Automatic Classification of Anuran Sounds Using Convolutional Neural Networks

Juan Colonna  
Universidade Federal do Amazonas  
Manaus, Brazil  
jcolonna@inesctec.pt

Alípio M. Jorge  
Universidade do Porto  
Porto, Portugal  
amjorge@fc.up.pt

Tanel Peet  
Estonian Information Technology College  
Tallinn, Estonia  
tpeet@itcollege.ee

Elsa Ferreira Gomes  
Instituto Superior de Engenharia do Porto  
Porto, Portugal  
efg@isep.ipp.pt

Carlos Abreu Ferreira  
Instituto Politécnico do Porto  
Porto, Portugal  
cgf@isep.ipp.pt

João Gama  
Universidade do Porto  
Porto, Portugal  
joao.jgama@gmail.com

ABSTRACT

Anurans (frogs or toads) are closely related to the ecosystem and they are commonly used by biologists as early indicators of ecological stress. Automatic classification of anurans, by processing their calls, helps biologists analyze the activity of anurans on larger scale. Wireless Sensor Networks (WSNs) can be used for gathering data automatically over a large area. WSNs usually set restrictions on computing and transmission power for extending the network’s lifetime. Deep Learning algorithms have gathered a lot of popularity in recent years, especially in the field of image recognition. Being an eager learner, a trained Deep Learning model does not need a lot of computing power and could be used in hardware with limited resources. This paper investigates the possibility of using Convolutional Neural Networks with Mel-Frequency Cepstral Coefficients (MFCCs) as input for the task of classifying anuran sounds.

CCS Concepts

• Computing methodologies → Neural networks;

Keywords

Convolutional Neural Networks; Machine Learning; MFCC; Wireless Sensor Networks; Anurans

1. INTRODUCTION

Amphibian populations are directly affected by environmental changes and therefore are closely related to ecosystem [4]. In addition, it is shown that there is a clear relationship between climate change and mortality in amphibian populations [3]. Therefore it can be concluded that anomalies in their behavior can be used as early indicators of ecological stress. Amphibian monitoring systems may help to estimate long-term changes in amphibian populations and determine the causes of changes.

Anurans belong to amphibians class and produce sounds (calls), which contain enough information to identify each species [12]. A human expert can be used to classify anurans by their calls, but this approach is slow and error-prone. One of the alternative approaches is to use Wireless Sensor Networks (WSNs) with automatic classification methods. WSNs consist of low cost nodes which could be spread all over the desired areas [1]. This offers a good solution for automatically monitoring anurans over a large area, but low cost hardware also sets restrictions on computing power [22]. The cost of energy in data processing is much less compared to data communication. Therefore, local data processing is good option for minimizing power consumption [1]. Depending on the setup of the WSN and the complexity of the classification model, it may be more efficient to do the classification in low-level nodes and transmit only the results of classification to the sink nodes.

Deep Learning algorithms are eager learners, which means that the model can be trained on fast computers and the trained model could be used in WSN nodes with less computing power and therefore less power consumption [16]. Deep Learning has shown promising results in different research fields, being especially popular in image recognition tasks.

Studies in sound processing show that Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most popular features for presenting audio signals in speech recognition [14] and music information retrieval tasks [13]. MFCCs mimic some parts of the human speech production and perception, including logarithmic perception of loudness and pitch of human auditory system [19]. Anuran call consist of short sounds, called syllables, which are a single blow of air from the lung of the anuran [12]. These syllables can be compared to phonemes in human speech and therefore MFCC features have shown good results in anuran classification tasks [5]. In the process of MFCC extraction, the signal is divided into fixed length frames and a user defined amount of MFCC features are extracted for every frame [19].

This paper investigates the possibility of using Convolu-
MFCC Extraction Preprocessing Deep Learning ID

Figure 1: Anuran classification stages

2. STATE OF THE ART

The approach for classifying one anuran species is done by Hu et. al. [11] and Taylor et. al [9]. In their proposal, all audio recording are collected and transmitted to a central node, where spectrogram is generated. The C4.5 Decision tree classifier is then used based on power features of the spectrogram. This approach requires a lot of resources due to the audio recordings which are sent to a central node.

WSNs are used for monitoring the habitats of the birds by Cai et. al. [2]. A noise reduction algorithm is proposed, based on minimum mean square error. MFCC and MFCC delta features were used for each frame of the bird calls. Context Neural Network was used for classification, which uses frames from "future" and "past", taking into account the dynamic nature of the bird songs.

Vaca-Castao and Rodriguez [21] used MFCCs to classify birds and anurans with k-Nearest Neighbors classifier. The Principal Component Analysis was used in their approach for dimensionality reduction, which increased the computational cost of the whole process.

Yen and Fu [23] classify anuran species by using the results of Discrete Wavelet Transform as features. All samples were then classified by Multilayer Perceptron. Huang et. al. [12] used kNN and SVM classifiers to automatically recognize frog species.

Potamitis [20] and Oliveira et. al. [8] used spectrograms of bird call recordings as images for classifying the bird species. Unlike the methodology proposed in this paper, they did not segment the raw audio into syllables. Instead they turned audio signal into spectrogram and used morphological filtering to find patterns from spectrograms, which correspond to the calls of the birds. Potamitis extracted region of interests (ROIs) and used a random forest under the multi-label formulation of one vs. all for classification. Oliveira et. al. algorithm extracts only the frames of the spectrogram which correspond to one certain bird species.

3. METHODOLOGY

The main ideas behind the methods used for classifying the anuran sounds using Convolutional Neural Networks are explained in this chapter. Signal processing and the algorithm for audio signal segmentation is described in Section 3.1. In Section 3.2, the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) from audio segments is explained in detail. Section 3.3 describes the principles and building blocks of Convolutional Neural Networks.

3.1 Preprocessing

Pre-processing of the anuran calls used in this paper follows the steps described by Colonna et. al. [5]. First step is to standardize the signal, which includes normalization of the signal and making it have zero mean. Signal is padded with zeros with the width equal to the length of the half segment - $\frac{1}{2}$. After standardization and zero padding, the signal is segmented into syllables. The segmentation is done by the following algorithm, where $\alpha$ is the threshold for the normalized signal (ranging between 0 and 1) and $\beta$ is half of the length of the segment:

1. Find the maximum absolute value $S(t)$ of the signal;
2. If $S(t) < \alpha$ go to step 5;
3. Select $\beta$ seconds to the right and to the left of the peak, which gives us one syllable;
4. Extract the syllable from step 3 and replace the values in the signal with randomized small numbers to simulate noise. Go to step 1;
5. End.

The syllable extraction is illustrated in Figure 2.

3.2 MFCC extraction

For the experiments in this paper, the features extracted from the audio signal are Mel-Frequency Cepstral Coefficients (MFCCs), which are engineered for human speech recognition tasks and contain information about timbral features [19]. For the extraction of MFCCs, the audio signal
Figure 2: Syllable extraction from a Hyla Minuta species recording

Figure 3: MFCC extraction from a Hyla Minuta syllable

is divided into short frames. For each frame the Fourier transform is taken, resulting in the spectrum of the frame. The spectrum is mapped onto the mel scale. The logs of the mel frequency spectrum are taken, followed by discrete cosine transform. The amplitudes in the resulting spectrum are MFCCs [6].

The extraction of MFCCs divides signal into multiple frames, which size is called window length. The division of audio signal to MFCC frames can be seen in Figure 3. The parameter named window step shows how much the window moves before calculating the features for the next frame, meaning the frames can overlap each other. For each frame an user defined amount of features are calculated, usually between 8 and 14 for optimal amount of information [15].

3.3 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are variants of Multilayer Perceptron, inspired by biology and by animal visual cortex. The visual cortex of an animal contains a complex arrangement of cells which are sensitive to small regions of the visual field. These small regions are tiled to cover the entire visual area [18]. The similar approach is used in CNNs.

The layers in a CNN have neurons arranged in 3 dimensions: width, height and depth. In convolution layer the neurons are scanned by moving a fixed width, length and depth sized selection, called kernel, over the width and depth dimensions of the layer. The amount of displacement of the kernel after each step is called stride. The output of each kernel has a user defined depth and a unit width and length. The behavior in the edges of the input data is defined by zero-padding parameter. Pooling layers are often used between convolutional layers for controlling overfitting. The last layer of convolutional layer is usually fully connected layer, meaning that neurons in a fully connected layer have full connections to all activations in the previous layer [18]. The structure of the CNN used in the experiment is shown on Figure 4.

4. EXPERIMENT

The following chapter describes the setup and results of the experiments on anuran sound classification. Section 4.1 explains how the dataset is defined and built. Two binary classification problems are defined in Section 4.2. The configuration for MFCC extraction, different Machine Learning approaches for anuran classification and the evaluation of results is described in Section 4.3. The results of the two experiments can be found in Section 4.4, followed by the analysis of these results in Section 4.5.

4.1 Dataset

The dataset used in our experiments contains ten anuran species from four different families. The sound recordings were collected in the anuran habitat, under real noise conditions. The recordings for Ameerega Trivittata, Hyla Minuta, Scinax Ruber and Adenomera Hylaedactyla were extracted from three different regions: Mata Atlantica, Brazil [10], Bolivia [7] and French Guiana [17]. The rest of the recordings were collected on the campus of the Federal University of Amazonas in Manaus, Brazil [5]. Species used in our experiments can be seen in Table 1.

The recordings in the dataset have different length and therefore contain a different number of syllables. The number of syllables extracted from one recording varies from 3 syllables (Osteocephalus Oophagus recording) to 1071 syllables (Rhinella Granulosa recording). The recordings have different kind of background noises, including natural ones but also noises specific to the recording device and its settings. Hence, it is important to separate training and dataset by recordings, not by syllables. The audio recordings are stored in wav format with a sample rate of 44.1 kHz and 32 bits per sample.

In the preprocessing phase the value of $\alpha$ was 0.5 and $\beta$ was 0.101 seconds. The results of the syllables extraction
can be seen in Table 1. In addition to species label, each syllable has also a label containing the record identifier to distinguish from which recording the syllable was extracted. Each syllable has a length of 202 ms.

4.2 Problem

Two different binary classification experiments are performed using Deep Learning algorithms. The obtained results are compared against five other classifiers. In the first experiment the *Leptodactylidae* family of anurans forms one class and all the other *families* form the second class. The *Leptodactylidae* class consists of *Adenomera Andraea*, *Adenomera Hylaedactyla* and *Leptodactylus Fuscus* syllables. The *other families* class consists of the syllables from *Ameerega Trivittata*, *Hyla Minuta*, *Hypsiboas Cinerascens*, *Hypsiboas Cordobae*, *Osteocephalus Oophagus*, *Scinax Ruber* and *Rhinella Granulosa*. The *Leptodactylidae* family class has 3812 syllables from 23 recordings and *other families* class has 3972 syllables from 32 recordings.

The goal of the second experiment is to recognize one species from *Leptodactylidae* family - *Adenomera Hylaedactyla*. Hence, it is a binary classification where *Adenomera Hylaedactyla* species forms one class and all the other *species* form the second one. The *other species* class contains therefore all the syllables from *Adenomera Andraea*, *Leptodactylus Fuscus*, *Ameerega Trivittata*, *Hyla Minuta*, *Hypsiboas Cinerascens*, *Hypsiboas Cordobae*, *Osteocephalus Oophagus*, *Scinax Ruber* and *Rhinella Granulosa* species. The *Adenomera Hylaedactyla* class has 3084 syllables from 11 recordings compared to *other species* class with 4700 syllables from 44 recordings.

4.3 Configuration of the experiments

After the preprocessing phase the MFCCs are extracted for each syllable. The experiments done with Convolutional Neural Network (CNN) use MFCCs with 20 features. The window length and window step size are chosen to get 20 frames for each syllable, resulting in 20x20 matrix. The values for window length and window step in MFCC extraction are accordingly 20 ms and 10 ms. For all the other classifiers, the window length and step size are the same as the length of a syllable. This results in a 20 dimensional vector for each syllable.

The Convolutional Neural Network has two convolutional layers, followed by two fully connected layer, as described in Figure 4. Rectified Neural Units (ReLUs) were used as an activation function for each convolutional layer for increasing the nonlinear properties of the network. No pooling layers were used. The output layer is one-hot encoded array with two elements, where each element represents the probability of the corresponding class.

The kernel size of 3x3 is used with a stride of 2. The depth of the convolutional layers is 16. The number of neurons in the fully connected hidden layer is 128. Stochastic gradient descent was used for training with a learning rate of 0.05. The training takes 200 steps, each using a batch of 16 syllables.

Leave-one-out cross-validation (LOOCV) is used on recordings for measuring the performance of the classification algorithms. The syllables are divided into two classes based on the definition of the problem. Each step of the LOOCV takes all the syllables from one recording and separates them from the training set for later testing. These steps are repeated until every recording has been used as a test set. In every step, the predictions for every syllable are saved. After LOOCV is completed, accuracy is calculated by dividing the number of correctly predicted syllables by the total number of syllables. In addition, F1 score and area under receiver operating characteristic curve (AUC) are used with the predictions from LOOCV as a performance measure of the algorithms.

The results from the CNN are compared to the results of different classification algorithms, such as k-Nearest Neighbors (kNN), Decision Tree (DT), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM). The k of kNN was set to 50, which yielded the best results, compared to other values of k. A radial basis function kernel was used in SVM training. QDA and Decision Trees were used with their default settings. 1

Python was used for segmentation, MFCC extraction and CNN. Scikit-learn2 library was used for several data processing tasks. MFCC feature extraction was done by the help of

1http://scikit-learn.org/

![Figure 4: Structure of the CNN used in experiments](image-url)
Table 1: The dataset description containing species of anurans grouped by their families

<table>
<thead>
<tr>
<th>Family</th>
<th>Species</th>
<th>Recordings</th>
<th>Syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leptodactylidae</td>
<td>Adenomera Andraea</td>
<td>8</td>
<td>487</td>
</tr>
<tr>
<td></td>
<td>Adenomera Hylaedactyla</td>
<td>11</td>
<td>3084</td>
</tr>
<tr>
<td></td>
<td>Leptodactylus Fuscus</td>
<td>4</td>
<td>241</td>
</tr>
<tr>
<td>Dendrobatidae</td>
<td>Ameerega Trivittata</td>
<td>5</td>
<td>366</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hylidae</td>
<td>Hyla Minuta</td>
<td>11</td>
<td>241</td>
</tr>
<tr>
<td></td>
<td>Hypsiboa Cinerascens</td>
<td>2</td>
<td>343</td>
</tr>
<tr>
<td></td>
<td>Hypsiboa Cordoba</td>
<td>4</td>
<td>1168</td>
</tr>
<tr>
<td></td>
<td>Osteocephalus Oophagus</td>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Scinax Ruber</td>
<td>4</td>
<td>182</td>
</tr>
<tr>
<td>Bufonidae</td>
<td>Rhinella Granulosa</td>
<td>3</td>
<td>1569</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>55</td>
<td>7784</td>
</tr>
</tbody>
</table>

4.4 Results

The results for evaluation of anuran classification algorithms can be seen in Table 2. Problem I is the classification of whether the anuran call is from Leptodactylidae family or not. The second problem II is the classification of whether the anuran call is from Adenomera Hylaedactyla species or not.

4.5 Analysis of results

Convolutional Neural Networks outperformed other classifiers in both problems and all the evaluation metrics, as seen in Table 2.

To get a better understanding of what the CNN was able to learn, the results were analyzed by looking at the performance of each recording in LOOCV steps. For the first problem, where anurans from Leptodactylidae family were being identified, the accuracy for each species identification was calculated for the purpose of further analysis. The identification accuracy for Adenomera Andraea, Adenomera Hylaedactyla and Leptodactylus Fuscus was 66.55%, 99.17% and 40.40% accordingly. Correlation can be seen between the number of syllables for a species and their identification accuracy. This means that due to the larger amount of training data available for the Adenomera Hylaedactyla species (3084 syllables), the algorithm learns to recognize this species a lot better than other members of the Leptodactylidae family - Leptodactylus Fuscus (241 syllables) and Adenomera Andraea (487 syllables).

5. CONCLUSIONS

In this study, the usage of Deep Learning image classification algorithms in anuran sounds classification task was proposed for the usage in Wireless Sensor Networks. The approach included the segmentation of the audio recording to syllables and extracting suitable sized MFCC features, followed by Deep Learning classification algorithm. The performance of the model was evaluated and compared to other classifiers. The proposed approach outperformed all the other classifiers, achieving 91.41% accuracy in binary classification task of identifying Leptodactylidae family members and 99.14% accuracy in binary classification for identifying the members of Adenomera Hylaedactyla species.

The future work will investigate the usage of different algorithms and hyper-parameters for increased performance. The analysis of results showed that the unbalanced dataset might be problematic for identifying some species and that results could be improved by collecting more training data. The usage in WSNs should also be validated in future works by comparing the cost of classification and cost of transmitting one number instead of MFCC features.

6. ACKNOWLEDGMENTS

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. Authors JC and JG acknowledge that in addition, the project was partly supported by the European Commission through MAESTRA (ICT-2013-612944).

7. REFERENCES


python_speech_features library\(^3\). For the creation of CNN the Tensorflow\(^4\) library was used. The classification with other classifiers was done in Matlab\(^5\).

4.4 Results

The results for evaluation of anuran classification algorithms can be seen in Table 2. Problem I is the classification of whether the anuran call is from Leptodactylidae family or not. The second problem II is the classification of whether the anuran call is from Adenomera Hylaedactyla species or not.

4.5 Analysis of results

Convolutional Neural Networks outperformed other classifiers in both problems and all the evaluation metrics, as seen in Table 2.

To get a better understanding of what the CNN was able to learn, the results were analyzed by looking at the performance of each recording in LOOCV steps. For the first problem, where anurans from Leptodactylidae family were being identified, the accuracy for each species identification was calculated for the purpose of further analysis. The identification accuracy for Adenomera Andraea, Adenomera Hylaedactyla and Leptodactylus Fuscus was 66.55%, 99.17% and 40.40% accordingly. Correlation can be seen between the number of syllables for a species and their identification accuracy. This means that due to the larger amount of training data available for the Adenomera Hylaedactyla species (3084 syllables), the algorithm learns to recognize this species a lot better than other members of the Leptodactylidae family - Leptodactylus Fuscus (241 syllables) and Adenomera Andraea (487 syllables).

5. CONCLUSIONS

In this study, the usage of Deep Learning image classification algorithms in anuran sounds classification task was proposed for the usage in Wireless Sensor Networks. The approach included the segmentation of the audio recording to syllables and extracting suitable sized MFCC features, followed by Deep Learning classification algorithm. The performance of the model was evaluated and compared to other classifiers. The proposed approach outperformed all the other classifiers, achieving 91.41% accuracy in binary classification task of identifying Leptodactylidae family members and 99.14% accuracy in binary classification for identifying the members of Adenomera Hylaedactyla species.

The future work will investigate the usage of different algorithms and hyper-parameters for increased performance. The analysis of results showed that the unbalanced dataset might be problematic for identifying some species and that results could be improved by collecting more training data. The usage in WSNs should also be validated in future works by comparing the cost of classification and cost of transmitting one number instead of MFCC features.

6. ACKNOWLEDGMENTS

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. Authors JC and JG acknowledge that in addition, the project was partly supported by the European Commission through MAESTRA (ICT-2013-612944).

7. REFERENCES

Table 2: The comparison of different classification algorithm results. Problem I - classifying *Leptodactylidae* family; II - classifying *Adenomera Hylaedactyla* species

<table>
<thead>
<tr>
<th>Prob.</th>
<th>Metric</th>
<th>kNN</th>
<th>DT</th>
<th>QDA</th>
<th>SVM</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Accuracy</td>
<td>0.890</td>
<td>0.863</td>
<td>0.909</td>
<td>0.901</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>0.964</td>
<td>0.880</td>
<td>0.941</td>
<td>0.962</td>
<td>0.970</td>
</tr>
<tr>
<td>II</td>
<td>F1 Score</td>
<td>0.926</td>
<td>0.912</td>
<td>0.937</td>
<td>0.954</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>0.988</td>
<td>0.945</td>
<td>0.980</td>
<td>0.990</td>
<td>0.999</td>
</tr>
</tbody>
</table>


