International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 11 | November - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Automated Bird Species Identification using AudioSignal Processing and Neural Networks

Avinash Tatar, Kashyap Bhamare, Bhushan Chavan, Snehal Shirode

Prof. Abhay Gaidhani

Department of Computer Engineering SITRC, Nashik

Abstract – In this paper, an automatic bird species recognition system has been developed and methods for their identification has been investigated.Automatic identification of bird sounds without physical intervention has been a formidable and onerous endeavor for significant research on the taxonomy and various other sub fields of ornithology.In this paper, a two-stage identification process is employed. The first stage involved construction of an ideal dataset which incorporated all the soundrecordings of different bird species. Subsequently, the sound clips were subjected to various sound pre- processing techniques like pre-emphasis, framing, silence removal and reconstruction. Spectrograms were generated for each reconstructed sound clip. Thesecond stage involved deploying a neural network to which the spectrograms were provided as input. Basedon the input features, the Convolutional Neural Network (CNN) classifies the sound clip and recognizes the bird species. A Real time implementation model was also designed and executedfor the above-described system.

***Key Words*: Bird species identification, bird sound, sound pre-processing techniques, ConvolutionalNeural Network, Spectrograms**

INTRODUCTION:

According to the International Union for Conservation of Nature (IUCN), there are nearly 10,000 known species of birds dispersed among a vast range of ecosystems, from the rainforests of Brazil to the icy shores of Antarctica [1]. These species exhibit amazing diversity in terms of behavior and morphology and are quintessential for the normal functioning of an ecosystem. But this magnificent biological diversity has been threatened the recent human activities which range from intrusion into their habitats to complete annihilation of their habitats and this coupled with natural phenomenon like global warming and climate change has driven many species to extinction. It is estimated that nearly 1,370 species are threatened with extinction which amounts to nearly 13% of the entire bird population. Although a number of bird species are commonly seen, they are difficult to identify by the people. The aim of this study is to develop automated techniques for identifying bird species from their sounds.

Automatic identification of bird calls from continuous recordings gathered from the environment would be significant addition to the research methodology in ornithology and biology, in general. Often these recordings are clipped or contain noise due to which reliable methods of automated techniques have to be used instead of manual conventional methods. Manual inspections of the spectrograms are often error prone and usually the techniques are esoteric in nature and involves multiple experts which makes it unreliable, hence there is a need for automated systems. There is a significant commercial potential for such systems because bird watching is a popular hobby in many countries. Substantial international programs are also invigorating the ventures into the area of bioacoustics signal processing and pattern recognition.

In this paper, a technique has been proposed which involves the use of sound processing and convolutional neural networks to automate the entire process of bird sound identification. The first stage involves the creation of a database consisting of all the sound recordings. Subsequently, these recordings are subjected to sound pre-processing techniques like pre-emphasis, framing, silence removal and reconstruction. Spectrograms are generated for the sound clips and these spectrograms were given as input to a CNN which was trained on a GPU. A real time implementation model was designed for the trained CNN model. A Graphical User Interface (GUI) was also designed for the system. An application can be designed in the future and deployed for mobile devices, which would enable users to use their smartphones as handheld devices for prediction and analysis of bird sounds.

DATA ACQUISITION:

Dataset is of paramount importance and is a critical deciding factor for any machine learning based approach to a problem. A dataset should have certain crucial characteristics for it qualify as a good dataset. It should be accurate and precise, devoid of flawed elements and misleading information. It should be reliable and consistent, there shouldn't be contradictions in the dataset between different data elements irrespective of the source they have been collected from. The dataset must be complete, comprehensive and relevant, because fragmented data provides an inaccurate overall picture. Dataset should be granular and unique to prevent confusions which arise due to aggregated, summarized and manipulated data.
One of the main constraints for our study was the availability and accessibility of the bird sound dataset due to the low level of research activity being done in this field. So, we manually built a dataset comprising of bird sound recordings from the surrounding environment. We also downloaded bird sound recordings from xeno-canto.com which is a website dedicated for sharing bird sounds from around the world [2]. Each sound clip had a length ranging from 5 seconds to 20 seconds. The dataset was created for 4 birds namely cuckoo, sparrow, crow and laughing dove where each bird had a sample space of 100

recordings each, which in total amounted to 400 samples of bird sound recordings. We also decided to build a dataset encompassing 100 samples of human voice clips and surrounding environmental ambient noise, which were sourced from Google Audio-set and Libri Speech ASR Corpus datasets. The decision to include ambient noise and human voice samples are justified because in real time recordings, the bird sound clips are usually interlaced with surrounding noise and the human voice samples are used to build a fool proof network. Now, our dataset has 500 sound clips, which must be divided into a Train dataset, Validation dataset and Test dataset before being given as input to the CNN in the ratio of 70:10:20. The Train dataset is used to train the network and fit the model. Validation dataset is used to tune the hyperparameters of a model during iterative training. Test dataset is used to provide an equitable evaluation of the terminal model fit on the training dataset. Finally, the dataset can be divided into several segments and cross validation can be used to ensure that the sound clips present in each dataset have equal data representation and distribution from all classes.

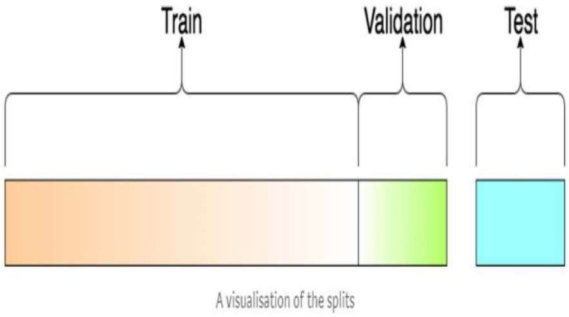
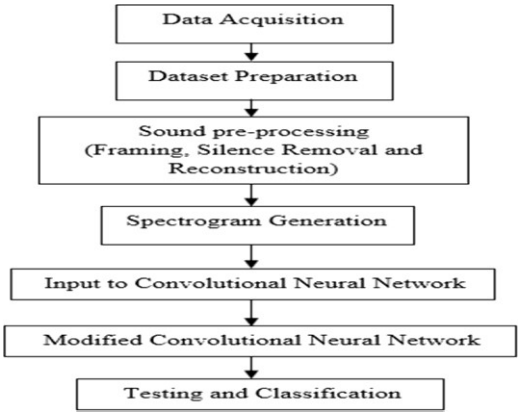


Fig. Division of the Dataset

METHODOLOGY:



A. Pre-Emphasis

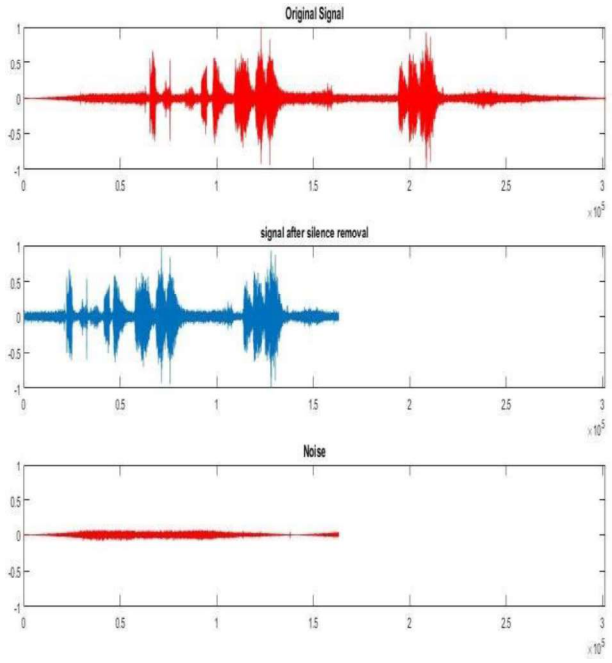
An audio signal that is recorded by a microphone usually contains a large range of frequencies. Speech signal and the bird chirps typically consists of higher frequency energy. In other words, pre-emphasis boosts or increases the higher frequency components. Hence, it is vital to emphasize this higher frequency energy that is of our interest and reduce other frequencies. This is done using a simple first order high pass filter. All the data points in the signal is passed through this filter which is given by the mathematical equation:

$$y[n] = x[n] - S*x[n-1]$$

S can be changed to vary the degree of emphasis needed. Inthis paper 0.95 is taken as the value of S.

Framing and Silence removal

An audio signal is not a stationary signal. It consists of various statistical properties that can vary over the duration thetime. It is necessary to first divide the recorded audio



signal into a number of frames based on its length and then extract the signal which is devoid of any unwanted silence period. The frame length is decided based taking into consideration the total length of the signal and also the sampling period used. In this paper 2.5 % of the entire lengthof the audio clip is taken as the length of one single frame. After framing is completed, the silence removal is carried outusing a thresholding function. The threshold function is chosen such that the audio signal above it is of our interest and the signal which falls below the threshold is considered as silence period or background noise. This silence removal is repeated for all the frames and dynamically changing the threshold value to 7% of the maximum amplitude present in that frame.

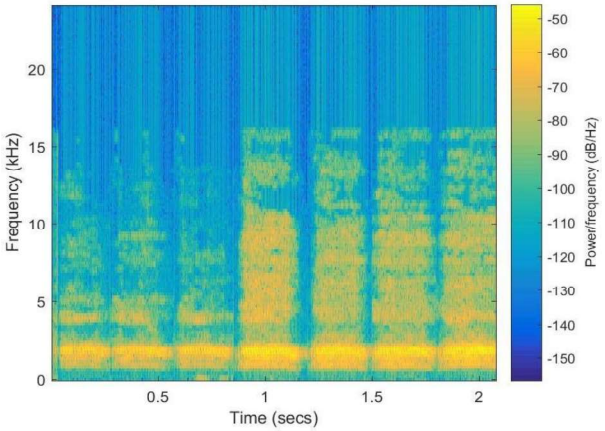
B. Reconstruction

Reconstruction is the process of combining orconcatenating all the frames that were obtained after the process of framing and

silence removal. The result of this process is generation of a signal that is void of any noticeable silence period and still includes most of the information that is of our interest. The final step is to now procure the best sample in the pre-processed audio signal by considering 1 second of the clip that contains the highest amplitude in the total duration of the reconstructed signal. Fig.3 shows the result of reconstruction the signal after silence removal.

domain using Fourier Transforms and then plotting the frequencies [3],[4]. In this paper, the spectrogram is generated using an inbuilt MATLAB function. The spectrogram is generated only for the data that was obtained after

reconstruction and not the entire signal. This process is repeated for all the audio-clips in the dataset and the respective spectrograms are stored in the labelled folders. Each of the spectrograms are unique and have their own characteristics. For example, a spectrogram of sparrow'schirp will usually contain relatively low frequencies of high intensity whereas the spectrogram of a crow's chirp will havehigh intensities over a range of frequencies. These differences are easily picked up by a Neural Network when it is trained. In this paper, AlexNet is chosen as the Neural Network because of its high accuracy and easy implementation in the MATLAB. The spectrograms generated are scaled down to the resolution of 227 by 227 by 3 which can then be used to train the AlexNet Neural Network. Fig.4 shows an example of a spectrogram generated for an audio signal consisting of the chirps of a crow.



C. Introduction to AlexNet

AlexNet is a pretrained convolutional neural networkwhich is used in image classification applications. It is a fast GPU implementation of a CNN [5]. This network was trainedusing over a million images to classify multiple categories ofimages. The input is an image of resolution 227*227*3 and acategorical output is obtained. The network contains 5 convolutional layers and 3 fully connected layers. ReLU is applied after every convolutional and fully connected layer. Dropout is applied before the first and the second fully connected layers. The image classification is done based on feature extraction which is one of the fastest ways of training an image classifier. AlexNet has a high accuracy rate and is available on multiple platforms like Python, MATLAB and Tensorflow. We used AlexNet for this application because the primary aim was to classify the bird species using spectrogram image of the bird sounds of different species. AlexNet was preferred because it had a very good accuracy for classifying multiple categories of data.

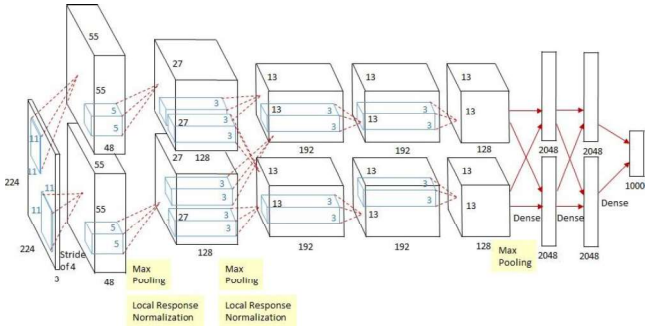


Fig. AlexNet Architecture

D. Introduction to Transfer Learning

Transfer learning is a commonly used machine learning technique primarily used when the available dataset is limited. It is a process where the network is trained on a largedataset with similar data and then modifying the same to work well on the target data. In this particular application, we identify the bird species based on their Spectrogram images of the bird sounds. AlexNet was trained using transfer learning to recognize new categories of images which are thespectrogram images of the sounds of various bird species.

E. Transfer learning using AlexNet

A pre-trained convolutional neural network, AlexNet wasused for classification. In this case, we had to retrain AlexNet to recognize new categories of objects (Spectrogram images of the sounds of different bird species) for our application domain. Transfer Learning provided the advantage of not having to build an effective neural network model fromscratch.

Initially, we gathered sounds of two different bird species namely, Crow and Sparrow. The spectrogram images of these bird sounds were generated and saved. Since the classification objects are different from the ones AlexNet wastrained on, the network is not trained from the top of the network, instead the last few layers are tweaked. The fully- connected layer, which holds the number of classes of images, is changed from one thousand to three and the output layer is frozen to hold the categorical output of the classification. Once the training was done, a random set of spectrogram images were used for validating the model. In order to enhance the performance metrics, parameters such as learning rate, number of epochs, batch size and the split up oftraining and test data were varied. The same model was applied for recognizing human voice samples after successfulvalidation of bird species detection.

Number of species	Data Split	Epoch	Accuracy
2	80:20	20	92%
2	70:30	20	90%
4	80:20	20	88%
4	70:30	20	85.25%
4	80:20	35	97%
4	70:30	35	94%

Table 1 Training Results

REAL-TIME IMPLEMENTATION:

Now, that we have trained the convolutional neural network (AlexNet), we can predict the bird species for a given input sound recording. But these input sound recordings have been collected in ideal environments where the noise is virtually non-existent and in some cases the noisehas been removed by some pre-processing techniques. Hence, the network has trained on a dataset which can predicta given species when the input data is devoid of any perturbances or noise. Unfortunately, this isn't the case in realtime recording as there is the presence of noise in ambient environment. The noise can be due to many factors like vehicular sounds, overlapping human voices and natural phenomenon. It must be made sure that the network works as intended and performs to the same level as it used to performin a simulated environment as in the case of the prediction ofbird species on a noiseless dataset. To ensure that the network works

in real time, the CNN must be retrained on a dataset containing the ideal dataset as well as sound samples which have been collected from an ambient environment.

The audio clips of the different species of birds present inthe dataset are chosen randomly is converted into a fixed sampling rate of 44100Hz or 48000Hz to preserve the diversity and to prevent overfitting. Also, the bit rate is set to128kbps and 320kbps as they are the standard bit streams thatare used in audio applications because it ensures a clear audiorecording with a low file size. After all the audio clips are converted into the desired sampling rates and bit streams, spectrograms are generated for each audio clip. These spectrograms are used to retrain the Neural Network. After transfer learning is completed, the model can be saved and reused to classify the real time audio signal by converting it into spectrograms. It is recommended to configure the microphone to have a sampling rate of 44100 Hz and a bit rate of 128 kbps or 320 kbps. The system was tested in a realtime environment which produced classification results up toan accuracy of 91%. A Graphical User Interface (GUI) was designed to operate the above system which involved all the processes to be executed, right from the recording of sound in a real time environment to processing the data and displaying the results.

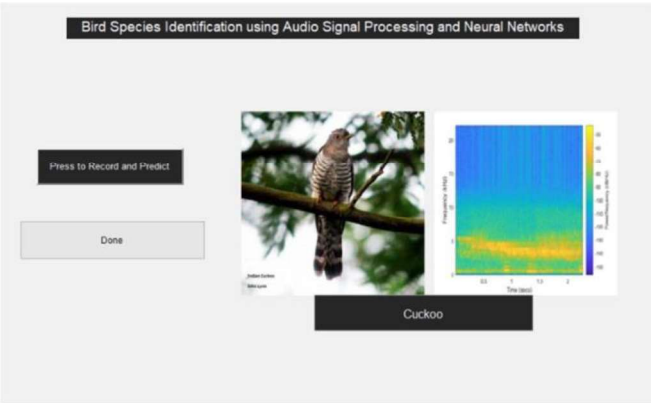


Figure 6 Real-Time Output

FUTURE SCOPE:

This project has an exponential scope for improvements in the future in terms of economic as well as scientificopportunities. An application can be designed and deployed for mobile devices, which would enable users to use their smartphones as handheld devices for prediction and analysisof bird sounds. The bird sound can be recorded by the user, the app then processes the recording on the device and returnsthe result along with the images, description and population distribution of the bird. The recording can also be sent to a cloud server running a sophisticated CNN for meticulous analysis and assessment to obtain more accurate results. TheCNN can also be deployed on a hardware setup like a NeuralCompute Stick or a Raspberry Pi. These hardware setups can Compute Stick or a Raspberry Pi. These hardware setups can be installed in ecological parks, conservation parks and bird sanctuaries. The data obtained can either be stored locally oron the cloud. The data thus obtained will be of huge significance in studies relating to bird migration patterns, population distribution, biodiversity and bird demographics in a given area.

CONCLUSION:

In this project, four different bird species were identified.The approach involved pre-processing of the bird sounds followed by the spectrogram generation of the same and these were used to train the model for classification. The data usedfor training consisted of real bird sounds recorded in their natural habitat amidst all other sounds. The outputs were observed for different values of learning rates, number of epochs and data split. The system was able to classify bird species based on the spectrogram image generated from their sounds with an accuracy of 97%. This accuracy was obtained considering human voices along with bird sounds. The accuracy can be further enhanced by fine tuning the performance parameters.

REFERENCES:

[1] www.iucn.org/theme/species/our-work/birds

[2] www.xeno-canto.org/

[3] Sujoy Debnath, Partha Protim Roy , Amin Ahsan Ali, M Ashraful Amin" Identification of Bird Species from Their Singing", 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016

[4] Rong Sun, Yihenew Wondie Marye, Hua-An-Zhao "FFT Based Automatic Species Identification Improvement with 4-layer Neural Network", 2013 13th International Symposium on Communications and Information Technologies (ISCIT)

[5] www.mathworks.com/help/deeplearning/ref/alexnet.html