Audio Classification of Underwater Mammals using Convolutional Neural Network

Monika Aggarwal[1,#], Shivangi Roy[2]

[1]*CARE Department, IIT Delhi, India*

[2]*Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi*

[1,#]Corresponding Author, Email: maggarwal@care.iitd.ac.in

***Abstract*— This paper focuses on the identification and classification of unwater mammals based on sound produced by them. It describes the approach taken to classify the different audio samples of 5 classes of underwater mammals namely, Atlantic Spotted Dolphin, Bearded Seal, Bottlenose Dolphin, Humpback Whale, and Walrus. Detecting the presence of different underwater mammals in the sea provides a great level of help in preventing collision of huge ships with the underwater mammals thus saving their lives and also preventing any possible damage to the ship fans. The motivation behind this work is to save aquatic mammals using deep learning to avoid manual overheads. The model has been trained to work successfully even in cases where high noise levels are present in the audio samples. The paper presents the use of training a convolutional neural network model to achieve high levels of accuracy of audio classification of sound produced by marine mammal even with different noise levels.**

***Keywords*— CNN: Convolutional Neural Network, ROC: Region of Convergence, MFCC: Mel Frequency Cepstrum Coefficients, SNR: Signal to Noise Ratio, AWGN: Additive White Gaussian Noise**

1. INTRODUCTION

Different underwater mammals produce sound at different frequencies. These sounds can be captured [1] to identify the marine mammal that produced it. These sounds, along with the high level of noise naturally present underwater, specifically in oceans, are captured and a series of such audio samples are stored in multiple audio files. These audio files have been used in this project as input training audio sets for training the neural network. The dataset is cleaned before the training process. Features need to be extracted from the dataset for the classification [2] purpose. Fast Fourier Transform has been used in the feature extraction process to extracts the MFCCs. The methodology followed uses a CNN [3] model which is trained to predict the type of underwater mammal and classify [4][5] them into various classes. Further, AWGN of highly negative SNR levels has also been added to the dataset and the model is rigorously trained to achieve good accuracy values.

2. EXPERIMENTAL DATA

The Watkins Marine Mammal Sound Database [6] is used. The dataset contains a variety of audio files for various marine mammals collected over the years. These data files consist of audios that are identified to be produced by underwater mammal species in bounded areas during specific seasons. They can also be used as reference datasets for underwater mammal detections from the increasing amounts of Passive Acoustic Monitoring (PAM) data that are being collected worldwide. The dataset used here consists of 1225 audio files of different lengths wherein there are 231 audio files of Atlantic Spotted Dolphin, 134 audio files of Bearded Seal, 187 audio files of Bottlenose Dolphin. 410 audio files of Humpback Whale and 263 audio files of Walrus. Training (60%), validation (20%) and test (20%) are the 3 sets that the dataset has been divided into.

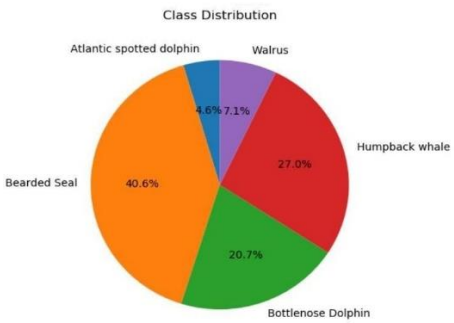

Fig 1. Class Distribution Diagram of Dataset

3. METHODOLOGY

To classify the audio data of various marine mammals, a 7 layered CNN model is used. CNN is a type of deep neural network which is used in image processing and recognition. A CNN can consist of one or more convolutional layers. There can also be fully connected layers in the CNN. It includes a number of algorithms that try to compute underlying relationships in a set of data through a process that works in a way the human brain processes data. It is made up of layers consisting of various neurons. The neurons are made up of learnable weights and biases. Each neuron receives inputs from the neuron of the previous layer, passes it through an activation function and computes the output given to the neuron in the next layer. The input is a 3- dimensional vector.

4. EXPERIMENTAL SETUP AND ANALYSIS

4.1 Data Processing and Sampling

The audio dataset is processed before the training process. The cleaning process of the audio data is done so as to remove very low frequencies that are below the value of 0.01 Hz. The first step in any speech recognition algorithm is to extract features i.e. identify the various components of the audio signal that can be used to distinguish between different classes. The linguistic content of the audio data is identified and all the other stuff like background noise, etc. is discarded.

4.2 Feature Extraction

Feature Extraction [7] is a very critical part of analyzing the audio data. It is required for prediction and classification algorithms. The feature extraction of the audio is done for the classification process. We need to take the Fast Fourier Transform [8][9][10] of all the classes in order to plot the Mel Cepstrum. Cepstrum provides information about the rate of change in spectral bands. During the analysis of time signals, the features can be extracted using Fast Fourier Transform. The discrete Fourier transform of a series or its inverse can be calculated using Fast Fourier Transform. This analysis converts a signal from its time domain to its frequency domain and vice versa. An audio signal is changing with the change in time so we assume that in a fixed time frame, the audio signal doesn't change much.

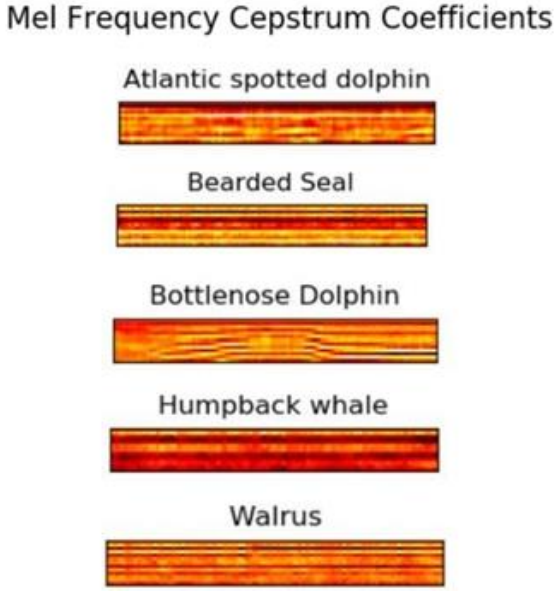

Fig 2. Mel Frequency Cepstrum Coefficients Plot

4.3 CNN Model Training and Testing

The CNN model consists of various layers with different number of neurons. The model has a combination of CNN as well as dense layers. The training as well as validation loss is minimized in this process. Increasing the number of epochs eventually increased training and validation accuracy. The training accuracy comes out to be 98.67% and the validation accuracy comes out to be 98.23%. Further, AWGN of SNR levels like -10dB, -20dB and -30dB is added to the dataset using MATLAB. AWGN is a primary noise model. Various random processes are occurring in nature. The effect of these random processes is imitated using AWGN. The ratio of the power of the signal to that of noise is the SNR. The model is trained and tested with noise. The accuracies obtained are satisfactory even at highly negative SNR levels. The accuracy decreases with decrease the SNR levels from -10dB to -30dB.

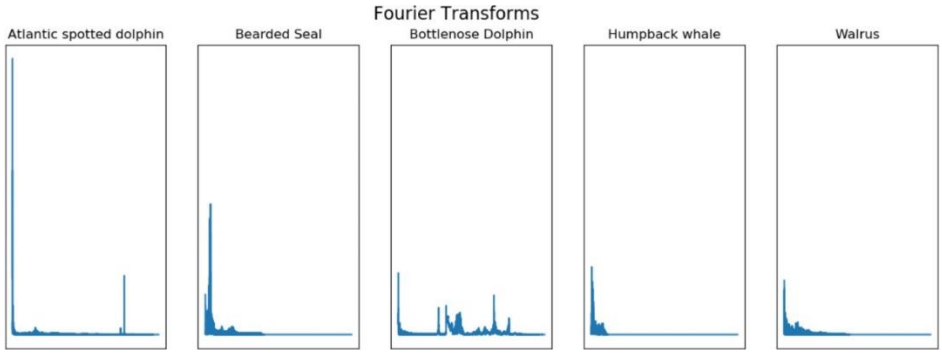

Fig 3. Fourier Transform

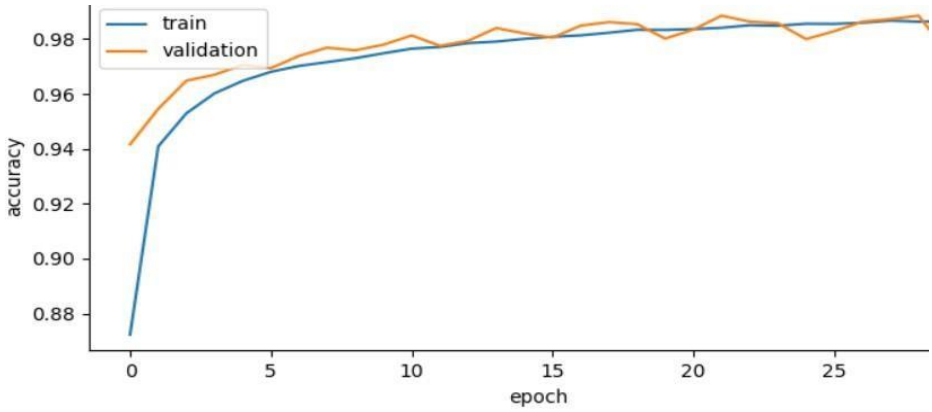

Fig 4. Training Vs Validation Plot

5. Results and Analysis

The samples are provided with labels, i.e., they are classified into the label which has the maximum prediction percentage value. The accuracy of the CNN model was found to be 98.63%. The following is the accuracy obtained when AWGN of different levels of SNR is added to the dataset.

Table I

SNR	Model accuracy
0 dB	98.63%
-10 dB	90.12%
-20 dB	78.59%
-30 dB	71.65%

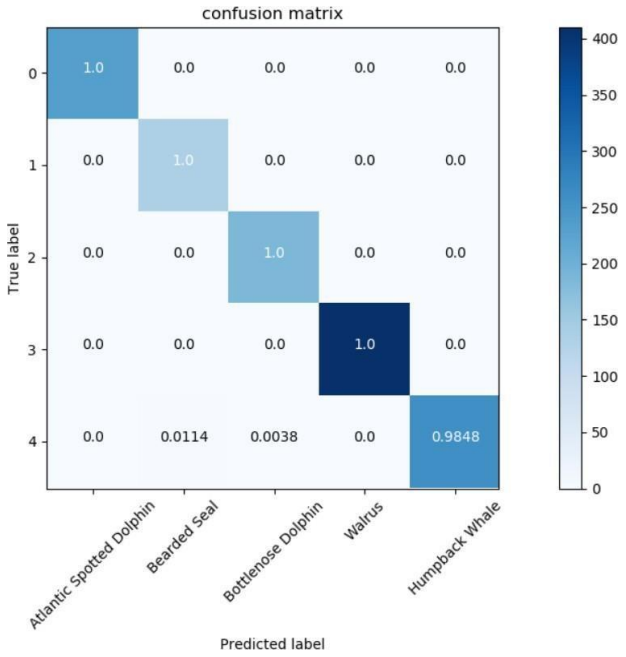

Fig 5. Confusion Matrix

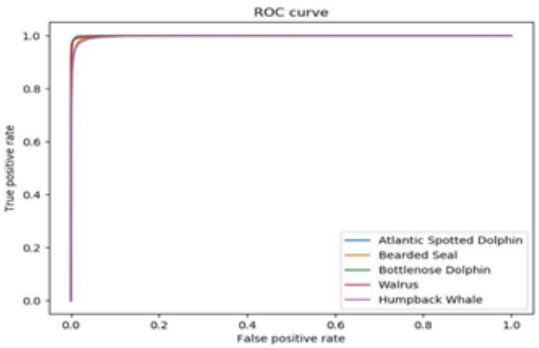

Fig 6. ROC Curve

REFERENCES

[1] Lepper, P. A., Simon, L., & Dufrechou, L., "Autonomous recording system for simultaneous long-term ambient noise and marine mammal monitoring", In OCEANS 2016 MTS/IEEE Monterey, 2016.

[2] Rong, F., "Audio classification method based on machine learning", In 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2016.

[3] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M., "CNN architectures for large-scale audio classification", In 2017 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 131-135), 2017.

[4] Wang, D., Zhang, L., Lu, Z., & Xu, K., "Large-scale whale call classification using deep convolutional neural network architectures", In 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2018.

[5] Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., ... & Wallin, J., "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls", The Journal of the Acoustical Society of America, 135(2), 2014

[6] Watkins, W. A., Fristrup, K., & Daher, M. A, "Marine animal Sound database (No. WHOI-91-21)", WOODS HOLE OCEANOGRAPHIC INSTITUTION MA, 1991

[7] Patel, N. P., & Patwardhan, M. S.," Identification of Most contributing features for Audio Classification", In 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013

[8] Groutage, D., Schempp, J., & Cohen, L., "Characterization and analysis of marine mammal sounds using time-frequency and time-prony techniques", In Proceedings of OCEANS, 1994.

[9] Ma, Y., & Chen, K., "A time-frequency perceptual feature for classification of marine mammal sounds", In 2008 9th International Conference on Signal Processing, 2008.

[10] Lambrou, T., Kudumakis, P., Speller, R., Sandler, M., & Linney, A., "Classification of audio signals using statistical features on time and wavelet transform domains", In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), 1998.