right whale contact call as well as many other noises. The noise rate increases rapidly in vessel-busy areas such as by the Boston harbor. This paper presents and studies the problem of detecting the North Atlantic right whale contact call with the presence of noise and other marine life sounds. A novel algorithm was developed to preprocess the sound waves before a tree based hierarchical classifier is used to classify the data and provide a score. The developed model was trained with 30,000 data points made available through the Cornell University Whale Detection Challenge program. Results showed that the developed algorithm had close to 85% success rate in detecting the presence of the North Atlantic right whale.

1. Introduction

The dependency on ship transportation for goods has increased the ocean congestion alongside the eastern side of the United States and Canada. The North Atlantic right whale (NARW) suffers from this increase (Jensen et al.) (2004). NARW is a mammal and thus requires air to breath. Heading towards the ocean surface can be dangerous as impacts with one of these large vessels may be deadly. Ship crews seldom notice the presence of the whale(s) and most times there is nothing to do after a hit has occurred. This problem is further escalated due to the fact the NARW is an endangered species with only a couple hundred of these mammals survive nowadays Kraus et al. (2005), Caswell et al. (1999), and Fujiwara et al. (2001). Countless efforts have been done to conserve and study the behavior of the NARW Matthews et al. (2001), Clark et al. (2007), Vanderlaan et al. (2003), and Parks et al. (2005). To help solve this problem an autonomous near-real-time buoy system for automatic detection of NARW has been developed Spaulding et al. (2009). The NARW makes a unique call known as 'contact call' or 'up-call' these calls are used as a communication method between the whales as a way of letting each other know of their presence. This distinguished call has unique characteristics that will be used as the benchmark for the NARW call detection. This paper will propose a method to analyze the sound recording acquired by the deployed buoys Spaulding et al. (2009) and then describe the developed algorithm that will automate the detection process. The rest of the paper is organized as follows: the next paragraph will list the related work section. The characteristics of the up-call with emphasis on the significant features will follow in the next section. Details on the implemented algorithm with its parameters and flowchart will then be described followed by the results and discussion. The paper is then ended with a future work section and a conclusion.

2. Related Work

Many work has been done to detect the presense of certain animal species based on their sounds or calls. Whales and birds are mostly studied because of their unique communication capabilities. Some studies were done to classify birds. These studies help in reducing birdstrikes with airplanes especially near airports Kwan et al. (2006). The authors used a Hidden Markov Model and Guassian Mixture Models to do suitable real-time monitering for a large number of birds. Marcarini et al. (2008) studied specific kinds of migrating birds using Guassian Mixture Models alongside with spectogram correlation. In fact, many studies use spectogram correlation when it comes to acoustic analysis Mellinger et al. (2000), Chabot (1988), and Mellinger et al. (2000). The study of the NARW call is not that common. The rest of this section will specifically discuss the NARW contact call classification. Dugan et al. (2010) used classification and regression trees (CART) and artificial neural networks (ANN) methods. The results were then compared with a feature vector testing (FVT) approach. The CART had the highest assignemnt rate. The FVT had low false possitive rates, however, it also had an overall assignemnt rates less than the ANN method. Dugan et al. (2010) improved on the results that were generated in their previous work. The proposed method features multiple algorithms running in parrallal. The output of the individual algorithms is then fed into a decesion algorithm that provides the final output of the system. The developed algorithm had a better detection probability than the FVT. Side by side comparision between the FVT and the developed method showed that the developed method had lower number of false positive rates.The authors howerever did not mention the performance when it comes to missing a positive call. Urazghildiiev el al. (2010) developed a multistage, hypothesis testing technique that involves the generalized likelihood test detector. The proposed algorithm was able to detect approximaty 80% of the contact calls that where detected by a human operator. The algorithm had about 26 false alarms per day.The method implemented in this paper uses a novel algorithm that preprocess the sound waves before theyare passed into a tree based classifier to output the final result. The next section will discuss the distinctive features that uniquely identify the contact call of the NARW.

3. Contact Call Features

The unique characteristics of the NARW are the key components that help identify the presence of a NARW even in the existence of noise as well as in the presence of other kinds of whales. Figure 1 shows a sample NARW contact call. The duration, minimum frequency, maximum frequency, and bandwidth are some of the features that uniquely distinguish the NARW from a pool of sounds. For example, 99% of these calls are within 0.3 to 1.5 seconds of range Gillespie et al. (2004). Another key feature of the signal is that the upsweep part comprises of 30 to 100% of the total signal duration. A list of features was identified by Dugan et al. (2010). Table 1 lists these features; ultimately these features were used to classify the NARW up call.

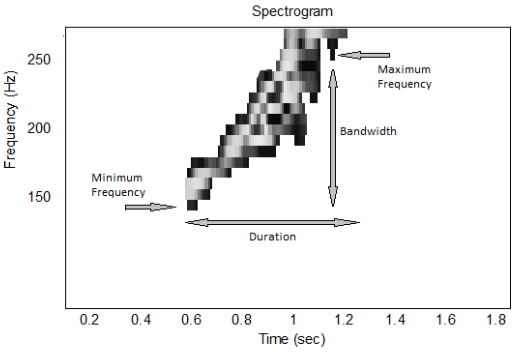


Figure 1. Sample North Atlantic Right Whale Contact Call

Table 1. List of Features used to Classify the North Atlantic Right Whale Up-Calls

Feature	Description
f_1	Signal Duration
f_2	Minimum Frequency
f_3	Maximum Bandwidth
f_4	Start-end Bandwidth
f_5	Duration of Upsweep
f_6	Local Noise Level
f_7	Segmentation thresholds
f_8	Mean Value of the Instantaneous Bandwidth
f_9	Percentage of holes in the object
f_{10}	Percentage of down sweeps in object
f_{11}	Percentage of harmonics in the object

4. Developed Algorithm

Given the features and the signal properties provided in the previous section, an algorithm was developed to process the sound signals and determine whether a NARW is present or not. The first developed step is the signal pre-analysis. Signal pre-analysis will be discussed in the next section followed by the description and analysis on the implemented algorithm.

4.1 Signal Pre-Analysis

Before the developed algorithm can be implemented a pre-analysis is done on the sound signals. Pre-analysis constitutes of the 3 steps listed below:

1. Read the sound wave and then use fast fourier transform to convert the signal from the time domain to the frequency domian.
2. Ignore the weakest 80% of the data matrix
3. Clear data islands using the weakest neighborhood points

The weakest neighborhood method is a developed algorithm that aims at clearing data islands. It does not eliminate 100% of the data islands but it does significant improvements on the data matrix and makes it easier to handle. Algorithm 1 shows the weakest neighborhood method. Figure 2 shows the return parameters of the $getNeigherhood(i)$ function that is called in line 14.

Algorithm 1 Weakest Neighborhood Method

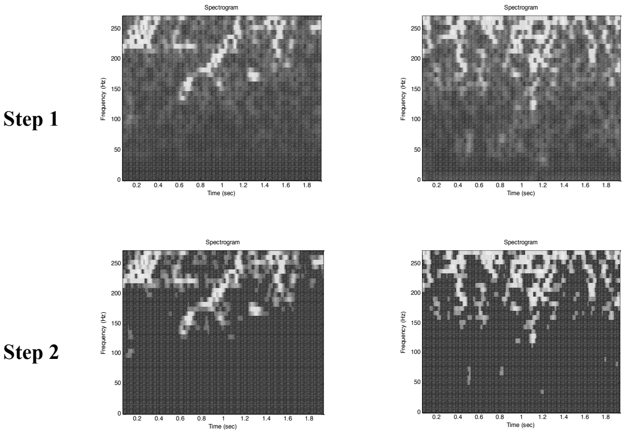
```
Input: DataMatrix
Output: $DataMatrix_{MODIFIED}$
1  StopCondition1 ← 0
2  while ¬ StopCondition1do
3      temp = DataMatrix
4      DataMatrix = clearDataIslands(DataMatrix)
5      if temp = DataMatrix then
6          $DataMatrix_{MODIFIED}$ = temp
7          StopCondition1 = TRUE
8      end
9  end
10 return $DataMatrix_{MODIFIED}$

DataMatrix = clearDataIslands(DataMatrix)
11 StopCondition2 ← 0
12 while¬ StopCondition2 do
13     $\forall$ $non - edge\ point\ i\ \neq 0\ \in\ DataMatrix$ do
14     $[a\ b\ c\ d\ e\ f\ g\ h] =\ getNeigherhood(i)$
15     if a >0 then a=1end
16     if b >0 then b=1 end
17     if c >0 then c=1 end
18     if d >0 then d=1 end
19     if e >0 then e=1 end
20     if f >0 then f=1  end
21     if a >0 then g=1 end
22     if a >0 then h=1 end
23     $SUM = a + b + c + d + e + f + g + h$
24     if  SUM < 4
25         update DataMatrixby setting
```

```
26          $(non - edge\ point\ i = 0)$
27      end
28      if $\nexists\ non - edge\ point$
29      |   StopCondition2 = TRUE
30      end
31  end
32  return DataMatrix
```

a	b	c
d	***i***	h
e	f	g

Figure 2. The Return Parameters of the $getNeigherhood$ Function

Lines 1 to 10 in Algorithm 1 show the main method. Lines 11 to 32 illustrate the *clearDataIslands* function that is called in line 4. Lines 1 and 11 initiate the terminate conditions for the routine and the subroutine respectively. The routine in lines 1 to 10 keeps calling the *clearDataIslands* function until there is no improvement on the results. Another termination criterion might also be added in case there is a computational concern. For example, a limit might be set on the number of improvements rather than waiting until no improvement is possible. The subroutine *clearDataIslands* gets the neighborhood of each non-zero non-edge point in the *DataMatrix* and then finds if less than 4 of the neighboring points has a value of 0 (line 24) if this is the case then that specific point is replaced with zero otherwise it is left unchanged. The subroutine is terminated when all the points in the *DataMatrix* have been tested. The results of the signal preprocessing section are significant and it sets the path towards easier processing in the next section. Figure 3 shows 2 examples of different sound recording after the implementation progresses through the 3 steps described earlier.



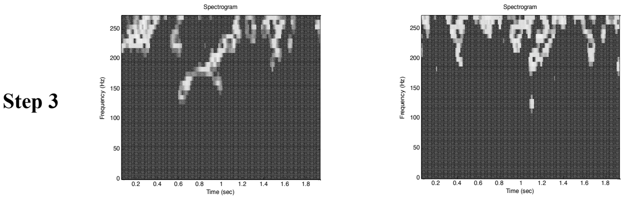


Figure 3. Spectogram Changes after Implementing Steps 1, 2 and 3 as Described in the Signal Pre-Analysis Section

4.2 Developed Algorithm

The output of the process that was done in the previous section is then used as the input to the developed algorithm. The developed algorithm consists of 8 steps. These steps are listed in their sequence of execution in the following order:

1. Initiate 10 equally spaced particles $p_1, p_2, \ldots, p_{10}$ on the time axis (For example, if the sound wave spectrogram is 2 seconds in length then the particles will be located at [0.00 0.22 0.44 0.66 0.88 1.11 1.33 1.55 1.77 2.00]).
2. These particles are set to explore and determine the first non-zero terms on the frequency matrix (Figure 4 shows the selected particles for a sample sound recording example)

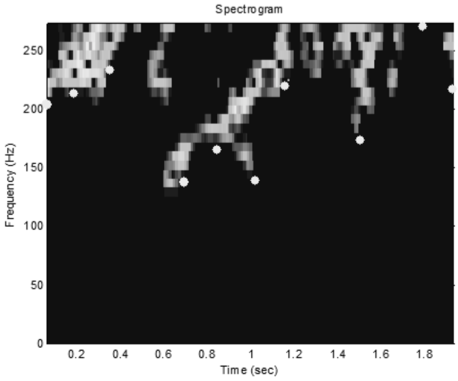


Figure 4. Circles Representing the Particles (Step 2)

3. Particle $p_1$then travels to particle p_2, particle p_2 travels to particle p_3, etc…The particles travel only on non-zero roads (i.e. when the frequency >0). The equation of motion is expressed as follows: Roulette wheel selection is used to pick the next point of motion. The probability of picking a point j in the neighborhood of i is given by a deterministic factor and a random factor. The equation that determines that factors is expressed by (1)

$$\alpha\left(\frac{1}{1+(f(p_n)-\mu)^2}\right) + \beta x \qquad (1)$$

Where α and β are the learning parameters ($\alpha + \beta = 1$). x is a random number with the range from 0 to 1. $f(p_n)$ is the frequency value at point p_n. μ is the mean value (of the frequency values) between p_n and p_{n+1}. Figure 5 shows the result of executing step 4. If there is no route between the points then the path is not generated and it will be eliminated in the next step.

Equation 1 is inspired by the Particle Swarm Optimization technique and by how the particles move from a random point to the minimum or maximum in an optimization problem Kennedy and Eberhart (1995) and Venter et al. (2003). The values of α and β can be adjusted to favor the deterministic factor V/S the stochastic factor, if β increases then α must decrease thus favoring the second term in (1). For example, setting α to 0 will cause the motion between p_n and p_{n+1} to be completely stochastic.

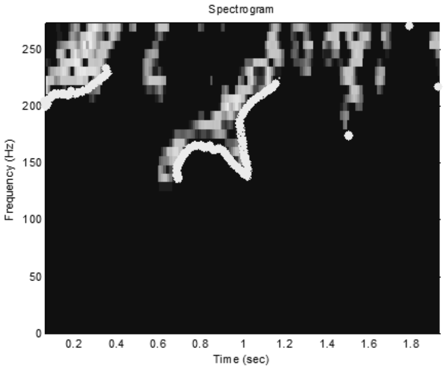


Figure 5. The Result of Executing Step 4

4. All found paths that are < 0.3 seconds are ignored (The 0.3 value came from the fact that no NARW contact call should be less than 0.3 seconds in duration). Figure 6 shows the outcome of this step.
5. Repeat step 1 (initiate 10 points but this time on the new time axis as shown in Figure 7)

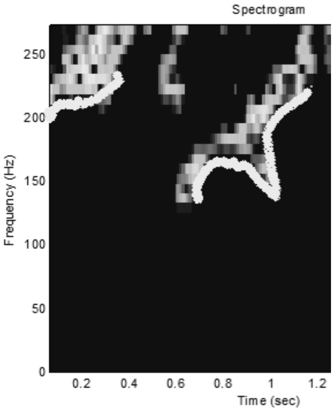


Figure 6. Outcome of Step 4

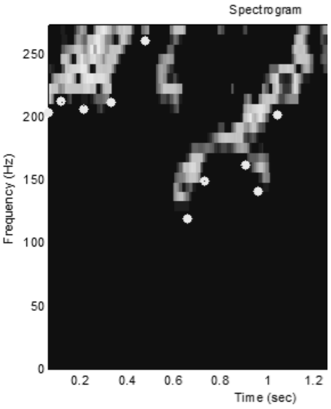


Figure 7. Outcome of Step 5

6. Repeat step 3 (particles travel from one point to the other) until there are no paths that are <0.3 seconds in length. Figure 8 shows the outcome of step 6.

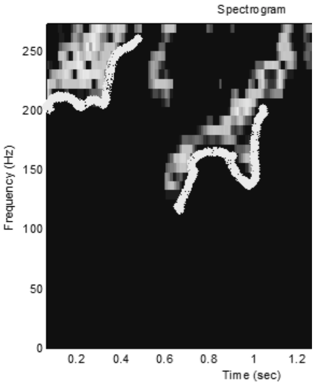


Figure 8. Outcome of Step 6

7. Let the particles propagate in the y-axis (frequency) until a zero element is observed. Figure 9 shows the outcome of this step.
8. Any particle that fails in observing a zero element shall be ignored. Furthermore any isolated particle shall also be ignored. Figure 10 shows the outcome of step 8.

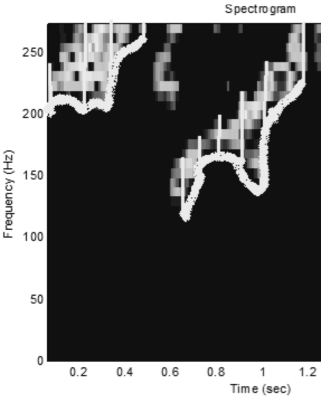


Figure 9. Outcome of Step 7

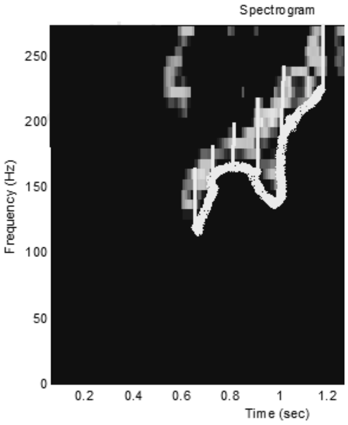


Figure 10. Outcome of Step 8

4.3 Feature Recognition Process

The feature recognition process starts with the output that was generated from step 8 from the previous section. All the found path(s) are used in the feature recognition process. The features listed in Table I are identified and a score is given to the path(s). Gaussian based assignment is used for this process. For example, the length of the path (f_1 in Table I) should be between 0.3 seconds to 2 seconds. On the other hand, 99% of the signal lengths are between 0.3 to 1.5 seconds. Thus it makes more sense to give more score to the signals that fall within this range. Figure 11 shows an example of how would the score be distributed for f_1. The same is done to all of the features listed in Table 1. The Gaussian mean and variance, that defines how the curve deviates, change for every feature according to the trained data. Once the score for every feature is calculated, these scores will be used as the split decision factor in the tree based classifier until a decision is made.

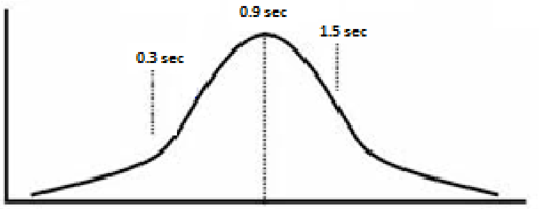


Figure 11. Gaussian Distribution for Feature f_1

5. Results

The algorithm was trained with a sample of 30,000 points. The success rate was at 84.7%. For a sample of 1000 sound signals, 112 were identified as false positive calls. On the other hand, 41 calls were mistakenly identified as a non NARW contact call.

6. Future Work

Many improvements can be done on the designed algorithm to enhance the results. αandβ can be better estimated to provide the optimal combination between the deterministic V/S the stochastic factor. Another improvement is to optimally design the Gaussian based assignment described earlier. The design should adhere to each of the specific feature vectors. The weakest neighborhood method can also be enhanced to have better coverage by deleting the unnecessary data islands that might still exist after the method had executed. Finally, the algorithm should improve, to guarantee a no-miss of the whale up call. Although the miss rate is low but getting the number closer to zero will have a huge improvement on the solution.

7. Conclusion

In this paper, an algorithm was presented to detect the presence of an up-call of the North Atlantic right whale. 30,000 recording were used to train the model that was based on a tree classifier. The algorithm proved to successfully work by detecting the contact calls with a success rate close to 85%.

References

D. Chabot, "A quantitative technique to compare and classify humpback whale (magapter novaeangliae) sounds," Ethology, vol. 77, pp. 89-102, 1988.

Caswell, Hal, Masami Fujiwara, and Solange Brault. "Declining survival probability threatens the North Atlantic right whale." *Proceedings of the National Academy of Sciences* 96.6 (1999): 3308-3313.

Clark, Christopher W., et al. "Listening to their world: acoustics for monitoring and protecting right whales in an urbanized ocean." *The urban whale: North Atlantic right whales at the crossroads (SD Kraus and RM Rolland, eds.). Harvard University Press, Cambridge, Massachusetts* (2007): 333-357.

D.K. Mellinger and C.W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am., vol. 107, no. 6, pp. 3518-2529, June 2000.

Dugan, P. J., et al. "North Atlantic right whale acoustic signal processing: Part II. Improved decision architecture for auto-detection using multi-classifier combination methodology." *Applications and Technology Conference (LISAT), 2010 Long Island Systems*. IEEE, 2010.

Dugan, Peter J., et al. "North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms." *Applications and Technology Conference (LISAT), 2010 Long Island Systems*. IEEE, 2010.

Fujiwara, Masami, and Hal Caswell. "Demography of the endangered North Atlantic right whale." *Nature* 414.6863 (2001): 537-541.

Gillespie, Douglas. "Detection and classification of right whale calls using an'edge'detector operating on a smoothed spectrogram." *Canadian Acoustics* 32.2 (2004): 39-47.

Jensen, Aleria S., Gregory Keith Silber, and John Calambokidis. *Large whale ship strike database*. US Department of Commerce, National Oceanic and Atmospheric Administration, 2004.

Kennedy, James, and Russell Eberhart. "Particle swarm optimization." *Neural Networks, 1995. Proceedings., IEEE International Conference on*. Vol. 4. IEEE, 1995.

Kraus, Scott D., et al. "North Atlantic right whales in crisis." *SCIENCE-NEW YORK THEN WASHINGTON-* 5734 (2005): 561.

Kwan, C., et al. "An automated acoustic system to monitor and classify birds." *EURASIP Journal on Advances in Signal Processing* 2006 (2006).

Marcarini, M.; Williamson, G.A.; de Sisternes Garcia, L., "Comparison of methods for automated recognition of avian nocturnal flight calls," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* , vol., no., pp.2029,2032, March 31 2008-April 4 2008.

Matthews, J. N., et al. "Vocalisation rates of the North Atlantic right whale (Eubalaena glacialis)." *Journal of Cetacean Research and Management* 3.3 (2001): 271-282.

Mellinger, David K., and Christopher W. Clark. "Recognizing transient low-frequency whale sounds by spectrogram correlation." *The Journal of the Acoustical Society of America* 107 (2000): 3518.

Parks, Susan E., and Peter L. Tyack. "Sound production by North Atlantic right whales (Eubalaena glacialis) in surface active groups." *The Journal of the Acoustical Society of America* 117 (2005): 3297.

Spaulding, Eric, et al. "An autonomous, near-real-time buoy system for automatic detection of North Atlantic right whale calls." *Proceedings of Meetings on Acoustics*. Vol. 6. 2009.

Urazghildiiev, Ildar R., et al. "Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise." *Oceanic Engineering, IEEE Journal of* 34.3 (2009): 358-368.

Vanderlaan, Angelia SM, Alex E. Hay, and Christopher T. Taggart. "Characterization of North Atlantic right-whale (Eubalaena glacialis) sounds in the Bay of Fundy." *Oceanic Engineering, IEEE Journal of* 28.2 (2003): 164-173.

Venter, Gerhard, and Jaroslaw Sobieszczanski-Sobieski. "Particle swarm optimization." *AIAA journal* 41.8 (2003): 1583-1589.

The Orchive : Data mining a massive bioacoustic archive

Steven Ness SNESS@UVIC.CA
Department of Computer Science, University of Victoria, Canada

Helena Symonds INFO@ORCALAB.ORG
OrcaLab, P.O. Box 510 Alert Bay, BC, Canada

Paul Spong INFO@ORCALAB.ORG
OrcaLab, P.O. Box 510 Alert Bay, BC, Canada

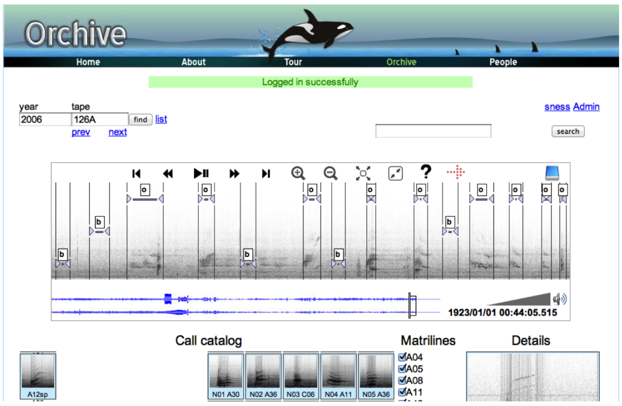
George Tzanetakis GTZAN@CS.UVIC.CA
Department of Computer Science, University of Victoria, Canada

Abstract

The Orchive is a large collection of over 20,000 hours of audio recordings from the OrcaLab research facility located off the northern tip of Vancouver Island. It contains recorded orca vocalizations from the 1980 to the present time and is one of the largest resources of bioacoustic data in the world. We have developed a web-based interface that allows researchers to listen to these recordings, view waveform and spectral representations of the audio, label clips with annotations, and view the results of machine learning classifiers based on automatic audio features extraction. In this paper we describe such classifiers that discriminate between background noise, orca calls, and the voice notes that are present in most of the tapes. Furthermore we show classification results for individual calls based on a previously existing orca call catalog. We have also experimentally investigated the scalability of classifiers over the entire Orchive.

1. Introduction

The Orchive is a large archive containing over 20,000 hours of recordings from the Orcalab research station. These recordings were made using a network of hydrophones and originally stored on analog cassette tapes. OrcaLab is a research station on Hanson Island which is located at the north part of Vancouver Island on the west coast of Canada. It has been in continuous operation since 1980. It was designed as a land based station in order to reduce the impact on the orcas under study, as the noise and disturbance from boats affects the orcas in observable but currently unquantified ways. In collaboration with OrcaLab, we have digitized the tapes and have made these recordings available to the scientific community through the Orchive website (http://orchive.cs.uvic.ca).



Logged in successfully

year 2006

tape 126A

1923/01/01 00:44:05.515

☑A04
☑A05
☑A08
☑A11

Figure 1. Annotated audio from from the Orchive

Over the past 5 years, a number of orca researchers using our website have added over 18,000 clip annotations to our database. A small section of annotated audio from the Orchive is shown in Figure 1. These clip annotation are of two main types: The first is clips that differentiate background noise from orca calls and from the voice notes of the researchers that collected

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

the data. The second type of clip annotations classify orca vocalizations into different calls. Orcas make three types of vocalizations, echolocation clicks, whistles and pulsed calls. The pulsed calls are highly conserved stereotyped vocalizations which have been classified into a catalog of over 52 different calls by John Ford (Ford, 1987). Of the 18,000 annotations currently in the Orchive, 3000 are of these individually classified calls. In addition, we have a curated call catalog containing 384 different recordings of different calls vocalized by a variety of different pods and matrilines. This catalog is used for training the annotators.

Many parts of the recordings contain boat noise which makes identifying orca calls both difficult and tiring. In addition, the size of the Orchive makes full human annotation practically impossible. Therefore we have explored machine learning approaches to the task. One data mining task is to segment and label the recordings with the labels background, orca, voice. Another is to subsequently classify the orca calls into the classes specified in the call catalog.

2. Related Work

Audio feature extraction is the first step in classifying audio using machine learning algorithms. Mel-Frequency Cepstral Coefficients (Logan, 2000) (MFCC) have been widely used for this purpose. MFCCs have also been used in bioacoustics, and have been used to classify bird songs (Lee et al., 2006) and orca calls (Ness et al., 2008). In this work we also use MFCCs, but supplement them with other audio features including Centroid frequency, Rolloff frequency, Flux, and Zero Crossings.

Our system uses two types of web based interfaces. The first are tools aimed at expert users, and the second are simpler interfaces designed for crowdsourcing the annotation. There are a number of tools that experts use to segment and analyze audio and specifically bioacoustic data. One of the most popular is Raven (http://www.birds.cornell.edu/raven), a toolkit developed at the Cornell Lab of Ornithology. The biggest difference our system compared to systems such as Raven is that our system web-based, can more easily view and analyze large amounts of data.

3. System Overview

We have developed a collaborative web interface that allows expert researchers to listen to, view and annotate large collections of audio data. The system also supports a variety of audio feature extraction and machine learning algorithms, and enables users to view the results of these algorithms. A diagram of this system is shown in Figure 2.

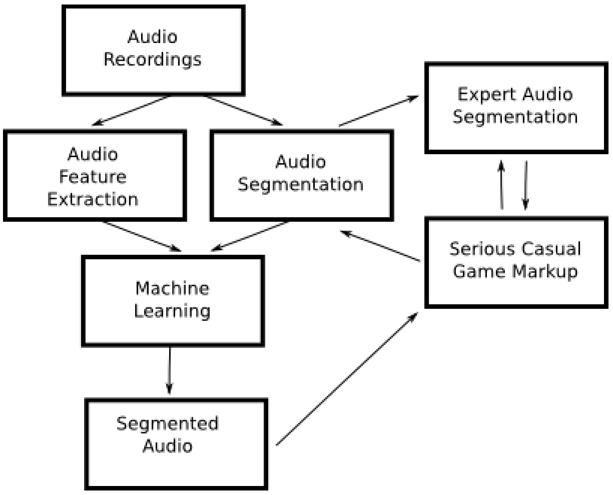


Figure 2. System Diagram

For audio features we use the Marsyas(Tzanetakis, 2008) Music Information Retrieval system. Marsyas allows us to perform both audio feature extraction and machine learning on audio data directly.

In order to efficiently analyze large audio archives we utilize distributed computing. There are many systems for distributing computation. We currently use the Portable Batch System (PBS) (Henderson, 1995), a grid-computing system where similar data can be processed in parallel by a large number of computers.

4. Experimental Results

4.1. Audio Feature Extraction Parameters

The first set of parameters that needed to be optimized were the Window Size and Hop Size of the Digital Signal Processing (DSP) algorithms that take the input audio and calculate spectral information from them, the fundamental basis for which is the Fast Fourier Transform (FFT) algorithm. The length of time over which to calculate the statistical properties of the features, this is known in bextract as the “memory” and corresponds to the number of frames of features that are accumulated. We ran this on a 600 second audio dataset labeled as orca, background and voice with equal lengths of each label. In this dataset, the voice was trimmed by hand, the orca consisted of the middle 0.023 seconds of approximately 10,000 clips, and the background consisted of 0.15 seconds of approximately 1300 clips. The results for this are shown in Table 1. From this we can see that as we go to longer window sizes, the classification performance increases, and as

winsize	hopsize	memory	# correct
20	512	256	70.16
20	1024	512	71.88
20	2048	1024	74.17
20	4096	2048	73.38
40	512	256	72.94
40	1024	512	75.67
40	2048	1024	78.29
40	4096	2048	80.58
80	512	256	76.53
80	1024	512	78.39
80	2048	1024	81.88
80	4096	2048	85.72

Table 1. In this table results of a systematic parameter search through different DSP parameters is shown. winsize is the window size of the FFT in samples, and hop size is the number of samples skipped between each successive application of the FFT. memsize refers to the number of FFT frames on which the mean and standard deviation are determined.

we go to longer accumulation window sizes, the performance also increases. For the remaining experiments we use these optimal settings.

4.2. Orca/Background/Voice Classification

The first task we investigate is the classification of audio into three classes: orca, background, and human voice. In order to test the different distributed audio classification systems we first generated a set of training and testing data, one of these was a set of calls from the curated call catalog with silence removed, and the other was an entire 45 minute recording from the Orchive which had been annotated by an orca researcher. In a previous paper (Ness et al., 2008), we were able to obtain a classification performance of 82% when using a SVM classifier on hand labeled data. We looked in more detail at the training data, and found that there was a small amount of silence before and after the vocalization. The results can be found in the first line of table 2 and had 93.5% of the instances classified correctly. This large jump in performance was unexpected but easily understood, because if feature vectors of silence are labeled as orca, this will cause issues for the classifier. We then took a 4 minute region of orca calls and voice notes and removed all the silences from both of them, for this we obtained a classification accuracy of 96.1% when looking at the call catalog dataset, and 95.0% when looking at the annotated recording.

However, this process of hand trimming recordings would be unfeasible to do on the entire 18,000 current annotations. For this, we instead tested a procedure where we extracted a small section of audio from the middle of each clip where it was most probable that the orca call would be found.

Training dataset	length (sec)	% corr. 10-fold	% corr. (calls)	% corr. (442A)
hand-10sec	30	99.4	93.5	93.1
hand-4min	720	99.9	96.1	95.0
ms 100	300	99.9	96.5	93.4

Table 2. Classification results with hand trimmed orca vocalizations using bextract using an SMO SVM classifier.

We then extracted audio features from these sections of audio using Marsyas. Marsyas has a wide variety of audio features that it can calculate, including MFCCs, number of zero crossings per window and various high level descriptions of the spectrum including the centroid (center of mass of the spectrum), rolloff (the frequency for which the sum of magnitudes of its lower frequencies are equal to percentage of the sum of magnitudes of its higher frequencies) and the flux (the norm of the difference vector between two successive magnitue/power spectra). We tried different combinations of these, and found that using all of these features gave the best performance. All subsequent results in this paper use all of these features.

To classify these features, we used a Sequential Minimal Optimization implementation of a Support Vector Machine classifier (Platt, 1998), an algorithm which had shown its effectiveness in our previous work (Ness et al., 2008) in this problem domain.

The results for this procedure for a clip of 0.023 seconds from the middle of each orca call was 96.5% and for the recording from the Orchive, the accuracy was 93.4%.

4.3. Call classification

Using the Orchive interface we created a collection of 197 calls of 6 classes, these included the common calls “N1”, “N3”, “N4”, “N7”, “N9” and “N47”. Audio features for each 20ms audio frame of these files were generated, these included the MFCC coefficients, Centroid, Rolloff, Flux and Zero crossings as described and justified in the previous section. The mean and standard deviation for each of these features were then calculated and were output as a .arff file. The SMO SVM classifier produced gave an accuracy of 98.5% accuracy on this set of calls, and the confusion matrix for this is shown in Table 3.

	N1	N4	N7	N9
N1	1726	0	0	0
N4	12	2858	0	0
N7	0	2	1297	59
N9	0	0	70	3231

Table 3. Confusion matrix for 10-fold crossvalidation with SVM classifier on labelled calls from Orchive.

Training data (sec)	% of Orchive	Run time (d:h:m:s)
30	1	00:00:05:18
30	5	00:00:25:20
30	10	00:00:50:58
30	100	00:09:01:05
240	1	00:06:16
240	5	00:00:31:21
240	10	00:04:47:12
240	100	02:04:18:32

Table 4. Performance results of timing on subsets of the entire Orchive dataset.

4.4. Performance

In order to investigate the performance of the classification of recordings into Orca, Background and Voice, we trained a SVM with a section of 30 and 240 seconds of hand trimmed data using the bextract program in Marsyas. We then used the sfplugin program in Marsyas to classify all the recordings in the Orchive on the Hermes/Nestor cluster, part of the Westgrid computational resource. For this we divided the data into sets of 1%, 5%, 10% and 100% of the Orchive. The timing results of these datasets run on 10 computers are shown in Table 4. From this we can see that the classifier that had more data took longer to classify, and that the speedup from taking samples of the data was almost linear.

5. Conclusion

In this paper we described a system that allows orca researchers to listen to, view and annotate the large amount of audio data in the Orchive. The system also allows researchers to run and view the results of audio feature extraction and machine learning algorithms on this data.

We investigated the performance of different parameters for the audio feature extraction process and showed that in general, large window sizes were beneficial, and that increasing the length of time that statistics were taken over the data was also beneficial. We showed that by carefully hand editing clips to remove silence was very useful, and boosted performance from around 90% to 96% on actual recordings. We then used these classifiers on a cluster to classify all the recordings in the Orchive into the classes, Orca, Background and Voice. The performance of call classification was also good, with a classification accuracy of 98.5% using a collection of 197 calls culled from the Orchive. The calls most often misclassified were the N7 and N9 calls, and these are also difficult for non-experts in orca vocalizations to differentiate.

References

Ford, J.K.B. A catalogue of underwater calls produced by killer whales (orcinus orca) in british columbia. Technical Report 633, Canadian Data Report of Fisheries and Aquatic Science, 1987.

Henderson, R. Job scheduling under the portable batch system. In Feitelson, Dror and Rudolph, Larry (eds.), *Job Scheduling Strategies for Parallel Processing*, volume 949 of *Lecture Notes in Computer Science*, pp. 279–294. Springer, 1995.

Lee, C.H., Lien, C.C., and Huang, R.Z. Automatic recognition of birdsongs using mel-frequency cepstral coefficients and vector quantization. In *IMECS*, pp. 331–335, 2006.

Logan, B. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

Ness, S.R., Wright, M, Martins, L.G., and Tzanetakis, G. Chants and orcas: semi-automatic tools for audio annotation and analysis in niche domains. In *Proc. of the 2nd ACM workshop on Multimedia semantics*, pp. 9–16, 2008.

Platt, John C. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances In Kernel Methods - Support Vector Learning, 1998.

Tzanetakis, G. *Marsyas-0.2: A case study in implementing music information retrieval systems*, chapter 2, pp. 31–49. Intelligent Music Information Systems: Tools and Methodologies. Information Science Reference, 2008. Shen, Shepherd, Cui, Liu (eds).

Physeter catodon localization by sparse coding

Sébastien PARIS SEBASTIEN.PARIS@LSIS.ORG
DYNI team, LSIS CNRS UMR 7296, Aix-Marseille University

Yann DOH YANNDOH.M2@GMAIL.COM
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Hervé GLOTIN GLOTIN@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Xanadu HALKIAS HALKIAS@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Joseph RAZIK RAZIK@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Abstract

This paper presents a spermwhale' localization architecture using jointly a bag-of-features (BoF) approach and machine learning framework. BoF methods are known, especially in computer vision, to produce from a collection of local features a global representation invariant to principal signal transformations. Our idea is to regress supervisely from these local features two rough estimates of the distance and azimuth thanks to some datasets where both acoustic events and ground-truth position are now available. Furthermore, these estimates can feed a particle filter system in order to obtain a precise spermwhale' position even in mono-hydrophone configuration. Anti-collision system and whale watching are considered applications of this work.

1. Introduction

Most of efficient cetacean localisation systems are based on the Time Delay Of Arrival (TDOA) estimation from detected[1] animal's click/whistles signals (Nosal & Frazer, 2006; Bénard & Glotin, 2009). Long-base hydrophones'array is involving several fixed, efficient but expensive hydrophones (Giraudet & Glotin, 2006) while short-base version is requiring a precise array's self-localization to deliver accurate results. Recently (see (Glotin et al., 2011)), based on Leroy's attenuation model versus frequencies (Leroy, 1965), a range estimator have been proposed. This approach is working on the detected most powerful pulse inside the click signal and is delivering a rough range' estimate robust to head orientation variation of the animal. Our purpose is to use i) these hydrophone' array measurements recorded in diversified sea conditions and ii) the associated ground-truth trajectories of spermwhale (obtained by precise TDAO and/or Dtag systems) to regress both position and azimuth of the animal from a third-party hydrophone[2] (typically on-board, standalone and cheap model).

We claim, as in computer-vision field, that BoF approach can be successfully applied to extract a global and invariant representation of click's signals. Basically, the pipeline of BoF approach is composed of three parts: i) a local features extractor, ii) a local feature encoder (given a dictionary pre-trained on data) and iii) a pooler aggregating local representations into a more robust global one. Several choice for encoding local patches have been developed in recent years: from hard-assignment to the closest dictionary basis (trained for example by Kmeans algorithm) to

[1] As click/whistles detector, matching filter is often preferred

[2] We assume that the velocity vector is collinear with the head's angle.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

a sparse local patch reconstruction (involving for example Orthogonal Maching Pursuit (OMP) or LASSO algorithms).

2. Global feature extraction by spare coding

2.1. Local patch extraction

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. Each matrix $\boldsymbol{C}^j$ is defined by $\boldsymbol{C}^j \triangleq \{\boldsymbol{c}_i^j\}$, $i = 1, \ldots, N^j$ where $\boldsymbol{c}_i^j \in \mathbb{R}^n$ is the i^{th} click of the j^{th} hydrophone. For our *Bahamas2* dataset (Giraudet & Glotin, 2006), we choose typically $n = 2000$ samples surrounding the detected click. The total number of available clicks is equal to $N = \sum_{i=1}^{H} N^j$.

As local features, we extract simply some local signal patches of $p \leq n$ samples (typically $p = 128$) and denoted by $\boldsymbol{z}_{i,l}^j \in \mathbb{R}^p$. Furthermore all $\boldsymbol{z}_{i,l}^j$ are ℓ_2 normalized. For each $\boldsymbol{c}_i^j$, a total of 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 retrieved (see Fig. 1). All local patches associated with the j^{th} hydrophone are denoted by $\boldsymbol{Z}^j \triangleq \{\boldsymbol{Z}_i^j\}$, $i = 1, \ldots, N^j$ while $\boldsymbol{Z} \triangleq \{\boldsymbol{Z}^j\}$ is denoting all the local patches matrix for all hydrophones. A final post-processing consists in uncorrelate local features by PCA training and projection with $p' \leq p$ dimensions.

2.2. Local feature encoding by sparse coding

In order to obtain a global robust representation of $\boldsymbol{c} \subset \boldsymbol{C}$, each associated local patch $\boldsymbol{z} \subset \boldsymbol{Z}$ are first linearly encoded *via* the vector $\boldsymbol{\alpha} \in \mathbb{R}^k$ such as $\boldsymbol{z} \approx \boldsymbol{D}\boldsymbol{\alpha}$ where $\boldsymbol{D} \triangleq [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k] \in \mathbb{R}^{p \times k}$ is a pre-trained dictionary matrix whose column vectors respect the constraint $\boldsymbol{d}_j^T \boldsymbol{d}_j = 1$. In a first attempt to solve this linear problem, $\boldsymbol{\alpha}$ can be the solution of the Ordinary Least Square (OLS) problem:

$$l_{OLS}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \right\}. \quad (1)$$

OLS formulation can be extended to include regularization term avoiding overfitting. We obtain the ridge regression (RID) formulation:

$$l_{RID}(\boldsymbol{\alpha}|\mathbf{z}; \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \beta \|\boldsymbol{\alpha}\|_2^2 \right\}. \quad (2)$$

This problem have an analytic solution $\boldsymbol{\alpha} = (\boldsymbol{D}^T \boldsymbol{D} + \beta \boldsymbol{I}_k)^{-1} \boldsymbol{D}^T \boldsymbol{z}$. Thanks to semi-positivity of $\boldsymbol{D}^T \boldsymbol{D} + \beta \boldsymbol{I}_k$, we can use a cholesky factor on this matrix to solve efficiently this linear system. In order to decrease reconstruction error and to have a sparse solution, this problem can be reformuled as a constrained Quadratic Problem (QP):

$$l_{SC}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \ s.t. \ \|\boldsymbol{\alpha}\|_1 = 1. \quad (3)$$

To solve this problem, we can use a QP solver involving high combinatorial computation to find the solution. Under RIP assumptions (Tibshirani, 1994), a greedy approach can be used efficiently to solve and eq. 3 and this latter can be rewritten as:

$$l_{SC}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1, \quad (4)$$

where λ is a regularization parameter which controls the level of sparsity. This problem is also known as basis pursuit (Chen et al., 1998) or the Lasso (Tibshirani, 1994). To solve this problem, we can use the popular Least angle regression (LARS) algorithm.

2.3. Pooling local codes

The objective of pooling (Boureau et al.; Feng et al.) is to transform the joint feature representation into a new, more usable one that preserves important information while discarding irrelevant detail. For each click signal, we usually compute L codes denoted $\boldsymbol{V} \triangleq \{\boldsymbol{\alpha}_i\}$, $i = 1, \ldots, L$. Let define $\boldsymbol{v}^j \in \mathbb{R}^L$, $j = 1, \ldots, k$ as the j^{th} row vector of $\boldsymbol{V}$. It is essential to use feature pooling to map the response vector $\boldsymbol{v}^j$ into a statistic value $f(\boldsymbol{v}^j)$ from some spatial pooling operation f. We use $\boldsymbol{v}^j$, the response vector, to summarize the joint distribution of the j^{th} compounds of local features over the region of interest (ROI). We will consider the ℓ_μ-norm pooling and defined by:

$$f_n(\boldsymbol{v}; \mu) = \left(\sum_{m=1}^{L} |v_m|^\mu \right)^{\frac{1}{\mu}} \quad s.t. \ \mu \neq 0. \quad (5)$$

The parameter μ determines the selection policy for locations. When $\mu = 1$, ℓ_μ-norm pooling is equivalent to sum-pooling and aggregates the responses over the entire region uniformly. When μ increases, ℓ_μ-norm pooling approaches max-pooling. We can note the value of μ tunes the pooling operation to transit from sum-pooling to max-pooling.

2.4. Pooling codes over a temporal pyramid

In computer vision, Spatial Pyramid Matching (SPM) is a technic (introduced by (Lazebnik et al.)) which improves classification accuracy by performing a more

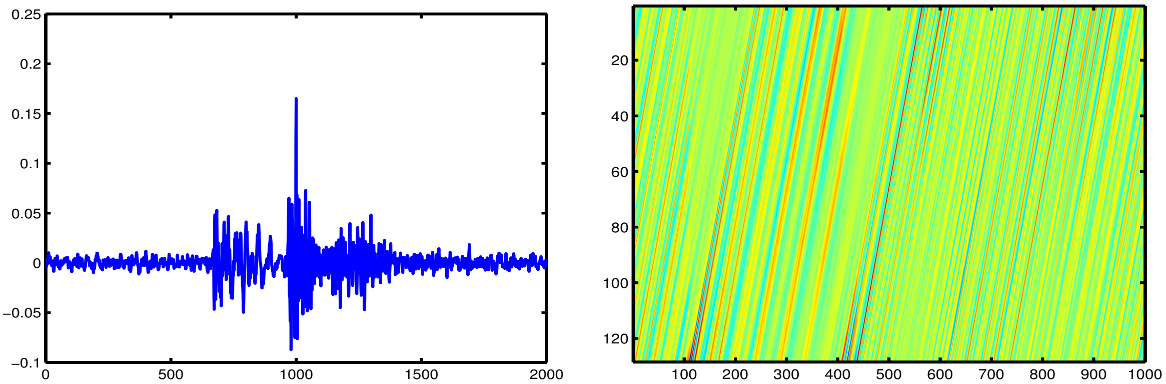


Figure 1. Left: Example of detected click with $n = 2000$. Right: extracted local features with $p = 128$, $L = 1000$ (one local feature per column).

robust local analysis. We will adopt the same strategy in order to pool sparse codes over a temporal pyramid (TP) dividing each click signal into ROI of different sizes and locations. Our TP is defined by the matrix $\mathbf{\Lambda}$ of size $(P \times 3)$ (Paris et al.):

$$\mathbf{\Lambda} = [\boldsymbol{a}, \boldsymbol{b}, \mathbf{\Omega}], \tag{6}$$

where $\boldsymbol{a}$, $\boldsymbol{b}$, $\mathbf{\Omega}$ are 3 $(P \times 1)$ vectors representing subdivision ratio, overlapping ratio and weights respectively. P designs the number of layers in the pyramid. Each row of $\mathbf{\Lambda}$ represents a temporal layer of the pyramid, *i.e.* indicates how do divide the entire signal into sub-regions possibly overlapping. For the i^{th} layer, the click signal is divided into $D_i = \lfloor \frac{1-a_i}{b_i} + 1 \rfloor$ ROIs where a_i, b_i are the i^{th} elements of vector $\boldsymbol{a}$, $\boldsymbol{b}$ respectively. For the entire TP, we obtain a total of $D = \sum_{i=1}^{P} D_i$ ROIs. Each click signal $\boldsymbol{c}$ $(n \times 1)$ is divided into temporal ROI $\boldsymbol{R}_{i,j}$, $i = 1, \ldots, P$, $j = 1, \ldots, D_i$ of size $(\lfloor a_i.n \rfloor \times 1)$. All ROIs of the i^{th} layer have the same weight Ω_i. For the i^{th} layer, ROIs are shifted by $\lfloor b_i.n \rfloor$ samples. A TP with $\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$ is designing a 2-layers pyramid with $D = 1 + 4$ ROIs, the entire signal for the first layer and 4 half-windows of $\frac{n}{2}$ samples with 25% of overlapping for the second layer. At the end of pooling stage over $\mathbf{\Lambda}$, the global feature $\boldsymbol{x} \in \mathbb{R}^d$, $d = D.k$ is defined by the weighted concatenation (by factor Ω_i) of L pooled codes associated with $\boldsymbol{c}$.

2.5. Dictionary learning

To encode each local features by sparse coding (see eq. 4), a dictionary $\boldsymbol{D}$ is trained offline with an important collection of $M \leq N.L$ local features as input. One would minimize the regularized empirical risk $\mathcal{R}_M$:

$$\mathcal{R}_M(\boldsymbol{V}, \boldsymbol{D}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \tag{7}$$

$$s.t.\ \boldsymbol{d}_j^T \boldsymbol{d}_j = 1.$$

Unfortunately, this problem is not jointly convex but can be optimized by alternating method:

$$\mathcal{R}_M(\boldsymbol{V}|\hat{\boldsymbol{D}}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \hat{\boldsymbol{D}}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1, \tag{8}$$

which can be solved in parallel by LASSO/LARS and then:

$$\mathcal{R}_M(\boldsymbol{D}|\hat{\boldsymbol{V}}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}}_i\|_2^2 \quad s.t.\ \boldsymbol{d}_j^T \boldsymbol{d}_j = 1. \tag{9}$$

Eq. 9 have an analytic solution involving a large matrix $(k \times k)$ inversion and a large memory occupation for storing the matrix $\boldsymbol{V}$ $(k \times M)$. Since M is potentially very large (up to 1 million), an online method to update dictionary learning is preferred (Mairal et al.). Figure 2 depicts 3 dictionary basis vectors learned *via* sparse coding. As depicted, some elements represents more impulsive responses while some more harmonic responses.

3. Range and azimuth logistic regression from global features

After the pooling stage, we extracted unsupervisly N global features $\boldsymbol{X} \triangleq \{x_i\} \in \mathbb{R}^{d \times N}$. We propose to regress *via* logistic regression both range r and azimuth az (in $x - y$ plan, when animal reach surface to breath) from the animal trajectory groundtruth denoted $\boldsymbol{y}$. For the current train/test splitsets of the

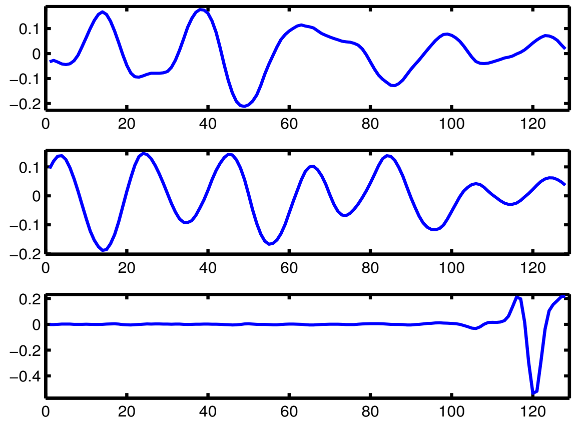


Figure 2. Example of trained dictionary basis with sparse coding.

data, such as $\boldsymbol{X} = \boldsymbol{X}_{train} \bigcup \boldsymbol{X}_{test}$, $\boldsymbol{y} = \boldsymbol{y}_{train} \bigcup \boldsymbol{y}_{test}$ and $N = N_{train} + N_{test}$, $\forall \; \{\boldsymbol{x}_i, y_i\} \in \boldsymbol{X}_{train} \times \boldsymbol{y}_{train}$, we minimize:

$$\widehat{\boldsymbol{w}}_\theta = \arg\min_{\boldsymbol{w}_\theta} \left\{ \frac{1}{2}\boldsymbol{w}_\theta^T \boldsymbol{w}_\theta + C \sum_{i=1}^{N_{train}} \log(1 + e^{-y_i \boldsymbol{w}_\theta^T \boldsymbol{x}_i}) \right\}, \tag{10}$$

where y_i denotes r_i and az_i for $\theta = r$ and $\theta = az$ respectively. Eq. 10 can be efficiently solved for example with Liblinear software (Fan et al., 2008). In the test part, range and azimuth for any $\boldsymbol{x}_i \in \boldsymbol{X}_{test}$ are reconstructed linearly by $\widehat{r}_i = \widehat{\boldsymbol{w}}_r^T \boldsymbol{x}_i$ and by $\widehat{az}_i = \widehat{\boldsymbol{w}}_{az}^T \boldsymbol{x}_i$ respectively.

4. Experimental results

4.1. bahamas2 dataset

This dataset (Giraudet & Glotin, 2006) contains a total of $N = 6134$ detected clicks for $H = 5$ different hydrophones (named H^7, H^8, H^9, H^{10} and H^{11} and with $N^7 = 1205$, $N^8 = 1238$, $N^9 = 1241$, $N^{10} = 1261$ and $N^{11} = 1189$ respectively).

To extract local features, we chose $n = 2000$, $p = 128$ and $L = 1000$ (tuned by model selection). For both the dictionary learning and the local features encoding, we chose $\lambda = 0.2$ and fixed 15 iterations to train dictionary on a subset of $M = 400.000$ local features drawn uniformly. We performed $K = 10$ cross-validation where training sets represented 70% of the total of extracted global features, the rest for the testing sets. Logistic regression parameter C is tuned by model selection. We compute the average root mean square error (ARMSE) of range/azimuth estimates per hydrophone: $ARMSE(l) = \frac{1}{K}\sum_{i=1}^{K}\sqrt{\sum_{j=1}^{N^l_{test}}(y^l_{i,j} - \widehat{y}^l_{i,j})^2}$ where $y^l_{i,j}$, $\widehat{y}^l_{i,j}$ and N^l_{test} represent the ground truth, the estimate and the number of test samples for the l^{th} hydrophone respectively. The global ARMSE is then calculated by $\overline{ARMSE} = \frac{1}{H}\sum_{l=1}^{H} ARMSE(l)$.

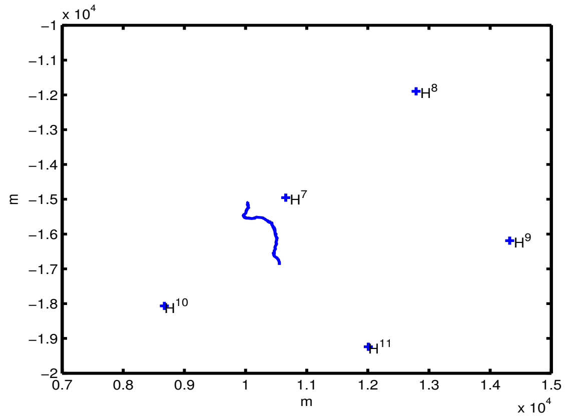


Figure 3. The 2D trajectory (in xy plan) of the single sperm whale observed during 25 min and corresponding hydrophones positions.

4.2. ℓ_μ-norm pooling case study

For preliminary results, we investigate the influence of the μ parameter during the pooling stage. We fix the number of dictionary basis to $k = 128$ and the temporal pyramid equal to $\boldsymbol{\Lambda}_1 = [1, 1, 1]$, *i.e.* we pool sparse codes on whole the temporal click signal. A value of $\mu = \{3, 4\}$ seems to be a good choice for this pooling procedure. For $\mu \geq 20$, results are similar to those obtained by max-pooling. For azimuth, we observe also the same range of μ values.

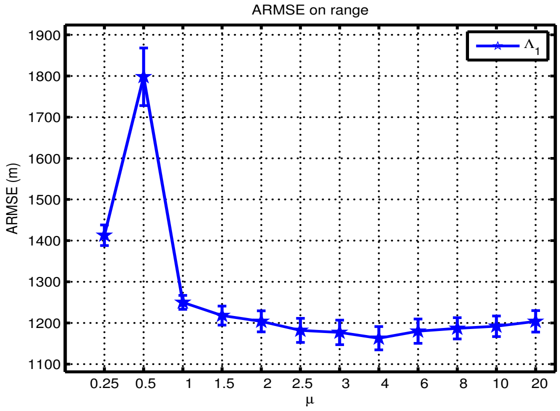


Figure 4. $\overline{ARMSE}$ vs. μ for range estimation.

4.3. Range and azimuth regression results

Here, we fixed the value of $\mu = 3$ and we varied the number of dictionary basis k from 128 to 4096 elements. We also investigated the influence of the temporal pyramid and we give results for two particulary choices: $\mathbf{\Lambda}_1 = [1, 1, 1]$ and $\mathbf{\Lambda}_2 = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{bmatrix}$. For $\mathbf{\Lambda}_2$, the sparse are first pooled over all the signal then pooled over 3 non-overlapping windows for a total of $1 + 3 = 4$ ROIs. In order to compare results of our presented method, we also give results for an hand-craft feature (Glotin et al., 2011) specialized for spermwhales and based on the spectrum of the most energetic pulse detected inside the click. This specialized feature, denoted *Spectrum feature*, is a 128 points vector.

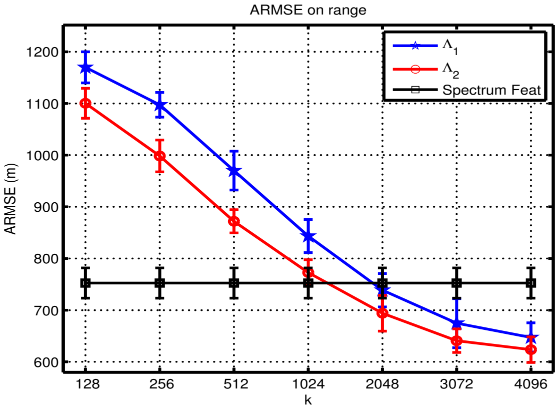


Figure 5. $\overline{ARMSE}$ vs. k for range estimation with $\mu = 3$.

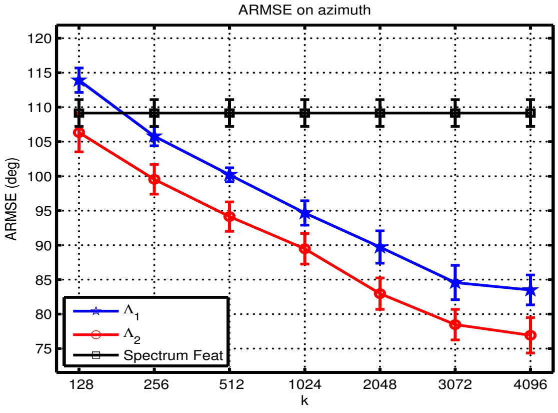


Figure 6. $\overline{ARMSE}$ vs. k for azimuth estimation with $\mu = 3$.

For both range and azimuth estimate, from $k = 2048$, our method outperforms results of the *Spectrum feature* and particulary for azimuth estimate. Using a temporal pyramid for pooling permits also to improve slightly results.

5. Conclusions and perspectives

We introduced in the paper, for spermwhale localization, a BoF approach *via* sparse coding delivering rough estimates of range and azimuth of the animal, specifically towarded for mono-hydrophone configuration. Our proposed method works directly on the click signal without any prior pulses detection/analysis while being robust to signal transformation issue by the propagation. Coupled with non-linear filtering such as particle filtering (Arulampalam et al., 2002), accurate animal position estimation could be perform even in mono-hydrophone configuration. Applications for anti-collision system and whale watching are targeted with this work.

As perspective, we plan to investigate other local features such as spectral features, MFCC (Davis & Mermelstein, 1980; Rabiner & Juang, 1993), Scattering transform features (Andén & Mallat). These latter can be considered as a hand-craft first layer of a deep learning architecture with 2 layers.

References

Andén, Joakim and Mallat, Stéphane. Multiscale scattering for audio classification. In *ISMIR, 11.*

Arulampalam, M. Sanjeev, Maskell, Simon, and Gordon, Neil. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. SP*, 50:174–188, 2002.

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

Boureau, Y-Lan, Ponce, Jean, and Lecun, Yann. A theoretical analysis of feature pooling in visual recognition. In *ICML' 10.*

Chen, Scott Shaobing, Donoho, David L., Michael, and Saunders, A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20: 33–61, 1998.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, 28:357–366, 1980.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

Feng, Jiashi, Ni, Bingbing, Tian, Qi, and Yan, Shuicheng. Geometric l_p-norm feature pooling for image classification. In *CVPR '11*.

Giraudet, Pascale and Glotin, Hervé. Real-time 3d tracking of whales by precise and echo-robust tdoas of clicks extracted from 5 bottom-mounted hydrophones records of the autec. *Applied Acoustics*, 67:1106–1117, 2006.

Glotin, H., Doh, Y., Abeille, R., and Monnin, A. 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*, 2011.

Lazebnik, Svetlana, Schmid, Cordelia, and Ponce, Jean. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06*, pp. 2169–2178.

Leroy, C. Sound attenuation between 200 and 10000 cps mesured along single paths. Technical Report 43, Saclant ASW Research Center, 1965.

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

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

Paris, Sébastien, Halkias, Xanadu, and Glotin, Hervé. Efficient bag of scenes analysis for image categorization. In *ICPRAM' 13*.

Rabiner, L. and Juang, B.H. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

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

Bioacoustical Periodic Pulse Train Signal Detection and Classification using Spectrogram Intensity Binarization and Energy Projection

Marian Popescu CP478@CORNELL.EDU
Peter J. Dugan PJD78@CORNELL.EDU
Mohammad Pourhomayoun MP749@CORNELL.EDU
Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850

Denise Risch DENISE.RISCH@NOAA.GOV
Northeast Fisheries Science Center, Woods Hole, MA, USA, 02543

Harold W. Lewis III HLEWIS@BINGHAMTON.EDU
Department of Systems Science and Industrial Engineering, Binghamton University, NY, USA, 13850

Christopher W. Clark CWC2@CORNELL.EDU
Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850

Abstract

The following work outlines an approach for automatic detection and recognition of periodic pulse train signals using a multi-stage process based on spectrogram edge detection, energy projection and classification. The method has been implemented to automatically detect and recognize pulse train songs of minke whales. While the long term goal of this work is to properly identify and detect minke songs from large multi-year datasets, this effort was developed using sounds off the coast of Massachusetts, in the Stellwagen Bank National Marine Sanctuary. The detection methodology is presented and evaluated on 232 continuous hours of acoustic recordings and a qualitative analysis of machine learning classifiers and their performance is described. The trained automatic detection and classification system is applied to 120 continuous hours, comprised of various challenges such as broadband and narrowband noises, low SNR, and other pulse train signatures. This automatic system achieves a TPR of 63% for FPR of 0.6% (or 2.2 FP/h), at a Precision (PPV) of 84% and an F1 score of 71%.

1. Introduction

Passive acoustic monitoring allows the exploration of marine mammal acoustic ecology at diverse temporal and spatial scales. While this technique is effective in understanding and characterizing habitats (Clark et al., 1996), it can often generate large acoustical data volumes. Furthermore, the acoustical signal domain presents various challenges such as: non-stationary and non-Gaussian noise, low signal to noise ratio (SNR), self-induced broadband and narrowband sensor noise, abiotic, environmental noise such a rain fall, ice and wind (Martin et al., 2012), and anthropogenic noise caused by vessels (Parks et al., 2009) or seismic airgun exploration activities (Guerra et al., 2011). Therefore, the current research is focused on creating efficient, robust automatic algorithms that can mine, identify, and classify marine mammal sounds across highly variable, large data sets.

Machine learning is an important step in the development of automatic acoustic species detection. Early automatic detection techniques used matched filters, hidden Markov model, and spectrogram cross-correlation (Clark et al. 1987). These methods were later improved through the use of machine learning approaches such as a feed-forward neural network classifier (Mellinger and Clark, 1993; Potter et al., 1994; Deecke et al., 1999; Mellinger, 2004; Mazhar et al., 2007; Pourhomayoun et al., 2013). Other machine learning algorithms, such as classification and regression tree classifiers (CART), have also been implemented in recognizing contact calls made from the

Work made possible by National Oceanic Partnership Program (NOPP), Office of Naval Research (ONR) - N000141210585, National Fish and Wildlife Foundation (NFWF) -0309.07.28515.

North Atlantic Right Whale (Dugan et al., 2010). Improvements over single recognition methods have been shown by using an advanced technique, which combines several recognition methods running in parallel (Dugan et al., 2010; Pourhomayoun et al., 2013).

In this paper we discuss an automated approach, for detecting and classifying periodic, broadband, pulsed signals using machine learning techniques. In particular, we will focus on the detection and classification of minke whale (*Balaenoptera acutorostrata)* songs, and the development of a system that can be applied to other datasets without re-training.

1.1 Minke whale (*Balaenoptera acutorostrata)*

The minke whale is a marine mammal species within the suborder of baleen whales and is found throughout the North Atlantic Ocean. Like all whales, minkes use sound to feed, breed, navigate and communicate (Richardson et al., 1995). Recent studies have shown that their perception of sound (Brkic et al., 2004) can be influenced by various environmental conditions such as wind and ice, but also anthropogenic noises (Martin et al., 2012). Therefore, quantifying large-scale biological phenomena such as seasonal occurrence and season distribution is critical for understanding the potential influences of natural and manmade factors on population dynamics. While various minke whale studies have been conducted (Schweder et al., 1997; Oswald et al., 2011), little information is available regarding the North Atlantic minke whale's seasonal distribution and occurrence off the U.S. East Coast. The methodology described here was developed to analyze large data sets collected by Cornell University using Marine Autonomous Recording Units (MARUs) during 2006-2010 (Calupca et al., 2000). The multi-channel data, continuously recorded at 2 kHz, was captured off the coast of Massachusetts, in the Stellwagen Bank National Marine Sanctuary (SBNMS). The algorithm was applied to 895 continuous days in order to analyze the seasonal distribution and occurrence of minke whales (Risch Calupca et al., 2000) in the SBNMS.

1.2 Signal characteristics and challenges

The minke whale vocalizations are characterized as pulse trains that can last somewhere between 40-60 sec, typically within the 100-1400 Hz frequency band. The pulse trains are comprised of individual pulses lasting 40-60 msec, and can exhibit variable pulse rates ranging from 2.8 pulses/sec to 4.5 pulses/sec (Mellinger et al., 2000). While our proposed methodology can be used for any pulse train series, here we focused on pulse trains contained within the 75-350 Hz frequency band, with variable length Inner Pulse Interval (IPI) described above. Figure 1 depicts the spectrogram of a minke whale pulse train song, as well as additional sources of noise and energy. The challenge is to detect and classify these pulse train signatures as they occur within a continuous stream of acoustic data.

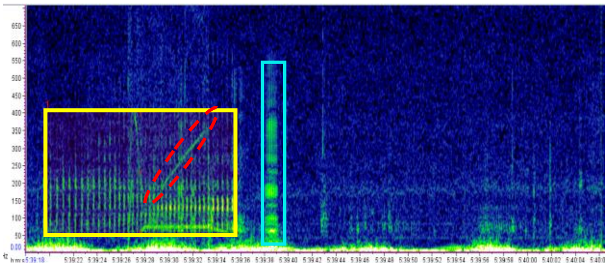
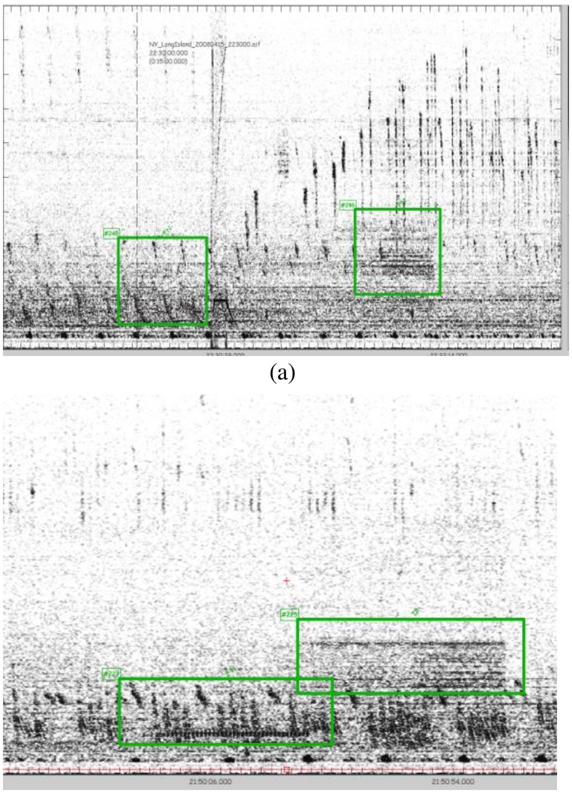


Figure 1. The spectrogram of a minke whale vocalization lasting ≈ 17 secs. The yellow box indicates the minke pulse train signature with the variable IPI. The noise generated by hard disk drive (red dotted ellipse) can be seen clearly within the minke pulse train. The spectrogram also reveals energy from an additional species known as Haddock (blue box), constant narrowband noises between 70-200 Hz, other sources of short impulse broadband and low-frequency noises. These noise characteristics change from sensor to sensor and sometimes on a minute by minute basis.

1.3 Train and Test Datasets

Since the signal of interest contains such broad variability, a training dataset was created in order to capture the parameter space. The dataset contains 2429 minke pulse trains from each of the 10 sensors. The minke pulse trains were identified, by an expert human biologist, by manually hand browsing randomly chosen subsets of the recordings. Additionally, a total of 2788 noise events that ranged from ambient noise, to shipping vessel noise, sensor hard-drive noise, and other cross species, was added. Overall, the train dataset consists of 112 continuous hours recording and is used in designing the detector and qualitatively analyzing the performance of various classifiers.

Furthermore, in order to analyze the performance of the trained system, a test dataset was created. The test dataset consists of 120 continuous hours, containing 729 total minke vocalizations. The dataset is constructed by using 3 days from Stellwagen Bank National Marine Sanctuary recording and 2 days from other external sensors from the Long Island, New York area. This will allow us to measure how well the methodology can be generalized using the trained model. The test dataset also contains various challenges, including very low SNR vocalizations and as well as additional species know has haddock which also has broadband pulse signals. Figure 2 presents some of the challenges in the test dataset.



(a)

(b)

Figure 2. The spectrogram of minke whale vocalizations in the test dataset: (a) low SNR minke vocalization in the left green box, and minke vocalization influenced by other species and broadband pulses in the right box. Other sources noise can be also observed. (b) minke vocalizations superimposed by pulse train signatures created by the haddock species.

2. Methods

Previous methods for detecting pulse type vocalizations are based on: (1) cross-correlation with a pre-designed kernel, or (2) auto-correlation of a given signal block (Mellinger and Clark, 1993). However, their performance is highly depended on choice of kernel and threshold. The implementation can also suffer from high computational complexity. The proposed methodology for automatic detecting and classifying of minke pulse trains in a continuous dataset consists of a two-stage approach. In the first stage, we try to identify the pulse train signatures based on a set of rules that match a description of the minke whale signal. In the second stage, we extract a set of features from the detected events, which will be later used to recognize the events using a previously trained classifier.

2.1 Stage I – Detecting pulse train signatures

The proposed detection stage consists of several steps. First, since the acoustical data are continuous, a sliding window of duration equal to 30 sec was applied to create the time–domain signal slices s(t). Secondly, since the signals of interest are located within the 75-300 Hz frequency band, s(t) is conditioned using a type II, Chebyshev bandpass FIR filter; with -30 dB attenuation, 40 Hz roll-off, and 0.1 dB of ripple in the passband. The filter is implemented in order to reduce the energy outside the desired frequency bands and to improve the intensity-based spectrogram binarization step. Next, a spectrogram is computed for the filtered s(t) signal using a Blackman window, 8% overlap, 512 point FFT, to yield 20.5 ms time and 3.89 Hz frequency bins. The spectrogram is then cropped to match the frequency band bounds of the bandpass filter. Once the spectrogram is obtained, a binarization based on image intensity is applied in order to denoise the signal and remove the ambient noise, and place the signal in the same basis across all the sensors. First, we convert the spectrogram matrix to a gray-scaled intensity image.

We then compute an intensity mask using:

$$\gamma = 1.75 * \sigma_s + \mu_s \qquad (1)$$

where μ_s is the mean intensity of the image and σ_s is the standard deviation of the zero-mean intensity image. The level was derived based on the idea that the signal is not wide-sense stationary, which implies a different mean for each signal slice s(t), and that any acoustical signatures above the mean ambient noise level is captured within the standard deviation. Applying the level masking produces a binarized image, in which all pixels of the gray-scaled image with luminance greater than the level γ have a value of 1 (white), and replaces all other pixels with the value 0 (black). Using the *N*x*M* binarized image matrix, an image energy project function, *P*(*n*) is created as:

$$P(n) = \sum_{m=1}^{M} BW\left(\mathrm{n,m}\right) \quad \text{for} \quad n = 1, 2, \dots, N \qquad (2)$$

This process will place emphasis on broadband signatures, since pulse spectrogram time slices will contain a large number of vertical pixels (i.e. energy). Next, we find the local maxima of the energy projection function and apply the following set of rules, which have been designed for the minke vocalization pulse train, but can be generalized to any other pulse train signature: (1) local maxima above a threshold; (2) minimum and maximum number of local maxima above the threshold; (3) a range for the local maxima spacing (based on IPI). Any events that meet these criteria are then identified as minke pulse trains and sent to the next stage for feature extraction and classification. Figure 3 illustrates the detection process.

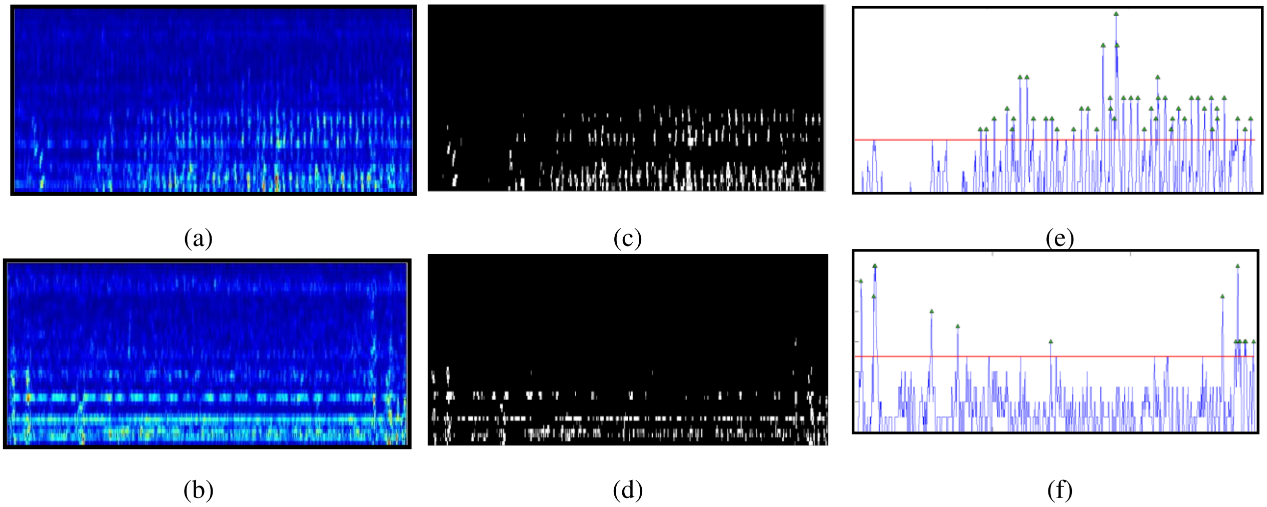


Figure 3. The detection process for a minke pulse train (top) and noise event (bottom),respectively; (a) and (b), spectrogram after bandpass filtering and cropping, respectively; (c), (d) the intensity-based binarization of the spectrograms, respectively. (e), (f) the energy projection function $P(n)$ with the same applied threshold.

2.2 Feature extraction

A set of 18 features is extracted for each detected event. The features are designed and chosen with the intent to distinguish the detected minke pulse trains from the ambient noise events (detector errors). A summary of the selected features is shown in Table 1.

Table 1. Features used to train and evaluated the classifiers.

Feature Number	Feature Name	Description (of Pulse Train)
F1	delta time	Duration of pulse train
F2-F3	frequency pair min-max	Frequency bounds
F4	number of clicks	Number of pulses
F5	average bandwidth	Average bandwidth of pulse train
F6	center frequency	Center bandwidth of the pulse train
F7	average sharpness	F4 / F1
F8	CEC for signal	LEQ of the detected pulses within the pulse train
F9	Mean Leq	Mean LEQ of the detected pulses
F10	DeltaT- mean	The mean of the IPI of detected clicks
F11	DeltaT- mode	The mode of IPI of detected clicks
F12	DeltaT- max	The max IPI of detected clicks
F13	DeltaT- min	The min IPI of detected clicks
F14	SNR	Signal to Noise Ratio of the detected pulse train
F15 -18	SNR: x^th^ percentile	SNR of pulse train using the 5th , 10th , 20th and 25th percentile of slice as noise

2.3 Classification

The detection method, discussed above, identifies areas of energy that meet the criterion presented in figure 3; we will refer to these as regions of interest (ROI's). Many of the ROI's which are recognized by the detection stage result from various noise conditions such as vessel noise, or additional marine mammal vocalizations, and thus a classification stage is implemented to increase the overall performance of the system. This stage is designed to reduce the false positive rate of the detector, since in bio-acoustical applications, the analysts have to manually verify the output results. In order to analyze the performance of various classifiers, a feature vector is extracted after applying the detection stage on the train data. Our analysis investigates the performance of the following classifiers: (1) grafted C4 tree with a confidence factor of 0.25 (Webb, 1999), (2) a Random Forest with 10 random trees in the forest and 5 features used in random selection (Breiman, 2001), (3) a Bayesian network via a Simple Estimator with alpha equal to 0.5 and K2 search algorithm (Cooper and Herskovits, 1992), a ripple-down rule learner with 3 fold used for pruning and 2 minimum weights of the instances in a rule (Gaines and Compton ,1992) and a functional tree that did not use binary split and used 15 instances for node splitting (Gama, 2004; . The methods are evaluated using at a 66%, 33% split on the training data. The performance of the classifiers is shown in Figure 4. It can be seen that the random forest classifier has the best area under the curve (AUC).

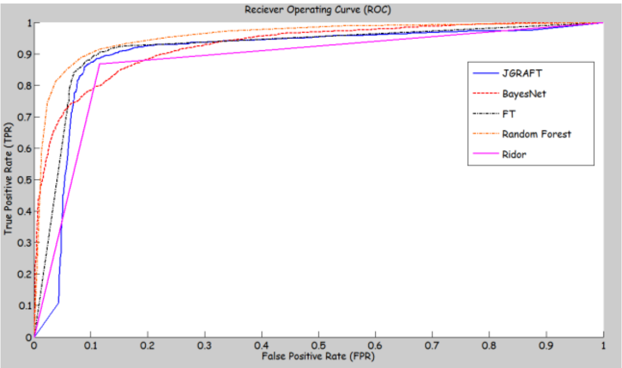


Figure 4. The performance of various classifiers across the training dataset, using a 66% split for training and a 33% split for test. The figure and area under the curve indicate that the random forest classifier is the best option for our given feature space.

3. Results and Conclusion

The proposed technique was applied on a test dataset using an energy projection function with threshold equal to 6. A total number of 28820 signal slices, of which 3158 were minke vocalizations, were analyzed by the detector. The detection stage produces a True Positive Rate (TPR) of 79%, a False Positive Rate (FPR) of 11% or 15.48 False Positives per hour (FP/h), at a Precision (PPV) of 40% and an F1 score of 53%. In order to reduce the number of false positives generated by the detector, a random forest classifier is applied on the testing dataset. The performance of the proposed classifier on the testing dataset is shown below in Table 2.

Table 2. The performance of the trained classifier on the challenge test data without further training.

TPR	FPR	Precision	F1	AUC	Class
94%	36%	84%	0.89	85%	Non-Minke
79%	6%	84%	0.72	85%	Minke

It can be seen that the performance of the classifier diminished when applied to the new testing dataset. This was due to the low SNR conditions, and other interfering broadband signatures that were being detected. If increased performance in true positive is required, the signal should either be further de-noised, additional features should be added to the training data, or the training vector size should be increased to include detection events from the test data. When the detector and trained classifier system is applied to the test data, it produced a TPR of 63% for FPR of 0.6% (2.2 FP/h), at a PPV of 84% and an F1 score of 72%. It should be noted that while the TRP went from 79% to 63%, the FPR went from 11% to 0.6%.

In this paper we have shown the design and implementation of an automatic detection and classification system, used to mine and identify minke whale pulse trains within a continuous stream of acoustic data. The results show that the proposed method can achieve high performance even in the presence of high noise conditions.

References

Breiman, L. Random Forests. *Machine Learning*, 45(1): 5-32, 2001.

Brkic, I., Jambrosic, K. and Ivancevic, B. Perception of sound by animals in the ocean. *Electronics in Marine,2004.Proceedings Elmar. 46th International Symposium*, 258-264, 2004.

Calupca, T.A., Fristrup, K.M., and Clark, C.W. A compact digital recording system for autonomous bioacoustic monitoring. *J. Acoust. Soc. Am.*, 108:2582(A), 2000.

Clark, C.W., Marler, P. and Beeman, K. Quantitative analysis of animal vocal phonology: an application to swamp sparrow song. *Ethology*. 76:101-115, 1987

Clark, C. W., Mitchell, S. G., and Charif, R. A. Distribution and behavior of the bowhead whale, *Balaena mysticetus*, based on preliminary analysis of acoustic data collected during the 1993 spring migration off Point Barrow, Alaska, Report, Intl. Whal. Commn. 46:541–554, 1996.

Cooper G., and Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9: 309-347, 1992.

Deecke, V.B., Ford, J.K.B. and Spong, P. Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (Orcinus orca) dialects. *J.Acoust.Soc.Am.,* 105(4): 2499-2507, 1999.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., and Clark, C. W. North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms. *IEEE Proceedings of the 2010 Long Island Systems*, Applications and Technology Conference, 1-6, Farmingdale, NY, 2010.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., and Clark, C. W. "North Atlantic right whale acoustic signal processing: Part II. Improved decision architecture for auto-detection using multi-classifier combination methodology," *IEEE Proceedings of the 2010 Long Island Systems*, Applications and Technology Conference, 1–6, Farmingdale, NY, 2010.

Gaines, B. R., and Compton, P. Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information System,* 5(3): 211-228, 1992.

Gama, J. Functional Trees, *Machine Learning*, 55(3): 219-250, 2004.

Guerra, M., Thode, A. M., Blackwell, S. B., and Macrander, A. M. Quantifying seismic survey reverberations off the Alaskan North Slope. *J. Acoust. Soc. Am.,* 130: 3046–3058, 2011.

Landwehr, N., Hall, M, and Eibe, F, Logistic Model Trees, *Machine Learning,* 59(1): 61-205, 2005.

Martin, B., Delarue, J., and Hannay, D. Soundscape of the North-Eastern Chukchi Sea. *J. Acoust. Soc. Am.,* 132(3):1948, 2012.

Mazhar, S., Ura, T. and Bahl, R. Vocalization based Individual Classification of Humpback Whales using Support Vector Machine. *IEEE OCEANS 2007*, 1-9. 2007

Mellinger, D. K. A comparison of methods for detecting right whale calls. *Canadian Acoustics*, 32:55–65. 2004.

Mellinger, D. K., and Clark, C. W. A method for filtering bioacoustics transients by spectrogram image convolution. *Proc. IEEE,* 3:122–127, 1993.

Mellinger, D. K., Carson, C., and Clark, C. W. Characteristics of minke whale (Balaenoptera acutorostrata) pulse trains recorded near Puerto Rico. *Marine Mammal Science* 16: 739–756, 2000.

Mellinger, D. K. and Clark, C. W. Methods for automatic detection of mysticete sounds. *Mar. Freshwater Behav. Physiol*. 29(3): 163–181, 1997.

Mellinger, D. K., and Clark, C. W. Recognizing transient low-frequency whale sounds by spectrogram correlation. *J. Acoust. Soc. Am.,* 107: 3518–3529, 2000.

Oswald, J.N., Au, W.W.L., and Duennebier F. Minke whale (Balaenoptera acutorostrata) boings detected at the Station ALOHA Cabled Observatory. *J.Acoust.Soc.Am.*, 129(5): 3353-3360, 2011.

Parks, S. E., Urazghildiiev, I., and Clark, C.W. Variability in ambient noise levels and call parameters of North Atlantic right whales in three habitat areas. *J. Acoust. Soc. Am.,* 125(2): 1230-1239, 2009.

Potter, J. R., Mellinger, D. K., and Clark, C. W. Marine mammal call discrimination using artificial neural networks. *J. Acoust. Soc. Am.,* 96: 1255–1262, 1994.

Pourhomayoun, M., Dugan P., Popescu M., and Clark C., Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network, *ICML 2013 Workshop on Machine Learning for Bioacoustics,* 2013 (*submitted for publication*).

Pourhomayoun, M., Dugan P., Popescu M., Risch D., Lewis H., and Clark C., Classification for Big Dataset of Bioacoustic Signals Based on Human Scoring System and Artificial Neural Network, *ICML 2013 Workshop on Machine Learning for Bioacoustics,* 2013 (*submitted for publication*).

Richardson, W. J., Greene, C. R., Jr., Malme, C. I., and Thomson, D. H. *Marine Mammals and Noise*, Academic Press, 1995.

Risch, D., Siebert, U., Dugan, P., Popescu, M. & Van Parijs, S.M. Acoustic ecology of minke whales in the Stellwagen Bank National Marine Sanctuary. *Marine Ecology Progress Series*, 2013 (submitted).

Schweder, T., Skaug, H.J., Dimakos, X., Langaas, M., and Øien, N. Abundance estimates for Northeastern Atlantic minke whales. Estimates for 1989 and 1995. Rep. Int. Whal. Comm. 47:453–484, 1997.

Webb, G. I. Decision tree grafting from the all-tests-but-one partition. *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 702–707, San Francisco, CA, 1999 Morgan Kaufmann

Classification for Big Dataset of Bioacoustic Signals Based on Human Scoring System and Artificial Neural Network

Mohammad Pourhomayoun MP749@CORNELL.EDU
Peter J. Dugan PJD78@CORNELL.EDU
Marian Popescu CP478@CORNELL.EDU
Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850

Denise Risch DENISE.RISCH@NOAA.GOV
Northeast Fisheries Science Center, Woods Hole, MA, USA, 02543

Harold W. Lewis III HLEWIS@BINGHAMTON.EDU
Department of Systems Science and Industrial Engineering, Binghamton University, NY, USA, 13850

Christopher W. Clark CWC2@CORNELL.EDU
Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850

Abstract

In this paper, we propose a method to improve sound classification performance by combining signal features, derived from the time-frequency spectrogram, with human perception. The method presented herein exploits an artificial neural network (ANN) and learns the signal features based on the human perception knowledge. The proposed method is applied to a large acoustic dataset containing 24 months of nearly continuous recordings. The results show a significant improvement in performance of the detection-classification system; yielding as much as 20% improvement in true positive rate for a given false positive rate.

1. Introduction and Background

Passive acoustic monitoring is one of the primary and popular methods used to help scientists investigate and understand animal behavioral patterns (Potter, Mellinger, & Clark, 1994). The acoustic modality is particularly appropriate for marine mammals, because all of those studied are known to produce sounds for foraging, navigation or communication. Furthermore, acoustic monitoring methods are not subject to visual sighting limitations imposed by weather, daylight and ocean environmental conditions (Norris, Oswald, & Sousa-Lima, 2010). In passive acoustic monitoring, fixed acoustic sensor systems record the underwater sounds. These systems collect huge amounts of acoustic data. Exploring the data can be done by the human. The inspection however, is inefficient and slow. Instead, advanced computer algorithms have been designed to identify various animal sounds which tend to augment the humans ability, making this process more efficient for data analysis.

For decades, scientists have been actively recording and archiving marine bioacoustic data, and have devoted significant effort at designing automated algorithms for processing these data for sounds of interest. Processing the data poses many challenges including highly variable ambient noise conditions and a host of biological and anthropogenic sound sources.

The process of bioacoustic signal identification usually includes three main stages; signal detection, feature extraction, and classification. One of the most widely used detectors for acoustic signals is the Energy Detector (Ichikawa et al., 2006; Jarvis, DiMarzio, Morrissey, & Morretti, 2006; Ura et al., 2004). However, it suffers greatly when the signal to noise ratio (SNR) is low. One solution to address this problem is to denoise the incoming data before the detection step (Datta & Sturtivant, 2002; Gillespie et al., 2008; Ichikawa et al., 2006; Niezrecki, Phillips, Meyer, & Beusse, 2003; Popescu et al. 2013). Adjusting thresholds remains a common issue when balancing between false positive (false alarm) and false negative (missed detection) errors. For example, with endangered animals, populations are typically very low and the detection requirement aims to minimize the false negative rate. For animals that are relatively abundant, reducing the false positive rate may be a more optimal detection requirement (Dugan, Rice,

Work made possible by National Oceanic Partnership Program (NOPP), Office of Naval Research (ONR) - N000141210585, National Fish and Wildlife Foundation (NFWF) -0309.07.28515.

Urazghildiiev, & Clark, 2010). The detection threshold offers the ability to vary these conditions.

The second stage in the process of bioacoustic signal identification is feature extraction from the detected sound events. These features can then be used as input for the classifier. Classification is typically the final stage for a typical computer algorithm. Since classification is highly dependent on prior information, provided by detection and feature extraction, it is often the most critical step for marine mammal identification. Similar sounds produced by various species, overlapping vocalization and interfering noise are common reasons that make marine mammal acoustic classification difficult. However, several effective and promising methods are being developed and used by researcher in addressing this issue (Afifi & Clark, 1996; L. A. Clark & Pregibon, 1991; Deecke, Ford, & Spong, 1999; Mazhar, Ura, & Bahl, 2007; Mellinger, 2004, Mellinger, D.K; Murray, Mercado, & Roitblat, 1997; Norris et al., 2010; Dugan, Rice, Urazghildiiev, & Clark, 2010). Fig. 1 shows a sample time-frequency spectrogram including four different types of acoustic events (i.e. objects) overlapping with each other in a short period of time.

Experiments show that for small datasets, researchers have been successful at finding combinations of detector-classifier pairs that produce satisfactory results. However, when acoustically sampling over long time periods (months to years) and geographical areas (10^5 km^2), artifacts in the sound environment tend to color the spectrum and make automatic recognition less successful (especially increases in rates of false positive errors) (Clark et al., 2010; Dugan et al., 2010).

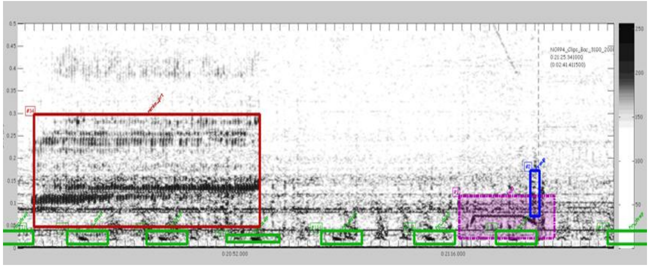


Figure 1. Example spectrogram including a minke whale pulse train song (red), fin whale song notes (green), a right whale up-call (blue) and hard disk drive noise (pink).

In this paper, we propose a technique for improving the algorithm recognition results for use with large datasets. The technique herein is particularly useful when human operators want to visualize seasonal activity or diel patterns over long time periods. This approach uses human knowledge along with information from the recognition system, such as signal features or recognition parameters, to improve the overall performance of the detection-classification system. The proposed method is used as a post-processing stage for existing algorithms, and works by classifying based on human knowledge along with recognition parameters. For this work, classification is done through the use of the Artificial Neural Network (ANN). Combining human perception along with recognition parameters improves the performance for seasonal activity of a marine mammal species commonly known as the minke whale (*Balaenoptera Acutorostrata*). We refer to our algorithm as the human knowledge-ANN, or HK-ANN. Since minke whale populations are broadly distributed throughout an ocean basin, for this work we are not interested in finding every call. Instead, producing a quantitatively based, geographic map of seasonal occurrence and distribution with minimal human involvement is the primary biological goal. For this work, we will show how the HK-ANN is first trained and tested on a smaller dataset using labeled events and quality scores. By allowing the neural net to utilize the human knowledge during the training, the HK-ANN is then applied to 24 months of nearly continuous sounds (from 2008 to 2010) recorded in the Stellwagen Bank National Marine Sanctuary (SBNMS), Massachusetts Bay, United States (Morano et al., 2010). The results show a significant improvement in performance of the detection-classification system in identifying minke whale sounds. The outcome also shows a remarkable improvement in eliminating the false alarm errors during the 24-month period by using the proposed method. Results are described by comparing the diel activity pattern for the 24-month case. We will show that the HK-ANN achieves a true temporal/seasonal minke whale song distribution pattern, with minimal false positive errors.

2. Bioacoustic Signal Processing and Classification

2.1 ANN Classifier for Bioacoustic Signals

The Artificial Neural Network (ANN) is an effective method for acoustic signal classification (Mellinger, 2004; Murray et al., 1997; Potter et al., 1994) (Deecke et al., 1999; Dugan et al., 2010). There are several reasons that make ANN a promising method. First, ANN is a non-linear estimator. Thus, it can be well-suited for noisy inputs with arbitrary distributions, especially when the interfering noise is not statistically independent of the desired signal (Potter et al., 1994). Second, ANN is an adaptive classifier. Feed-forward ANN can be trained by interactive methods that adjust the weighing matrix to minimize the cost function and to guarantee achieving an (at least a local) optimal weighing network model (Potter et al., 1994). Moreover, ANN can take metrics from a wide range of acoustic representations (e.g. spectrogram, waveform, frequency contour) as input. This flexibility helps in a variety of applications and supports designing different classifier topologies for several different signal types. For example, Deecke *et al.* (1999) used a standard back-propagation trained ANN to classify killer whale dialects to nine different categories by using the extracted pulse-rate contours of killer whale signals as input to an ANN. Potter *et al.* (1994) used a feed-forward ANN to distinguish bowhead whale song endnotes from interfering noises, and they used the signal spectrogram as the input of an ANN.

There are several types of ANNs that can be applied for the purpose of bioacoustic signal classification. Feed-forward networks are more preferred for our purpose because they have less complexity and a relatively lower number of neurons and connections when compared to other networks (Potter et al., 1994). However, feed-forward networks usually need a larger training set for the learning phase (Potter et al., 1994). In our case, this is not a problem because we have a large amount of data available for training. Thus, we preferred to choose a standard feed-forward, back-propagation trained network for our marine mammal sound post-classification task. A feed-forward neural network usually includes at least one hidden layer. The output of each hidden layer is the non-linear function of the linear combination of its input data coming from the previous layer. The coefficients in each combination (called weights) are adaptive parameters adjusted during the training step. The i^{th} output of the first hidden layer is calculated as the sum of weights and a bias term,

$$y_i^{(1)} = f(\sum_{n=1}^{N} w_{in}^{(1)} x_n + b_i^{(1)}) \tag{1}$$

where $f(.)$ is the non-linear function, x_n is the n^{th} input element, $w_{in}^{(1)}$ is the weight element of the first layer weight matrix and $b_i^{(1)}$ is the bias. Similarly, for the second layer we have,

$$y_i^{(2)} = g\left(\sum_{m=1}^{M} w_{im}^{(2)} f(\sum_{n=1}^{N} w_{mn}^{(1)} x_n + b_m^{(1)}) + b_i^{(2)} \right) \tag{2}$$

where $g(.)$ is the nonlinear function again, and $w_{in}^{(2)}$ is the weight element of the second layer. The network may be trained in a variety of ways including simple gradient descent, where we represent the collection of synaptic weights as the vector $\vec{w}$, the gradient of the error function as $\vec{g}$, and a learning rate ρ, iteratively calculating $\vec{w} = \vec{w} - \rho \cdot \vec{g}$ until an appropriate level of convergence is attained.

2.2 ANN Post Classification Based On Human Scoring

Minke whales are known to sing, and singing occurs seasonally in different regions of an ocean. Minke songs consist of 40-60 sec sequences of short duration (40-60 msec), broadband (ca. 100-1400 Hz) pulses, referred to as a pulse train (as shown in Figure 2-(a)). To achieve a higher performance in identifying the minke whale calls in a large dataset, we designed and implemented a post-classifier process based on human expertise, and applied it on the output of an existing detection-classification system. Note that the goal of this paper is not to compare the performance of ANN-based classification to other types of classifiers. Instead, we consider the existing detection-classification system as a black box, and we aim to improve its overall performance using human-knowledge post-processing approach.

In the detection stage, we used a simple energy approach (Datta & Sturtivant, 2002; Gillespie et al., 2008; Ichikawa et al., 2006; Jarvis et al., 2006; Niezrecki et al., 2003; Ura et al., 2004) in addition to color compression and image processing methods (Witten, 2011) to detect acoustic events in the time-frequency domain. Each event can be either a minke pulse train or other sounds, which include ambient noise, anthropogenic noise, or the acoustics of other marine mammals.

Afterwards, a feature set (feature vector) for each event was extracted from the original time-domain signal as well as the signal spectrogram. We used the signal feature vector as the ANN's input and consider a score assigned to each event as the ANN's expected output. The input feature vector includes 18 features such as event duration, event minimum and maximum frequencies, number of pulses in the pulse train, average bandwidth, center frequency, equivalent continuous sound pressure level (L_{eq}), mean, mode, maximum and minimum of the pulse duration and pulse intervals, as well as SNR with respect to 5th, 10th, 20th and 25th percentile of the signal.

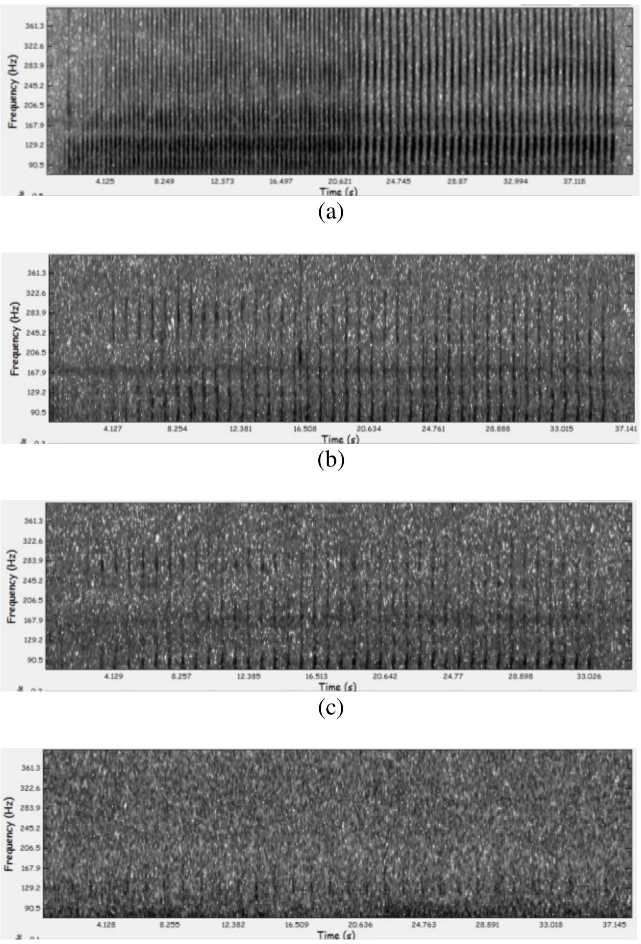
For this work, we used a heuristic approach to derive the training parameters. We used a training dataset containing 2625 events. Note that all of these events have been already identified as minke pulse trains (i.e. songs) by the existing detection-classification system; however, they include a large amount of false positive errors. The main goal of using the post-classifier is to improve the performance by eliminating these false positives by exploiting a combination of human expertise and machine-learning techniques.

After selecting the training set, expert biologists assigned a score to each one of the 2625 events. Scores were assigned by evaluating the spectrogram of the sound signal. The scores varied from 0 to 4 and were defined as following; 0: *Not target species,* 1: *Unsure of target species,* 2: *Faint target species,* 3: *Mediocre target species,* 4: *Strong target species*. According to the scores assigned by the human expert, only 981 of the events were likely to be minke pulse trains with scores of greater than or equal to 3. Figure 2 shows the spectrogram of four sample events scored from 1 to 4 by an expert biologist. Quality scores based on human intuition were added to the standard training set as shown in (3),

$$TV_{i,j}^{HK} = \left\langle FV_{i,j=1\cdots 18}, S_{k\in[0,4]} \right\rangle \tag{3}$$

where $FV_{i,j}$ is the j^{th} feature, ranging from 1 to 18 for the i^{th} object in the training set. The output class is given by scores $S_{k\in[0,4]}$ as mentioned above.

It was discovered that an acceptable convergence was obtained by using three hidden layers. Hidden layers used a *sigmoid* activation function, and the output layer applied *softmax* activation normalized to the interval shown in equation (3). The ANN was initialized with random weights, and training was accomplished by correcting the weights iteratively using backpropagation rule of steepest descents and minimizing the mean squared error. After around two hundred training iterations, the mean square error stabilized to less than 0.01, which was acceptable for our purpose.



(a)

(b)

(c)

(d)

Figure 2. Spectrogram of the 4 sample events scored by human expert. (a) score = 4, (b) score = 3, (c) score = 2, (d) score = 1.

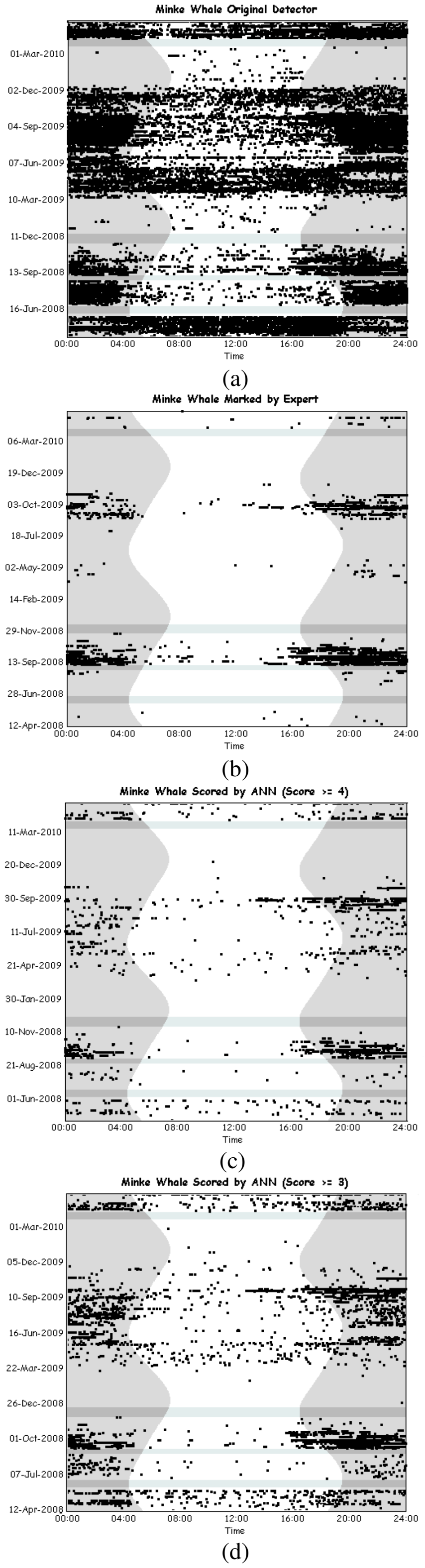
3. Results

The designed HK-ANN was tested on a big dataset including 24 months of nearly continuous data recorded in the Stellwagen Bank National Marine Sanctuary (SBNMS), Massachusetts Bay, United States. The testing dataset contains 41560 sound events detected as minke pulse trains by the primary detector. To evaluate the performance of the proposed method, we asked the human expert to mark the minke pulse trains in the entire dataset. We then compared this marked dataset against HK-ANN classifier output using a temporal/seasonal pattern (diel pattern).

The diel pattern shows the distribution of the bioacoustic events in a date-time plane, as shown in Figure 3. The shaded area represents nighttime, while the white area shows day time (the length of day time changes along the calendar). Horizontal greyish strips illustrate periods when the recording sensors were being recovered and redeployed , so no data were available for those time periods. Figure 3-(a) shows the classification results of the primary *decision tree classifier* using the same 18-feature set. However, given our prior knowledge on minke whale seasonal distribution in the sampling area, we expected a large portion of these primary detections to be false alarms. Figure 3-(b) demonstrates the true results identified by the human expert. Comparing figures (a) and (b), we can observe a significant rate of false positives in the classification stage. Figure 4-(c) shows the proposed HK-ANN classification results when we consider the output events, with a score equal to 4, as minke vocalization. Comparing this figure to traditional decision tree classifier (figure (a)) and the truth set (figure (b)) shows a big improvement in the identification of minke vocalization and a reduction in false positives. Figure 4-(d) represents the proposed HK-ANN classification results when we consider the output events, with a minke vocalization score greater than or equal to 3.

Biologists are usually interested in investigating the behavior of the animals during a specific time period. Marine mammals typically exhibit a seasonal pattern, and thus one of the most important parameter showing the performance of a marine mammal detection/classification method is to achieve the correct temporal/seasonal animal distribution. As we see in figure 3, the proposed method is able to eliminate a big portion of the false alarms and achieve a fairly accurate animal temporal/seasonal distribution pattern.

Figure 4 shows the Receiver Operating Curve (ROC) performance of the various common classifiers including Bayesian Network Classifier (Duda, Hart, & Stork, 2001; Witten, 2011), Grafted Decision Tree Classifier (Webb, 1999), and Classification/Regression Tree Classifier (Breiman, Friedman, Olshen, & Stone, 1984), on a dataset containing 4474 highly noisy sound samples. The blue curve shows the performance of the system after applying the proposed post-classifier HK-ANN on the output of above classifiers for the same dataset. We can see a significant improvement in overall performance (especially at low False Positive Rates), by using the proposed method. For example, at a FPR of 6% we have an improvement in TPR of approximately 20%. This improvement can help biologists make better informed and more accurate decisions about marine mammal seasonal occurrences and distributions.



(a)

(b)

(c)

(d)

Figure 3. Date versus time diel patterns for test dataset. (a): Original detection/classification by existing decision tree classifier. (b): True detections by human expert. (c): Detection by ANN with score=4. (d): Detection by ANN with score > 3.

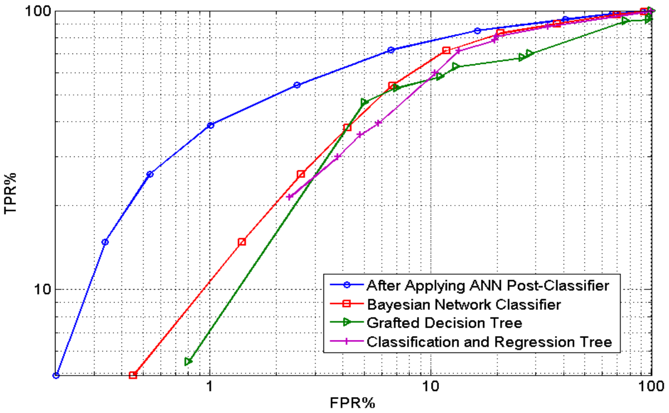


Figure 4. ROC of various common bioacoustic signal classifiers, and the effect of applying HK-ANN (the ANN Post-Classifier).

4. Discussion and Conclusion

The work herein considered a novel method, combining human intuition with an ANN classification stage for processing large amounts of passive acoustic data.

Since marine mammals typically exhibit seasonal patterns of occurrence and distribution, the automated algorithms are also expected to provide trends, as shown by a diel plot graphic. However, errors due to ambient noise and other conflicting acoustics events can pose significant challenges for automated algorithms, especially with larger datasets. Often times, developers do not have access to large amounts of data when developing recognition tools. Furthermore, background noise and other conditions can color recordings, offering bias and making pre-trained recognition algorithms prone to high error rates when running on large scale datasets. Results show that in highly noisy environments, training a basic classifier using a fixed feature set was not sufficient for building an effective automated classification stage for studying seasonal patterns for minke song activity. For this situation, excessive numbers of false positives destroys the basic seasonal migration pattern in the diel graph.

The proposed approach, referred to as the HK-ANN, was used to augment an ANN with a post-classifier stage by incorporating human knowledge. Using human scoring measure, along with the same feature set, provided a significant improvement in the automatic detection and classification of the signals of interest. Based on the results for the seasonal patterns, the post processing stage properly recognized minke whale songs and provided a seasonal pattern as shown in the diel plots.

References

Afifi, A. A., & Clark , V., *Computer-aided multivariate analysis, fourth edition* Chapman & Hall/CRC, 1996.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J., Classification and regression trees, 1984.

Clark , L. A., & Pregibon, D. Tree-based models. In J. M. Chambers, & T. J. Hastie (Eds.), (pp. 377), 1991.

Clark, C. K., Ellison, W. T., Hatch, L. T., Merrick, R. L., Van Parijs, S. M., & Wiley, D. N. *An ocean observing system for Large-scale monitoring and mapping of noise throughout the stellwagen Bank National Marine Sanctuary,* 2010.

Datta, S., & Sturtivant, C. Dolphin whistle classification for determining group identities *Signal Processing, 82*(2), 251- 258. doi:10.1016/S0165-1684(01)00184-0, 2002.

Deecke, V. B., Ford, J. K. B., & Spong, P. Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (orcinus orca) dialects *The Journal of the Acoustical Society of America, 105*(4), 2499. doi:10.1121/1.426853, 1999.

Duda, R. O., Hart, P. E., & Stork, D. G. (Eds.). *Pattern classification*, 2001.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., & Clark, C. W. North atlantic right whale acoustic signal processing: Part II. improved decision architecture for auto-detection using multi-classifier combination methodology. [*Applications and Technology Conference (IEEE-LISAT), Long Island Systems*, 2010.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., & Clark, C. W. North atlantic right whale acoustic signal processing: Part I. comparison of machine learning recognition algorithms. *Applications and Technology Conference (IEEE -LISAT), Long Island Systems,* 2010.

Gillespie, D., Gordon, J., Caillat, M., Claridge, D., Moretti, D., & Boyd, I. Detection of beaked whales using near surface towed hydrophones: Prospects for survey and mitigation *The Journal of the Acoustical Society of America, 123*(5), 3774. doi:10.1121/1.2935394, 2008.

Ichikawa, K., Tsutsumi, C., Arai, N., Akamatsu, T., Shinke, T., Hara, T., & Adulyanukosol, K. Dugong (dugong dugon) vocalization patterns recorded by automatic underwater sound monitoring systems *The Journal of the Acoustical Society of America, 119*(6), 3726. doi:10.1121/1.2201468, 2006.

Jarvis, S., DiMarzio, N., Morrissey, R., & Morretti, D. Automated classification of beaked whales and other small odontocetes in the tongue of the ocean, bahamas. *OCEANS 2006,* 1-6, 2006.

Mazhar, S., Ura, T., & Bahl, R. Vocalization based individual classification of humpback whales using support vector machine. *OCEANS 2007,* 1-9, 2007.

Mellinger, D. K. (Mellinger, D.K). A comparison of methods for detecting right whale calls. *Can. Acoust, 32,* 2004.

Morano, J. L., Rice, A. N., Tielens, J. T., Estabrook, B. J., Murray, A., Roberts, B., and Clark, C. W. Year-Round Presence of Right Whales in Massachusetts Bay: Acoustically Monitoring for a Critically Endangered Marine Mammal in a Highly Urbanized Migration Corridor. *Con. Bio*. 26:698-707, 2012.

Murray, S. O., Mercado, E., & Roitblat, H. L. The neutral network classification of false killer whale (pseudorca crassidens) vocalizations. *The Journal of the Acoustical Society of America, 102*(5), 3122. doi:10.1121/1.420590, 1997.

Niezrecki, C., Phillips, R., Meyer, M., & Beusse, D. O. Acoustic detection of manatee vocalizations *The Journal of the Acoustical Society of America, 114*(3), 1640. doi:10.1121/1.1598196, 2003.

Norris, T., Oswald, J., & Sousa-Lima, R. *A review and inventory of fixed installation passive acoustic monitoring methods and technologies*, 2010.

Popescu M., Dugan P., Pourhomayoun M., Risch D., Lewis H., Clark C., Bioacoustical Periodic Pulse Train Signal Detection and Classification using Spectrogram Intensity Binarization and Energy Projection, *ICML 2013 Workshop on Machine Learning for Bioacoustics,* 2013 (*submitted for publication*).

Potter, J. R., Mellinger, D. K., & Clark, C. W. Marine mammal call discrimination using artificial neural networks *The Journal of the Acoustical Society of America, 96*(3), 1255. doi:10.1121/1.410274, 1994.

Ura, T., Bahl, R., Sakata, M., Kojima, J., Fukuchi, T., Ura, J., Yanagisawa, M. (2004). Acoustic tracking of sperm whales using two sets of hydrophone array. *Underwater Technology, 2004. UT '04. 2004 International Symposium on,* 103-107.

Webb, G. Decision tree grafting from the all-tests-but-one partition, 1999.

Witten, I. H. (Ed.). *Data mining: Practical machine learning tools and techniques* (3rd ed.), 2011.

Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network

Mohammad Pourhomayoun MP749@CORNELL.EDU
Peter J. Dugan PJD78@CORNELL.EDU
Marian Popescu CP478@CORNELL.EDU
Christopher W. Clark CWC2@CORNELL.EDU
Bioacoustics Research Program (BRP), Cornell University, Ithaca, NY, USA, 14850

Abstract

In this paper, we develop a novel method based on machine-learning and image processing to identify North Atlantic right whale (NARW) up-calls in the presence of high levels of ambient and interfering noise. We apply a continuous region algorithm on the spectrogram to extract the regions of interest, and then use grid masking techniques to generate a small feature set that is then used in an artificial neural network classifier to identify the NARW up-calls. It is shown that the proposed technique is effective in detecting and capturing even very faint up-calls, in the presence of ambient and interfering noises. The method is evaluated on a dataset recorded in Massachusetts Bay, United States. The dataset includes 20000 sound clips for training, and 10000 sound clips for testing. The results show that the proposed technique can achieve an error rate of less than FPR = 4.5% for a 90% true positive rate.

1. Introduction and Background

Bioacoustic signal detection and classification is one of the most common and effective techniques used by scientists to explore marine bioacoustics and understand marine mammal behavioral patterns. For passive marine acoustic research, hydrophone sensor systems collect huge amounts of underwater sound data, thereby placing a premium on automated computer algorithms, including machine-learning methods, for detecting and classifying sounds of interest in the data (Sousa-Lima et al. 2013).

Right whales produce frequency-modulated upsweeps, referred to as up-calls, for long-range communication in the 50-250Hz frequency band (Clark 1982), and detection of up-calls has been shown to be the most effective mechanisms for determining whale presence in critical habitats (Clark et al. 2010). In this paper, we develop novel methods based on image processing and machine-learning to detect North Atlantic Right Whales (NARW) up-calls in the presence of high environmental noise, using a fairly small feature set (5, 15 or 20 features). The NARW is one of the world's most highly endangered whales (Clapham *et al.* 1999). Therefore, there is an urgent need to develop efficient techniques to detect the presence of NARWs so as to determine their seasonal occurrences and protect them from possible harm (Kraus et al. 2005).

For decades, researchers have been working to design effective automated algorithms for identifying marine mammal vocalizations including NARW up-calls (Mellinger 2004; Mohammad *et al.* 2011). Mellinger compared different methods for up-call detection, including spectrogram correlation and an artificial neural network. He evaluated the spectrogram correlation method for two different cases; manually selected parameters, and parameters selected based on an optimization procedure. In another approach, he used spectrogram frames, consisting of 252 cells, as inputs to a feed-forward neural network with 10 hidden layers. He used standard gradient-descent back-propagation with 5000 epochs to train the neural network (Mellinger 2004). However, the size of the selected neural network and the large number of inputs lead to extremely high computational complexity and a long training time. Dugan *et al.* also developed two new approaches for NARW sound identification based on artificial neural networks and decision tree classifiers, and compared their performance to a multi-stage feature vector testing (FVT) method (Dugan *et al.* 2010).

Gillespie applied an edge detection algorithm on the smoothed spectrogram to determine the boundary of the sound. He then extracted features, including duration, bandwidth and details of the frequency contour, which were used in an up-call classification stage (Gillespie 2004). Sánchez-García *et al.* also used a spectrogram region-based segmentation technique to identify the sound

Work made possible by National Oceanic Partnership Program (NOPP), Office of Naval Research (ONR) - N000141210585, National Fish and Wildlife Foundation (NFWF) -0309.07.28515.

signal, and then extracted the mean values of a fixed number of radial basis function (RBF) coefficients. These coefficients were later used to classify the sound signals (Sánchez-García *et al.* 2009). Mohammad *et al.* (2011) developed a region-based active contour model and support vector machine classifier to identify the NARW up-call in shallow water.

It is important to note that in environments with high ambient noise levels and various amounts of acoustic clutter, including sounds from other species, NARW up-calls can be extremely difficult to detect. Thus, regular region growing techniques, or methods based on using the maximum spectrogram value as the initial point of contour segmentation usually fail to find the up-call in cases of low SNR or high clutter.

In this paper, we develop a new method, based on continuous region processing and grid masking methods, to detect NARW up-calls in the presence of ambient noise, interfering noise or other non-NARW sounds. We apply a continuous region algorithm on the de-noised and normalized spectrogram to extract continuous regions of interest that might represent portions of an up-call. Then, we use grid masking techniques to generate two sets of features that are used as inputs to an artificial neural network classifier to identify the NARW up-calls.

2. Methods

In this section, we describe the details of the proposed method for up-call identification. Figure 1 shows the block diagram and different steps of the proposed approach.

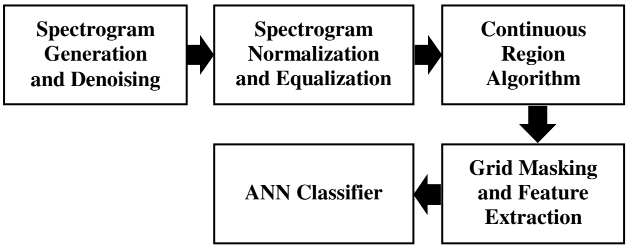

Spectrogram Generation and Denoising
Spectrogram Normalization and Equalization
Continuous Region Algorithm
Grid Masking and Feature Extraction
ANN Classifier

Figure 1. Block diagram of the proposed method.

2.1 Spectrogram Normalization and Equalization

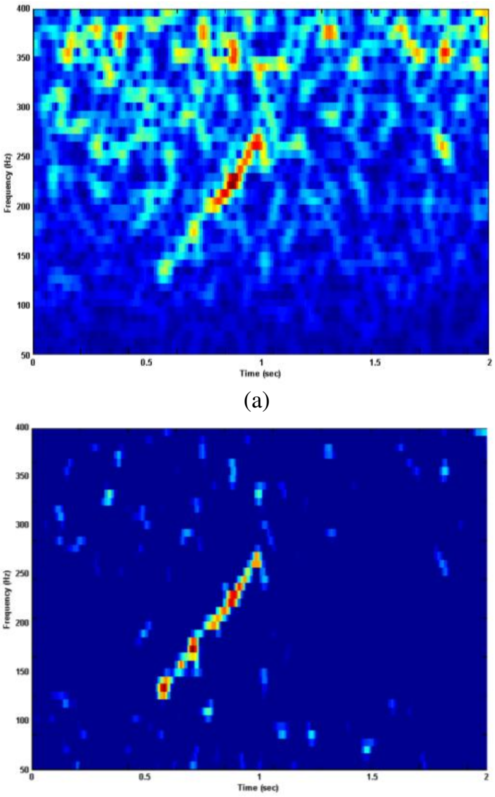
The sound signals, sampled at 2 kHz, are clipped into 2 sec slices, which are used to generate time-frequency spectrograms. To produce the spectrograms, we apply a STFT with window size of 128 ms (Hann window, 256 samples, 50% overlap).

After producing a spectrogram, we apply a two-dimensional wiener filter in order to denoise and smooth the spectrogram, using a 5x5 window around each pixel to estimate the local variance (Lim 1990). For each frequency band, we zero-mean the denoised spectrogram to remove the effects of constant narrowband noise, such as ship tonals, wind noise or electrical device noise, and to emphasize short-duration FM sounds such as NARW up-calls (Mellinger 2004).

The next step is hard-limiting the upper and lower bounds of spectrogram amplitudes to remove the influence of extreme values (Mellinger 2004):

$$\hat{S}(t,f) = \max(S_{floor}, \min(S_{ceiling}, S_N(t,f))) - S_{floor} \qquad (1)$$

where $S_N(t,f)$ is the normalized spectrogram, and S_{floor} and $S_{ceiling}$ are the desired lower and upper bounds on spectrogram values. Figure 2 shows the effect of denoising, normalization and equalization on a sample spectrogram.



(a)

(b)

Figure 2. Spectrogram examples with and without denoising, normalization and equalization: (a) original spectrogram. (b) denoised, normalized and equalized spectrogram.

2.2 Continuous Region Processing

After the denoising, normalization and equalization steps, we convert the spectrogram into a binary image for the purpose of continuous region processing. Note that this binary image will only be used to find regions in the spectrogram that are considered part of an up-call; however, the output of the algorithm will be a non-binary spectrogram including the regions of interest. Since some of the up-calls are extremely faint, we set the threshold value very low (e.g 10% of the image mean) to reduce the chances that we would miss an object of interest.

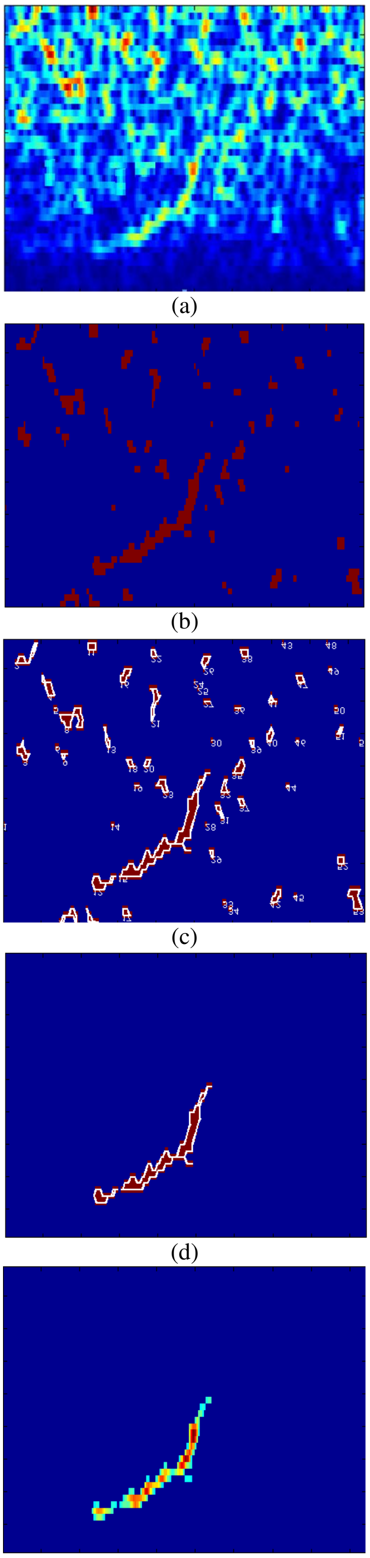
We use the Moore-Neighbor tracing algorithm modified by Jacob's stopping criteria (Gonzalez, 2004) to determine objects (i.e. continuous regions) in the image. After that, we extract the properties of each object and compare them to a set of thresholds to find an up-call. As mentioned before, under conditions of low SNR, NARW up-calls can be completely buried in noise, and typical region-growing techniques fail to find the up-call. Furthermore, as shown in Figure 5, sometimes up-calls are extremely faint compared to other objects in the spectrogram (e.g. interfering noise or sounds of other species). In such situations, methods that use the maximum spectrogram value to identify an initial point contour segmentation are not able to detect and classify the up-call.

Table 1 shows the continuous region parameters and thresholds used for detecting up-call segments in spectrograms given the specifications mentioned in 2.1. Figures 3 and 4 illustrate the continuous region algorithm process, and the elimination of noise and other possible sound objects in the frame that do not meet the NARW up-call criteria. The spectrogram in Figure 3 (top panel) includes high ambient noise, while the spectrogram in Figure 4 (top panel) contains other sound objects in the same frequency band.

Figure 5 demonstrates how the proposed technique performs under very challenging conditions when the NARW up-call is very faint and SNR is very poor.

Table1. The continuous region parameters and the thresholds used for detecting and north right whale up-call piece.

Parameter	Threshold
Minimum Perimeter	15 pixel
Minimum Area	15 pixel
Minimum Height (frequency band)	14 Hz
Maximum Height (frequency band)	250 Hz
Minimum Width (duration)	0.1 sec
Maximum Width (duration)	2 sec
Minimum orientation of the surrounding Ellipse	1°
Maximum orientation of the surrounding Ellipse	88°
Minimum Height/Width ratio	0.05
Maximum Height/Width ratio	3
Minimum Frequency	50 Hz
Maximum Frequency	400 Hz
Maximum surrounding Ellipse axes ratio	3.5


(a)

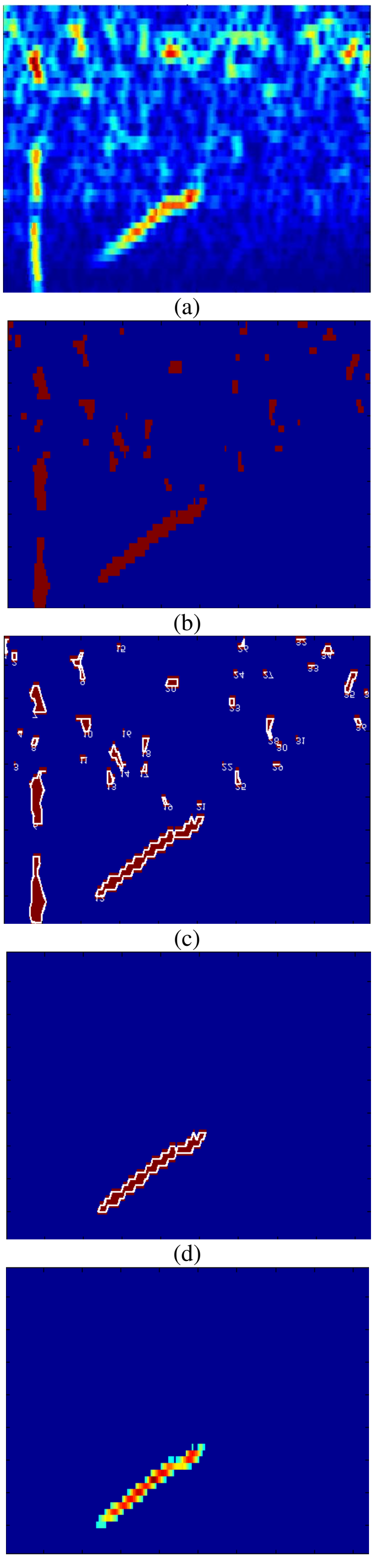
(b)

(c)

(d)

(e)

Figure 3. Continuous Region Processing: (a) original Spectrogram; (b) spectrogram after denoising, normalization, equalization and binarization; (c) continuous region detection; (d) detected region of interest; and (e) the algorithm's output.


(a)

(b)

(c)

(d)

(e)

Figure 4. Continuous Region Processing: (a) original Spectrogram; (b) Spectrogram after denoising, normalization, equalization and binarization; (c) continuous region detection, (d) detected region of interest; and (e) the algorithm output.

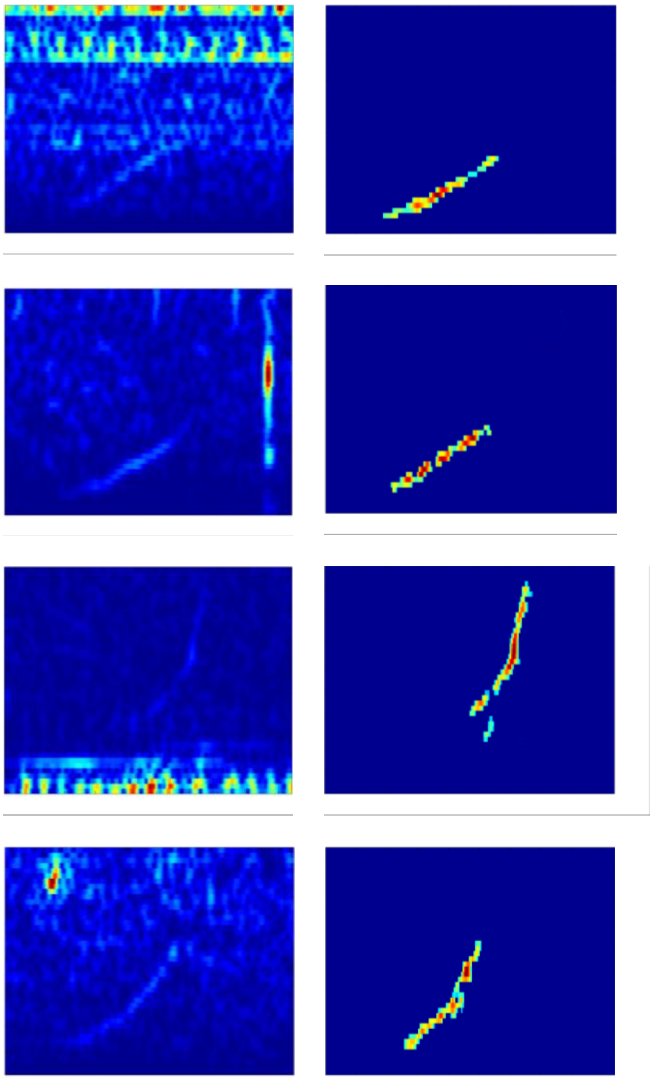
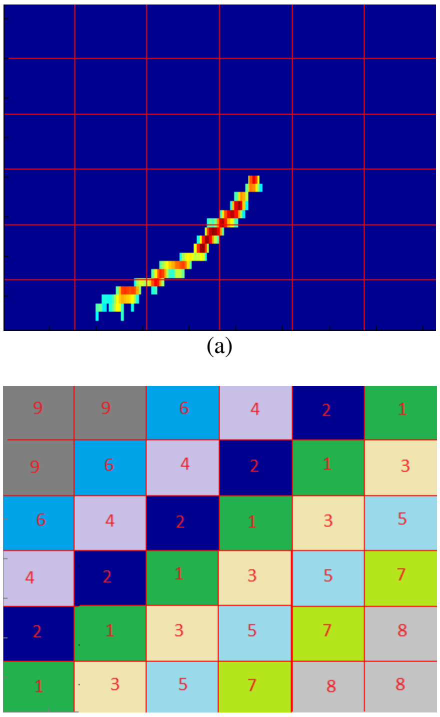


Figure 5. Four examples of faint NARW up-calls and outputs from the continuous region algorithm process. Left: Original spectrogram. Right: Proposed continuous region algorithm output.

2.3 Grid Masking and Feature Extraction

After continuous region processing and generation of the new spectrogram including the regions of interest, we divide the new spectrogram into equally spaced grids. As shown in Figure 6-(a), we used a 6X6 grid for the spectrogram with the specifications mentioned in 2.1.

The first set of features includes the means of spectrogram values over minor diagonals of the grid plane. Figure 6-(b) shows the grid pattern used to extract the diagonal features. In this figure, the diagonal grid cells are distinguished using colors and numbers. For example, in Figure 6-(a), the grid cells over diagonal #3 have a significant mean value compared to other diagonals.



(a)

(b)

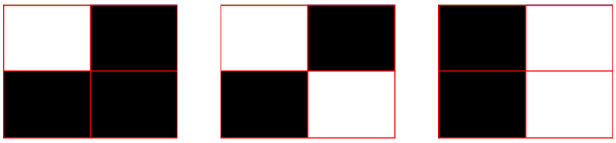
Figure 6. Spectrogram gridding: (a) sample spectrogram gridding, (b) the grid pattern used to extract the diagonal features. In this case (6x6 gridding), the diagonal feature set includes 9 features corresponding to the means of spectrogram values over the diagonal grid cells shown above.

The second set of features is generated using the sliding masks shown in Figure 7. These are binary masks having the value of one for black cells and zero for white cells. Each mask slides over the grid plane and calculates the averages of the spectrogram mean values located in black cells. The feature for each grid cell is determined as the maximum value of the three masking results as following,

$$f(x,y) = \max\{M_1(x,y), M_2(x,y), M_3(x,y)\} \tag{2}$$

$$\begin{aligned} M_1(x,y) &= (mean(x+1,y) + mean(x,y+1) + mean(x+1,y+1))/3 \\ M_2(x,y) &= (mean(x+1,y) + mean(x,y+1))/2 \\ M_3(x,y) &= (mean(x,y) + mean(x,y+1))/2 \end{aligned} \tag{3}$$

where *f(x,y)* is the feature allocated to the each grid cell (*x,y*); $M_1(x,y)$, $M_2(x,y)$, and $M_3(x,y)$ are the masking results for grid cell (*x,y*) corresponding to the three masks shown in Figure 7; and *mean*(*x,y*) is the mean value of spectrogram points located inside the grid cell (*x,y*).



(a) (b) (c)

Figure 7. The three masks used to extract the grid features

2.4 Artificial Neural Network

The Artificial Neural Network (ANN) is a popular and effective technique for bioacoustic signal classification (Potter et al. 1994; Mellinger 2004). ANNs can accept a wide range of feature variables as input, such as those from a spectrogram, frequency contour, or waveform. This flexibility allows the application of ANNs to different detection conditions and various types of signals. For example, Potter *et al.* used a feed-forward ANN to distinguish bowhead whale endnotes from interfering noises. They used the signal spectrogram as the input to the ANN (Potter *et al.* 1994). In another example, Deecke *et al.* used a standard back-propagation trained ANN to classify killer whale dialects to nine different categories (Deecke *et al.* 1999). They used the extracted pulse-rate contours as the input to the ANN.

There are several types of ANNs that can be applied for the purpose of classification. Feed-forward networks are more preferred for our purpose because they have less complexity, and relatively fewer numbers of neurons and connections compared to feedback networks (Potter, et al. 1994). However, feed-forward networks usually need a larger training dataset for backpropagation training (Potter, et al. 1994). In our case, this is not a problem because we have a large enough dataset available for training purpose. Therefore, we chose a standard, feed-forward, back-propagation trained network for classification. In this problem, we use a network with two hidden layers that receives the feature vectors extracted from the spectrograms (as described in 2.1, 2.2 and 2.3) as input.

3. Results and Conclusion

The proposed method was evaluated on a dataset recorded in Massachusetts Bay, United States. The dataset includes 20000 sound clips for training (containing 4473 NARW up-calls, and 15527 non-up-calls), and 10000 sound clips for testing (containing 2554 NARW up-calls, and 7446 non-up-calls). After spectrogram denoising, normalization, and equalization, we applied the proposed continuous region processing to detect the regions of interest, and extract the features as presented in 2.3. In this case, we extracted and used only 20 features including the 5 diagonal features corresponding to diagonals 1-5 in Figure 6-b, and 15 masking features generated by sliding the masks shown in Figure 7 over the spectrogram except for the first and last columns. We used an ANN classifier with only 2 hidden layers, with sizes of 32 and 16 neurons, trained using standard gradient-descent back-propagation with 100 epochs. The proposed method was evaluated for three different cases: with only 5 diagonal features as the feature set, with only 15 masking features as the feature set, and with the total 20 features as the feature set for ANN training, testing and classification. Figure 8 demonstrates the Receiver Operating Curve (ROC) of the proposed method evaluated on the testing dataset for the three different cases. Table 2 also shows the error rate (False Positive) for a fixed 90% True Positive Rate (e.g FPR=4.5% for TPR=90% using all 20 features). As we see in Figure 8 and Table 2, the proposed method can achieve high performance even using fewer numbers of features (5 features). This can be very beneficial when we aim to reduce the computational complexity of the classification stage.

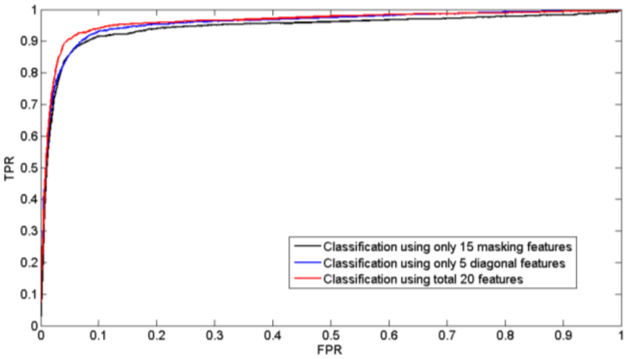


Figure 8. Receiver Operating Curve (ROC) of proposed method for the testing dataset

Table 2. The performance of the proposed method for fixed 90% True Positive Rate using diagonal and masking features.

Features	FPR%	TPR%
Using total 20 features	4.5	90
Using only 5 diagonal features	7.0	90
Using only 15 masking features	8.1	90

As shown in Figures 3 and 4, the proposed technique is effective for detecting up-calls under conditions of high ambient noise and/or interfering sounds. Figure 5 also illustrates that this method is also capable of capturing even very faint up-calls (i.e. low SNRs). Furthermore, this approach performs very well despite the relatively low number of features (Table 2). Future directions for this work include applying this algorithm to large continuous archival sounds streams and investigating the performance for recognizing NARW calls within context of accurate seasonal information.

References

Clapham, P.J., Young S.B., and Brownell R.L., Baleen whales: conservation issues and the status of the most endangered populations. Mamm. Rev. 29:35-60, 1999.

Clark C. W., The acoustic repertoire of the Southern right whale, a quantitative analysis, *Anim. Behav.,* vol. 30, pp. 1060-1071, 1982.

Clark, C. W., Brown, M. W. and Corkeron, P. Visual and acoustic surveys for North Atlantic right whales, *Eubalaena glacialis*, in Cape Cod Bay, Massachusetts, 2001-2005: Management implications, *Mar. Mamm. Sci*. vol. 26, pp. 837-854. 2010.

Deecke V.B. , Ford J.K.B. and Spong P., Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (Orcinus orca) dialects, *J.Acoust.Soc.Am.*, vol. 105, no. 4, pp. 2499-2507, 1999.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., and Clark, C. W., North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms, *in IEEE Proceedings of the 2010 Long Island Systems, Applications and Technology Conference,Farmingdale*, NY, pp. 1–6.

Gillespie D., Detection and classification of Right Whale calls using ’edge’ detector operating on a smoothed spectrogram, *Can. Acoust.*,vol. 32, no. 2, pp. 39–47, 2004.

Gonzalez, R. C., Woods, R. E. and Eddins S. L., Digital Image Processing Using MATLAB, New Jersey, Pearson Prentice Hall, 2004.

Kraus, S., M. W. Brown, H. Caswell, C. W. Clark, M. Fujiwara, P. K. Hamilton, R. D. Kenney, A. R. Knowlton, S. Landry, C. A. Mayo, W. A. McLellan, M. J. Moore, D. P. Nowacek, D. A. Pabst, A. J. Read and R. M. Rolland, North Atlantic right whales in crisis, *Science* vol. 309, pp. 561-562, 2005.

Lim, Jae S., Two-Dimensional Signal and Image Processing, Englewood Cliffs, NJ, Prentice Hall, 1990, p. 548, equations 9.44 - 9.46.

Mellinger D. K., A comparison of methods for detecting Right Whale calls, *Can. Acoust.*, vol. 32, no. 2, pp. 55–65, 2004.

Mohammad B., McHugh R., Automatic Detection and Characterization of Dispersive North Atlantic Right Whale Up-calls Recorded in a Shallow-Water Environment Using a Region-Based Active Contour Model, *IEEE journal of oceanic eng.,* 2011.

Norris T., J. Oswald and R. Sousa-Lima, A review and inventory of fixed installation passive acoustic monitoring methods and technologies, Technical report, 2010.

Potter J. R., Mellinger D.K. and Clark C.W., 'Marine mammal call discrimination using artificial neural networks, *J.Acoust.Soc.Am.*, vol. 96, pp. 1255, 1994.

Sánchez-García A., Muñoz-Esparza P., and Sancho-Gomez J. L., A novel image-processing based method for the automatic detection,extraction and characterization of marine mammal tonal calls, *J. Mar. Biol. Assoc. U.K.*, pp. 1–18, Jul. 9, 2009.

Sousa-Lima, R. S., Norris, T. F., Oswald, J. N., and Fernandes, D. P., A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals, *Aquat. Mamm.* Vol. 39, pp. 23-53.

Song-based Classification techniques for Endangered Bird Conservation

Erick Stattner ESTATTNE@UNIV-AG.FR
LAMIA Lab. University of the French West Indies and Guiana, France

Wilfried Segretier WSEGRETI@UNIV-AG.FR
LAMIA Lab. University of the French West Indies and Guiana, France

Martine Collard MCOLLARD@UNIV-AG.FR
LAMIA Lab. University of the French West Indies and Guiana, France

Philippe Hunel PHUNEL@UNIV-AG.FR
LAMIA Lab. University of the French West Indies and Guiana, France

Nicolas Vidot NVIDOT@MARTINIQUE.UNIV-AG.FR
LAMIA Lab. University of the French West Indies and Guiana, France

Abstract

The work presented in this paper is part of a global framework which long term goal is to design a wireless sensor network able to support the observation of a population of endangered birds. We present the first stage for which we have conducted a knowledge discovery approach on a sample of acoustical data. We use MFCC features extracted from bird songs and we exploit two knowledge discovery techniques. One relies on clustering-based approaches and highlights the homogeneity in the songs of the species. The other one is based on predictive modeling and demonstrates the good performances of various machine learning techniques for the identification process. The knowledge elicited provides promising results to consider a widespread study and to elicit guidelines for designing a first version of the automatic approach for data collection based on acoustic sensors.

1. Introduction

In last decades, due to the exponential growth of global commercial and industrial activities, numerous scientists have focused on environmental and social troubles that are becoming increasingly worrying: global warming, pollution, disease spreading, exhaustion of energy resources or biodiversity assessment.

This last phenomenon is one of the most challenging problem for geologists, ecologists, biologists and ethologists. Indeed, it is now apparent that changes occurring on an environment, as small as they are, can have significant impacts on the equilibrium of an ecosystem, and especially on the survival of animal species that depend on it.

New information and communication technologies and devices provide powerful tools for collecting useful and wide scale information on a variety of factors potentially involved in biodiversity loss. For example, GPS devices have been used for tracking individual movements in situations in which a human presence is not possible (Rumble et al., 2001; Ryan et al., 2004). Fixed devices such as sensors are considered for detecting the presence of individuals (Fagerlund, 2007; Cai et al., 2007) and studying their behavior (Stattner et al., 2010; 2011).

In this paper, a case study is presented on an endangered bird species endemic to the site of *La Caravelle* on the Martinique Island, French West Indies, called the *Moqueur Gorge Blanche* (MGB), for which local scientists have initiated a work of both study and protection (see Figure 1). Since it is quite tedious and unproductive to manually collect data by visual ob-

ICML 2013 Workshop on Machine Learning for Bioacoustics, Atlanta, Georgia, USA, 2013. Copyright 2013 by the author(s).

Figure 1. An individual of Moqueur Gorge Blanche in its habitat (©Vincent Lemoine)

servation during programs conducted on the natural area of the species, the final goal is to design an efficient methodology to automate the collection. Indeed although the manual technique has helped to obtain a first view of the species behavior, it also raises many questions about the reliability of the data collected, since (i) data are collected only during short periods, (ii) the human presence may, itself, affect the bird behavior.

Our long-term objective is to design an automatic solution for data collection, based on a wireless acoustic sensor network, able to capture widespread information about the bird population. Each sensor of the network should be fitted with a microphone and be able to detect the presence of the species by analysing the songs.

In this paper, we address the latter issue to evaluate how song analysis is relevant for predicting the presence of the species with a good precision on the basis of their songs. In this preliminary step, we have considered a set of pre-recorded songs of the Moqueur Gorge Blanche (Roché et al., 2009). We have trained different knowledge discovery techniques on the corresponding signals and we have obtained encouraging results to consider an efficient knowledge extraction from acoustical data and an optimal design of the sensor network.

The papers is organized as follows. Section 2 presents standard recognition processes on audio signals. Section 3 is devoted to the description of our case study data and to their pre-processing. In Section 4 we discuss the knowledge discovery approach conducted on pre-processed data. In Section 5 after a short synthesis, we present further works to lead on the project.

2. Related works

Last years, the analysis of acoustic signals has been a very active research area that have found applications in various domain such as speech recognition (Sakoe & Chiba, 1978), speaker identification (Reynolds, 1995), source localization (Valin et al., 2003). Globally, the recognition process performs in two key steps.

(i) A parametrization process, that summarizes the recorded audio signal through a characteristic fingerprint. This fingerprint is designed so that if two records are similar, their associated fingerprints should also be very close. More specifically, the parametrization process allows representing the audio signals by a series of coefficients that describe it (Eisele et al., 1996). Several parametrization techniques have been proposed such as LPCC (Linear Prediction Cepstral Coefficients) or MFCC (Mel-Frequency Cepstral Coefficients). As our final objective is to implement our recognition process on sensors of a network, we use in this work the MFCC (Mel Frequency Cepstral Coefficient) technique, since it has been shown that this technique has good performances on this kink of devices (Levy et al., 2003).

(ii) A classification process, that aims to determine if the generated fingerprint belongs to a known class (Rabiner & Wilpon, 1979). For instance, in the particular case of the recognition of bird songs, Fagerlund (Fagerlund, 2007) uses a support vector machine to identify songs of a given bird species, while Cai et al. (Cai et al., 2007) suggest the use of a neural network.
As our objective is to perform the recognition on sensors, that are known to have limited capacities in terms of calculation and memory, we use in this work standard classification algorithm with the objective to obtain a predictive model easy to implements on the motes of the network.

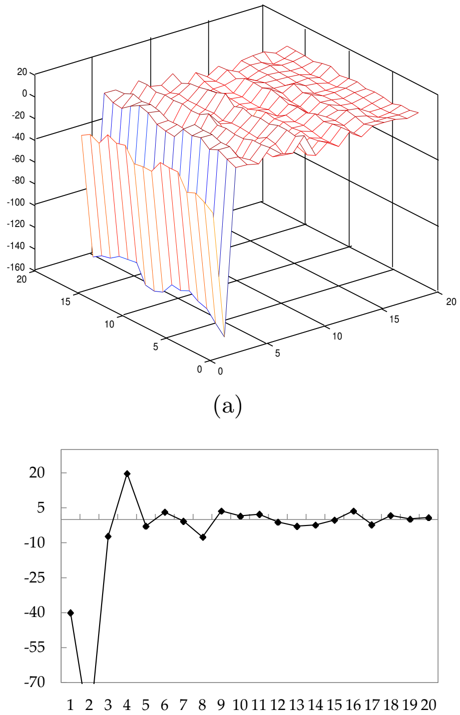
3. Acoustical data

In this section, we explain how original audio raw data have been preprocessed to produce acoustical datasets on which knowledge discovery techniques have been applied.

The initial audio data or *raw data* are songs of Moqueur Gorge Blanche collected in 2009 on the site of La Caravelle (Roché et al., 2009). We have extracted seven song examples. Each song is sampled at 44.1 kHz and stored as 16-bit signed mono .wav format files.

For each song example, MFCC features have been computed with the Java framework CoMIRVA (Schedl et al., 2007) developed and maintained by Markus Schedl (see Figure 2(a)). In order to simplify the computation phase, we have generated a mean fingerprint for each song, by averaging the different MFCC

values provided for each coefficient (see Figure 2(b)). Thus we have a mean fingerprint vector of 20 averaged MFCC coefficients for each of the 7 songs. Each of the 20 averaged MFCC coefficients corresponds to a time window of an original song.
In this way, the problem is to determine, for such a fingerprint, whether it belongs to the species or not. In other words, we would like to predict with a good accuracy that a new occurring song belongs to the Moqueur Gorge Blanche.



(a)

(b)

Figure 2. Example of (a) MFCC and (b) average fingerprint, obtained from a song of the Moqueur Gorge Blanche

In order to conduct classification tasks, we have complemented these examples with counter-examples to obtain a training dataset of mean fingerprints obtained on other bird species. Thus, 17 new examples have been added for species such as Common blackbird, Accentor, Heron, Woodcock, etc.

Traditionally the data mining *preprocessing* step is very important since choices have to be done for preparing, cleaning and transforming the raw data that may be often noisy, imbalanced and not in the adequate format that learning algorithms requires as input. In the current study, as described just above, the original audio data are transformed to produce the dataset composed of 24 averaged MFCC fingerprints. For predictive modelling, it is necessary to label each data sample with a corresponding class name. In this work we address a typical binary classification problem to predict if a data sample is an actual Moqueur Gorge Blanche song (”MGB” class) or another bird species song (”Other” class). In the following, the ”MGB” class will be considered as the positive class while the ”Other” class will be considered as negative. As said before, the ”MGB” class contains 7 examples while the ”Other” contains 17 examples. Our dataset is thus clearly imbalanced since there are more than twice as much ”Other” examples than ”MGB”. In order to overcome the imbalance between classes in datasets, two strategies are commonly used:

1. assign distinct costs to class examples (usually higher costs for the minority class) (Pazzani et al., 1994)

2. re-sample the source dataset, either by over-sampling the minority class and/or under-sampling the majority class (Kubat & Matwin, 1997)

Since the dataset is rather short, we have followed the second approach to create balanced training sets. We have thus over-sampled the ”MGB” class using the *Synthetic Minority Over-sampling Technique* (SMOTE) algorithm (Chawla et al., 2002) which consists in generating synthetic examples by randomly selecting points along the lines that join a minority class original sample and some of its nearest neighbors. Since this over-sampling technique is based on random properties, we have generated 100 extended balanced datasets (with 16 ”MGB” and 17 ”Other”) and we have averaged the performances obtained on each of them in order to have statistically significant results.

In the following, we have conducted experiments on both datasets without and with over-sampling. In the following, we refer to them as Simple_MFCC and Extended_MFCC dataset respectively.

4. Classification-based automatic detection

We describe in this section the knowledge discovery method applied that covers both descriptive and predictive approaches. On the *descriptive* axis, we have conducted a clustering based on dynamic time warping (DTW), a time series alignment algorithm developed originally for the purpose of speech recognition (Sakoe & Chiba, 1978), and on the *predictive* axis, we have trained classical algorithms. On the one hand, the clusters extracted that have a very good matching with the classes "MGB" and "Other", tend to show the homogeneity of MGB song signals. On the other hand, *predictions* obtained by training on the Simple_MFCC and Extended_MFCC datasets provide good performances that allow us to consider further wide scale training and testing experiments.

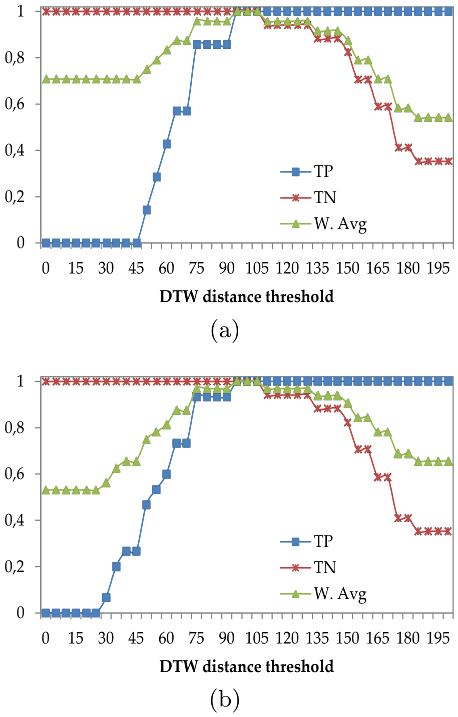
4.1. Clustering

In the area of the analysis of acoustic signals, the underlying assumption is that it exists a high degree of similarity between the elements of a same class. In other words, we can expect that the examples of a same class also belong to a same cluster.

Thus, in a first approach, we have studied the matching between the clusters and the classes. The matching is based on the notion of distance between the examples. For evaluating the distance between two examples, we use the classical DTW algorithm commonly use in the area of signal processing.
More precisely, let s_1 and s_2 be two sequences, the objective of DTW is to align these sequences by warping the time axis iteratively until an optimal match is found. The originality of this algorithm is its capacity to evaluate the similarity between two sequences that may vary in size, time or speed.

Our approach was as follows:
(i) A reference fingerprint has been created by averaging all fingerprints of examples of Moqueur Gorge Blanche.
(ii) For a given song fingerprint, we measure, by using the DTW algorithm, the distance to the reference fingerprint that characterizes the species. When this distance is under a given threshold β, the fingerprint is closed to the reference and we suppose that the associated song belongs to the species. Otherwise, the song is identified as different from that species.

By using both datasets (simple and extended), we have measured the performances of the identification when using the distance with DTW. Figure 3 shows these results according to the threshold β. In this clustering context, for a given cluster, we call TP, or True Positive rate, the fraction of "MGB" class examples into this cluster for which the DTW distance is under the β threshold. Similarly, we call TN, or True Negative rate, the fraction of "Other" class examples into the cluster for which the DTW distance is above the β threshold. *W. Avg* gives information about the performances of the detection by measuring the following ratio $\frac{|MGB|\times TP+|Other|\times TN}{|MGB|+|Other|}$.



(a)

(b)

Figure 3. Performances on the identification process based on DTW distance with (a) simple (b) extended datasets (TP: True positive, TN: True negative, *W. Avg* performances)

When the DTW distance threshold β is very low, all examples are detected as belonging to the “Other” cluster. This explains why TN is very high and TP very low. Inversely, when β is high, a lot of examples are detected as belonging to the MGB cluster, which explains the high degree of TP and the low degree of TN.
However, we can observe that it exists a DTW distance threshold for which the detection is maximal. Indeed, when $\beta \in [95..110]$, we observe that TP, TN and $W.\ Avg$ are all equal to 1. These results suggest that the MGB songs tend to be very homogeneous, since we can find some distance thresholds for which all examples of a same class belong to the same cluster.

4.2. Predictive modelling

In this section we present the results that we obtained in the area of supervised machine learning with the following commonly used machine learning techniques: C4.5 (Quinlan, 1993) and Random Forest (RF) (Breiman, 2001) decision tree approaches, Naive Bayes (NB) (Rish, 2001) and Multi-layer Perceptron (MLP) Artificial Neural Network (ANN). Since the total number of examples was quite low in order to generate typical 66%/33% train and test datasets, we used a leave-one-out cross validation (loocv) scheme (i.e. k-

Table 1. Performances of standard machine learning algorithms averaged on 100 leave one out cross validation experiments

Technique		Simple_MFCC			Extented_MFCC		
		TN	TP	W.Avg	TN	TP	W.Avg
C4.5	Avg	88.20	71.40	83.33	91.66	88.2	89.88
	Std Dev	-	-	-	2.98	0.00	1.42
RF	Avg	86.99	90.33	89.58	98.32	90.22	94.14
	Std Dev	9.29	5.06	4.23	2.89	5.39	2.99
NB	Avg	88.20	85.7	87.5	96.21	100.00	98.17
	Std Dev	-	-	-	3.02	0.00	1.46
MLP	Avg	82.51	100	87.58	100	83.84	91.65
	Std Dev	0.00	0.81	0.59	0.00	3.12	1.61

fold cross validation with k set as the total number of available examples minus one) to estimate the performances of each experiment conducted in this section.

Table 1 presents the performances obtained by each technique with and without the synthetic examples generated by the SMOTE algorithm. First column refers to the technique used, while columns TN, TP and W.Avg stands for True Negative, True Positive and Weighted Average ($\frac{|MGB|\times TP+|Other|\times TN}{|MGB|+|Other|}$) rates. In this context of supervised classification, true positive rates correspond, as usually, to the percentages of "MGB" examples correctly classified as "MGB" while true negative rates stand for the percentages of "Other" examples correctly classified as "Other". Since each algorithm has been launched 100 times to overcome either its own random side and/or the randomness introduced by SMOTE, we give averaged performances rates and their associated standard deviation as a confidence[1].
We can see that for the Simple_MFCC dataset, the best weighed average performances are obtained with Random Forest (89.8%) while for the extented datasets best results are given by Naive Bayes (98.52%). Even if no clear tendency can be observed concerning the repartition of TP and TN rates in both types of datasets, the overall performances obtained on extented data seem to be significantly better. Figure 4 shows an example of decision tree learned with the C4.5 algorithm. The leaves show the final classification decisions taken according to the values observed on selected attributes (represented by the nodes). For instance, in this case, if the value observed on the attribute "C02" is lower or equal than -52.33 and the value observed on "C06" is greater than 5.69, the "Other" class is predicted.

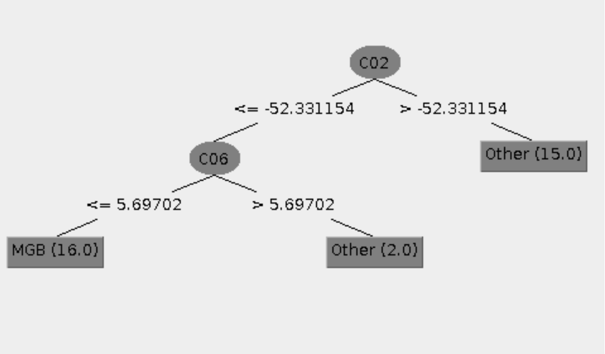


Figure 4. Example of decision tree obtained with C4.5.

5. Conclusion and future directions

Our overall project is to design a wireless sensor network able to support the observation of an endangered birds species endemic to the Martinique island, called the Moqueur Gorge Blanche.
This paper has focused on the first stage of the work, by evaluating how song analysis is relevant for predicting the presence the species with a good precision.

For this purpose, we have shown how knowledge discovery techniques can be used for recognize the songs of the species. The results obtained, that highlight the good performances of these techniques for the recognition of the Moqueur Gorge Blanche songs, allow to consider a real deployment on the ground

In our future works, our objective is to implement the recognition process on a wireless sensor network, in which sensors are fitted with microphone. In a first step, we plan to use the data collected on given periods for optimizing the network configuration.

A first track should be to use the sensor to identify the regions in which the population density is high. More sensors could thus be positioned in such regions, while less sensors would be allocated to the regions less frequented by the species.

[1] Only deterministic techniques C4.5 and Naive Bayes have been ran (with loocv) once on the original data and thus do not have std dev values.

At long terms, the data collected on the presence of the species could be used to search for correlations between the features of the habitat and the presence of individuals. Such a knowledge could have very relevant applications for the preservation of the species. For instance, it could be used for recreating a favorable habitat for the species.

References

Breiman, Leo. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

Cai, Jinhai, Ee, Dominic, Pham, Binh, Roe, Paul, and Zhang, Jinglan. Sensor network for the monitoring of ecosystem: Bird species recognition. In *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*, pp. 293–298. IEEE, 2007.

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., and Kegelmeyer, W. Philip. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 2002.

Eisele, Thomas, Haeb-Umbach, Reinhold, and Langmann, Detlev. A comparative study of linear feature transformation techniques for automatic speech recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pp. 252–255. IEEE, 1996.

Fagerlund, Seppo. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

Kubat, Miroslav and Matwin, Stan. Addressing the curse of imbalanced training sets: One-sided selection. In Fisher, Douglas H. (ed.), *ICML*, pp. 179–186. Morgan Kaufmann, 1997. ISBN 1-55860-486-3.

Levy, Christophe, Linares, Georges, and Nocera, Pascal. Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems. In *Proceedings of the Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan*, 2003.

Pazzani, Michael J., Merz, Christopher J., Murphy, Patrick M., Ali, Kamal, Hume, Timothy, and Brunk, Clifford. Reducing misclassification costs. In *ICML*, pp. 217–225, 1994.

Quinlan, J. Ross. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN 1-55860-238-0.

Rabiner, L and Wilpon, J. Considerations in applying clustering techniques to speaker independent word recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, volume 4, pp. 578–581. IEEE, 1979.

Reynolds, Douglas A. Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17(1):91–108, 1995.

Rish, I. An empirical study of the naive bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3(22):4146, 2001. URL http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf.

Roché, Jean C., Benito-Espinal, Edouard, and Hautcastel, Patricia. Oiseaux des antilles. Audio CD, 2009.

Rumble, M.A., Benkobi, L., Lindzey, F., and Gamo, R.S. Evaluating elk habitat interactions with gps collars. *Tracking Animals with GPS*, 2001.

Ryan, P. G., Petersen, S. L., Peters, G., and Gremillet, D. Gps tracking a marine predator: the effects of precision, resolution and sampling rate on foraging tracks of african penguins. *Marine Biology*, 145(2), 2004.

Sakoe, Hiroaki and Chiba, Seibi. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

Schedl, Markus, Knees, Peter, Seyerlehner, Klaus, and Pohle, Tim. The comirva toolkit for visualizing music-related data. In *Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization*, pp. 147–154. Eurographics Association, 2007.

Stattner, Erick, Collard, Martine, Hunel, Philippe, and Vidot, Nicolas. Detecting movement patterns with wireless sensor networks: application to bird behavior. In *MoMM*, pp. 251–258, 2010.

Stattner, Erick, Collard, Martine, Hunel, Philippe, and Vidot, Nicolas. Wireless sensor networks for social network data collection. In *IEEE Conference on Local Computer Networks (LCN)*, pp. 867–874, 2011.

Valin, J-M, Michaud, François, Rouat, Jean, and Létourneau, Dominic. Robust sound source localization using a microphone array on a mobile robot. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 2, pp. 1228–1233. IEEE, 2003.

5.2.Workshop Short papers

Birds calls classification in field recordings: analysis of challenges and difficulties in comparison with speech recognition

Ales MISHCHENKO ales.mishchenko@mail.ru
SATT sud-est / DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var

Herve GLOTIN glotin@univ-tln.fr
DYNI team, LSIS CNRS UMR 7296, Universite Sud Toulon-Var

Abstract

The automatic classification of sounds, produced by different biological species, is becoming more and more important task with a growing number of recordings. Among the most challenging problems is the birds calls recognition in the nature. The paper analyses the difficulties of bird calls recognition and its differences in comparison with automatic speech recognition (ASR). The particular focus in this analysis is made on a neural networks (NN) approach, some conclusions on preferable optimization methods (NN learning) are made.

1. Introduction

The importance of ecological monitoring along with availability of in-field recording systems resulted in increasing interest in automatic detection and classification of biological sounds, such as bird calls. The choosing and adaptation of methods from vast variety of automatic speech recognition (ASR) approaches is an important question. This paper analyses applicability of different ASR methods to the birds calls classification. Modern ASR systems are usually designed to recognize speech in the range from the isolated words recognition (e.g. command recognition used by call-center systems) to continuous speech recognition (e.g. computer dictation). The recognition of spontaneous speech in an audiostream (e.g. audiosurveliance systems) is a more difficult task, not yet matured to the robust commercialtype applications. The biological sounds recognition is close to the latter problem with the following additional difficulties: high signal-to-noise ratio (SNR problem) and high spectral variety along with poor repetitions variety (referred below as spectral-variety-to-temporal-variety, or SVTV problem). The SNR problem referres to the fact that the SNR for bioacoustics field recordings is usually below 5 dB (such as for ICML bird-recognition challenge), whereas the requirement to signal quality for ASR systems (callcenter or computer dictation) have a possibility to impose the SNR requirements higher than 20 dB. In addition, ASR systems usually rely on a single-type, or even single-known-type of noise, such as telephone line noise for call-center command recognition systems. In biological environment, on the contrary, there exist many different unknown noise sources. The problem of subtraction of different types of noise in bioacoustics recordings is not well-studied yet. However, some studies show that type of noise recognition is a separate difficult task than in ASR (X.Haitian, 2005).

1.1. The SVTV problem

The typical ASR systems require recognition of approximately 50 phonemes (for English language), whereas animal calls in a given location exhibit much more variability. The important feature of this is high variability in spectrum and in frequency of call repetitions along with a low temporal variability of call patterns. In addition to spectrum variability, the variance of biological calls depends on the distance to the microphone, which is also assumed to be a constant for the ASR applications. The evolutionary reason for the SVTV problem is that the human speech developed differently from the most of other animal calls evolution, including birds calls. The goal of the human speech evolution was to produce as many as possible different sequences of sound patterns, using the same articulation apparatus. Therefore, different

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

words share the same spectral characteristics and same frequencies of ”soundchanging”. Birds calls, on contrary, evolved with the goal to change the articulation apparatus in order to change spectral characteristics of emitted sound and increase the frequency of ”sound-changing” in the song. The result of this evolution is observable in human words and birds calls spectrograms. The latter frequently displays similar temporal patterns using different spectral characteristics and different frequency of ”sound-changing”. In some training examples from ICML bird call recognition challenge, the spectrogram of one bird-call can be approximately transformed to the spectrogram of another bird-call using scaling along time and frequency axes. Whereas the speech recognition is invariant to corresponding operations (that is the reason we recognize the same words pronounced at different pitch and speed). The following chapter discusses the applicability of ASR technologies to the biological sounds recognition.

2. Applicability of ASR methods to the birds calls classification.

The SNR and SVTV problems make it difficult to apply many popular ASR algorithms to biological sounds recognition. The obvious consequence of the SVTV problem is inapplicability of one of the earliest approach to ASR (acoustic-phonetic approach) (S.King, 2007), (A.). For the similar reasons, the Pattern Recognition and Statistical based approaches perform worse on bioacoustics data, as , in most cases, they rely on pattern comparison by statistical transition model such as HMM. Moreover, HMM require a lot of training samples, usually unavailable for biological sounds. Note, that some Pattern Recognition approaches were applied to birds, producing poor spectral-variety, but rich pattern-variety of calls (similar as humans do). The examples are Ground Parrot calls detection by event pattern recognition (Kirschel A. N. G, 2009), (M. Towsey, 2012) and event pattern recognition (Towsey M) applied to 2-pattern sequences of Eastern Whipbird call, emitting a call consisting from a whistle followed by a whip. Some other ASR approaches turn also to be less successful in biological sounds recognition. The example is the Dynamic time warping, which aims to measure similarity between two temporally different sequences and therefore successfully deals with temporal variation in human speech. The template-based approaches are among those that can be applied both to ASR and biological sounds recognition. Similarly to speech, biological sounds can be compared against a set of pre-recorded templates (usually local averages) in order to find the best match, followed by using some form of dynamic programming to temporarily align patterns. The binary template matching is used in (M. Towsey, 2012) to recognize the Currawong and Curlew calls. Another approach, equally applicable to ASR as to biological sounds recognition is neural networks (NN), which usually outperforms the template-based approaches, especially on a complicated and/or multispecies recognition tasks. This approach is analysed below.

3. Characteristics of a Bird call recognition with a Neural network

The typical stages of NN-based recognition are transforming parts of spectrogram into more abstract and more compact form (e.g. using one or more convolutional layers as first layers of NN), followed by one or more recognition layers (in a simplest case, one fully-connected layer). It also possible to use recurrent NN (RNN) to recognize temporal sequences. The important direction of theoretical investigation of NN-based biological sounds recognition is estimation the properties of the target optimization function (TOF) for a particular NN architectures. The SVTV problem leads to intuitive conclusion that higher statistical variation in bird calls spectrograms entails more complicated form for the TOF and, therefore, necessity of more elaborate optimization (learning) procedures. Let us illustrate this with a simple, but easy generalizable example. Below we are using simple architecture for the recognition part of NN (M input, N hidden and 1 output neuron) to show that with the same number of training inputs and outputs the weights of NN as well as the TOF may exhibit more variance near the optimum point (near the correctly learned weights). The output $O(\boldsymbol{x_j})$ for each training sample $\boldsymbol{x_j}$ of the recognition part of NN may be written as

$$O(\boldsymbol{x_j}) = \sum_{i=1}^{N} g(\boldsymbol{w_j}\boldsymbol{x_j} + b_j)u_j \qquad (1)$$

where $\boldsymbol{w} = (\boldsymbol{w_1}\boldsymbol{w_2}...\boldsymbol{w_N})$ and $\boldsymbol{u}$ are the weight-matrix and a weight-vector of, correspondingly, the N hidden neurons and one output neuron, $\boldsymbol{b}$ is a bias vector and $g(.)$ is an activation function. The sum of the squared difference between the NN output $O(\boldsymbol{x_j})$ and the target values t_j represent the TOF to be minimized over weights $\boldsymbol{w_j}$ and u_j. Supposing the recognition MLP is big enough to potentially learn all training samples, we write the zero of the TOF in the matrix form $\boldsymbol{Hu} = \boldsymbol{t}$, where $H = \begin{pmatrix} g(\boldsymbol{w_1}\boldsymbol{x_1} + b_1) & ... & g(\boldsymbol{w_N}\boldsymbol{x_1} + b_N) \\ ... & ... & ... \\ g(\boldsymbol{w_1}\boldsymbol{x_N} + b_1) & ... & g(\boldsymbol{w_N}\boldsymbol{x_N} + b_N) \end{pmatrix}$

The linear approximation of perturbation of this equation with respect to weights $\boldsymbol{w_j}$ and u_j, resulting in the same classification results t_j, leads to the $\boldsymbol{\delta H u} + \boldsymbol{H \delta u} = 0$ matrix equation or, in more detail,

$$\sum_{i=1}^{N} g'(\boldsymbol{w_i x_j} + b_i)\boldsymbol{\delta w_i x_j} u_i = -\sum_{i=1}^{N} g(\boldsymbol{w_i x_j} + b_i)\delta u_i \tag{2}$$

leading to the $\frac{\sum_{i=1}^{N} g'(\boldsymbol{w_i x_j}+b_i)\boldsymbol{\delta w_i} u_i}{\sum_{i=1}^{N} g(\boldsymbol{w_i x_j}+b_i)\delta u_i}\boldsymbol{x_j} = 1$

Or, in a more compact form,

$F(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{b}, \boldsymbol{x_j}, \boldsymbol{\delta w}, \boldsymbol{\delta u})\boldsymbol{x_j} = 1$

The input x_j and a function $F(.)$ are not independent as random variables and, therefore, it is impossible to connect their standard deviations with a simple formula, applicable in such a general case. Nevertheless, it is clear that the variance in distribution of training samples $\boldsymbol{x_j}$ should be compensated by the complexity of a function $F(.)$, reflecting, among others, the complexity of a TOF in the vicinity $(\boldsymbol{\delta w}, \boldsymbol{\delta u})$ of an optimal point. Therefore, increased variation in sub-spectrograms of a bird calls (as compared with human speech) entails increased variation in the function $F(.)$ and, probably, in the TOF.

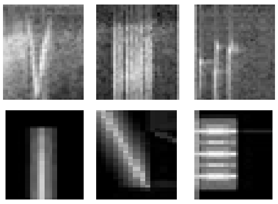


Figure 1. Examples of natural(upper row) and artificial (lower row) spectrograms of bird calls.

The theoretical analysis on connection between variability of spectrograms and neural network weights was verified in Torch machine learning environment. The MLP with a single hidden layer was trained to recognize 12x12, 24x24, 48x48 and 72x72 images of typical elements of bird calls. These elements were chosen as whistle (single tone, close to one or several horizontal segments in the spectrogram), chirps(slowly modulated tone), whips (rapidly modulated tone) and clicks (one or more vertical segments in a spectrogram). The training dataset was manually extracted from spectrograms of natural bird calls (ICML bird recognition challenge), as well as generated (artificial spectrograms) with different call duration, different ascending/descending slope (for chirps or whips), as well as different SNR with artificially added noise. The resulting dataset consisted from 1000 images of one bird call element, sometimes together with neighboring elements. Typical examples are shown in a Figure 1).

Training MLP on 90% of dataset with SGD resulted in 79% recognition rate on the remaining 10% of samples. As it may be expected intuitively, in the case of processing only artificial spectrograms, the robustness of recognition with respect to small random changes in the weights of a trained MLP was higher than in the case of both artificial and natural spectrograms. This may be understood as increasing the complexity of TOF function with complexity of input spectrogram. Therefore, having a more complicated shapes of a TOF for bird calls classification (as compared with ASR classification) may require more robust optimization methods. The 1st order methods, such as SGD perform poorly with complicated TOFs, having twists and pathological curvatures. Therefore, 2nd order methods, such as Quasi-Newton methods are, most probably, even more advantageous for bird call classification as they are for the ASR. This is a consequence of high variety in spectral characteristics as well as in frequencies of bird-calls repetitions (SVTV problem). Another side of the SVTV problem is low temporal variety of bird calls. This makes the simple RNN networks advantageous in bird call recognition. Unlike human speech, bird calls rarely consist from long, temporally-varying sequences. In most cases the call is a repetition of one sound pattern or alternation of two different sound patterns (such as whistle followed by a whip, mentioned earlier). As a general conclusion, RNN networks capable of recognizing short sequences together with Quasi- Newton optimization methods, may be advised for a bird call recognition.

4. Future work

One direction of future work is a theoretical study of application of different quasi-Newton methods for optimization of different neural architectures with the goal of adaptation to a particular cases of bioacoustics data. Another direction is using theoretical and experimental study of these methods to find an optimal set of parameters (such as damping, learning rate, stopcondition, etc) for a particular tasks of training the birds call recognition.

References

A., Zissman M. Predicting, diagnosing and improving automatic language identification performance. In *Eurospeech 97 Proceedings*, volume 1, pp. 51–54.

Kirschel A. N. G, Earl D. A, Yao Y I.A. Escobar E. Vilches Vallejo E. E Taylor C. E. Using songs to identify individual mexican antthrush formicarius moniliger:comparison of four classification methods. *Bioacoustics*, 19:1–20, 2009.

M. Towsey, B.Planitz, A.Nantes J.Wimmer P.Roe. A toolbox for animal call recognition. *Bioacoustics : The International Journal of Animal Sound and its Recording*, 21(2):107–125, 2012.

S.King, J.Frankel, K.Livescu E.McDermott K.Richmond M.Wester. Speech production knowledge in automatic speech recognition. *Acoust Soc Am.*, 121(2):723–42, February 2007.

Towsey M, Planitz B. Technical report: Acoustic analysis of the natural environment. URL http://eprints.qut.edu.au/41131/.

X.Haitian, Z.H. Tan, P. Dalsgaard B.Lindberg. Robust speech recognition based on noise and snr classification - a multiple-model framework. In *Interspeech Eurospeech*, Lissabon, Portugal, 2005.

North Atlantic Right Whale Call Detection with Convolutional Neural Networks

Evgeny Smirnov EVGENY.VERSUS.SMIRNOV@GMAIL.COM
Saint-Petersburg State University, Universitetskii prospekt 35, Petergof, Saint-Petersburg, Russia 198504

Abstract

Convolutional Neural Networks (CNN) have shown success in many image processing and speech recognition tasks. In this paper we propose to apply Convolutional Neural Network to the bioacoustic task of whale call detection. We trained a CNN to detect whale calls in 2-second audio clips in the Marinexplore and Cornell University Whale Detection Challenge. On the test data we achieved 2.4% error rate. Our best neural network consists of three convolutional layers followed by max-pooling layers, one fully-connected layer and final 2-way softmax layer. We used maxout hidden units with dropout to improve accuracy and reduce overfitting.

1. Introduction

The North Atlantic right whale, *Eubalaena glacialis*, is in danger of extinction (Kraus et al., 2005). One of the main threats to whale survival is high human activity in the areas of their migration. One third of all right whale mortalities are caused by collisions with ships and entanglement in fishing gear.

One way to reduce whale mortality is monitoring for the occurrences of whales by detecting their sounds on data recordings (Spaulding et al., 2009). Right whale species produce many different sounds, but most frequent and distinct one is a contact call ("up-call"). Automatic detection of such calls became a popular method of detecting right whales, and now there is a need for good algorithms of call detection in raw audio data.

There are already several different approaches to this task (Mellinger & Clark, 2000), (Urazghildiiev & Clark, 2006), (Urazghildiiev & Clark, 2007), (Urazghildiiev et al., 2009), (Dugan et al., 2010a). One of them is neural network approach (Dugan et al., 2010b). In this paper we try to improve it by using state-of-the-art type of neural networks (Convolutional Neural Networks with maxout hidden units) in the Marinexplore and Cornell University Whale Detection Challenge [1].

2. Dataset

The Marinexplore and Cornell University Whale Detection Challenge team provided us with a dataset of 30,000 training samples and 54,503 testing samples. Each sample is a 2-second .aiff sound clip with a sample rate of 2 kHz. Dataset contains mixture of right whale calls, non-biological noise and other sounds. The task was to create an algorithm for detecting right whale calls and to beat the existing whale detection algorithm of Cornell University.

For our experiments we compute Mel-frequency cepstral coefficients (MFCCs) along with their first and second temporal derivatives, and Fourier-transform-based filter-banks for all sound clips.

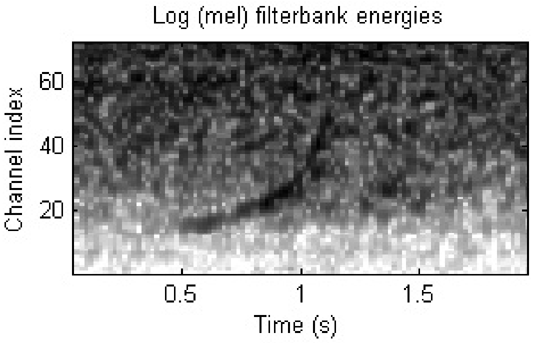


Figure 1. Example of filter-bank representation of a sound clip, containing right whale call

ICML 2013 Workshop on Machine Learning for Bioacoustics, Atlanta, Georgia, USA, 2013. Copyright 2013 by the author(s).

[1] http://www.kaggle.com/c/whale-detection-challenge

MFCCs were calculated with Hamming window, frame length of 25 ms and frame shift of 10 ms, for whole 2-second sound clip, so there were 2010 total input values (MFCCs with first and second derivatives) for each example. Filter-banks were calculated in range of 50 - 650 Hz, and include 72 coefficients, distributed on mel scale, for each of the 97 time steps.

3. Model

We used two different types of neural networks in our experiments: fully-connected Neural Network (NN) with sigmoid hidden units (Rumelhart et al., 1986), and Convolutional Neural Network (CNN) (LeCun et al., 1998) with maxout hidden units (Goodfellow et al., 2013).

3.1. Fully-connected Neural Network

This kind of neural networks was already used for whale call detection (Dugan et al., 2010b). We tried to improve its performance by using larger neural network and new regularization technique called "dropout" (Hinton et al., 2012). We tried several architectures with different parameters, and our best one consisted of 2010 units in input layer, 2000 sigmoid units in first and second hidden layers, and 2-way softmax layer. We used MFCC-based vector as input, and trained neural network for 500 epochs with backpropagation with batch size of 100, starting learning rate of 1 (reduced linearly for 300 epochs to the final value of 0.01) and dropout fraction of 0.5 for both hidden layers.

3.2. Convolutional Neural Network

After using fully-connected neural networks, we decided to try Convolutional Neural Network (CNN), other type of neural network, which uses some extra concepts like local filters, max-pooling and weight sharing (LeCun & Bengio, 1995). Convolutional Neural Networks already demonstrated good performance in several speech- and music-related tasks (Dieleman et al., 2011) (Abdel-Hamid et al., 2012), so they seem to perform well with sound and can be useful in bioacoustic tasks too.

Main difference between CNN and fully-connected NN is that CNN is aware of 2D structure of the input data. It can be very helpful if there are some local correlations between spatially adjacent input values. In image recognition tasks CNN uses local receptive fields to extract local features like oriented edges and corners, and then combine them in higher layers to get more complex features. Since in our whale detection task we have 2D filter-bank input data, which contains local correlations between energy values both in time and frequency domain, we can use CNN in image-like manner.

For preventing overfitting and for using highly-optimized implementation of 2D-convolution (cuda-convnet[2], made by Alex Krizhevsky), we cropped out three overlapping square patches of size 72 x 72 from our 72 x 97 filter-bank input data. Due to the lack of time, memory and fast GPU, we rescaled 72 x 72 patches to the size of 36 x 36.

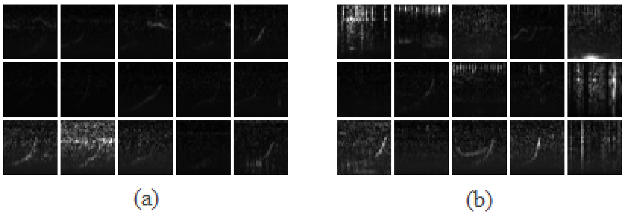


Figure 2. Examples of filter-bank-based patches (a) with right whale call, (b) without right whale call

We used recently proposed maxout units (Goodfellow et al., 2013) as hidden units. Given an input $x \in \mathbb{R}^d$, a maxout hidden layer implements the function

$$h_i(x) = \max_{j \in [1,k]} z_{ij}$$

where

$$z_{ij} = x^T W_{\dots ij} + b_{ij}$$

for learned parameters $W \in \mathbb{R}^{d \times m \times k}$ and $b \in \mathbb{R}^{m \times k}$. In the context of convolutional networks, a maxout feature map can be constructed by taking the maximum across k affine feature maps. A single maxout unit can be interpreted as making a piecewise linear approximation to an arbitrary convex function. So, training algorithm learns not just the relationship between hidden units, but also the activation function of each hidden unit.

Due to the lack of time, we didn't perform proper hyper-parameter search, and just used the same parameters, as in (Goodfellow et al., 2013) for MNIST and CIFAR-10 datasets. Our first CNN architecture consisted of 36 x 36 input layer, three convolutional layers, followed by max-pooling layers, and final 2-way softmax layer. First and second convolutional layers had 48 kernels of size 8 x 8, followed by max-pooling with pool size of 4 x 4. Third layer had 24 kernels

[2] https://code.google.com/p/cuda-convnet/

of size 5 x 5 and followed by max-pooling with pool size of 2 x 2. Learning rate at the start was 0.05, and then decreased by dividing by 1.00004 after each epoch. Dropout was used on the first convolutional layer, with dropout rate of 0.8. At the testing time, when all of the 36 x 36 patches were already classified, we averaged the results for each three patches, cropped from single testing 2-second sample.

Our second CNN architecture consisted of three convolutional layers, followed by max-pooling layers, one fully-connected layer with maxout hidden units and final 2-way softmax layer. First convolutional layer had 48 kernels of size 8 x 8 followed by max-pooling with pool size of 4 x 4. Second convolutional layer had 128 kernels of size 8 x 8 followed by max-pooling with pool size of 4 x 4. Third layer had 128 kernels of size 5 x 5 followed by max-pooling with pool size of 2 x 2. Fourth layer was fully-connected and had 240 maxout hidden units. Learning rate at the start was 0.1, and then decreased by dividing by 1.00004 after each epoch. Dropout was used on the first convolutional layer, with dropout rate of 0.8.

4. Results

Table 1. Test set AUC performance of different whale call detection methods

Method	AUC
NN with sigmoid units (this paper)	0.954
First CNN with maxout units (this paper)	0.971
Second CNN with maxout units (this paper)	0.976
Gradient Boosting Classifier	0.984
Cornell University Benchmark	0.721

Our best fully-connected neural network got Area under the ROC curve (AUC) performance of 0.954, our best CNN with maxout units got AUC performance of 0.976. Cornell University algorithm before the challenge got AUC performance of 0.721. Winner team of the Marinexplore and Cornell University Whale Detection Challenge used two averaged gradient boosting classifiers with complex feature engineering, and got AUC performance of 0.984.

5. Discussion

Our results show that fully-connected and convolutional neural networks are capable of achieving good performance in whale call detection task. We also used new type of hidden units - maxout units - and show that they can perform well in audio processing tasks. Our results can be easily improved with more careful parameter tuning, using better GPU and training for longer time. Also it must be useful to pre-train neural network on unlabeled data with some unsupervised feature learning model like CDBN (Lee et al., 2009).

References

Abdel-Hamid, O., Mohamed, A., Jiang, Hui, and Penn, G. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4277–4280, 2012.

Dieleman, Sander, Brakel, Philé mon, and Schrauwen, Benjamin. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th international society for music information retrieval conference : Proc. ISMIR 2011*, pp. 669–674. University of Miami, 2011.

Dugan, Peter J., Rice, Aaron N., Urazghildiiev, Ildar R., and Clark, Christopher W. North Atlantic right whale acoustic signal processing: Part II. improved decision architecture for auto-detection using multi-classifier combination methodology. In *Systems, Applications and Technology Conference IEEE Long Island*, 2010a.

Dugan, P.J., Rice, A.N., Urazghildiiev, I.R., and Clark, C.W. North atlantic right whale acoustic signal processing: Part i. comparison of machine learning recognition algorithms. In *Applications and Technology Conference (LISAT), 2010 Long Island Systems*, pp. 1–6, 2010b.

Goodfellow, Ian J, Warde-Farley, David, Mirza, Mehdi, Courville, Aaron, and Bengio, Yoshua. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

Hinton, Geoffrey E., Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

Kraus, Scott D., Brown, Moira W., Caswell, Hal, Clark, Christopher W., Fujiwara, Masami, Hamilton, Philip K., Kenney, Robert D., Knowlton, Amy R., Landry, Scott, Mayo, Charles A., McLellan, William A., Moore, Michael J., Nowacek, Douglas P., Pabst, D. Ann, Read, Andrew J., and Rolland, Rosalind M. North atlantic right whales in crisis. *Science*, 309(5734):561–562, 2005.

LeCun, Yann and Bengio, Yoshua. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, Honglak, Pham, Peter T., Largman, Yan, and Ng, Andrew Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS*, pp. 1096–1104, 2009.

Mellinger, David K. and Clark, Christopher W. Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, 107(6):3518–3529, 2000.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Spaulding, Eric, Robbins, Matt, Calupca, Tom, Clark, Christopher, Tremblay, Tremblay, Waack, Amanda, Warde, Ann, Kemp, John, and Newhall, Kristopher. An autonomous, near-real-time buoy system for automatic detection of North Atlantic right whale calls. *The Journal of the Acoustical Society of America*, 125(4):2615, 2009.

Urazghildiiev, Ildar R. and Clark, Christopher W. Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test. *The Journal of the Acoustical Society of America*, 120(4):1956–1963, 2006.

Urazghildiiev, Ildar R. and Clark, Christopher W. Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics. *The Journal of the Acoustical Society of America*, 122(2):769–776, 2007.

Urazghildiiev, I.R., Clark, C.W., Krein, T.P., and Parks, S.E. Detection and recognition of north atlantic right whale contact calls in the presence of ambient noise. *Oceanic Engineering, IEEE Journal of*, 34(3):358–368, 2009. ISSN 0364-9059.

5.3. Workshop Bird Challenge worknotes

Acoustic Identification of Bird Species Using Probabilistic Latent Component Analysis

Emmanouil Benetos EMMANOUIL.BENETOS.1@CITY.AC.UK
Department of Computer Science, City University London, London, UK.

Abstract

This submission for the ICML 2013 Bird Challenge uses the Probabilistic Latent Component Analysis (PLCA) method for identifying bird species in continuous audio recordings. A birdsong dictionary is created using pre-extracted spectral templates from provided training set. Sparsity constraints are also enforced in the symmetric PLCA model in order to lead to more meaningful solutions.

1. Introduction

These working notes for the ICML 2013 Bird Challenge[1] present a submitted system for identifying bird species in continuous audio recordings using Probabilistic Latent Component Analysis (PLCA). Section 2 presents the PLCA method while section 3 presents the proposed bird identification system. Finally, possible model extensions are discussed in section 4.

2. PLCA

Probabilistic latent component analysis (PLCA) is a spectrogram factorization technique that was first proposed in (Smaragdis et al., 2006). It approximates an input spectrogram $V_{\omega,t}$ as a bivariate probability distribution $P(\omega,t)$, where ω is the frequency index and t the time index, and attempts to factorize $P(\omega,t)$ as a series of spectral components and component activations. It is closely related to non-negative matrix factorization (NMF) (Lee & Seung, 1999), where PLCA can be viewed as a special case of NMF using the Kullback-Leibler cost function. However, contrary to NMF, PLCA provides a probabilistic framework that is extensible as well as easy to interpret. PLCA and related spectrogram factorization techniques have been used extensively in audio and image signal processing research, namely for source separation, multi-pitch detection, acoustic event detection, and action recognition.

The symmetric PLCA model can be formulated as:

$$V_{\omega,t} \approx P(\omega,t) = \sum_z P(z)P(\omega|z)P(t|z) \qquad (1)$$

where $P(\omega|z)$ are the spectral templates corresponding to component z, $P(t|z)$ are the time-varying component activations, and $P(z)$ is the prior probability for the components. For estimating $P(z)$, $P(\omega|z)$, and $P(t|z)$, iterative update rules are employed, which are derived from the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

3. Proposed Method

3.1. Time-frequency Representation

As a time-frequency representation, the constant-Q transform (CQT) with a spectral resolution of 60 bins/octave is used (Schörkhuber & Klapuri, 2010). The lowest frequency bin is at 330Hz and the highest bin is at 12.5kHz, while the time step is 10ms. Afterwards, a simple noise suppression procedure is applied to the log-frequency spectrogram $V_{\omega,t}$ using a $\frac{1}{3}$-octave span median filter.

3.2. Extracting Spectral Templates

For each species recording in the training set, a dictionary of 20 atoms is extracted using PLCA ($z = 1,\cdots,20$). The resulting spectral templates $P(\omega|z)$ are stacked together for all species, resulting in a matrix of dimensions $\Omega \times 700$ (where $\Omega = 295$ is the number of log-frequency bins and $700 = 20\cdot 35$, with 35 being the number of bird species in the challenge). An example of extracted templates is given in Fig. 1.

[1] http://sabiod.univ-tln.fr/icml2013/

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

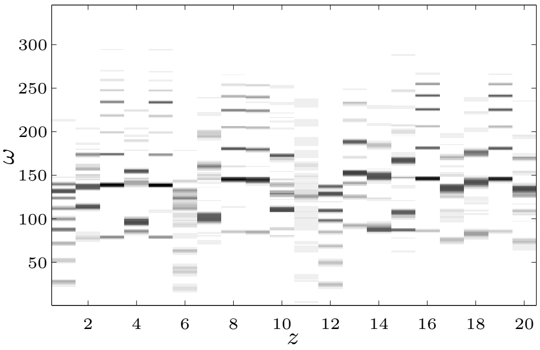


Figure 1. Pre-extracted spectral templates for the *Branta canadensis* species.

3.3. Bird Species Identification

For each recording from the test set, its normalized log-frequency spectrogram $V_{\omega,t}$ is fed into the model of (1). This time, the number of components $Z = 700$ and the update rules are only applied to $P(z)$ (which represents the mixture probability of the spectral components) and $P(t|z)$ (which represents the activation of each component over time), while $P(\omega|z)$ are the pre-extracted templates, which remain fixed. 70 iterations are used per test recording for estimating the unknown parameters.

Since the proposed model is overcomplete (it contains more information than in the input), it can converge to non-meaningful solutions. To that end, sparsity is enforced using the entropic prior proposed in (Smaragdis, 2009). In specific, sparsity constraints are applied to $P(t|z)$ (implying that only few dictionary components are active at a given time frame) and to $P(z)$ (implying that only few components should be present in the whole recording). The sparsity constraint in $P(z)$ also implies that only few bird species should be present in the recording.

Finally, the output of the PLCA model is $P(z)$, which is used to compute the probability of a bird species being present in a test recording:

$$P(bird_C) = \sum_{j \in C} P(z_j) \qquad (2)$$

where $bird_C$ denotes a bird class and C the set of components that belong to that class.

4. Model Extensions

The proposed model is fairly simple yet so far has reached solid results (with AUC=0.625), surpassing the baseline system by more than 9%. Its biggest drawback is the lack of any temporal modeling, which however can be supported by PLCA-based methods. One such example is Shift-invariant PLCA (Smaragdis & Raj, 2007) which supports time-frequency patches instead of spectral templates. Another option would be to add temporal constraints to the one-dimensional spectral templates, using the Non-negative Hidden Markov Model (Mysore, 2010), which combines PLCA with Hidden Markov Models. Finally, the constant number of basis can become variable, e.g. by performing segmentation on the training data and extracting one basis per segment.

Acknowledgments

E.B. is supported by a City University London Research Fellowship.

References

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

Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788–791, October 1999.

Mysore, G. *A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures.* PhD thesis, Stanford University, USA, June 2010.

Schörkhuber, C. and Klapuri, A. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conf.*, Barcelona, Spain, July 2010.

Smaragdis, P. Relative-pitch tracking of multiple arbitary sounds. *Journal of the Acoustical Society of America*, 125(5):3406–3413, May 2009.

Smaragdis, P. and Raj, B. Shift-invariant probabilistic latent component analysis. Technical report, Mitsubishi Electric Research Laboratories, December 2007. TR2007-009.

Smaragdis, P., Raj, B., and Shashanka, Ma. A probabilistic latent variable model for acoustic modeling. In *Neural Information Processing Systems Workshop*, Whistler, Canada, December 2006.

ICML 2013 Bird Challenge – Tech Report

Forrest Briggs BRIGGSF@EECS.OREGONSTATE.EDU
Oregon State University, Corvallis, OR, 97333, USA

Abstract

This technical report details our entry into the ICML 2013 Bird Challenge, in which the goal is to predict which birds are singing in a given audio recording. Our approach is based on 2D-supervised time-frequency segmentation, a histogram-of-segments representation, and binary relevance with a Random Forest.

1. Summary of Methods

The method we used for the challenge is very similar to our approach in (Briggs et al., 2013), which was also applied to multi-bird recordings collected with Songmeters (at a different location). The main steps in this system are:

1. Split into Chunks – Longer audio files are split into 10-second chunks.
2. Spectrograms – A spectrogram is computed for each chunk.
3. Noise-Filtering – A noise-filter reduces stationary noise in the spectrograms.
4. Segmentation – We manually annotate a collection of spectrograms with examples of correct segmentation (as a 2D mask of bird-sound regions). These annotations are used to train a Random Forest which classifies individual pixels of a spectrogram as bird-sound or background noise. The segmentation classifier is used to find 2D segments in the spectrograms of recordings in the test set.
5. Segment Features – Each 2D segment of the spectrogram is described by a 38D feature vector (see (Briggs et al., 2012) for a description of this feature vector).
6. Histogram-of-Segment Features – Segment features from the entire dataset are clustered to form a codebook. Then each 10-second chunk is described by a histogram of segments based on this codebook. This gives a representation of each 10-second chunk as a fixed-length feature vector (regardless of the number of segments it contains).
7. Multi-Label Classification – Each chunk is associated with a feature vector, and paired with a set of species. From here, we treat the problem using the multi-label classification framework, and apply binary relevance (with a Random Forest as the base classifier).

2. Isolating Bird Calls in Rain

Many of the recordings from the challenge have birds singing amidst background noise from rain. On interesting differences in the methods we used for this challenge compared to (Briggs et al., 2013) is that the segmentation algorithm is modified to better handle rain. In the original implementation from (Briggs et al., 2013), the segmentation algorithm associates each pixel the spectrogram with a feature vector consisting of the pixel intensities in a box around it, the average intensity in the box, the y-coordinate of the center of the box. Recall that this supervised segmentation approach requires manually annotated spectrograms to train. Rain drops are easily recognized by a human eye as vertical lines in the spectrogram. However, even though we provided examples with rain drops marked as negative, the segmentation algorithm could not differentiate between these regions and bird sound (both look like bright lines in the spectrogram). Therefore, we modify the feature vector to additionally include the intensities of all pixels in the same column. This allows the segmentation algorithm to "see" enough of the spectrogram to differentiate between rain and bird sound. Figure 1 shows an example result from the segmentation algorithm– it avoids rain drops (top of the spectrogram), but isolates bird sounds (bottom).

To form the training set for segmentation (pairs of

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

pixel feature vectors with bird/no-bird label), we sample pixels from the manually annotated training spectrograms. For this challenge, we annotated 9 ten-second chunks of audio from the validation set (Fig. 2), with red=bird sound and blue=rain drop. There are hundreds of thousands of pixels in each spectrogram, so it is expensive to use every pixel. Therefore, we sample 30% of red pixels as positive examples, 30% of blue pixels as negative examples, and 4% of uncolored pixels as negative examples.

3. Aggregating Predictions on Chunks

The recordings in the test set that we are asked to make predictions about are 2 minutes and 30 seconds. However, our system views each of these as a collection of 10-second chunks. Multi-label classification predictions are made based on the histogram-of-segments feature for each chunk, then predictions are aggregated.

Let each recording R_i be represented as n_i 10-second chunks, each described by a feature vector $\mathbf{x}_{i,j}$ for $j = 1, \ldots, n_i$. The classifier score for each class c on chunk $\mathbf{x}_{i,j}$ is $f_c(\mathbf{x}_{i,j})$. We compute the score for class c on recording R_i as $\max\{f_c(\mathbf{x}_{i,1}), \ldots, f_c(\mathbf{x}_{i,n_i})\}$.

4. Training and Val Sets

The data provided consists of a collection of training recordings, which are made with directional microphones of a single bird. The test data is instead collected with omnidirectional microphones, and contains multiple simultaneous birds, echoes, and rain. Additionally, 3 recordings are provided as a “validation” set, which are from the same collection as the test set; these val recordings are labeled with a set of species present.

We originally tried two methods using the training set (and not the val set), based on the MFCC features provided by the competition organizers, and also based on our own segmentation and features. However, both of these approaches achieved an AUC lower than 0.5. Based on these results, and inspection of the data, we suspect that the mismatch between training and test data is a major problem for our basic supervised classifiers. The proposed method in this document does not use the test set at all for training. Instead, it only uses the 3 labeled examples in the validation to train (which are split into 45 ten-second chunks with 3 distinct label sets). Surprisingly, the proposed method achieves much better AUC using only this training set than using the full training set. We suggest that this outcome is because the validation set comes from the same distribution as the test set.

5. Parameters

Table 1 states each of the parameters used in our proposed method.

6. Discussion

The most interesting part of our proposed method is the modification of the segmentation algorithm to better handle rain. Qualitatively, it appears to work much better than the original segmentation for recordings with light rain, and somewhat better with strong rain. However, in some cases it can introduce a new kind of segmentation error where a bird syllable is split in half by a rain drop (whereas the original segmentation would make a single segment encompassing both the rain drop and syllable). Further work is needed to explore methods for segmentation of bird sound in audio with rain.

References

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J.K., Hadley, A.S., and Betts, M.G. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131:4640, 2012.

Briggs, F., Fern, X. Z., and Irvine, J. Multi-Label Classifier Chains for Bird Sound. *ArXiv e-prints*, April 2013.

Table 1. Parameters used in our bird species classification system.

Parameter	Value	Explanation
frame-size	512	The number of samples in one frame of the spectrogram
frame-overlap	75%	Amount of overlap between successive frames of the spectrogram.
seg-num-trees	21	Number of trees in the RF used for segmentation
seg-feature-box	11	Width and height of the box of pixels used for features in segmentation
probmap-blur-radius	2	Radius of Gaussian blur applied to segmentation probabilities
seg-prob-threshold	0.5	Threshold applied to probabilities for segmentation
codebook-k	100	Number of clusters for k-means codebook / histogram dimension
BR-num-trees	100	Number of trees in the RFs for each class in Binary Relevance

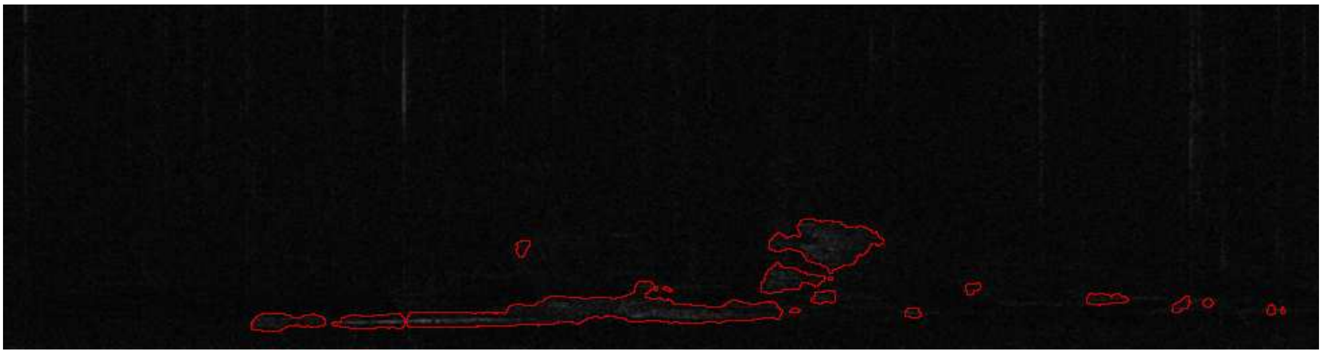


Figure 1. Example of automatic segmentation from our supervised segmentation algorithm. Note bird sound is isolated (bottom) and rain drops are ignored (top).

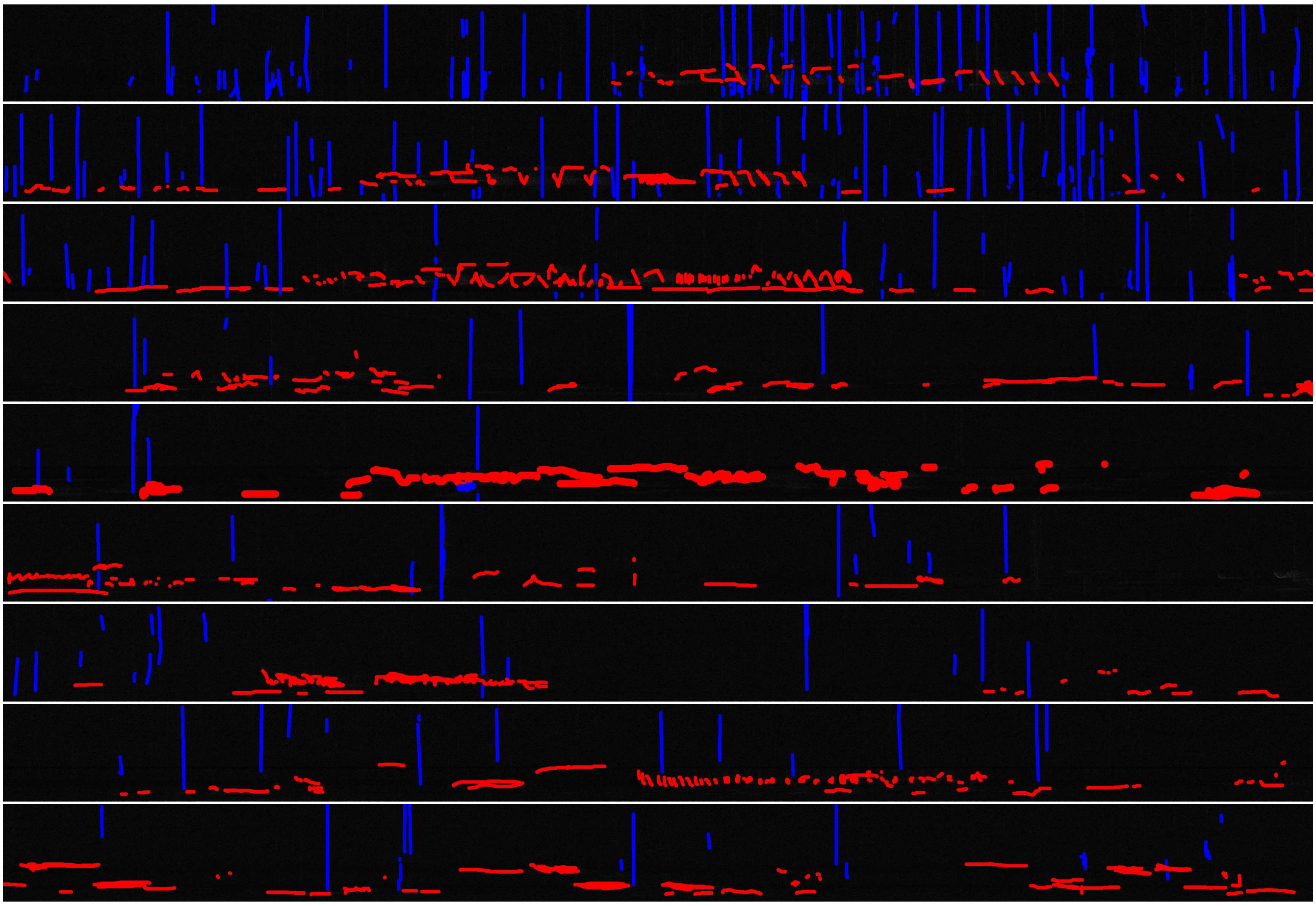


Figure 2. Our manual annotation spectrograms for 9 ten-second chunks from the validation set (for training supervised segmentation). Red = bird, blue = rain.

Clusterized Mel Filter cepstral coefficients and Support Vector Machines for bird song idenfication

Olivier Dufour OLIVIERLOUIS.DUFOUR@GMAIL.COM
LSIS, Université du Sud Toulon Var

Thierry Artieres THIERRY.ARTIERES@LIP6.FR
LIP6, Université Paris 6

Hervé GLOTIN GLOTIN@UNIV-TLN.FR
Aix-Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
Institut Universitaire de France, Bd St Michel, 75005 Paris

Pascale Giraudet GIRAUDET@UNIV-TLN.FR
Université du Sud Toulon Var

1. Introduction

We present here our contribution to the "Machine Learning for Bioacoustics" workshop technical challenge of 30th International Conference on Machine Learning (ICML 2013). The aim is to build a classifier able to recognize bird species one can hear from a recording in the wild. The method we present here is a rather simple strategy for bird songs and calls classification. It builds on known and efficient technologies and ideas and must be considered as a baseline on this challenge[1]. The method we present is dedicated to the particular setting of the challenge. It relies in particular on the fact that training signals are monolabel, i.e. only one species may be heard, while test signals are multilabel.

2. Description of the method

We present now the main steps of our approach. Figure 1 illustrates the main steps of the preprocessing and of feature extraction. We consider we want to learn a multilabel classifier from a set of N monolabeled training samples $\{(x^i, y^i)|i = 1..N\}$ where each input x^i is a audio recording and each y^i is a bird species $\forall i, y^i \in \{b_u|u = 1..K\}$ (in our case there are 35 species, $K = 35$). The system should be able to infer the eventually multiple classes (presence of bird species) in a test recording x.

2.1. Preprocessing

Our preprocessing is based on mfcc cepstral coefficients which have been proved useful for speech recognition (Chang-Hsing et al., 2006; Michael Noll, 1964). 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).

2.2. Windowing, silence removal and feature extraction

Windowing. We use windowing, i.e. computing a new feature vector on a window of n frames, to get new feature vectors that are representative of longer segments. The idea is close to the standard syllabe extraction step that is used in most of methods for bird identification (Neal et al., 2011; Briggs et al., 2012; 2009) but is much simpler to implement. In our case we considered segments of about 0.5 second duration (i.e. $n \approx$ few hundreds of frames) and used a sliding window with overlap (about 80%).

Silence removal. We first want to remove segments (windows) corresponding to silence since these would

[1] As we are also co-organizing this challenge, our participation aimed at defining a baseline system, with raw features, that all other participants could compare too. We did not look for optimizing each parameter of our system, and as any other participant, we conducted all the modeling and experimentation applying strictly the rules of the challenge.

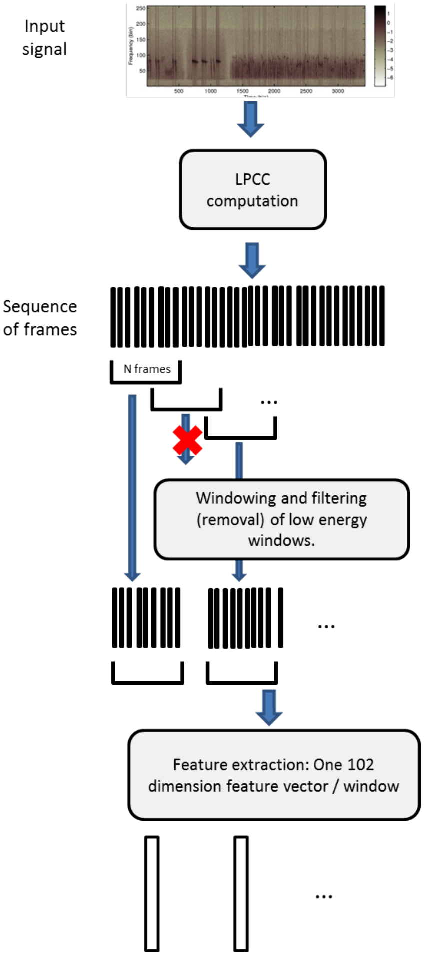

Input signal
LPCC computation
Sequence of frames
N frames
...
Windowing and filtering (removal) of low energy windows.
...
Feature extraction: One 102 dimension feature vector / window
...

Figure 1. Main steps of the preprocesing and of feature extraction.

perturbate the training and test steps. This is performed with a clustering step (learnt on training signals) that only considers the average energy of the frames in a window. Ideally this cluster makes that the windows are clustered into silence segments on the one hand, and calls and song segments on the other hand. Each window with low average energy is considered a silence window and removed from consideration. Our best results were achived when performing a clustering in three clusters and removing all windows in the lowest energy cluster.

Feature extraction. The final step of the preprocessing consists in computing a reduced set of features for any remaining segment / window. Recall that each segment consists in a series of n 17-dimensional feature vectors (with n in the order of hundreds). Our feature extraction consists in computing 6 values for representing the series of n values for each of the 17 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:

$$f_1 = \frac{\sum_{i=1}^{n}(|v_i|)}{n} \quad (1)$$

$$f_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v_i})^2} \quad (2)$$

$$f_3 = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(d_i - \bar{d_i})^2} \quad (3)$$

$$f_4 = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}(D_i - \bar{D_i})^2} \quad (4)$$

$$f_5 = \frac{\sum_{i=1}^{n-1}|d_i|}{n-1} \quad (5)$$

$$f_6 = \frac{\sum_{i=1}^{n-2}|D_i|}{n-2} \quad (6)$$

At the end a segment in a window is represented as the concatenation of the 6 above features for the 17 cepstral coefficients. It is then a new feature vector s_t (with t the number of the window) of dimension 102.

Each signal is finally represented as a sequence of feature vectors s_t, each representing a duration of about 0.5 second with 80% overlap.

2.3. Training

Based on the feature extraction step we described above the simplest strategy is to train a classifier (e.g. we used Support Vector Machines) on the feature vectors s_t which are long enough to include a syllabe or a call, with the idea of agregating all the results found on the windows of a test signal to decide which species are present (see section *Inference* below).

Yet we found that a better strategy was to first perform a clustering in order to split all samples (i.e. s_t) corresponding to a species into two different classes. The rationale behind this process is that call and song of a particular species are complety different sounds (Fagerlund, 2004) so that corresponding feature vectors s_t probably lie in different areas in the feature space. It is then probably worth using this prior to design classifiers (hopefully linear) with two times the number of species rather than using non linear classifiers with as many classes as there are species.

We implemented this idea by clustering all the frames s_t for a given species into two or more clusters. The two clusters are now considered as two classes that correspond to a single species. At the end, a problem of recognizing K species in a signal turns into a classification problem with $2 \times K$ classes. Note also that since the setting of the challenge is such that there is only one species per training signal, all feature vectors s_t of all signal of a given bird species b_u that fall into cluster one are labeled as belonging to class b_u^1 and all that fall into cluster 2 are labeled as belonging to class b_u^2.

The final step is to learn a multiclass classifier (SVM) in a one-versus-all fashion, i.e. learning one SVM to classify between the samples from one class and the samples from all other classes. This is a standard approach (named Binary Relevance) for dealing with multilabel classification problem where one sample may belong to multiple classes. It is the optimal method with respect to the Hamming Loss, i.e. the number of class prediction errors (either false positive and false negative).

2.4. Inference

At test time an incoming signal is first preprocessed as explained before in section 2.1, silence windows are removed (using clusters found on the train dataset), and feature extraction is performed for all remaining segments. This yields that an input signal is represented as a series of m feature vectors s_t.

All these feature vectors are processed by all $2K$ binary SVMs which provide scores that are interpreted as class posterior probabilities (we use a probabilistic version of SVM), we then get a matrix $m \times 2K$ of scores $P(c|s_t)$ with $c \in \left\{b_u^j | u = 1..K, j = 1, 2\right\}$ and $t = 1..m$.

We experimented few ways to agregate all these scores into a set of K scores, one for each species, enabling ranking the species by decreasing probability of occurence. Indeed this is the expected format of a challenge submission, from which a AUC score is computed (Area Under the Curve). First we compute $2K$ scores, one for each class, then we agregate the scores of the two classes of a given species.

Our best results were obtained by computing mean probabilities of all scores $\{P(c|s_t)|t = 1..m\}$ for each class c, using harmonic mean or trimmed harmonic mean (where a percentage of the lowest scores are discarded before computing the mean). This yields scores that we consider as class posterior probabilities of classes given the input signal x, $P(c|x)$.

The utimate step consists in computing a score for each species b_u given the scores of the the two corresponding classes b_u^1 and b_u^2. We used the following agregation formulae:

$$P(b_u|x) = 1 - (1 - P(b_u^1|x)) \times (1 - P(b_u^2|x)) \quad (7)$$

3. Experiments

3.1. Dataset

We describe now the data used for the “Machine Learning for Bioacoustics” technical challenge. Note that the training dataset (signals with corresponding ground truth) was available for learning systems all along the challenge together with the test set, without ground truth. Participants were able to design their methods and select their best models by submitting predictions on the test set which were scores on a subset only of the test set (33%). The final evaluation and the ranking of participants was performed on the full test set once all participants have selected 5 of all their systems submitted.

Training data consisted in thirty five 30-seconds audio recordings labeled with a single species, there was one recording per species (35 species overall). Yet, some train recording can include low signal-to-noise ratio (S.N.R) signals of a second bird species of bird. Moreover, according to circadian rythm of each species, other acoustically actives species of animals can be present such as nocturnal and diurnal insects.

Test data consisted in ninety 150-seconds audio recordings with possibly none or multiple species occuring in each signal.

The training and test data recordings have been performed with various devices in various geographical and climatological settings. In particular background and S.N.R are very different between training and test. All wav audio recordings have been sampled at 44 100 Hz with a 16-bits quantification resolution. Recordings were performed with 3 Song Meter SM2+ (Wildlife Acoutic recording device). Each SM2+ has been installed in a different sector (A,B and C) of a Regional Park of the Upper Chevreuse Valley. Every SM2+ recorded, at the same dates and hours (between 24 03 2009 and 22 05 2009), one 150-seconds recording per day between 04h48m00s a.m. and 06h31m00s a.m., which correspond to the acoustical maximum bird-activity period.

3.2. Implementation details

Frames and overlapping sizes. We computed Mel-frequency cepstral coefficients (MFCC) with the *melfcc.m Matlab* function from ROSA laboratory of Columbia university. This function propose 17 different input parameters. We tested numerous possible configurations and measured for each one the difference of energy contained in a given TRAIN file and a reconstructed signal of this recording based on cepstral coefficients. The difference was minimal with following parameters values:

window=512, *fbtype=mel*, *broaden=0*, *maxfreq=sr/2*, *minfreq*=0, *wintime=window/sr*, *hoptime= wintime/3*, *numcep=16*, *usecmp=0*, *dcttype=3*, *nbands=32*, *dither=0*, *lifterexp=0*, *sumpower=1*, *preemph=0*, *modelorder=0*, *bwidth=1*, *useenergy=1*

This process transformes a 30-seconds train audio recording (at 44 kHz sampling rate) into about 7 700 frames of 16 cepstral coefficients which we augmented with the energy computed by setting *useenergy=0.*

Next we computed feature vector s_t on 0.5 second windows with 80% overlap, which yields about $n = 300$ feature vectors per training signal (hence per species since there is only one training recording per species) and about $m =$ feature vectors per test signal.

LIBSVM settings. We used a multiclass S.V.M algorithm based on LIBSVM (Chang, 2008). We selected model parameters (kernel type etc) through two fold cross validation. Best scores have been obtained with C-SVC SVM type and linear kernel function.

3.3. Results

We report only our best results that correspond to the method presented in this paper for various computation for the class score at inference time. Table 1 shows how the way the mean score of a class is computed on the test set (see section 2.4) influences the final result. The table compares arithmetic mean, harmonic mean, and trimmed arithmetic mean (at 10, 20 et 30%). A trimmed mean at p% is the arithmetic mean computed after discarding p% extreme values, i.e. the $p/2$% lowest values and the $p/2$% largest values.

Table 1. Score Kaggle icml according to the way scores are aggregated.

mean aggregation	Kaggle score (AUC)
arithmetic mean	0.61362
harmonic mean	0.64234
trimmed mean 10%	0.64158
trimmed mean 20%	0.64639
trimmed mean 24%	0.64699
trimmed mean 30%	0.64614

Although our method is simple it reached the fourth rank over more than 77 participating teams at the Kaggle ICML Bird challenge with a score of 0.64639 while the best score was 0.69454. It is also worth noting that our system ranked about fifteen only on the validation set (one third of the total test set). This probably shows that our system being maybe simpler than other methods exhibits at the end a more robust behaviour and improved generalization ability.

4. Conclusion and perspectives

Although the method that we presented is simple it was to perform well on the challenge and to be much robust between validation step and test set. We believe this robustness comes from the simplicity of the method that do not rely on complex processing steps (like identifying syllables) that other participants could have used (Glotin & Sueur, 2013).

Possible improvements would consist in the integration in the model of additional information such as weather condition, or a taxonomia of species, allowing for more accurate hierarchical classification schemes.

5. Acknowledgments

This work is supported by the MASTODONS CNRS project Scaled Acoustic Biodiversity SABIOD and the Institut Universitaire de France which supports the "Complex Scene Analysis" project. We thank F. Jiguet and J. Sueur and F. Deroussen (Deroussen & Jiguet, 2011; Deroussen, 2001) who provided the challenge data.

PhD funds of 1st author are provided by Agence De l'Environnement et de la Matrise de l'Energie (mila.galiano@ademe.fr) and by BIOTOPE company (Dr Lagrange, hlagrange@biotope.fr, R&D Manager).

References

Briggs, F., Fern, X., and Raich, R. Acoustic classification of bird species from syllables: an empirical study. Technical report, Oregon State University, 2009.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X., Raich, R., Betts, M., Frey, S., and Hadley, A. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *Journal of the Acoustical Society of America*, 2012.

Chang, Chih-Chung. Libsvm. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2008. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chang-Hsing, L., Yeuan-Kuen, L., and Ren-Zhuang, H. Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications Vol. 1, pp.17-23*, 2006.

Deroussen, F. Oiseaux des jardins de france. Nashvert Production, Charenton, France, 2001. naturophonia.fr.

Deroussen, F. and Jiguet, F. Oiseaux de france, les passereaux, 2011.

Fagerlund, S. Acoustics and physical models of bird sounds. In *Seminar in acoustics, HUT, Laboratory of Acoustics and Audio Signal Processing*, 2004.

Glotin, H. and Sueur, J. Overview of the first international challenge on bird classification, 2013. URL http://sabiod.univ-tln.fr. online web resource.

Michael Noll, A. 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.

Neal, L., Briggs, F., Raich, R., and Fern, X. Time-frequency segmentation of bird song in noisy acoustic environments. In *International Conference on Acoustics, Speech and Signal Processing*, 2011.

Acoustic detection of multiple birds in environmental audio by Matching Pursuit

Dan Stowell and Mark D. Plumbley DAN.STOWELL@EECS.QMUL.AC.UK
Centre for Digital Music, Queen Mary, University of London

Abstract

We describe a submission to the ICML 2013 Bird Challenge, in which we explore the use of sparse representations as an advance on the standard technique of cross-correlation template matching in time-frequency representations. The Matching Pursuit algorithm is used to represent the signal as a sparse set of activations of templates derived from the challenge training audio.

Given an audio recording, it is a challenging task to detect automatically which bird species are represented, and a task that is relevant to practical applications in bioacoustics (Stowell & Plumbley, 2010). Recent research developments go beyond single-label classification and can identify multiple species simultaneously present in a recording (Briggs et al., 2012), or track multiple birds through an audio scene (Stowell & Plumbley, accepted). The ICML 2013 Bird Challenge stimulates developments in the field by challenging researchers to identify algorithmically which of 35 bird species are present in a public dataset of 90 audio recordings.[1] The present note describes a contribution to the challenge which explores the use of *sparse representations* in a multi-label classification problem.

In signal processing, a *sparse representation* is recovered by assuming that the signal is composed from some "dictionary" of atomic elements, with only a small number of those elements being active (nonzero) for any given signal of interest (Plumbley et al., 2010). This approach is motivated by the discovery that neural coding often makes use of such sparsity, and also by the engineering prospect of representing signals in highly compact form. Sparse representations are currently the subject of much research activity, and have been used in audio and music for tasks such as audio compression and transcription (Plumbley et al., 2010).

Our submission explores sparse representation to improve on the common technique of cross-correlation template matching in time-frequency representations (such as spectrograms). In the standard cross-correlation scenario, we have one or more templates per species, and each template is separately cross-correlated against the spectrogram in question. Peaks in the cross-correlation function are taken as detections for the corresponding species. However, when there is a large number of species to be detected, and some of these potentially have very similar templates, there is a problem: a single region of energy in the spectrogram (e.g. a single birdsong syllable) could independently match against multiple templates, giving spurious detection of many species from a single sound. Sparse decompositions can overcome this, by finding a representation of the signal as a sum of activations from all elements considered together as a dictionary.

1. Method

In the present case we use the simple and widespread *Matching Pursuit* (MP) algorithm to find a sparse representation of a sound spectrogram as a sum of templates. We use a fast optimised implementation provided by the free and open-source Matching Pursuit Toolkit (MPTK) (Krstulovic & Gribonval, 2006). Our script is implemented in Python, using the Python bindings recently available in MPTK 0.7.

Preprocessing: Each audio file is converted into a standard spectrogram representation (log-magnitudes of STFT, frame size 512). We apply median-based background subtraction to help counter stationary background noise.

Training: Training the system consists of creating a dictionary of spectrogram "patches" from the training audio files. Each file is divided into segments using

[1] `www.kaggle.com/c/the-icml-2013-bird-challenge`

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

simple power thresholding, and long segments are further divided into multiple segments of maximum duration one second. We then discard segments which do not contain much structure and are broadly flat, since these are not strongly discriminative and can match against any noise. This decision is based on the crest factor of pixels within the patch. Each segment is then normalised in magnitude (to unit L_2 norm).

Testing: To analyse an audio file, we apply MP to its spectrogram using the dictionary of time-frequency segments, via MPTK. This produces a list of activations associated with elements in the dictionary, which can be used to reconstruct the signal or for further analysis. In the present case the required output is a list of probabilities per species. We derive the probability for each species heuristically as proportional to the total energy that MP has allocated to activations associated with that species.

2. Results

At time of writing, the AUC score on the annotated development data is 70.3%, and the AUC score evaluated on $\frac{1}{3}$ of the full data (this is the method used on the Kaggle website to give results-in-progress) is 66.2%. This demonstrates that the approach generalises satisfacorily. However, results evaluated on the larger dataset at the close of the challenge give AUC of 53.8%, indicating scope for improvement.

3. Discussion

Note that this submission does not make use of any information other than the training and test audio provided by the challenge. In particular, for a real working system we would advocate the use of much larger audio collections to build the training data. However we wanted to explore how well the approach could make inferences from the provided data. Also, we do not perform adaptation of the system to the differing weather conditions, nor to the very different (reverberant) acoustic environment of the test audio.

In classical template-matching approaches, Dynamic Time Warping is commonly used to match templates against signals which have similar shape but with local differences in the length/speed of subregions. The MP approach we have deployed has no direct equivalent of this, which is a drawback when analysing sounds such as birdsong with natural variability in their production. The incorporation of such flexibility into sparse representations is an open topic; alternatively, novel signal representations may counter this issue.

We also note that template-matching approaches generally do not consider any time-sequencing of sounds at larger timescales, e.g. the grammatical sequencing of syllables. We are exploring how to combine syllable-by-syllable detection with methods which make inferences from temporal sequencing of birdsong, such as the Markov renewal process method we recently introduced (Stowell & Plumbley, accepted).

Future refinements of this approach could include more advanced approaches to creating the dictionary. For example, one might apply *dictionary learning*, a technique in sparse representations which directly optimises the dictionary so as to represent its inputs sparsely. We are also currently working with alternative signal representations which can provide detail of fine frequency modulations (Stowell et al., 2013).

Supported by EPSRC Fellowship EP/G007144/1.

References

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J. K., Hadley, A. S., and Betts, M. G. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *J Acoustical Society of America*, 131:4640–4650, 2012. doi: 10.1121/1.4707424.

Krstulovic, S. and Gribonval, R. MPTK: Matching Pursuit made tractable. In *Proc ICASSP 2006*, volume 3, pp. 496–499, Toulouse, France, May 2006.

Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. E. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2010. doi: 10.1109/JPROC.2009.2030345.

Stowell, D. and Plumbley, M. D. Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. Technical Report C4DM-TR-09-12, Centre for Digital Music, Queen Mary University of London, Aug 2010. URL `http://www.elec.qmul.ac.uk/digitalmusic/papers/2010/Stowell2010-C4DM-TR-09-12-birdsong.pdf`.

Stowell, D. and Plumbley, M. D. Segregating event streams and noise with a Markov renewal process model. *Journal of Machine Learning Research*, accepted. preprint arXiv:1211.2972.

Stowell, D., Muševič, S., Bonada, J., and Plumbley, M. D. Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering. In *Proc ICASSP*, 2013. preprint arXiv:1302.3642.

Bird identification from continuous audio recordings
The ICML 2013 Bird Challenge

Rafael Hernández Murcia RAFAEL.HERNANDEZ@UC3M.ES
Carlos III University of Madrid, Spain

Víctor Suárez Paniagua V.SUAREZ@ALUMNOS.UC3M.ES
Carlos III University of Madrid, Spain

Abstract

This working notes for the ICML 2013 Bird Challenge attempt to explain the method developed for identifying bird species in continuous audio recordings. Our approach uses a syllable segmentation procedure and calculates MFCC and Delta-MFCC as features. Then, we use a LDA projection and a neural network for classification.

1. Introduction

In this working notes we present a method to automatically decide whether a certain bird species sings in an audio clip recorded in a natural environment or not. The only training set used by this method were 35 clean audio clips provided, one for each bird species. First of all the audio is bandpass filtered keeping the frequencies corresponding to the birdsong spectra of most species. Then, we use a state-of-the-art syllable segmentation algorithm to select the relevant portions in the audio clip, eliminating most background noise in the process. For the windows corresponding to the syllables found, we extract the Mel-Frequency Cepstral Coefficients (MFCCs) and Delta-MFCCs, forming the set of available features. To further enhance temporal information, super-observations are constructed by concatenating variables over a sliding window of a fixed size. That is followed by a dimensionality reduction step, a projection using Linear Discriminant Analysis (LDA) is calculated. The final chosen classifier has been a bagger of artificial neural network, an approach which dampens the problem of local minima in common neural networks while still having an acceptable computational cost.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

2. Syllable Segmentation

Syllable segmentation is based on the amount of energy detected in parts of the spectrogram of the audio clips. In order to calculate the signal spectrograms, a kaiser-window ($\alpha = 8$) of size 256 samples is employed with an overlap of one-quarter of the window size.

Before applying syllable segmentation, we filter the signal of each audio clip with a 10th-order Butterworth band-pass filter. The lower normalized cutoff frequency is 0.03 and the higher is 0.5.

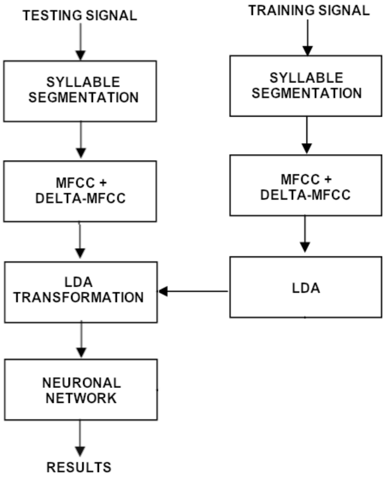


Figure 1. Block diagram of the entire process

The syllable segmentation algorithm, which is a simplified version of (1), can be described as:

1 Compute the spectrogram of the song segment using FFT, being $S(f,t)$ the spectrogram matrix where f is the frequency and t is time.

2 Find the maximum value in the spectrogram and the corresponding position in time. Then we move from the point of maximum forward and backward

while the spectrogram magnitude decreases until the decrease reaches a certain threshold U.

3 Observing the syllable extraction on the training set, the threshold was defined empirically.

4 Set to zero this part of the spectrogram to delete the segmented area.

5 We continue extracting syllables until the next maximum is below a certain threshold with respect to the global maximum of the clip.

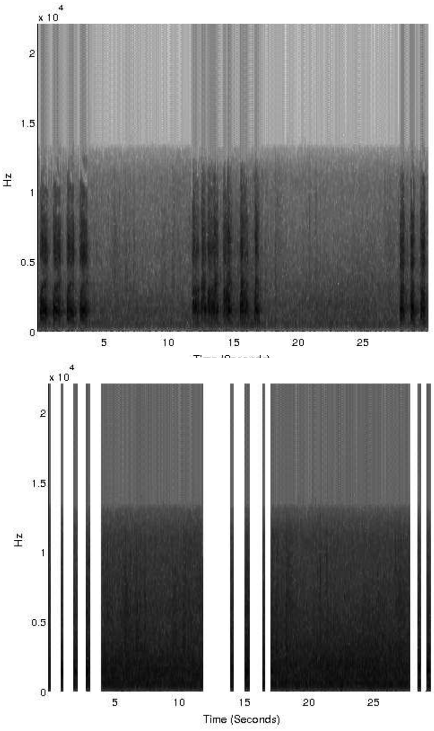


Figure 2. Spectogram and syllable segmentation

3. Feature Extracction

After obtaining the instants of time to the syllables of each bird, we extract the features using the MFCCs and Delta-MFCCs. Then, to exploit the temporal relationship between birds song of the same class, we group variables of adjacent samples into a vector with high dimensionality using a sliding window. The increase in the size of the dimensions is really significant.

4. Dimensionality Reduction

As we are working with high dimensionality vector, we decide to use a multiclass Linear Discriminant Analysis (LDA) for reducing the length of the vector. The projections obtained in the training phase are stored for using later in the test set, preventing the increase in dimensions suffered after applying the sliding window.

5. Classification

Even though we actually deal with 35 binary simultaneous classification problems, our method actually treats the problem as a set of many successive multiclass classification problems. That is, rather than directly answering “does bird X sing in clip Y?” we try to answer instead “which bird, if any, sings in instant T?”. By doing so, we will then combine a very high number of such answers per clip to actually solve the original binary classification decisions.

In order to solve the “instantaneous” multiclass classification problem, many different classifiers could be applied to this problem but, in practice, the high number of observations makes neural networks highly attractive for this task. We still need to handle the severe problem of convergence to local minima which appears in that kind of classifiers. Our approach to solve that has been to apply bagging to introduce diversity in both training set and initialization.

Given a pre-filtered test clip, we find the syllables according to the syllable extraction algorithm and then extract features for those syllables. Also, this feature vectors are projected using LDA. We use the soft output of the neural network bagger for each class. Because birds can sing at any given time instant in the clip, we thought that it was much more sensible to make a decision about the presence or absence of a birdsong during a test clip using the maximum score achieved. In this way, if a bird sings very clearly but only during a short time, we would still detect it.

Finally, before the end of the competition the score on the Public Leaderboard was 0.743, so the reduction in testing performance on the final test set (a score of 0.6954 in the Private Leaderboard) was not as high as the reduction suffered by other competitors.

References

[1] Harma, A., ”Automatic identification of bird species based on sinusoidal modeling of syllables,” Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03). 2003 IEEE International Conference on , vol.5, no., pp.V,545-8 vol.5, 6-10 April 2003

[2] Lee,Chang-hsing, Lee,Yeuan-kuen & Huang,Ren zhuang. Automatic Recognition of Bird Songs Using Cepstral Coefficients, Journal of Information Technology and Applications. Vol. 1 No. 1, May, 2006, pp. 17-23.

ICML 2013 Bird Challenge: 3-stage Mixed Supervised/Unsupervised Classification

Charles J. Turner CJTURNER@UCDAVIS.EDU
Academic Technology Services, UC Davis, Davis CA 95616 USA

Jennifer G. Turner

Abstract

This technical report details our approach to solving the ICML 2013 Bird Challenge, in which the goal was to predict which of 35 different species of birds were singing in a given audio recording. Our approach was based on a pipeline of unsupervised clustering, supervised classification for feature extraction, with a final supervised classification for bird identification.

1. Summary of Methods

1.1. Input Data

We worked exclusively with the pre-extracted MFCC (Mel Frequency Cepstral Coefficient) features extracted by the contest sponsors on both the training and test sets. We focused primarily on the 9 species represented in the first three files of the test data (6, 8, 12, 16, 18, 27, 32, 33, 34), plus two additional species deemed to be common and detectable (9, 22).

1.2. Processing Stages

Our method was based on a two-stage process of gathering accurate training data for use in the final stage of classification of all recordings. The main steps included:

1. For each species, the pre-extracted MFCC features (16 per sample) in the training data were passed through a K-means clustering algorithm with a small value of K. The first cepstral coefficient was used to isolate bird from background in each of the training files.
2. A three-layer backpropagation network was trained on the data from the training files for the 11 species plus all background data isolated from all of the training files. Each of the cepstral feature vectors was used as a training sample. There was no aggregation of adjacent vectors.
3. The first three recordings of the test set (also considered training data) were passed through the supervised classification algorithm to identify more representative samples of the 11 species and background. This step in the processing compensated for the lack of detailed annotation of the first three test records.
4. A second three-layer backpropagation network was trained on the combined training data from the training set and the classification results of processing the first three recordings in the test set.
5. Finally, the entire test set was passed through the second classifier, accumulating network activation values for each of the 11 species and background for each test recording. Our final output was based on a summation of these output activations for each species.

2. Discussion

We chose to work exclusively with the pre-extracted MFCC feature vectors as provided by the sponsors.

The upside of this approach was its simplicity - all records were treated the same, and computational load was kept to a minimum. As can be seen in Figure 1, there is some separability between species using just the cepstral features.

The downside of our approach is that it eliminates a great deal of higher-order information that can be obtained from both intra- and inter-call patterns, as could be obtained from a two dimensional spectrogram. Figure 2 shows a typical result of using K-means

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

to cluster the cepstral vectors of the 35 species into 40 clusters. Species are represented by columns; clusters by rows. Rows with multiple dark reddish squares indicate clusters that contain multiple species. This remained true in some cases even with 120 clusters, illustrating the difficulty of separating certain species based solely on the full set of MFCC features. One promising avenue that we did not have time to explore was the splitting of some species into their different calls.

Another serious limitation of our approach was difficulty in verifying the results of processing steps that identified individual cepstral records as belonging to a particular species. Working with such fine detail and without the inverse transforms prevented us from doing any form of audio verification of our processing. We were restricted to visually inspecting time series plots.

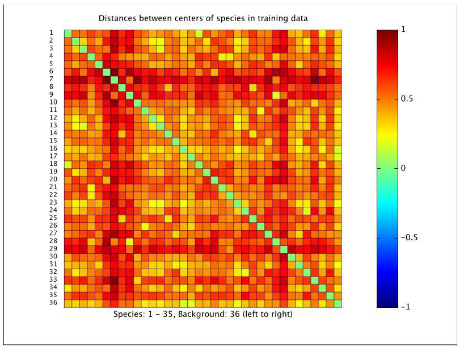


Figure 1. Distance matrix of centroids of MFCC vectors for each species in the training set with the addition of background noise from the first three test records. The species are in order by number, with background noise appearing in row and column 36. The diagonal elements represent a distance of zero. Darker reddish colors represent greater inter-cluster distances.

Acknowledgments

We would like to thank the sponsors of this competition for their dedication to the study of wildlife and for taking the time to provide the ground truth necessary to support machine-learning studies.

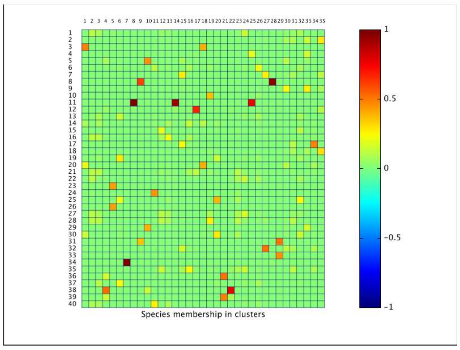
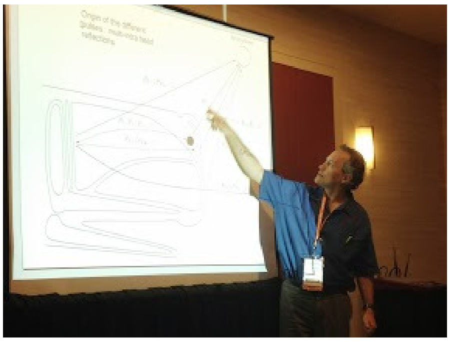
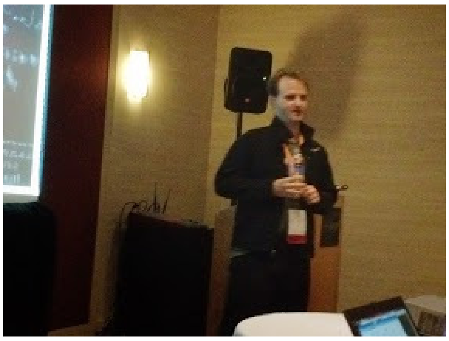
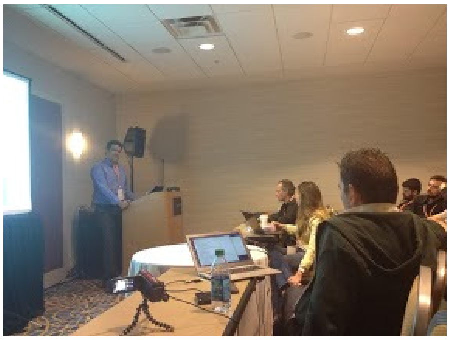
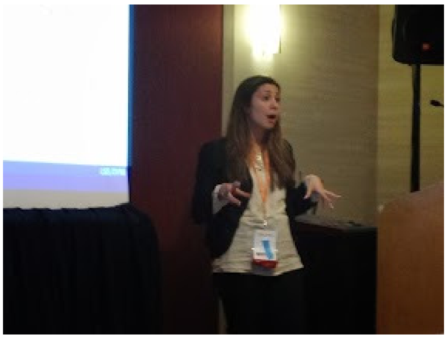


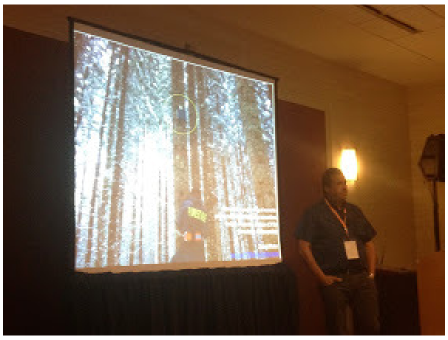
Figure 2. Results of using K-means to cluster the cepstral vectors of the 35 species into 40 clusters. Each species appears in one column and each cluster in one row. Green boxes represent zero membership of species in a particular cluster. Within each individual species column, colors represent percentage of the samples belonging to a cluster. So for example, a very dark reddish square indicates that the majority of a species' cepstral records belong to this cluster, whereas a species with multiple lighter squares indicates greater variety and splitting of the species calls.

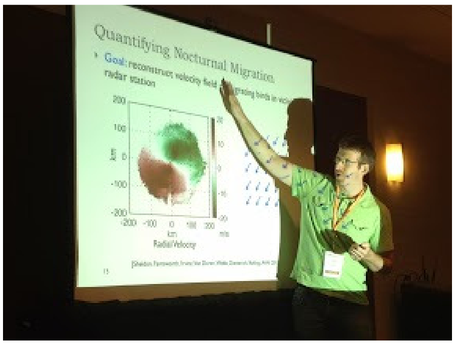

Origin of the different pulses : multi-intra head reflections










Quantifying Nocturnal Migration
Goal: reconstruct velocity field of migrating birds in vicinity of radar station
km
Radial Velocity

ISBN : 979-10-90821-02-6