Fish migration monitoring from audio detection with CNNs

Patrice Guyot, Fanny Alix, Thomas Guerin, Elie Lambeaux, Alexis Rotureau

Fish migration monitoring from audio detection with CNNs

Patrice Guyot
patrice.guyot@mines-ales.fr
EuroMov Digital Health in Motion,
Univ Montpellier, IMT Mines Ales,
Ales
France

Fanny Alix
f.alix@migrateursrhonemediterranee.org
Migrateurs Rhône Méditerranée,
Arles
France

Thomas Guerin
Elie Lambeaux
Alexis Rotureau
firstname.lastname@mines-ales.org

Figure 1: On spring nights, migratory fish produce splashes in rivers during spawning.

ABSTRACT

The monitoring of migratory fish is essential to evaluate the state of the fish population in freshwater and follow its evolution. During spawning in rivers, some species of alosa produce a characteristic splash sound, called "bull", that enables to perceive their presence. Stakeholders involved in the rehabilitation of freshwater ecosystems rely on staff to aurally count the bulls during spring nights and then estimate the alosa population in different sites. In order to reduce the human costs and expand the scope of study, we propose a deep learning approach for audio event detection from recordings made from the river banks. Two different models of Convolutional Neural Networks (CNNs), namely AlexNet and VGG-16, have been tested. Encouraging results enable us to aim for a semi-automatized and production oriented implementation.

CCS CONCEPTS

• **Information systems → Speech / audio search**; • **Computing methodologies → Neural networks**; • **Applied computing →** *Life and medical sciences.*

AM '21, September 1–3, 2021, virtual/Trento, Italy

ACM ISBN 978-1-4503-8569-5/21/09…$15.00
https://doi.org/10.1145/3478384.3478393

KEYWORDS

bioacoustics, water, deep learning, sound

ACM Reference Format:
Patrice Guyot, Fanny Alix, Thomas Guerin, Elie Lambeaux, and Alexis Rotureau. 2021. Fish migration monitoring from audio detection with CNNs. In *Audio Mostly 2021 (AM '21), September 1–3, 2021, virtual/Trento, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3478384.3478393

1 INTRODUCTION

In a context of global decline of wildlife population, numerous efforts aim at preserving biological diversity as well as conserving specific species. To be valued, and eventually validated, these actions rely substantially on measures of abundance and quantification of rate of change. Passive acoustic monitoring is a non-invasive way of reporting community information at the scale of the species (in the framework of bioacoustics [14]), but also at an higher level of organization, that is to say at an inter-species scale (in the ecoacoustics field [17]). Among different kind of ecosystems, freshwater produces a range of micro-habitats where terrestrial and aquatic worlds bleed gradually into each other. A broad diversity of organisms can be found in this environment, such as birds, frogs, fish and insects, that produce a rich and varied sonic environment.

Living mostly in the sea, migratory fish swim upstream in rivers at spring to spawn. This behavior is shared by different species, among them: *Alosa agone* in the Atlantic ocean, *Alosa fallax* (also known as *twait shad*) and *Alosa alosa* (also known as *allis shad*) in the Mediterranean Sea. We will refer thereafter to those species as *alosa.*

In different areas, such as the Rhone basin, an important number of infrastructures built in mid-20th century, such as power-plants and dam, hinder the migration. The alosa population has declined across Europe since the mid-20th century. Accordingly, it is a protected species since the Berne Convention of 1979.

Recent structures, such as sluices and fish passes, have been set up since then to create upstream and downstream fish passages and ensure longitudinal connectivity. The monitoring of the annual upstream migration of alosa assesses the effectiveness of these structures. Furthermore, it provides information on the abundance of a vulnerable population threatened by fishing, pollution and the deterioration of spawning grounds. Nonetheless, the monitoring of one of the biggest species of these fresh streams provides information about the ability of the overall underwater population to move upstream and downstream.

Surprisingly, aside various tools for fish detection based on image analyze such as photo traps, the migration of alosa is mostly monitored by sound. During spawning, at night, male and female come close to the surface and, half immersed, hit strongly the surface with their caudal fin while turning around each other (see Figure 1). These movements oxygenate the water and stimulate the development of eggs. It also produces a clearly audible and characteristic *splash sound* which lasts a few seconds, that is called "bull" [9]. In many locations, an important effort is provided by stakeholders during spring to aurally count these bulls throughout the night from different sites on the river banks.

The manual counting of bulls to monitor alosa migration has a significant cost. It currently involves two persons at each spot throughout the night. Hence, the automatic detection of these audio events through field recordings is a crucial issue. The automation of the migration monitoring would enlarge the study area thanks to more counting spots and more objective processes. This data would weigh on the policy of the rehabilitation of rivers for biodiversity conservation.

In this paper, we propose a bull detection approach with Convolutional Neural Networks (CNNs). We aim for a real implementation where detected bulls will be validated by human listing. Section 2 presents previous works in the larger scope of audio event detection. The audio material of this study is detailed in section 3. The last sections present our experiments and results.

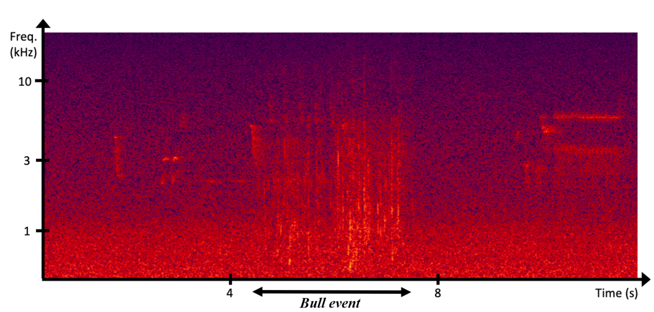

Figure 2: Spectrogram of a bull audio event (among other sounds like river stream) made by a fish during spawning.

2 RELATED WORKS

The task of automatic bull detection has been addressed in previous works in the last decade, mostly via a typical approach of shallow classification used in the 2010's for audio detection: MFCC features and GMM classifier [2, 3]. Other contributions aspired to detect water sounds, for instance in the context of activity recognition for elderly assistance [5], or related to the question of their auditory perception [4].

Aside these works, numerous studies address the broader issue of Audio Event detection (AED). This research area usually involves researchers from audio communities with related works in speech and Music Information Retrieval. AED usually refers to field recordings in uncontrolled environment [13], where many kinds of sound events may occur and overlap.

Significant progresses have been made in AED trough deep learning approaches. Research is stimulated by challenges such as the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [12]. In that context, many contributions deal with time-frequency representations of audio signals, and benefit from major advances in computer vision.

In the scope of bull detection from river bank recordings, similar tasks can be found in the field of bioacoustics. In the bird detection task of the DCASE 2018 challenge, the best results were achieved with a CNN approach and data augmentation [10]. Convolutional recurrent neural networks have also been proposed for bird audio detection [1], achieving very good results. In line with these results, this paper describes a first attempt to use CNNs for bull detection in a freshwater environment.

3 DATA

We focus on splash sounds, called bulls, made by fish at the surface with their caudal fin. They are broad-frequency noises with transients. The Figure 2 shows an example of this audio event. Bulls often overlap with other sounds of the freshwater environment such as bird vocalizations, and stream sounds.

Our dataset is composed of 20 recordings (mono, 16 bits, sr=44.1k) for a total duration of 68 hours (see Table 1). These files have been recorded at night from river banks in different parts of France, mostly from the Rhone Basin (Ceze and Vidourle rivers), and from the ocean side (Charente and Loire rivers). The recordings last in average between 3 and 4 hours. They have been manually labeled in bull events, resulting in 709 bull events in total, whose average duration equals to 4.5 seconds (sd: 2.2s).

Year	River / site	# Rec.	Duration	# Bulls
2009	Charente	2	46m	94
2012	Loire	6	30h59	73
2013	Charente	3	2h	251
2014	Ceze	2	3h54	208
2016	Vidourle	1	1h11	6
2017	Ceze & Vidourle	5	24h06	60
2018	Ceze	1	5h38	3

Table 1: Sites, number of recordings, total durations, and number of annotated bulls for each year.

4 METHODS

4.1 Pre-processing

We use as inputs $time \times frequency$ representations of the audio content. Recordings are split into 15 seconds segments with an overlapping of 5 seconds. This configuration enables to include most of the whole bull event in at least one segment. Each segment is labeled as *bull* if it contains a part of a bull event, and *no-bull* otherwise (see Figure 3). We compute a Mel-spectrogram from each audio segment (n_fft=4096, hop_length=1024, f_max=22050) and resize the output (to 128 x 646 bins). In that scope, our task becomes a binary classification of images. However the resulting dataset is unbalanced and contains about 45k segments labeled as *no-bull* but only 2.1k labeled as *bull*.

Figure 3: Audio segmentation with overlapping. In this example, the first four segments are labeled as *bull*.

4.2 Models

From the state-of-the-art CNNs AlexNet [8] and the more complex VGG-16 [16], two models, that we will call by extension AlexNet and VGG-16, have been implemented. We redesigned these networks in order to build two models adapted to our task (see Figure 4). In these two models, the first layer is used to normalize the input of the model. Its is followed by a convolutional layer to change the number of channels from 1 to 3, in line with the original inputs of AlexNet and VGG-16. In the second model, the third layer (group normalization layer) layer prepares the input data of the VGG-16 pretrained model and avoids memory issues.

In order to solve the problem of unbalanced data, our strategy is to balance out the losses coming from labeled segments so as to bring us back to a situation where the data would be perfectly balanced. To this end, we determine two weights w_{bull} and w_{no_bull} related to the proportion p of the labels *bull* and *no_bull* (in our case $p_{bull} \approx 5$ % and $p_{no\_bull} \approx 95$ %) in the audio segments:

$$w_{bull} \times p_{bull} = w_{no\text{-}bull} \times p_{no\text{-}bull} \tag{1}$$

To avoid the phenomenon of vanishing gradients, which is often caused by very high losses combined to certain activation functions, we use a second condition $w_{bull} + w_{no\text{-}bull} = 1$. We finally obtain:

$$w_{bull} = 1 - p_{bull} \quad \text{and} \quad w_{no\text{-}bull} = p_{bull} \tag{2}$$

4.3 Metric

Finally, as we strive for a semi-automatized approach where detected bulls will be validated by human hearing, our goal is to decrease the number of missed bulls (i.e. false negatives) while reducing the amount of audio segments that need to be listened by humans (predicted bulls). As our dataset is unbalanced, we choose to use as metric the average recall, which is defined as:

$$Average\ recall = (bull\ recall + no\text{-}bull\ recall)/2 \tag{3}$$

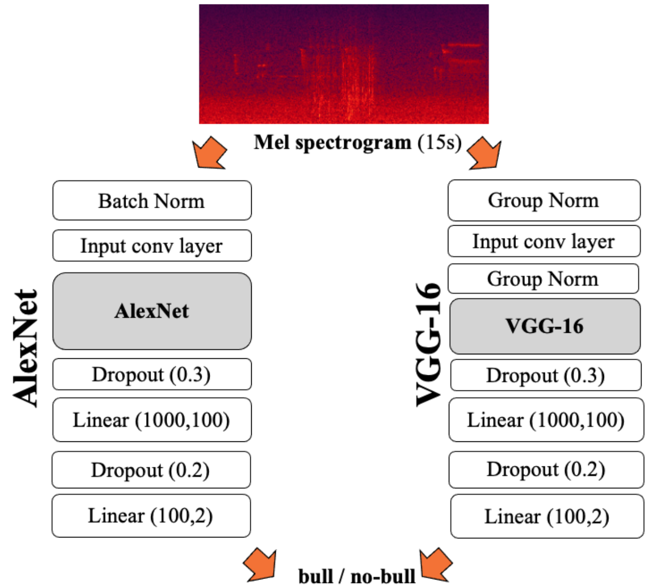

Figure 4: Architecture of the two implemented CNNs from the original AlexNet and VGG-16 models.

5 EXPERIMENTS

We implemented the pre-processing steps with the python libraries *librosa*[1] and *torchaudio* and our models with *PyTorch* [7]. We used pretrained AlexNet and VGG-16 versions, and then trained our models entirely (without freezing) on GPU (*Nvidia GeForce GTX 1080 Ti*).

Our dataset has been separated into training, validation and test sets. The sites between validation/test sets and training sets are different so as to extend our models to other sites.

We used a grid search strategy to test different hyper-parameters from the following values:

- audio segments duration: 10 and 15 seconds
- batch size: 32, 128, 512, 1024, 2028 and 8196
- learning rate: 10^{-3} and 10^{-4} (with the Adam optimizer)
- number of final dense layers: 1 and 2
- image pre-processing: input normalization, input standardization, batch normalization and group normalization [18]

The best results were obtained with the following configurations: audio segments of 15 seconds, batch size of 128 inputs, learning rate of 10^{-4}, two final dense layers at the end of the pipeline, as well as normalization by batch for AlexNet and by group for VGG-16. Moreover, the number of epochs has been chosen according to the best score on the validation set (5 epochs for AlexNet and 6 epochs for VGG-16). Furthermore, to optimize memory usage, we implemented the following methods:

- Data generation on-the-fly to load inputs only when they are required in the training step.
- Gradient accumulation to reduce the memory usage required by large batch sizes. The loss and the gradients are calculated

[1] https://librosa.org/

after each mini-batch but weights are updated every batch (i.e. less frequently).

- Group normalization layers, which require less memory usage than batch normalization layers, and remain quite performing for image recognition models [18].

Table 2 shows our results with those best configurations. As we can see, the model VGG-16 obtains better results than AlexNet, even if its precision is lower. Our experiments finally led to an average recall of 89.7%. This score is really interesting owing to the fact that most of the segments labeled as *bull* are detected (93.2% of bull recall) as we aimed to.

	AlexNet	VGG-16
Precision	41.4	21.3
Recall	81.4	93.2
Average recall	88.4	**89.7**

Table 2: Results of the bull detection on the test set.

.

In order to analyze these results in details, we introduce here a confusion matrix on the test set from the VGG-16 predictions that gave the best score (see Table 3). According to that table, 1550 segments were predicted as bulls by the model, whereas the whole test set includes 9166 segments. In a real implementation, this will lead to a reduction of human costs, in terms of listening duration, by a factor of approximately six. If we look over the missed events, 24 audio segments labeled as *bull* were not detected in this test set. However, as there is an important overlap between segments, some missed bulls have been detected in other adjacent segments (see Figure 3). If we consider bull events larger than 2 seconds (i.e. clearly audible events without contentious), only one bull event was totally missed on the test set. This result is very encouraging for a real implementation that would involve human listening of the detected segments.

	Bull predicted	No-bull predicted
Bull segment	**330**	24
No-bull segment	1220	**7592**

Table 3: Confusion matrix on the test set with the VGG-16 model. 330 segments labeled as *bull* are detected.

6 CONCLUSION

In this paper, we presented a deep learning approach for bull detection, in the context of migratory fish monitoring. We implemented two models based on the state-of-the-art CNNs AlexNet and VGG-16. As we aim for a semi-automatized approach, we tuned our models in order to minimize the number of missed bulls. Our method reaches almost 90% of average recall. This result is very encouraging for a real implementation of this semi-automatic approach, that would enable monitoring of more freshwater sites for a smaller human cost and a limited number of missed events.

This first implementation of a deep learning approach could be improved in the future. Regarding the data, we will collect new data each year in more sites to enlarge our dataset. We also intend to use an approach of data augmentation that proved to be effective in a bioacoustics context [10]. We could use effects as masking, shifting and stretching on the time and frequency dimensions of the Mel-spectrograms, and add different background noises, throughout the on-the-fly data generation phase.

Regarding the models, we may improve the processing of the temporal dimension of the events, by using Convolutional Recurrent Neural Networks [11] and/or attention [15]. Finally, we will also consider the use of strong labels to characterize the audio events, with using the information of the start and the end of an event, instead of a binary annotation of segments (i.e. weak labels for each segment) [6].

ACKNOWLEDGMENTS

The research presented in this article was funded by the French Association Migrateurs Rhône Méditerranée. This work was partially implemented through student projects at IMT Mines Ales. External data were provided by LOGRAMI and EPTB Charente. We thank Pierre Campton for his support, as well as Julien Denozi and Chayma Mefteh for their previous works.

REFERENCES

[1] Emre Cakir, Sharath Adavanne, Giambattista Parascandolo, Konstantinos Drossos, and Tuomas Virtanen. 2017. Convolutional recurrent neural networks for bird audio detection. In *Proc. EUSIPCO.* 1744–1748.

[2] Daniel Diep, Hervé Nonon, Isabelle Marc, Jonathan Delhom, and Frédéric Roure. 2013. Acoustic counting and monitoring of shad fish populations. In *AmiBio Workshop: Recent Progress in Computational Bioacoustics for Assessing Biodiversity.*

[3] Daniel Diep, Hervé Nonon, Isabelle Marc, Isabelle Lebel, and Frédéric Roure. 2016. Automatic acoustic recognition of shad splashing using a smartphone. *Aquatic Living Resources* 29, 2 (2016), 204.

[4] Patrice Guyot, Olivier Houix, Nicolas Misdariis, Patrick Susini, Julien Pinquier, and Régine André-Obrecht. 2017. Identification of categories of liquid sounds. *The Journal of the Acoustical Society of America* 142, 2 (2017), 878–889.

[5] Patrice Guyot, Julien Pinquier, and Régine André-Obrecht. 2013. Water sound recognition based on physical models. In *Proc. ICASSP.* IEEE, 793–797.

[6] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The Benefit of Temporally-Strong Labels in Audio Event Classification. In *Proc. ICASSP.* IEEE, 366–370.

[7] Nikhil Ketkar. 2017. Introduction to pytorch. In *Deep learning with python.* Springer, 195–208.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[9] MC Langkau, D Clavé, MB Schmidt, and J Borcherding. 2016. Spawning behaviour of Allis shad Alosa alosa: new insights based on imaging sonar data. *Journal of fish biology* 88, 6 (2016), 2263–2274.

[10] Mario Lasseck. 2018. Acoustic bird detection with deep convolutional neural networks. In *Proc. DCASE.* 143–147.

[11] Hyungui Lim, Jeongsoo Park, and Yoonchang Han. 2017. Rare sound event detection using 1D convolutional recurrent neural networks. In *Proc. DCASE.* 80–84.

[12] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2019. Sound event detection in the DCASE 2017 challenge. *IEEE/ACM TASLP* 27, 6 (2019), 992–1006.

[13] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *Proc. EUSIPCO.* 1267–1271.

[14] Martin K Obrist, Gianni Pavan, Jérôme Sueur, Klaus Riede, Diego Llusia, and Rafael Márquez. 2010. Bioacoustics approaches in biodiversity inventories. *Abc Taxa* 8 (2010), 68–99.

[15] Yu-Han Shen, Ke-Xin He, and Wei-Qiang Zhang. 2019. Learning How to Listen: A Temporal-Frequential Attention Model for Sound Event Detection. *Proc. Interspeech 2019* (2019), 2563–2567.

[16] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[17] Jérôme Sueur and Almo Farina. 2015. Ecoacoustics: the ecological investigation and interpretation of environmental sound. *Biosemiotics* 8, 3 (2015), 493–502.

[18] Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proc. ECCV.* 3–19.