Sounds Classification and Detection for Rain forest Species Using Custom CNN And Combinations Of Spectrograms

Shreejaa Talla
Department of Computer Science
Georgia State University
Georgia, USA
stalla1@student.gsu.edu

Sri Vinesh Pothana
Department of Computer Science
Georgia State University
Georgia, USA
spothana1@student.gsu.edu

***Abstract* --- Classification and detection of sounds with limited and complex datasets are pretty challenging. Real-time sound classification for rain forest species identifies the rarer species, which can help conservation management make effective decisions for their safety against human impacts on the environment [1]. The existing detection and classification algorithm needs enormous data sets to train the model but finding a large amount of data for rare species is difficult. Hence, the proposed solution concentrates on building a solid generalization model that can produce accurate results even when the training data is limited and complex. Additionally, extracting different combinations of spectrograms for those sounds can help the model have better learning capability. Therefore, this model uses convolutional neural networks with the leaky rectified linear unit as an activation function to improve the training ability. As a result, based on the research observations, a short-time Fourier transformation spectrogram combined with Mel-frequency coefficient has predicted different types of species with different sound ranges compared to other combinations.**

***Keywords* – Convolution Neural Networks, sound classification and detection, spectrograms.**

I. Introduction

Sound detection and classification can help conservation management to better understand the species present in the rainforest and help to protect the rarest ones, further helps in making decisions on maintaining these species in forests. Algorithms concerning convolutional neural networks have been widely used for automatic detection and classification in various fields. One of the latest involvements of CNN is in the field of Acoustic scene classification, which is implemented in various ways. One of the basic ways is to implement a direct CNN approach, but these had many implementations throughout the time.

A. *Related Works:*

In 2016 [14], baby cry detection in a domestic environment using deep learning has compared the logistic regression classifier with CNN and the result obtained by CNN had a considerable advantage over logistic regression classifier.

According to [5], the classification model was developed based on rectifier linear unit as the activation function in convolution layers to reduce the time taken in the training process is a much better and easy way but it still requires large sets of data for classification.

In 2017 [13], Infant baby cry classification was proposed based on SVM based neural network and MFCC, where MFCC is used to select the feature vector which improved the classification process. [4] Dilated convolution neural network with a Leaky ReLU algorithm was proposed which has achieved great accuracy and classification ability when compared to Convolution neural network.

In 2018 [2] approach was proposed which involves audio signal processing where MFCC and log-Mel were combined to extract the feature vectors which helps to enhance the abilities of the classifier much more when compared to other combinations or a single method.

In 2019 [3], a Deeper learning algorithm was proposed which has a strong generalization ability but needs a large set of data to train that algorithm. [5] For CNN along with a weighted filter helps to enhance the generalization ability, even while using weakly labeled data.

II. Methods

A. *Database:*

The dataset consists of test and train unlabeled audio files that contain various recordings of species. Here, the training data includes both false positive and true positive values [1]. Additionally, the train data provides 23 unique combinations of song type and species which are considered as labels.

B. *Data Preprocessing:*

The datasets contain train csv file for true positive and false positive values and labels for each recording based on their frequencies and timings. Hence, the data is sliced based on the minimum and maximum time when a particular species is identified. One of the audio files processed is shown in Figure 1 below.

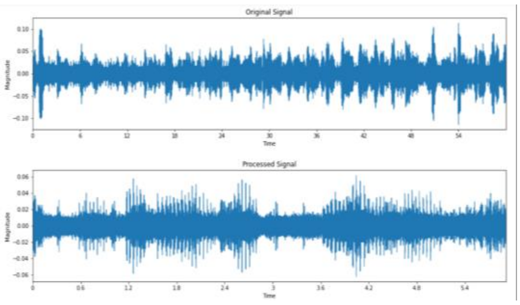


Figure 1 Data preprocessing.

C. *Audio Augmentation:*

According to [4], Audio Augmentation Method can help to reduce overfitting when the data is not large enough compared to the network size. Here, the given dataset is limited and complex, hence, data augmentation will increase the data size by some transformations. For audio waves, there are two types of widely used transformations.

One is Time stretching which aims to speeding up or slowing down the sample audio by linear interpolation between audio frames. The Time stretching formula is

$$y=(1-k).y0 + k.y1$$

where k= time-stretching factor controlling increasing or decreasing speed of samples, y0=previous audio frame, y1=next audio frame, and y=interpolated audio frame.

Another is Noise adding, where samples are mixed with other recordings with different background sounds. This noise adding can increase the complexity of the given dataset. Hence, the Time stretching method is used for data augmentation for the proposed model.

One of the recordings from the dataset is augmented using time stretching method and is shown in Figure 2 below.

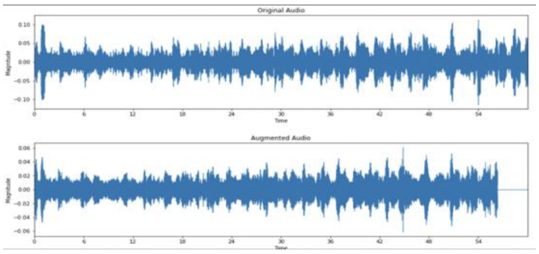


Figure 2 Audio Augmentation.

D. *Mel-Frequency Cepstral Coefficient(MFCC):*

According to [10], MFCC is popular in automatic speech and sound recognition as they were build based on human ear perception. The feature was extracted using the following steps:

- Splitting of a sound sample into frames
- Applying Discrete Fourier Transformation (DFT) on each framed window.
- Triangular overlapping frames are mapped into Mel-scale.
- At each frequency, the logarithm of the amplitude spectrum is calculated.
- Discrete cosine transformation is applied to log amplitudes resulting in MFCCs.

E. *Log-Mel Spectrogram:*

The Mel Scale is a logarithmic transformation of a audio signal's frequency. Log scaled Mel Spectrograms are spectrograms that visualize sounds on the Mel scale [16].

F. *Short-Time Fourier Transformation (STFT):*

STFT is nothing but an improved Fourier Transform (FT) version with both time and frequency information. STFT is used in a situation where we need to analyze the frequency of a signal which varies over time. In STFT the time signal is divided into many overlapping frames with the help of time-domain windows and for each frame, Fourier transform is applied, by this time and instantaneous frequency information [15].

G. *Combination of spectrograms:*

Audio signals are converted to mono by taking the mean values from the left to right channel with a sampling frequency of 44,100hz [2]. The combination of Log-Mel and MFCC is used to extract acoustic features.

According to the model in [2], this combination to generate acoustic features has increased the accuracy 8.7% when used individually and with other gammatone for sound classification that involves complex backgrounds. Hence, the proposed model uses different combinations of spectrograms, as shown in Figure 3a, 3b, 3c below are derived from one of the recordings.

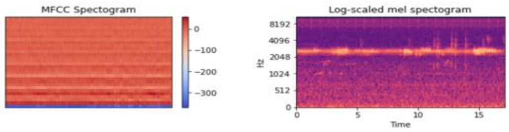


Figure 3a Combination of MFCC and Log-Mel.

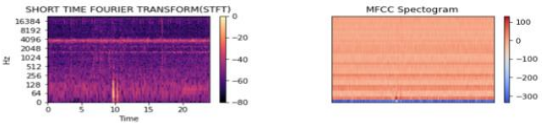


Figure 3b Combination of STFT and MFCC.

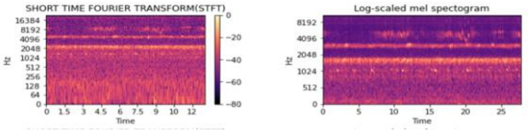


Figure 3c Combination of STFT and Log-Mel.

H. *Leaky Rectifier Linear Unit (Leaky ReLU):*

The Leaky ReLU is an enhanced activation function of the ReLU proposed method in [7]. ReLU helps to speed up the training process in a much better way when compared to classic sigmoid function, however, there is a 50% loss of information. As the data is limited and complex losing 50% of data on every run of activation function can lead to a lack of data, therefore, the Leaky ReLU function is introduced in the proposed architecture. Leaky ReLU, on one hand, inherits the non-linear and low computational complexity properties of ReLU, on other hand, all the feature information has remained to a certain extend.

I. *Convolution Neural Network Algorithm:*

Convolution Neural Network (CNN) is a quite powerful neural network for image recognition and recently, CNN is widely used in speech and acoustic classifications and detection.

III. PROPOSED ARCHITECTURE

The proposed model has training and testing modules, and this model is implemented to train and predict complex datasets. The following methods are used: Leaky ReLU, Data Augmentation layer, and inputs from different combinations of spectrograms to overcome the data limitation and complexity problem and train the model efficiently to predict unknown sounds.

The proposed architecture is shown in Figure 4 consists of a given unlabeled audio dataset with separate data files for testing and training modules. Firstly, the training module consists of a data preprocessing layer where audio files are trimmed based on minimum and maximum time, audio augmentation layer using the time stretching method, followed by combined spectrogram feature extraction to CNN-2D layers block1 as shown in Figure 6a, block2 as shown in Figure 6b, block3 as shown in Figure 6c and block4 as shown in Figure 6d, optimizer Nadam layer and categorical cross entropy loss function.

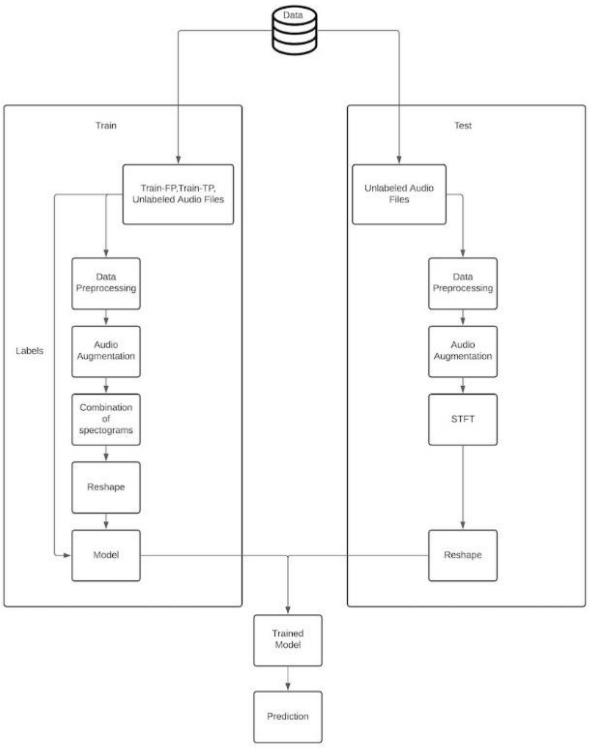


Figure 4 Proposed Architectural Diagram.

Training module workflow is explained as follows, Audio files from training data are sent as input to the data preprocessing layer where audio files are trimmed based on minimum and maximum time. This processed data is sent to audio augmentation layer which produces the audio file by stretching time. The output obtained in the above layer is send 50 percent to one spectrogram and 50 percent to another spectrogram based on different combinations as per Figure1a, 1b, 1c to generate audio features. Block-1 CNN-2D uses these features to arrange them and subsample using max pooling layer. Block-2 CNN-2D takes input from block1 CNN-2D where the convolutional layer will arrange the input as a set of linear activation using a Leaky ReLU and include batch normalization to speed up the training process ignoring the internal covariate shift phenomenon [9]. At the end of the block, max pooling is used to subsample the features obtained after CNN layer. This block is iterated three times with different filters and kernel size, as shown in Figure 5.

CNN-2D block-3 consists of batch normalization, dense layer, and dropout and process the features obtained from the second block of the CNN network, iterated for three times for various dense layers.

Finally, Block-4 contains the last dense layer, batch normalization and softmax. The final output is optimized based on Nadam and loss is calculated based on categorical cross entropy function.

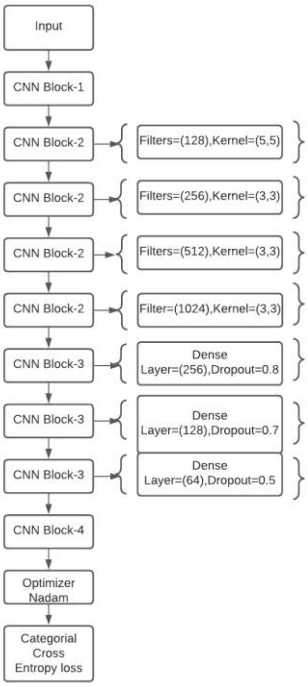

Input
CNN Block-1
CNN Block-2
Filters=(128),Kernel=(5,5)
CNN Block-2
Filters=(256),Kernel=(3,3)
CNN Block-2
Filters=(512),Kernel=(3,3)
CNN Block-2
Filter=(1024),Kernel=(3,3)
CNN Block-3
Dense Layer=(256),Dropout=0.8
CNN Block-3
Dense Layer=(128),Dropout=0.7
CNN Block-3
Dense Layer=(64),Dropout=0.5
CNN Block-4
Optimizer Nadam
Categorial Cross Entropy loss

Figure 5 Classification model.

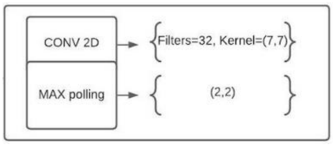
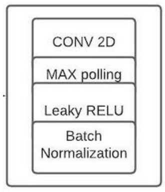


Figure 6a CNN-2D Block1.



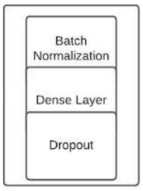


Figure 6b CNN-2D Block2. **Figure 6c** CNN-2D for Block3.

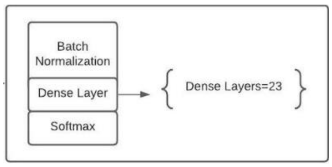


Figure 6d CNN-2D for Block4.

For prediction, the testing model consists of a data preprocessing layer where each audio file is split to multiple audio frames each of 10 seconds. The output obtained from data preprocessing is input to feature extraction where the test files are converted to STFT spectrograms and sent to a trained CNN model that predicts sounds of different species.

IV. Comparing Models Based On Spectrograms

Based on the above architecture, three combinations of spectrograms are used for feature extraction, and the following are the training- validation accuracy and loss graphs for the combination of log-mel and MFCC in Figure 7a, the combination of STFT and log-mel in Figure 7b, and combination of MFCC and STFT in Figure 7c. Although training accuracy and validation accuracy has a trivial difference in each graph, the loss graph is considerably unique for each.

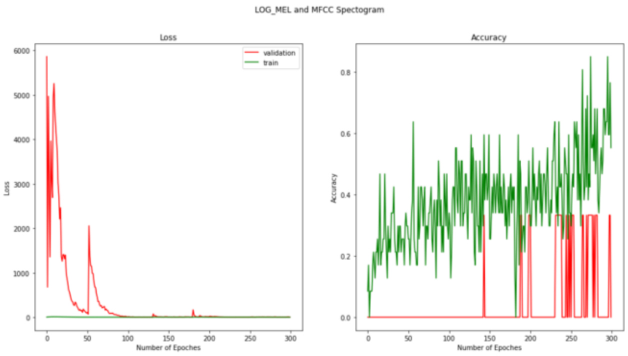


Figure 7a Accuracy and loss graph for log-mel and MFCC.

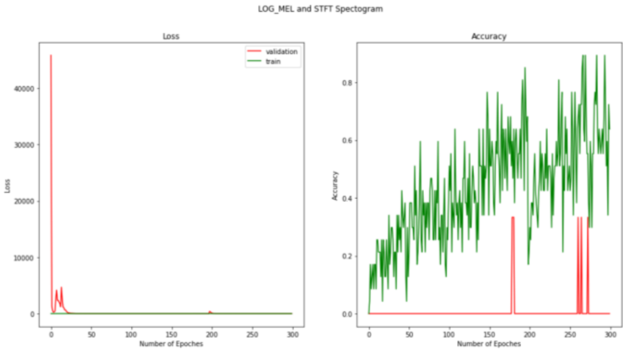


Figure 7b Accuracy and loss graph for log-mel and STFT.

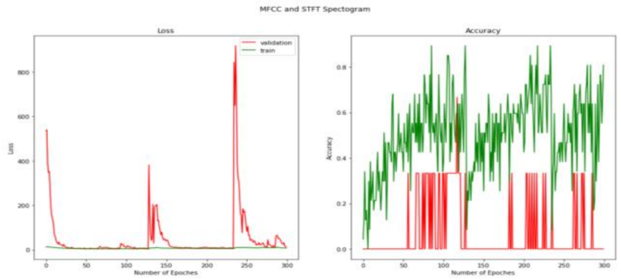


Figure 7c Accuracy and loss graph for STFT and MFCC.

After the training is finished for each model, the unlabeled test data are sent as input to the following trained models, and the sounds of various species are predicted. These predictions are plotted using a boxplot for the range of each species detected in all the test recordings. For the first combination of features, that is, log-mel and MFCC, the predicted species are quite a few, as shown in Figure 8a, whereas, for the combination of log-mel and STFT, results obtained are much better, as shown in Figure 8b. Overall, the range of species obtained using the combination MFCC and STFT has predicted the species more accurately for all recordings based on Figure 8c.

As a result, the STFT spectrogram combined with the other two spectrograms provided compelling observations. Hence, MFCC and STFT combination is considered best among others as they detect complex sounds with more certainty.

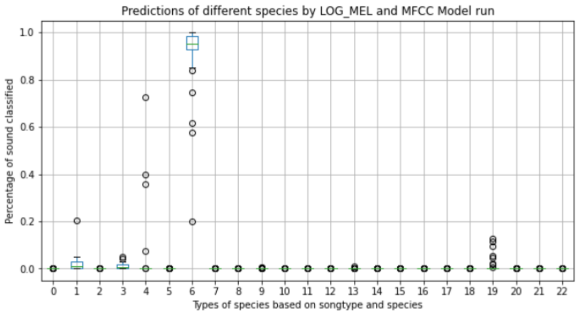


Figure 8a Range of each species detected from log-mel and MFCC.

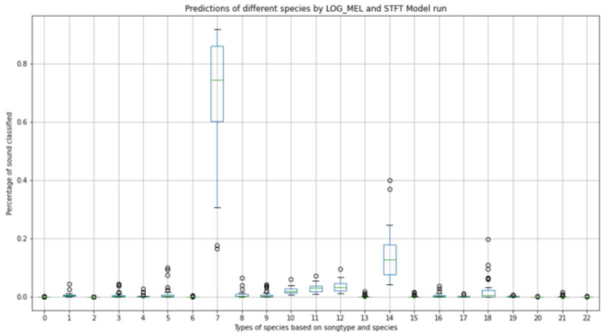


Figure 8b Range of each species detected from log-mel and STFT.

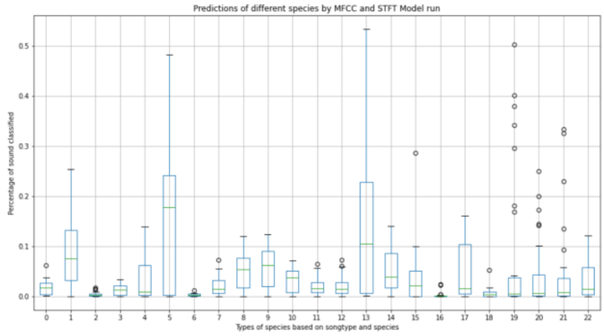


Figure 8c Range of each species detected from STFT and MFCC.

V. Conclusion

Based on the above boxplots, it is clear that for same samples of recording log-mel and MFCC combination was able to detect only one type of species and all other were almost null. However, Log-mel and STFT has provided slightly better predictions than the log-mel and MFCC. To summarize, STFT-MFCC combination-based model has predicted each species in wide ranges for different recordings.

Acknowledgment

The authors would like to thank Dr. Zhang, Data mining faculty at Georgia State University for his support.

References

[1] https://www.kaggle.com/c/rfcx-species-audio-detection/overview

[2] A. Dang, T. H. Vu and J. Wang, "Acoustic scene classification using convolutional neural networks and multi-scale multi-feature extraction," 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, 2018, pp. 1-4, doi: 10.1109/ICCE.2018.8326315.

[3] T. Doan, H. Nguyen, D. T. Ngo, L. Pham and H. H. Kha, "Acoustic Scene Classification Using A Deeper Training Method for Convolution Neural Network," 2019 International Symposium on Electrical and Electronics Engineering (ISEE), Ho Chi Minh, Vietnam, 2019, pp. 63-67, doi: 10.1109/ISEE2.2019.8921365.

[4] X. Zhang, Y. Zou and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," 2017 22nd International Conference on Digital Signal Processing (DSP), London, 2017, pp. 1-5, doi: 10.1109/ICDSP.2017.8096153.

[5] B. Tang, Y. Li, X. Li, L. Xu, Y. Yan and Q. Yang, "Deep CNN Framework for Environmental Sound Classification using Weighting Filters," 2019 IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, 2019, pp. 2297-2302, doi: 10.1109/ICMA.2019.8816567.

[6] M. Lin, Q. Chen, and S. Yan, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[7] Xu, Bing, et al. "Empirical evaluation of rectified activations in convolutional network." arXiv preprint arXiv:1505.00853 (2015).

[8] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[9] S.Ioffe and C.Szegedy, "Batch normalization: Accelerating deep network training be reducing internal covariate shift," in CoRR, vol.abs/1502.03167, 2015.

[10] S. Soni, S. Dey and M. S. Manikandan, “Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices,” 2019 7th Int. Conf. Digital Inform. Process. and Comm. (ICDIPC), Turkey, 2019, pp. 23-28.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, pp. 2278-2324, 1998.

[12] K. E. Schmidt and S. Löwel, "Long-range intrinsic connections in cat primary visual cortex," in The cat primary visual cortex, ed: Elsevier, 2002, pp. 387-vi.

[13] C. Y. Chang et al., ”DAG-SVM based infant cry classification system using sequential forward floating feature selection,” Multidim. Syst. Sign. Process., Vol. 28, No.3, pp. 961-976, 2017.

[14] Y. Lavner, R. Cohen, D. Ruinskiy and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, 2016, pp. 1-5, doi: 10.1109/ICSEE.2016.7806117.

[15] Rubayet-E-Azim, N. Karmakar and E. Amin, "Short Time Fourier Transform (STFT) for collision detection in chipless RFID systems," 2015 International Symposium on Antennas and Propagation (ISAP), 2015, pp. 1-4.

[16] N. Peng et al., "Environment Sound Classification Based on Visual Multi-Feature Fusion and GRU-AWS," in IEEE Access, vol. 8, pp. 191100-191114, 2020, doi: 10.1109/ACCESS.2020.3032226.