Animal Vocalization Recognition

Anushka Naik, Ayesha Sana, Aysha Pm, Mohammad Uvais, and Divya K K

Department of Computer Science and Engineering, P A College of Engineering, Mangalore574153

E-mail:

Abstract

The identification of animal vocalizations plays a vital role in bioacoustics and ecological studies. This project introduces a machine learning-based system that can accurately classify animal sounds, including bird chirps and marine mammal calls, by analyzing audio samples. Its central goal is to automate the detection of species-specific vocal patterns to aid in behavioral analysis and species recognition.

The approach employs sophisticated audio signal processing methods to extract relevant features, such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms, which are then fed into neural network architectures like CNNs and RNNs. These models are trained and evaluated on datasets encompassing diverse acoustic settings to ensure broad adaptability.

Potential uses for this technology include estimating wildlife population sizes, observing migration routes, and monitoring behavioral changes linked to environmental disruptions. This initiative aims to support global conservation and ecological efforts by offering a scalable, efficient tool for wildlife monitoring.

ISBN:97881-19905-39-3

1 Introduction

Animal sounds—such as bird melodies and whale vocalizations—offer valuable insight into wildlife communication and habitat interactions. These acoustic signals help researchers estimate species population sizes, understand behavioral patterns, and assess environmental impacts. However, manually analyzing these recordings is time-consuming, labor-intensive, and inefficient at scale. This project proposes an automated machine learning system designed to recognize and differentiate various animal vocalizations. By extracting distinct features like frequency and rhythmic structure from audio inputs, the model identifies the species responsible for the sound.

The system serves as a powerful tool for researchers and conservationists, simplifying the process of monitoring animal populations, migration behaviors, and responses to ecological changes such as climate shifts and human disturbances. Overall, the use of technology in this context enables more effective biodiversity preservation and environmental understanding.

2 Literature Survey

Researchers have long explored various methods to interpret and categorize animal vocalizations. Initially, this process was entirely manual—experts would analyze audio files, inspect spectrograms, and manually identify which animal made which sound. In the study by Qiang Yang[1], Xiuying Chen[1], Changsheng Ma[1](2024), the algorithm primarily relied on Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction and a Convolutional Neural Network (CNN)-based model for classification. While informative, this manual technique proved to be laborious and impractical when dealing with large-scale audio data from natural habitats.

To address these challenges, early automated systems incorporated signal processing techniques like Mel Frequency Cepstral Coefficients (MFCCs) and Zero Crossing Rate to extract sound characteristics in a machine-interpretable format. These features were input

ISBN:97881-19905-39-3

into traditional classifiers such as Support Vector Machines (SVM) and Random Forests. Although effective for small and clean datasets, these models often failed to perform in complex acoustic environments, particularly when overlapping sounds or background noise were present. For instance, in a 2024 study by Qiang Yang and colleagues, the authors proposed a feature optimization framework to enhance animal sound classification performance in challenging scenarios.

With the rise of deep learning, particularly Convolutional Neural Networks (CNNs), audio classification entered a new era. Unlike traditional models, CNNs automatically learn significant features from spectrograms, eliminating the need for manual feature selection. Studies by researchers like Stowell and Graving have demonstrated how CNNs can accurately classify bird songs and marine mammal calls, even in noisy environments.

More recent advancements include the integration of Internet of Things (IoT) devices for real-time data acquisition in remote environments and the use of self-supervised learning models that can learn patterns from unlabeled audio. These techniques are especially beneficial in wildlife monitoring, where labeled data is limited and new or rare animal calls remain unclassified.

3 Algorithms

In the study by Qiang Yang[1], Xiuying Chen[1], Changsheng Ma[1](2024), the algorithm primarily relied on Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction and a Convolutional Neural Network (CNN)-based model for classification. The process began with converting raw audio recordings into spectrograms and MFCC features, which effectively captured the unique frequency and time-based characteristics of bird vocalizations. These features were then fed into a CNN model, which used its layered architecture—comprising convolutional layers, pooling layers, and fully connected layers—to learn distinguishing patterns in the bird calls. The CNN was trained to identify and classify different species

ISBN:97881-19905-39-3

based on these vocal signatures. To enhance robustness, techniques like data augmentation and noise filtering were applied to handle variability in recordings caused by background sounds and environmental disturbances. The overall algorithm focused on mapping complex sound patterns to specific bird species while enabling potential extensions for estimating population density based on call frequency and distribution.

In the work of Qiang Yang and his team (2024), the researchers developed a classification pipeline that relies on MFCCs to capture critical acoustic properties and a Convolutional Neural Network (CNN) to perform species recognition. Their approach involves transforming raw audio signals into spectrograms and MFCC-based feature sets. These features are processed by the CNN, which includes layers designed to extract and generalize key audio patterns. Data augmentation techniques and background noise filtering were also applied to improve the model's robustness in natural conditions.

For our project, the algorithm follows a similar multistage process that begins with feature extraction. Using MFCCs, the audio signals are converted into numerical representations that reflect the texture and tonal characteristics of animal sounds. Spectrograms are also generated to visualize frequency changes over time.

These representations are passed into a CNN, a deep learning architecture adept at recognizing spatial patterns within images—making it ideal for spectrogram input. The CNN contains convolutional layers for pattern detection, pooling layers for dimensionality reduction, and fully connected layers for decision-making and classification.

4 Data Collection

The foundation of any effective vocalization recognition system lies in the quality and variety of its dataset. For this purpose, audio data is collected from a wide range of species in natural ecosystems. Public resources such as Xeno-Canto, Macaulay Library, the Animal Sound Archive, and the Cornell Lab of Ornithology offer extensive repositories of labeled

ISBN:97881-19905-39-3

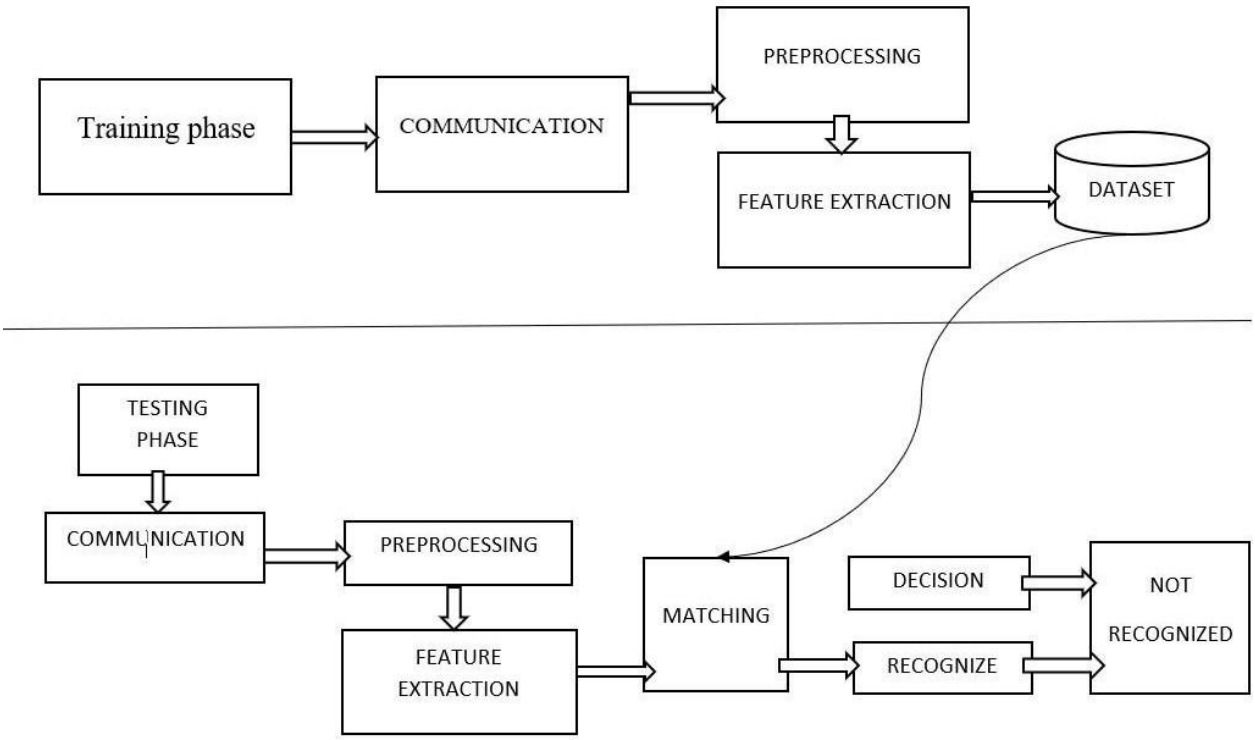

Figure 1: Architecture diagram

recordings, especially for birds.

In addition to these databases, field data can be collected using audio sensors or IoT-based recorders strategically placed in forests, oceans, or wetland areas. These devices continuously monitor and capture environmental audio, providing region-specific and real-time data.

Once gathered, the audio files are segmented into smaller clips for easier processing. Each segment is reviewed either manually or using annotation tools to label the corresponding species. Recordings are collected under diverse conditions—including various times of day and levels of background noise—to ensure that the model remains robust in different acoustic scenarios.

Preprocessing steps, such as silence removal, noise reduction, and class balancing, are

ISBN:97881-19905-39-3

applied to clean the dataset and prepare it for training.

5 Dataset Annotation:

Proper annotation is crucial to model training success. In this stage, the collected audio clips are tagged with labels indicating the species name, vocalization type (such as a call, song, or warning), and the exact timing of each sound.

While experts often carry out manual labeling to ensure precision, automated or semi-automated tools are increasingly used to assist the process. These tools generate preliminary labels by analyzing visual representations of the audio, like spectrograms, which can help in identifying distinct patterns unique to each animal.

Annotations also highlight areas with overlapping calls, background disturbances, or indistinct vocalizations. These marked segments allow the model to learn how to differentiate relevant sounds from background noise, significantly enhancing real-world performance.

Accurate annotation leads to better model generalization, enabling the system to recognize animal calls across various environments and conditions.

6 Training And Testing the Model

Developing a reliable animal sound recognition system requires a systematic training and evaluation process. After the dataset has been annotated, it is split into three subsets: training, validation, and testing. The training set is used to teach the model by exposing it to labeled examples of animal sounds. These sounds are typically converted into spectrograms or MFCC matrices to be compatible with the model input.

A Convolutional Neural Network (CNN) is then trained on these visual or frequency-based representations. As the model processes each training sample, it learns patterns by adjusting internal parameters to minimize errors, using optimization techniques like Adam or Stochastic Gradient Descent. The objective during training is to reduce classification loss,

ISBN:97881-19905-39-3

commonly measured by cross-entropy.

To fine-tune model behavior and avoid overfitting, the validation set is used. It provides feedback on how the model performs with data it hasn't been trained on, helping in optimizing parameters such as learning rate, number of layers, and dropout values.

The final model is tested on the separate test set to determine its practical effectiveness. Key performance metrics—such as accuracy, precision, recall, and F1-score—are computed to understand how well the model can generalize to unseen data.

In our approach, MFCC features are organized into matrices: $\mathbf{M=\{f_{11}...f_1N,...,fD_1...fDN\}}$. These matrices are reshaped to preserve temporal relations, considering both previous and future audio frames. To eliminate unnecessary information and enhance robustness, irrelevant features are filtered out. The processed MFCCs are then input into a hybrid deep learning network consisting of Bi-directional LSTM units, which are capable of analyzing forward and backward sequences. An attention layer is added to focus on the most significant parts of the input, enhancing the accuracy of species identification even in noisy conditions.

7 Result and Discussion

The ability to recognize animal vocalizations through machine learning represents a major breakthrough in ecological data collection. It shifts the paradigm from manual analysis—often slow and error-prone—to automated, efficient, and scalable monitoring of wildlife through sound. Using deep learning techniques such as CNNs and feature-rich audio inputs like spectrograms, this system provides detailed insights into species identification, behavioral trends, and population movements. These insights can support conservation policies, help detect endangered species, and assess habitat quality across regions.

Integration with IoT devices amplifies the system's power, allowing remote areas to be monitored in real-time. With low-energy sensors and smart devices capturing bioacoustic

ISBN:97881-19905-39-3

data around the clock, urgent threats like illegal poaching or deforestation can be identified early and addressed quickly. Emerging technologies, such as self-supervised learning and transformer-based models, offer new avenues for systems to learn from unlabeled audio—addressing one of the biggest challenges in wildlife research. Additionally, techniques like federated learning enable collaborative model training without centralized data storage, ensuring both data privacy and efficiency.

Edge computing further enhances system performance by allowing quick, local analysis of sound data. This minimizes reliance on cloud services, reduces latency, and enables real-time responses in the field. When deployed strategically, this combination of AI and hardware can transform how conservationists monitor biodiversity and protect vulnerable ecosystems.

Table 1:

Animal Category	Accuracy (%)
Bird Chirping	92.5
Cat	88.0
Dog Barking	95.0
Elephant Trumpet	90.0
Horse	85.5

Figure 1: Animal Recognition Classification Accuracy table

Table 2:

Shots	Noise Level 0.1	Noise Level 0.9
1-shot	0.865	0.982
2-shot	0.943	0.989
3-shot	0.976	0.993
4-shot	0.987	0.996

Highest Accuracy on Animal Prediction Test Set

8 Conclusion

The system presented offers a promising shift in wildlife conservation practices by automating the detection and interpretation of animal sounds. This technology eliminates the need for

ISBN:97881-19905-39-3

manual listening, enabling rapid and accurate analysis of ecological data.

The trained model can quickly identify shifts in animal behavior, detect patterns of migration, and even pinpoint disturbances in ecosystems, such as habitat loss or pollution. As climate change and human expansion continue to impact natural habitats, having a reliable acoustic monitoring tool becomes increasingly important.

By facilitating quicker decision-making and offering deeper ecological insights, this solution stands to play a vital role in long-term biodiversity management and environmental research.

ISBN:97881-19905-39-3