forum acusticum 2023

CRNN-BASED SPLASH AUDIO EVENT DETECTION FOR FISH MONITORING

Ardit Sota[1*] **Patrice Guyot**[2] **Fanny Alix**[3]
Klevisa Xhika[1]
[1] IMT Mines Ales. Ales, France.
[2] EuroMov Digital Health in Motion Univ. Montpellier, IMT Mines Ales. Ales, France.
[3] Migrateurs Rhône Méditerranée. Arles, France.

ABSTRACT

Monitoring migratory fish species provides a good indicator for rivers' health. Migratory fish as alosa (*Alosa fallax* also known as twait shad), swims up the rivers to reproduce if the dam infrastructure allows it. During spawning, some species of alosa produce during a few seconds a characteristic splash sound, that enables them 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 at different sites. In order to reduce the human costs and expand the scope of the analysis, we propose a deep learning approach for audio event detection from dozens of GB of audio files recorded from the riverbanks. An automatic detection system consisting of a Recurrent Convolutional Neural Network (CRNN) is presented. Encouraging results enable us to aim for an automated implementation on sites.

Keywords: *bioacoustics, deep learning, freshwater, audio event detection*

1. INTRODUCTION

In recent years, the conservation of biological diversity and specific species has become a global priority due to the decline in wildlife populations [1]. To support these initiatives, it is crucial to accurately measure species abundance and quantify the rate of change in order to assess their conservation status. In the field of bioacoustics [2] and ecoacoustics [3], passive acoustic monitoring has emerged as a non-intrusive method for gathering community-level information. Freshwater ecosystems, characterized by the gradual transition between terrestrial and aquatic environments, host a diverse array of organisms, including birds, frogs, fish, and insects, creating a rich and complex acoustic landscape.

One species of particular interest is the migratory fish known as alosa (*Alosa fallax* or twait shad), which breeds in rivers after spending most of its life in the sea. However, the construction of infrastructure such as power plants and dams since the mid-20th century has impeded the movement of alosa and contributed to its declining population throughout Europe. Consequently, alosa has been protected under the Berne Convention since 1979, and efforts have been made to improve longitudinal continuity and establish fish passes and sluices to facilitate upstream and downstream migration.

Monitoring the yearly upstream movement of alosa serves as an indicator of the efficiency of these infrastructure projects and provides insights into the abundance of this species, which is susceptible to overfishing, pollution, and the degradation of spawning habitats. Interestingly, the migration of alosa, one of the largest species in these freshwater streams, is primarily monitored through sound rather than visual observations. During spawning, male and female fish engage in a distinctive surface behavior, where they revolve around each other while making vigorous splashing sounds known as "bulls" [4]. These acoustic events are indicative of the fish's presence and breeding activity (see Figure 1).

*__Corresponding author__: ardit.sota@mines-ales.org.

Currently, manual counting of these acoustic events is labor-intensive and costly, requiring dedicated personnel to spend nights along the riverbanks. To address this challenge and enable automation, we propose a Convolutional Recurrent Neural Network (CRNNs) method for the automated detection of bull sounds in field recordings. By developing an efficient and accurate model, we aim to contribute to the automation of alosa migration monitoring, enabling the expansion of monitoring points and the implementation of more objective procedures. Ultimately, this automation has the potential to inform river rehabilitation policies and support the preservation of biodiversity.

In this paper, we present the details of our CRNNs approach for bull sound detection, discussing the architecture, training methodology, and performance evaluation.

Figure 1: Splashes called "bull" form migratory fish during spawning in rivers.

2. RELATED WORKS

In previous studies focused on alosa fish monitoring, limited attention has been given to the automated detection of the fish's specific spawning behaviors [5, 6]. In the field of automatic Audio Event Detection (AED), various approaches have been proposed utilizing supervised classifiers such as Gaussian mixture model hidden Markov models, fully connected networks, convolutional neural networks (CNN), and recurrent neural networks (RNN). For instance, in the context of bird audio detection (BAD) [7], a method was introduced for identifying bird sounds in audio recordings. This approach utilized CNNs to extract higher-level features and RNNs to capture longer-term temporal contexts within the audio signals. Similar approaches combining CNNs and RNNs have been successful in speech recognition (ASR) [8] and music classification [9]. In a previous work focused on fish migration monitoring using audio detection, deep learning models were employed [10]. Specifically, two different CNN models, namely AlexNet and VGG-16, were implemented and tested. Building upon these findings, this paper presents a novel approach utilizing CRNNs for bull sound detection in a freshwater environment. By leveraging the combined power of convolutional and recurrent neural networks, our method aims to capture both the spectral features and temporal dependencies inherent in the bull sounds produced during alosa fish spawning.

3. MATERIALS AND METHODS

3.1 Data

Audio recordings are made 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). Our dataset is composed of 73 recordings (mono, 16 bits, sr=44.1k) for a total duration of 354 hours.

3.2 Preprocessing

The audio recordings are split into 15 sec segments, with 5 sec overlap (see Figure 2). Each segment is labeled as bull if it contains a part of bull event and no bull otherwise. It results that the dataset is imbalanced, with 179k segments labeled as *no-bull* and 5k labeled as *bull*.

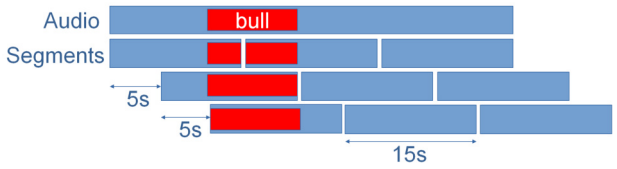

Figure 2: Audio segmentation with 5 sec overlapping. Here only four segments are labeled as bull.

3.3 Model

Inputs of the model are mel-spectrogram [1] that are resized into 128x646 images. One mel-spectrogram is time x frequency representation of a 15 sec segment of audio. In

[1] n_fft=4096 (Fast Fourier Transform size), hop_length=1024 (number of samples between consecutive frames), f_max=22050 (maximum frequency)

this way our sound event detection task becomes image classification task.

The architecture of the deep learning model was designed by combining CNN and RNN networks to create a CRNN. For CNN we used a pretrained model called VGG-16 [11] to extract complex features from images. A Bidirectional-Long-Short-Term-Memory (Bi-LSTM) layer is used for the recurrent part.

Figure 3: Architecture of CRNN model by combining VGG16 model and Bi-LSTM recurrent layer.

The proposed hybrid network for bull detection is illustrated in Figure 3. The network architecture is based on VGG-16 and Bi-LSTM layer. The first three layers are used to change the input shape of mel-spectrogram so that it is consistent with the VGG-16's initial inputs. To adapt it for our task, we modified the pretrained VGG-16 by removing the last dense layers. The modified model consists of 17 layers: 13 convolutional layers and 4 pooling layers. After VGG-16 a max-pooling layer, with 2x2 kernels, is employed to reduce the dimensions of the input image. The output shape is determined to be (8, 20, 4, 512), where 8 represents the batch size used for training and testing the model. In the latter part of the architecture, the feature map is passed to a bidirectional LSTM layer to extract temporal information. After the Bi-LSTM layer, the output is flattened and passed through a linear layer. A dropout layer is applied to prevent overfitting. Finally, another linear layer with 100 input features and 2 output features is used to produce the final classification scores.

3.4 Metrics

We use the classical precision (pre), recall (rec) and F1 score as evaluation metrics [12]. Moreover, as we worked toward a semi-automated method where detected bulls will be verified by human hearing, our main objective is to lower the number of missed bulls, while reducing the amount of audio segments that need to be listened by humans (predicted bulls). Therefore, we introduce a new metric that we call *Weighted Harmonic F1* (WF1) to favour recall over precision.

$$\begin{aligned} F1\_score &= 2 * (pre * rec)/(pre + rec) \\ WF1 &= (a + b)/((a/pre) + (b/rec)) \end{aligned} \tag{1}$$

We used a = 0.2 and b = 0.8 for the weighted harmonic coefficients, in order to give the recall greater weight.

4. EXPERIMENT

We conducted a series of experiments to evaluate the performance of the CRNN architecture for bull detection. Our experiments were implemented using the Pytorch framework [13], which provided a flexible and efficient platform for training and evaluating deep learning models. To process the audio data, we utilized the librosa library, a popular choice for audio processing tasks. Our dataset has been separated into training and test sets with a ratio of 90:10, respectively. We used a pretrained VGG-16 model. We adopted the Adam optimizer with a learning rate of 0.0001 and a cross-entropy loss function to optimize the CRNN model. The model was trained during 6 epochs.

4.1 Results

The evaluation of our model yielded the following results (Table 1). The recall score (63.74%), indicates that almost two third of the targeted bull events have been recovered by our model. The precision score (43.95%), indicating that the model had a lower ability to limit false positives. The F1 score (52.02%) combines the model's precision and recall values and provides an overall measure of its performance. Furthermore, we assessed the model's performance using our own metric, the weighted F1 score (58.47%) defined earlier, that favor recall over precision.

Table 2 shows the confusion matrix which provides a detailed overview of the model's predictions based on the actual labels.

These results suggest that further optimization and refinement of the model are required to improve its overall

accuracy and generalisability. Future listening to the results will enable us to identify the strengths and weaknesses of in-context prediction.

Table 1: Results of the bull detection on the test set for each metric respectively.

Metric\Model	CRNN
Recall	63.74
Precision	43.95
F1-score	52.02
Weighted-harmonic-F1	**58.47**

Table 2: Confusion matrix on the test set with CRNN model.

Actual\Predicted	Bull	No-bull
Bull	195	111
No-bull	249	17299

5. CONCLUSION

In this work, we proposed a convolutional recurrent neural network for audio event detection in the context of monitoring migratory fish called Alosa. By leveraging the power of PyTorch and the librosa library, we developed and evaluated a CRNN model on sound event dataset. Through our experiments, we achieved an accuracy of 58.47% for bull event detection. We conducted our confusion matrix, which provided insights into the model's performance for bull event. Even though the results are not as we expected, we can say that those are acceptable considering the challenges that we faced with a very imbalanced dataset (2.8% positive samples), also various background noises in field recordings (wind, rain, vocalization of birds and frogs, etc.) that leads in more false positive bulls detected.

To improve our model, we may consider in the future the use of strong labels for segments with considering the start and end of each bull instead of labels that tag all the segment [14], or incorporating attention mechanism [15] for CRNN network. Balancing the data by randomly selecting an equal number of negative samples as positive ones can also be beneficial. This ensures that the model receives sufficient exposure to both classes, preventing biases and promoting better generalization. Finally the achieved results and insights pave the way for further advancements in this field.

6. ACKNOWLEDGMENTS

We want to thank the French Association Migrateurs Rhône Méditerranée (MRM) for funding this research also their interns for the collection of data.

7. REFERENCES

[1] S. Díaz, J. Settele, E. S. Brondízio E.S., et al, *IPBES: Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. 2019.

[2] M. K Obrist, G. Pavan, J. Sueur, K. Riede, D. Llusia and R. Márquez, "Bioacoustics approaches," in *Biodiversity inventories*, pp. 206–212, 2010.

[3] J. Sueur and A. Farina, *Ecoacoustics: the ecological investigation and interpretation of environmental sound*. 2015.

[4] M. Langkau, D. Clavé, M. Schmidt, and J. Borcherding, *Spawning behaviour of Allis shad Alosa alosa: new insights based on imaging sonar data*. 2016.

[5] D. Diep, H. Nonon, I. Marc, J. Delhom, and F. Roure, *Acoustic counting and monitoring of shad fish populations*. 2013.

[6] D. Diep, H. Nonon, I. Marc, I. Lebel, and F. Roure, *Automatic acoustic recognition of shad splashing using a smartphone*. 2016.

[7] E. C¸ akır, S. Adavanne, G. Parascandolo, K. Drossos and T. Virtanen, *Convolutional recurrent neural network for bird audio detection*. 2017.

[8] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, *Convolutional, long short-term memory, fully connected deep neural networks*. 2015.

[9] K. Choi, G. Fazekas, M. Sandler, and K. Cho, *Convolutional recurrent neural networks for music classification*. 2016.

[10] P. Guyot, F. Alix, T. Guerin, E. Lambeaux, and A. Rotureau, "Fish migration monitoring from audio detection with cnns," in *Proceedings of the 16th International Audio Mostly Conference*, pp. 244–247, 2021.

[11] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* 2014.

[12] H. Dalianis, "Evaluation metrics and evaluation," in *Clinical Text Mining. Springer, Cham.*, pp. 45–53, 2018.

[13] N. Ketkar, "Introduction to pytorch," in *Deep learning with python*, p. 195–208, 2017.

[14] S. Hershey, D. PW Ellis, E. Fonseca, A. Jansen, C. Liu, R. Channing Moore, and M. Plakal, "The benefit of temporally-strong labels," in *Audio Event Classification*, p. 366–370, 2021.

[15] Y. Shen, K. He, and W. Zhang, *Learning How to Listen: A Temporal-Frequential Attention Model for Sound Event Detection*. 2019.