MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Deep Learning For Natural Sound Classification

Diez Gaspon, Itxasne[1], Saratxaga,Ibon[2]
University of the Basque Country. UPV/EHU
Aholab Signal Processing Laboratory.
Faculty of Engineering.
Urkixo zum. Z/g.
48013 Bilbao

Lopez de Ipiña, Karmele[3]
University of the Basque Country. UPV/EHU
Elekin Research Group
System Engineering and Automation
Faculty of Engineering. Europa Plaza, 1.
20018 Donostia.

ABSTRACT

Nowadays, it is very common to use sensors for controlling the population of different animal species in a natural environment. A large number of sensors can be deployed in wide areas and they will capture information relentlessly, producing a huge amount of data. However, analysing the collected data by humans is a big challenge and for that reason, it is necessary to develop automated technologies in order to help experts on that task.

Within this context, we present an automatic system to detect and classify sounds, especially those generated by birds and insects among other sounds that can be heard in a natural environment.

For the development of the system, it has been necessary to generate a sound database. The recorded database consists of field recordings in three different Natural Parks, with sounds of several bird and insect species, as well as background noises. The automated system employs state of the art neural networks for detecting and classifying sound frames. Experiments were done using several signal preprocessing and acoustic features.

The experiments show a good accuracy in detection and classification of sound frames and with results higher or comparable to other state of the art approaches.

Keywords: Environment, detection, classification, neural network
I-INCE Classification of Subject Number: 30
http://i-ince.org/files/data/classification.pdf

[1] itxasne.diez@ehu.eus
[2] ibon.saratxaga@ehu.eus
[3] karmele.ipina@ehu.eus

1. INTRODUCTION

Nowadays, the use of different sensors for monitoring animal species populations in a natural environment is becoming increasingly important. One of the major challenges that researchers have to face, is to detect segments of different sound events in large recordings obtained from continuously operating sensors deployed in the field. Since this is a very time consuming task, it is convenient to develop new technologies to automate this detection and classification process.

The aim of this paper is to present a natural sound detection and classification system. In particular, a system that detects and classifies bird and insects sounds among other sounds that could happen in a natural environment.

During the last years, some studies have been carried out in the field of automatic sound detection and classification in outdoor environments. Some authors have focused their research in the study of the environmental sounds (both natural and human produced) [1]–[4][5], while others have tried the more specific task of the detection and classification of different species of animals [6]–[10]

Machine learning requires gathering certain amount of data to train and test the models. In environmental sound classification the published works report databases ranging from 8723 recordings with a total length of 2 seconds for 10 classes [5], to 800 segments of 4 seconds for 4 classes[1]. In the classification of animal sounds the published works often report unbalanced databases ranging from 127 to 4641 segments depending on the with lengths between 2 to10 seconds [6], [8], [11]. In general, the total amount varies with the number of classes to consider, with a number of recordings ranging from 200 to 800 per class.

There are some publicly available databases used for the detection and classification of environmental and animal sounds. Some of the most used are: UrbanSound8K[4] database, a collection of 8732 short urban sound sources; Sound Event[5] database contains different types of events such as impacts, rolls, and pouring liquids; Freesound[6] database, with more than 230,000 sounds and effects. For birds there is a specialized database called XenoCanto[7] . It is also very common that authors create their own databases and complete them with material from public databases. For instance, some authors have used Voxforge database[4], which contains sentences from different speakers to complement the samples of the voice category. Finally, other authors like [1], [4], [8]carry out field recordings.

Raw audio data is not generally suitable as an input to a detection and classification system. Hence, different acoustic features are extracted from the audio. The most used features are spectrogram [6], [9], [12], mel-spectrogram [7], [9], [13] and Mel Frequency Cepstral Coefficients (MFCC) [9], [12]–[14]. There are other parameters such as entropy peak, spectral centroid, ratio of energy per band, high frequency content [10] and dominant frequency [7] that have been used as features.

Many supervised machine-learning algorithms have been used in the development of detection and classification systems. Comparing traditional algorithms like Decision Trees, Support Vector Machine (SVM), k-Nearest Neighbours and Hidden Markov Models (HMM), generally, the SVMs achieved the best results.

4 https://urbansounddataset.weebly.com/urbansound8k.html

5 http://www.psy.cmu.edu/~auditorylab/website/index/home.html

6 http:// www.freesound.org/

7 http://www.xeno-canto.org/

Since the introduction of Neural Networks (NN) for pattern recognition, they have outperformed the results obtained with traditional algorithms. For instance, in the system for urban sound classification [5], the performance of a SVM was compared with different configurations of neural networks such as a recurrent neural network (RNN), a deep neural network (DNN) and a Convolutional Neural Network (CNN), obtaining better results using a CNN or a DNN than using a SVM or a RNN. Adavanne proposed an architecture [7] that uses a bidirectional recurrent neural network for bird detection and obtains similar results to a CNN.

The objective of this study is to propose a convolutional neural network for detecting birds and insects in a natural environment, comparing three different features and evaluating different sampling frequencies for the training data. This approach is explained in the rest of the paper: in the following section the database is described, section 3 describes the detection and classification system based on a Convolutional Neural Network. The experimental results are shown in section 4. Finally, some conclusions are presented in section 5.

2. DATABASE

As it has been explained, the aim of the project is the detection and classification of animal sounds (distinguishing birds and insects) in a natural environment. To do that, we used part of a previously published database, called Akuoingoure [1], and we completed it with field recordings, to get more samples of each category.

2.2 Database Akuinguore

The Akuinguore [1] database was recorded in Costa Rica and the Natural Park of Doñana (south of Spain), with 2 hours and 800 samples for 4 categories.

The recorded sounds were bird calls, insects and natural environment sounds. All the recordings were made using a video camera and the audio was extracted with 44.1 kHz

The audio samples are homogenous, i.e. they contain only one kind of sound.

2.2 Database Bioinguru

The Bioinguru database gathers the recordings carried out in three different points within Urdaibai Biosphere Reserve (Kanala, Sukarrieta and Bermeo, Basque Country). Bioinguru database has a duration of 2 hours and was recorded using a manual audio recorder (Zoom H4nPro) at a sampling frequency of 44.1 kHz.

The recordings included sound from different birds and the background natural sounds consist mainly of voices, car passing noise, and natural murmur sound. The database is not homogeneous; meaning that in an audio sample different sound types can be heard.

2.3 Spectro-temporal characteristics

The different types of available sounds show distinct features both in the spectral and temporal domains. The sounds of birds are tonal, with an important presence of harmonic components reaching high frequencies (often over 14 kHz). Insect sounds show more variability. Some of them concentrate the energy at a given frequency, while others have the energy distributed in different bands, or are tonal sounds.

The natural background sounds usually have less energy and it is concentrated below the frequencies in which birds and insects emit. It can be said that birds and insects can emit sounds above 10 kHz while the energy of the other sounds is distributed below that frequency.

2.4 Labelling and Segmentation

Segmentation consists on marking the parts of the audio file where an event was heard and labelling implies assigning a description of the event (a label) to the segment. This process is manual and was carried out using an audio edition software.

The events that have been labelled are: sounds of birds, sounds of insects and the background natural sounds. The background sounds were also divided into sound of passing cars, voices and natural murmur noise, but they correspond to the same category for the experiment. There were also unwanted events that were labelled as noise for rejecting them in at a later stage.

After the cleaning and labelling of the signals, the number of segments of each label was:

Sounds	*Categories*	*Segments*	*Total Segments for Category*
Natural environmental noise	0	287	319
Passing Car	0	8	
Voice	0	24	
Bird	1	141	141
Insect	*2*	*50*	*50*

Table 1. Inventory of samples and categories of the database

3. DETECTION AND CLASSIFICATION SYSTEM

The detection and classification system is based on a Convolutional Neural Network. CNNs have been used since the 80s, but recently they have outperformed most of the traditional classifiers. Although primarily used in visual recognition, in the last years CNN have been applied in speech and music recognition tasks. Traditionally, classification of environmental sounds has been mainly based on statistical classifiers but in the last years deep learning techniques have been introduced in this context also.

3.1 Model Architecture

The developed neural network follows the architecture proposed by Karol Piczak in his work “Environmental Sound Classification with Convolutional Neural Networks” [2]. The CNN consists of an input layer, two pairs of convolutional and reduction layers, two hidden layers and an output layer with the next configuration:

- The features obtained from each fragment and its corresponding label, are used as input of the CNN. From each fragment, we obtain data in 2D input format. The number of columns is L=41 and the number of rows depends on the feature: the number of filters in the mel-spectrogram, the number of coefficients in the MFCC or the number of points in the spectrogram.
- The first convolutional layer, consists of 80 filters of two dimensions 57 x 6, with a stride of 1 x 1 followed by a reduction layer of 4 x 3 and 1x3 stride.
- The second convolutional layer, also consists of 80 filters of two-dimensions 1x3 and a stride of 1 x 1, followed also by a reduction layer with a max pooling of 1 x 3 and 1 x 3 stride.
- Two fully connected hidden layers, which have 500 neurons each with a dropout of 50%, i.e. a random draw of 50% of the neurons in the training phase.

- Finally, a softmax output layer with a number of neurons equal to the number of classes considered for classification purposes, namely 3.

All the layers use ReLU (Rectified Linear Units) as non-linear activation function.

3.2 Data Preparation

A detection and classification system has not only to be able to detect different events in a long signal, but also, it has to produce the category or type of those events. Typically, a Neural Network produces a classification decision for an input. To work as a detection system, the signal where we want to detect an event, has to be divided into smaller fragments. These fragments are used as an input of the CNN, which classifies each of them, thus allowing to detect the instant when the event is present.

For that reason, the audio segments were splitted in fragments with a length of 1 second, using a rectangular window. The windowing has an overlapping of 50% in the segments with no event and of 75% in the segments that have event. This way we obtain more fragments for the less represented classes. The label of the original segment is assigned to the fragments resulting from that segment.

3.2.1 Feature extraction

In this project, we compared the performance of three different features: log spectrogram (STFT), log mel filtered spectrogram (Mel Spectrogram) and mel frequency cepstral coefficients (MFCC).

For the log spectrogram, we used a windowing of 1024 points (32ms for 32 kHz and 23ms for 44.1 kHz) with a frame rate of 21ms and then we obtained the power in dB for each band. For the log mel-spectrogram computation, first, the Fourier Transform of the signal was calculated, then a mel filter of 60 bands is applied to the module of the signal and finally logarithm of each band is computed. A mel filter, is a set of triangular filters that tries to reproduce the non-linearity of the human ear perception; this filter has more resolution at low frequencies than at high ones. Finally, for the MFCC a discrete cosine transform is applied to the log-mel-spectrogram and only the first 60 coefficients are taken.

All these features are static, so we also computed the temporal difference between consecutive frames (the deltas) to consider the temporal evolution of the signal, and included them in the input parameters.

All the features and their deltas were obtained using *librosa*[8] library in Python.

3.2.2 Data set

The development of an automatic system can be divided into two steps: the training of the system including the adjustment of certain design parameters (the hyperparemeters) and the test of the system. During the training phase, the system is fed with samples from which a model is derived. Some data is reserved to evaluate different alternatives of the hyperparameters. When the system has learned the best model, it is evaluated with new, unseen samples to produce a classification decision.

Obviously, it is necessary to use different data in each phase get an accurate estimation of the actual performance of the system in real operation.

For that reason, the original dataset was divided in two parts, 80% for the training and 20% for the testing. Usually, data distribution is done in a random way. In our case, the categories are not balanced and as it is shown in Fig. 1. A random division of the data can

[8] Librosa:v0.5.1. DOI:10.5281/zenodo.591533

lead to a class being little or non represented in the test dataset. To solve this problem, the full segments are distributed randomly in the datasets. As the number of segments (see Table 1) are not so unbalanced, the resulting distribution will neither be.

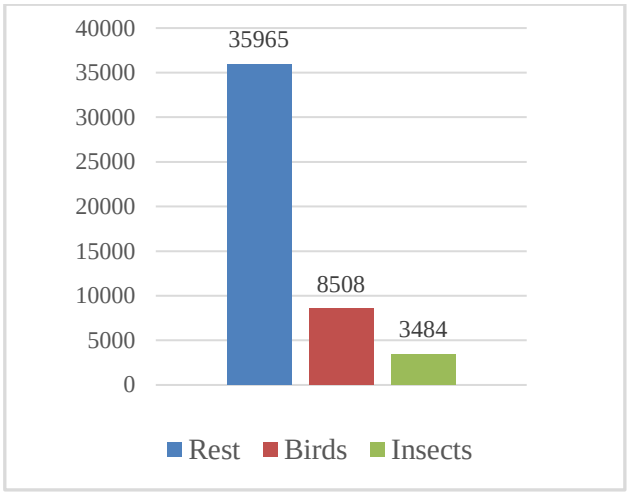

Fig. 1 Number of fragments in each category.

3.3 Training

Training a model means to adjust the weights and bias of the neurons of the NN in order to minimize the differences between the obtained classification and the real one. This is done minimizing a loss function, in our case, the cross entropy function.

The training was carried out using the Stochastic Gradient Descent that is an iterative algorithm for minimizing a loss function. This algorithm uses blocks of input data (mini-batches) that are processed at the same time to estimate the gradient. The mini batch size used for training was 100.

When optimizing a function it is possible to converge in a local minima and no to obtain the optimum solution. To avoid it, we tried different values of the learning-rate parameter, with values around 0.002 being the best one.

4. EXPERIMENTS

In the construction of a neural network, there are many design decisions that have to be taken. These decisions are always based in the results obtained from different evaluations during the developing stage. As it is impossible to show all the tests done for the adjustment of the model, we outline two of the most interesting ones in section 4.2. Then the final evaluation results are presented in section 4.3.

4.1 Metrics

There are different metrics for the evaluation of a classification system.

- Precision: for a class precision is the number of correctly classified positive divided by the total number of positive (correct and incorrect).
- Recall: the recall of each class is defined as the number of correctly predicted positives divided by the number of all the samples actually pertaining to that class.
- UAR: unweighted average recall is the average of the recall of all classes.
- Accuracy: is the number of correct classification divided by the total number of samples.

Accuracy is not a good metric when using a non balanced dataset as it overestimates the more representative class. For that reason, we will use the other three metrics in the evaluation of our model.

4.1 Hyperparameters adjustment

The first hyperparameter that we decided to adjust were the features. As indicated in section 2.4, we evaluated the model using different input data, a spectrogram, a mel spectrogram and MFCC.

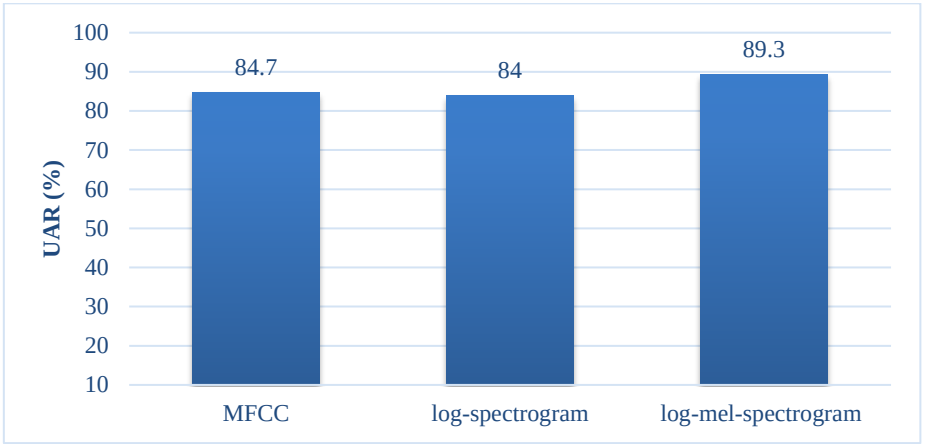

Fig. 2 Comparative of different parametrization

As it can be seen, in Fig. 2, the best performance is obtained for the log-mel-spectrograms and this feature will be used for the rest of the experiments.

The second parameter to adjust was the sampling rate of the input signal, which limits the bandwidth of the system. Audio signals were recorded at 44.1 kHz, but other sampling frequencies were evaluated, 12 kHz, 24 kHz, and 32 kHz.

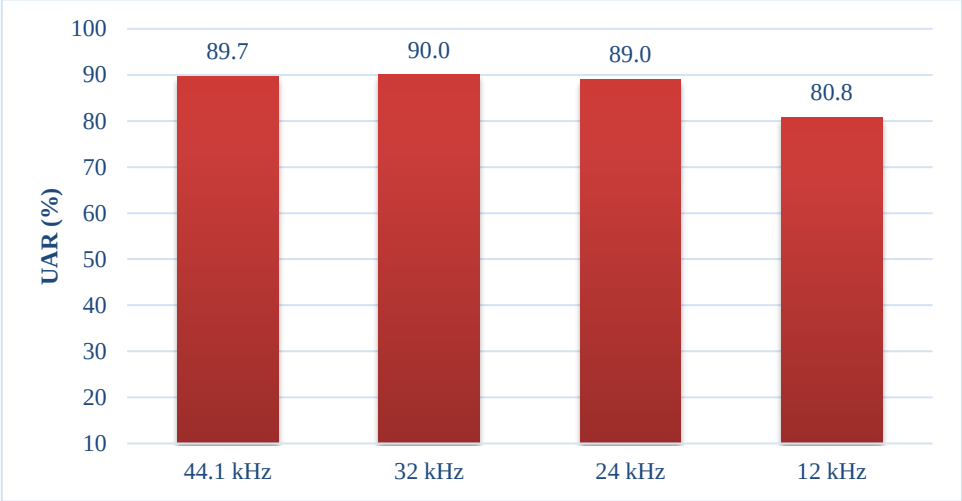

Fig. 3 UAR for different sample rates

The results in Fig. 3 show, that there is no much difference when using high sample rates.

As it can be seen, the best performances are obtained for 32 kHz and 44.1 kHz with an UAR of 90%. At the same time, it can be noticed that using a low sample rate (12 kHz) the UAR drops to 80.8%.That can be because the bandwidth is 6 kHz and the energy of the sound is also distributed at frequencies above 10 kHz.

4.2 Evaluation experiments

Once, the most important parameters were decided, we evaluated the performance of the best models (32 kHz and 44.1 kHz). For the evaluation experiment, we used a 5-fold cross validation. Cross-validation is a technique that involves partitioning the data into k folds; one of the folds is used for testing and k-1 for training. It is of common use in machine learning to estimate how the model will perform with unseen data.

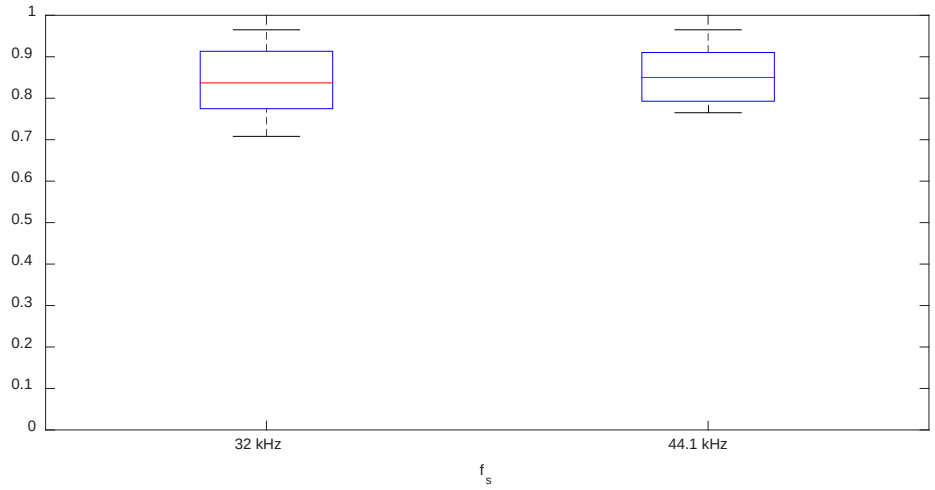

Fig. 4 UAR for 5-fold cross validation

As it can be seen, in Fig. 4, there is a slight difference between both models. The average UAR for the 44.1 kHz model is 84.6% while the UAR for the 32 kHz model is 85.4%. The variance of the UAR for the different iterations of the cross validation is high for both models, due to the reduced amount of data in the database.

We also analysed the average confusion matrix of the best model (32 kHz), and calculated the recall and precision for each category.

		Predicted Category		
		Backgrnd	Birds	Insects
Labelled Category	Backgrnd	5467	1604	167
	Birds	198	1470	48
	Insects	25	9	680

Table 2. Average Confusion Matrix for 32000Hz

	Recall Value (%)	Precision Value (%)
Backgrnd	76	73
Birds	86	47
Insects	95	76

Table 3. Evaluation Metrics

The total accuracy of the model is 79%.

If we evaluate the recall for each category, the values vary between 76%-95%. As it is shown in Table 3, the highest values correspond to those categories in which we have less training data. That high values could be due to overfitting and can cause both false positives and negatives with new data.

The system confuses the bird category with the background noise and this explains the low value in the precision. Undoubtedly, more data would be necessary to improve these results.

5. CONCLUSIONS

The aim of this research was to develop a system that could detect sound of birds and insects over other background sounds in a natural environment. For that porpoise, we used and two databases: Akuoingoure and Bioinguru. We implemented a detection and classification system based on deep neural networks, specifically a Convolutional network.

The experiments carried out showed that the proposed system is able to successfully detect birds and insects in a natural sound background, obtaining the best results using as input data the log-mel-spectrum and a sample rate of 32 kHz.

The overall accuracy of our system for the classification of the 3 categories is 79%, while the accuracy obtained by other environmental sound classification systems are 73%[5], 75%[2] for 10 categories and 75.4% for 5 categories [4]. Therefore, it can be said that the system performs at the level of the systems developed by other authors. Nevertheless, results suggest that it would be necessary to get more data to improve the results of the classifier in order to use it in a real application.

6. REFERENCES

[1] K. Lopez-De-Ipina, M. Iturrate, J. B. Alonso, and B. Rodriguez-Herrera, "Automatic acoustic analysis for biodiversity preservation: A multi-environmental approach," *IWOBI 2015 - 2015 Int. Work Conf. Bio-Inspired Intell. Intell. Syst. Biodivers. Conserv. Proc.*, pp. 43–48, 2015.

[2] K. J. Piczak, "Environmental Sound Classification With Convolutional Neural Networks," in *2015 Ieee International Workshop on Machine Learning for Signal Processing*, 2015.

[3] K. J. Piczak, "The Details That Matter: Frequency Resolution of Spectrograms in Acoustic Scene Classification," *DCASE 2017-Workshop Detect. Classif. Acoust. Scenes Events*, no. November, 2017.

[4] X. Valero, "Análisis de la señal acústica mediante Coeficientes Cepstrales Bio-inspirados y su aplicación al reconocimiento de paisajes sonoros.," *VIII Congr. Ibero-americano Acústica*, pp. 1–9, 2012.

[5] D. B. Chih-Wei Chang, "Urban Sound Classification: With Random Forest, SVM, DNN, RNN, and CNN Classifiers," 2016.

[6] M. A. Acevedo, C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecol. Inform.*, vol. 4, no. 4, pp. 206–214, 2009.

[7] S. Adavanne, K. Drossos, E. Çakir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," *25th Eur. Signal Process. Conf. EUSIPCO 2017*, vol. 2017–Janua, pp. 1729–1733, 2017.

[8] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez, "Real-time bioacoustics monitoring and automated species identification," *PeerJ*, vol. 1, no. October, p. e103, 2013.

[9] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," 2014.

[10] L. Zhang, M. Towsey, J. Xie, J. Zhang, and P. Roe, "Using multi-label classification for acoustic pattern detection and assisting bird species surveys," *Appl. Acoust.*, vol. 110, pp. 91–98, 2016.

[11] Z. Zhao *et al.*, "Automated bird acoustic event detection and robust species classification," *Ecol. Inform.*, vol. 39, no. April, pp. 99–108, 2017.

[12] J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," *Futur. Gener. Comput. Syst.*, vol. 29, no. 2, pp. 560–568, 2013.

[13] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? An introduction and overview of an emerging new science," *Landsc. Ecol.*, vol. 26, no. 9, pp. 1213–1232, 2011.

[14] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra, "Ecological Acoustics Perspective for Content-Based Retrieval of Environmental Sounds," vol. 2010, 2010.