Prediction Of Animal Vocal Emotions Using Convolutional Neural Network

Varun Totakura, Mohana Krishna Janmanchi, Durganath Rajesh, M. I. Thariq Hussan

Abstract: Language is always given paramount importance when it comes to communicating one's idea to another or transferring the information from one person to another. When we generalize this, we can say, communication is the medium of transferring information from one being to another. Being a human, we don't much consider other beings to get the idea of our own into their mind. Because it involves a lot of complexities. Even the beings of the same species as humans find difficulty when one needs to communicate with the person of different language. Considering this fact, the term of intentional language is introduced as the most useful mean to communicate. When someone or something shows intention against you, you can obviously say to the maximum extent what does it mean. It does not involve verbal communication mostly. But other than the verbal we have four more senses, which can play a vital role in translating the information with transparency. We have chosen one such sense: the sound. In our approach, we are very determined to decode detect the animal emotion with the help of the sounds produced by them. In general, animals' communication has four forms: visual communication, auditory communication, tactile communication, Chemical Communication. Because, we are not equipped with enough resources with the technology available (or if equipped we cannot make the process rapid with the present technology), out of these four forms of communication we felt auditory communication would be much helpful to understand animals easily.

Index Terms: Intentional communication, Mel-Frequency Cepstral coefficients, Sparse categorical cross entropy, CNN.

1. INTRODUCTION

Machine learning [1] is a cutting-edge digital technology which we are using these days. It has played a vital role in automating the things. Along with that has also been into classifying or categorizing the day which is an opted option these days in many commercial sectors too. And this is not only used for business purposed but also in the sectors of biology, chemical etc. When classifying the vocal abilities of the humans the machine learning has played an utmost important role. This in turn is used in speech recognition; this speech recognition [1] earned an important place in privacy too. When it comes to humans, we have a large enormous dataset but when it comes to animal's no one has ever thought about how deal with the animal sounds to decode it emotions [2]. We have come a conclusion based on the intentional behavioral action of animals that, animal sounds [3] also one of the factors which helps us to detect its emotions.

This idea made us to move forward on the things. We have made use of the mathematical carriers like the Fourier transformations indirectly. The Fourier transformations used in the reverse order to produce the Mel-Frequency Cepstral coefficients [4]. Along with this we also made use of neural networks in the further experimentation procedure. We used the layered architecture in the network which used rectified linear unit (relu) [5]. The rectified linear unit is a piecewise linear function that gives the output directly for the positive input, otherwise the output will be zero. It is easier to train and gives better performance. The relu is activated using the activation function. Activation function main purpose is to convert an input signal of a node in an artificial neural network to an output signal. That output signal now is used as an input in the next layer in the stack. At the end of the experiment, we have used Adam Optimizer to compile the model which we have built. Adam is an optimization algorithm that can use instead of the classical stochastic gradient descent procedure to update network weights iteratively in training data. Finally, the labels need to be assigned to the data which is acquired by the Sparse categorical cross entropy. Use sparse categorical cross entropy when your classes are mutually exclusive (e.g. when each sample belongs exactly to one class). We used this to classify images at the end.

- *Varun Totakura, Mohan Krishna Janmanchi are currently pursuing B. Tech Degree program in Computer Science & Engineering, Guru Nanak Institutions Technical Campus (Autonomous), Hyderabad, Telangana State, India. E-mail: totakura.varun@gmail.com, jmks.8008@gmail.com*
- *Durganath Rajesh is currently pursuing B. Tech Degree program in Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India. E-mail: rajeshdurganath@gmail.com*
- *Dr. M. I. Thariq Hussan is currently working as a Professor and Head in the Department of Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India. E-mail: thariqhussain@rediffmail.com*

2. RELATED WORK

An attempt has been made to describe some of the responses evoked by communication signals [3] in certain animals and to infer the kind of information which the signals transmit. Using the methods developed by CW. Morris (1946) for the logical analysis of human language, identifiers, designators, appraisers and prescriptions can be distinguished. Animal signals are rich in designative information, and five sub-categories are distinguished: species-specific, sexual, individual, motivational and environmental information. The influence of natural selection upon the form of a signal [4] will vary according to its information content. For example, the variable nature of some signals and the stereotypy of others can be related to the conveyance of different types of motivational information. A single signal often conveys several different items of information which are usually inherent in the whole signal and not represented by different parts of the signal. The form of some signals is arbitrary, but the physical structure is often directly related to information content, is an iconic manner, or in other ways. In a survey, it is showed that non-human animals have the ability to discriminate vocal expression of emotions. Vocalizations have the potential to influence the affective states of receivers through direct (e.g. acoustic startle reflex) or indirect effects (e.g. affective learning and learned affect, which could result in state matching. The evidence described in this paper suggests that in many cases, from zebra finches to dogs, vocalizations do play a role in emotion contagion and could also have an important function in triggering appropriate responses from caretakers. It also revealed that they might even facilitate higher, cognitive empathic processes (e.g. close-proximity calls of Asian elephants (Elephas maximus) for consolation). Therefore,

vocalizations are an important channel to focus on when investigating emotional contagion and its evolution. Since the acoustic channel is the main channel of communication in humans (speech), the study of vocal contagion of emotions across species should be encouraged in order to untangle the evolution of empathic processes using playback experiments to strengthen the evidence on vocal contagion of emotions. And finally in another survey it is showed that, emotions play a crucial role in an animal's life because they facilitate responses to external or internal events of significance for the organism. There has recently been a surge of interest in animal emotions in several disciplines, ranging from neuroscience to evolutionary zoology. Because measurements of subjective emotional experiences are not possible in animals, researchers use neuro physiological, behavioral and cognitive indicators. In this article Vocal expression of arousal has been extensively studied. It showed that the increase in vocalization/element rate, F0 contour, F0 range, amplitude contour, energy distribution, frequency peak and formant contour and the decrease in inter-vocalization interval are particularly good indicators of arousal.

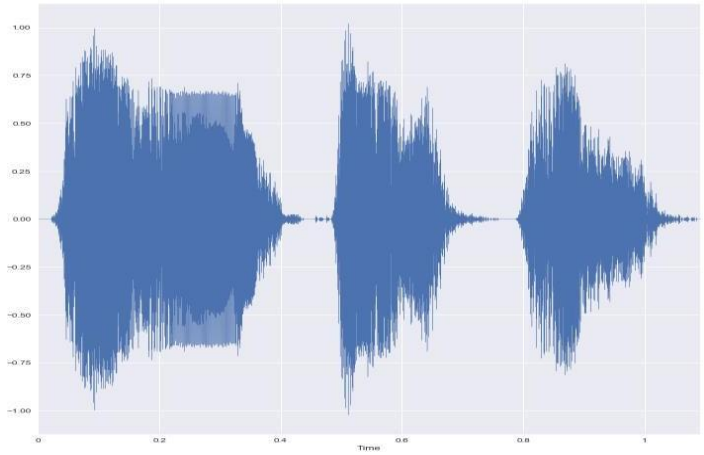
3. METHODOLOGY ADOPTED

The process of identifying the emotion of animals was done in four steps:

1. Get the audio of the sounds of animals and map to the MFCC then visualize the MFCC.
2. Collect the images for each audio file in form of visual MFCC.
3. Build the artificial neural network model and process the model with the images for classification.
4. Test the model with the images which we have not trained to the model before.

3.1 Importing Modules

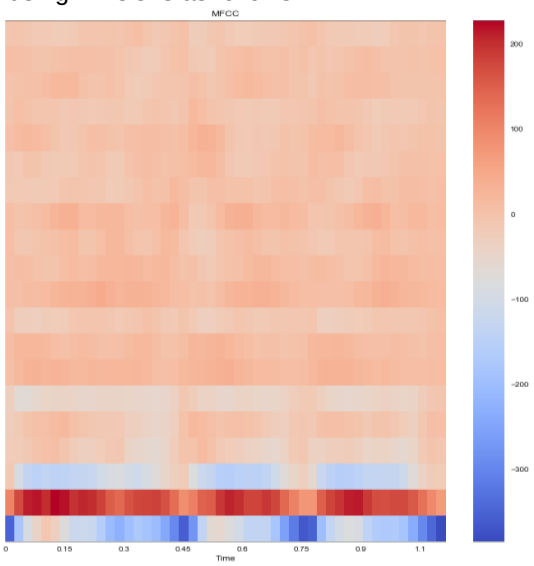
We have collected the various files from the internet. The audio files depict decent number of emotions of the dog. The emotions which we have collected includes: Angry, barking at the stranger, Crying, Howling, Hungry, Barking at the owner, Pain. All these files are referred from the different dog experts and dog lovers on the internet and we have invested our talks with them and collected the emotions accordingly. The emotions which we have chosen are little from the ocean of emotions spread over the dog's mind. The reason we have chosen above mention emotions is they are common, and people infer the animal opinion and there is need for the people to infer over these emotions which is tells the utmost interaction with the humans.



We have collected around 140 sounds which includes all the 7 emotions mentioned above. And we have classified them into 7 folders one for each emotion. Each folder in tum is converted into train set and test set. These training and testing sets are used in training and testing the model built in step 3. The visualization of audio is as follows:

3.2 Convert Audio into MFCC and Visualize

The collected audio files are exclusively in the .wav format. The audio file in train folder is now taken and we have found the Mel-frequency Cepstral coefficient for the audio. These coefficients are processed for the visualization. The visualization of the MFCC is done using the Cepstral representation of the audio clip, which in turn is the result of inverse Fourier transformation of the logarithm of estimated spectrum of signal. The spectrum of signal is plotted before we found the MFCC using the Librosa module from the module available in the python. The Librosa is the python package for the analysis of music and audio exclusively. The Librosa helps us to provide information from the audio files or we can say it acts as an information retrieval package for audio or music blocks. Every audio is examined and visualized, and these images are saved under the same training folder with respective emotions as the labels. The sample visualization of the audio using MFCC is as follows:



3.3 Build the ANN and Process the Model

3.3.1 Pre-Processing the Data

1. Primary step is to take all the images at once and put them in a folder.
2. A blank array is defined to store the images and respective classes which are multi-dimensional.
3. Read the images using Open CV, which converts the images into numerical arrays.
4. A NumPy file is been saved for every image by combining the [image_array, class].
5. Now, NumPy file is used to read all the images which exists along with their respective class.

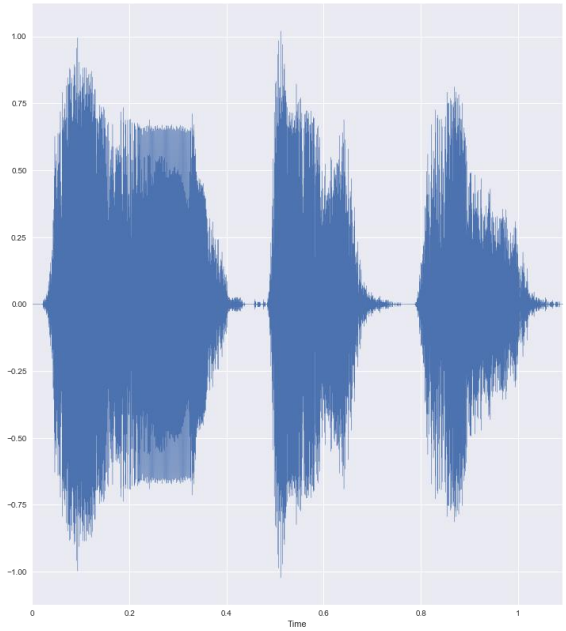
6. The images are in the order of audio files. This may impact the efficiency of the model, hence the NumPy file is shuffled to make the confused over an order, but not original data.
7. This NumPy file which is shuffled is saved as another file.

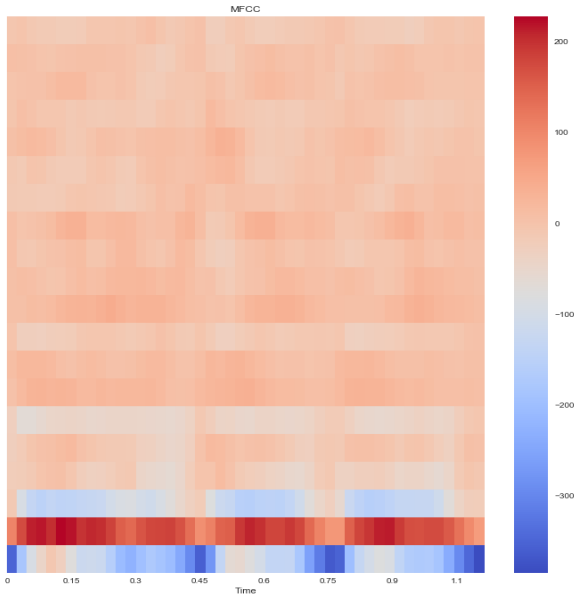
The same process is conducted for the train dataset and test dataset too.

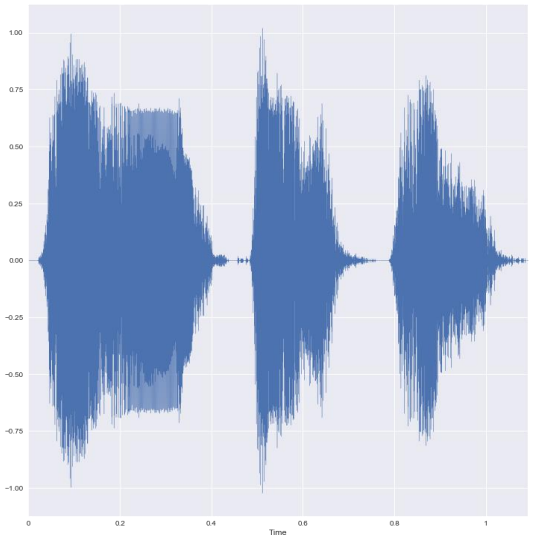
3.3.2 Building the Neural Network

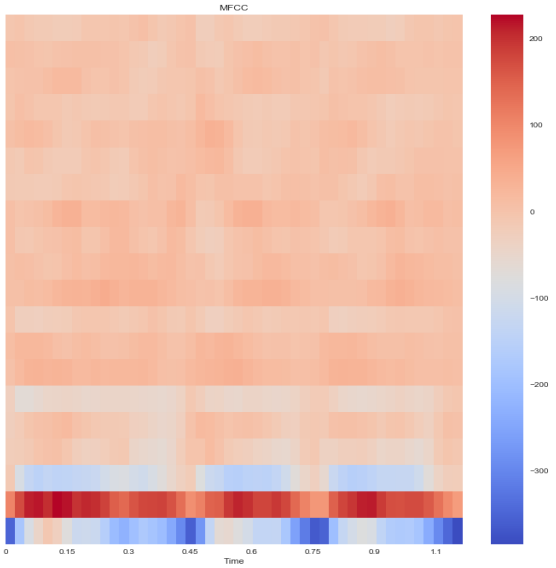
1. The Numpy file which is created above is imported and size adjustment is done from multidimensional array to single dimensional array.
2. A sequential network of 5 layers is built with 3 layers as the hidden layers.
3. First layer is known as the flatten layer, this layer used to resize the data.
4. Second layer, which is first hidden layer, is with neural network nodes and rectified linear activation function.
5. Third layer, which is second hidden layer, is built with 512 neural network nodes and relu activation function.
6. Fourth layer (third hidden layer) with 128 nodes neural network and relu as activation function.
7. The last layer is the layer with the same number of nodes as the classes that are available and activation function is soft max (normalized exponential function to normalizes k real numbers into a probability distribution consisting of K probabilities).
8. The model needs to be compiled using Adam Optimizer.
9. The loss in the network can be found with sparse categorical cross entropy.
10. The constructed model is trained with 300 epochs as a result it produces 100% accuracy.

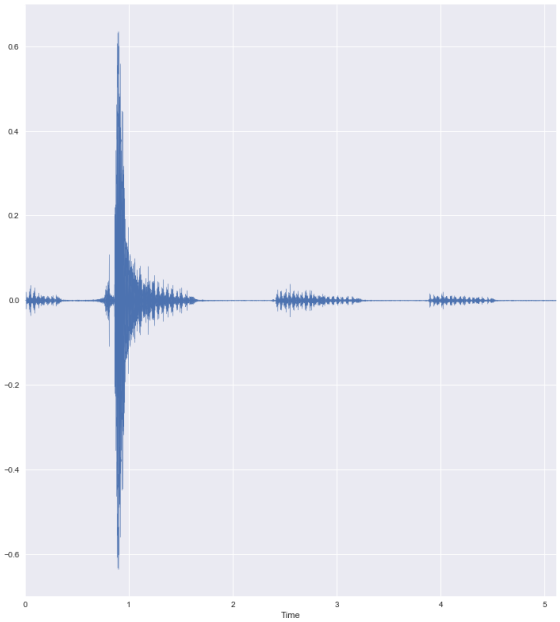
4. WAVE TO IMAGE MAPPING

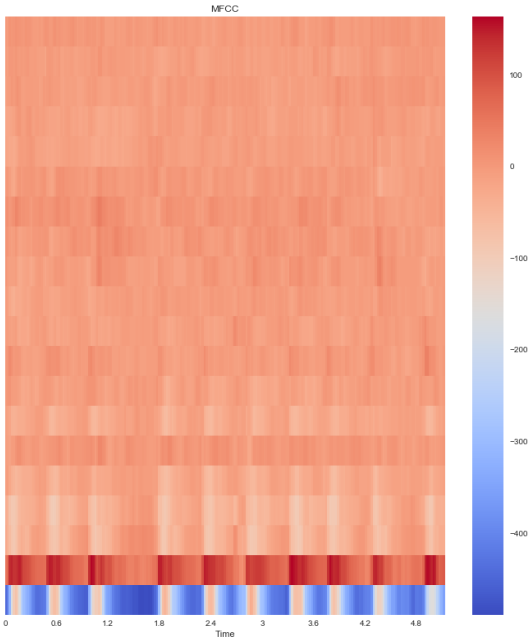


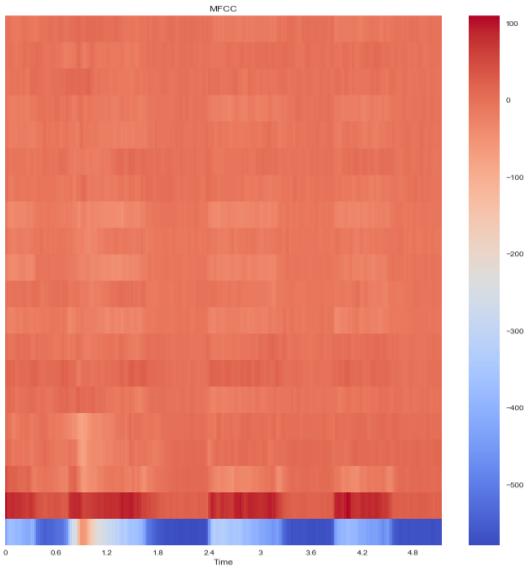


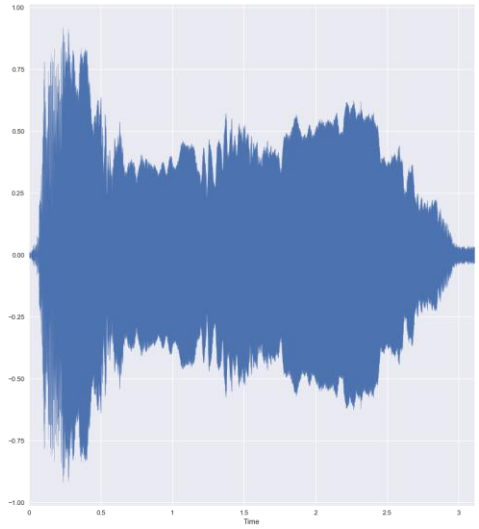


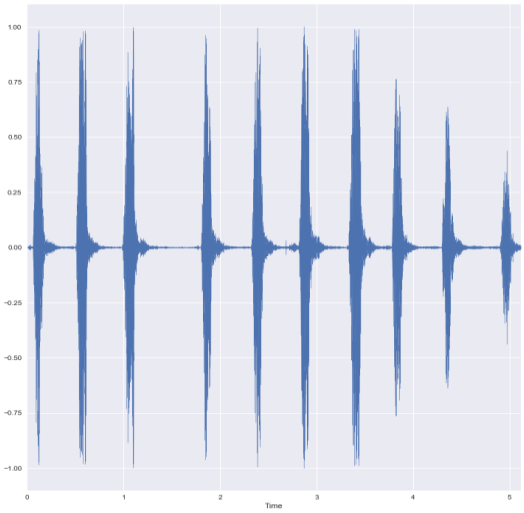


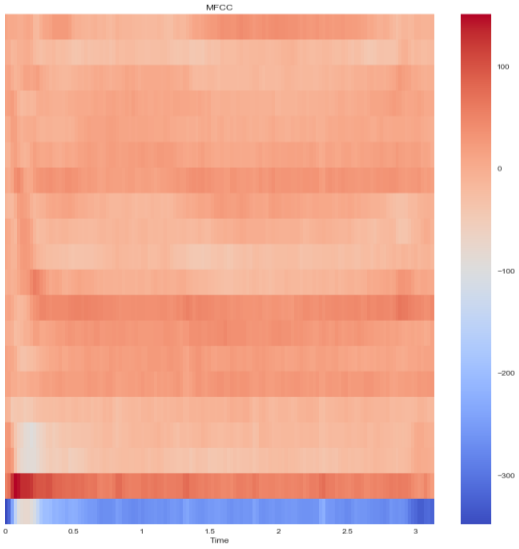

Time

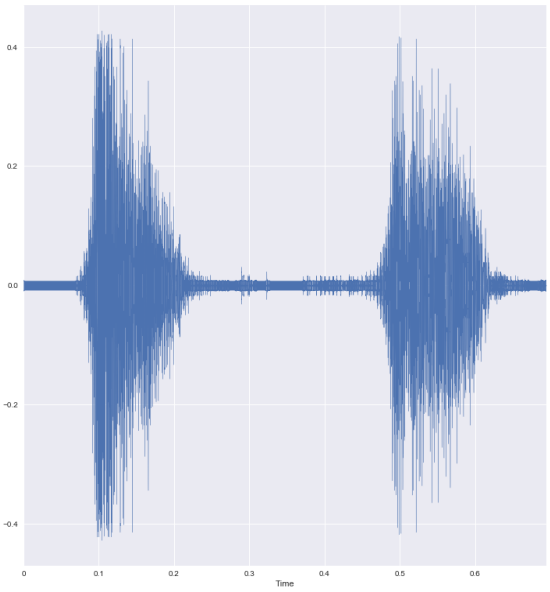

MFCC
Time

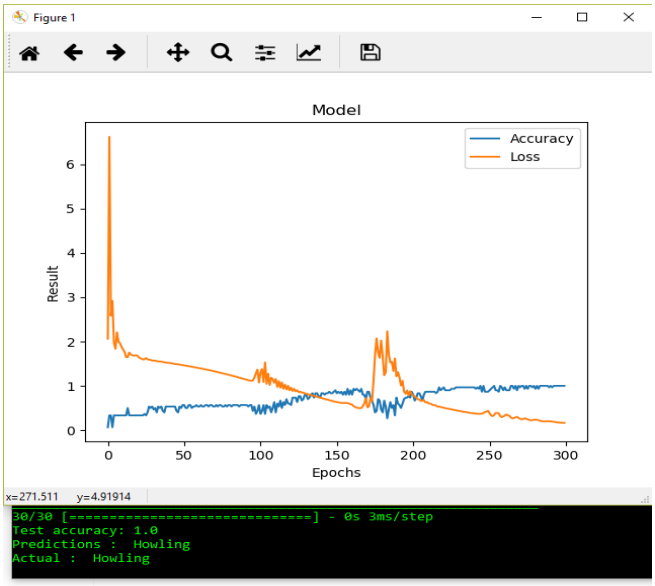

MFCC
Time

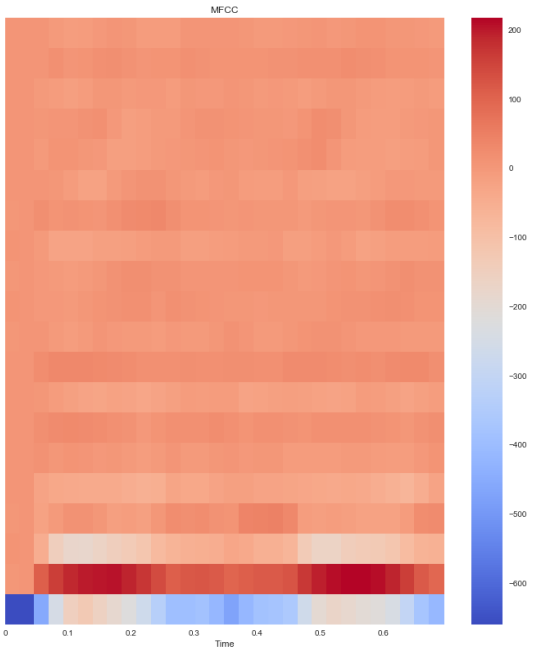

Time


Time


MFCC
Time







5. RESULTS AND DISCUSSION

The model is now ready to be tested which states the respective label of the image of classified with the existing label in the training set else the image is categorized as unclassified. After the test phase the predicted is checked and accuracy is calculated. The accuracy of the model is found to be 88%, which is appreciated. We have checked our model with 250epoch and prediction is being made. The results from the Sparse categorical cross entropy is as follows:

Total params: 1449895
Trainable params: 1449895
Non-trainable params: 0

6. CONCLUSION

The results are evaluated with the train data from the train dataset against the test dataset which earlier segregated in a separate folder. Those images are checked with the labelled images. The accuracy of the model is 88%. Our future work is going to emphasize on the application or device building using this algorithm, so that we can make use of this method in practical life.

REFERENCES

[1] Li Deng: Xiao Li, Machine Learning Paradigms for Speech Recognition: An Overview Publisher: IEEE, https://doi.org/10.1109/TASL.2013.2244083

[2] Robert M. Seyfarth, Dorothy L. Cheney, Signalers and Receivers in Animal Communication, https://doi.org/10.1146/annurev.psych.54.101601.145121

[3] [3] Fawaz S. Al-Anzi, DiaAbuZeina, The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition, waset.org/Publication/10008047

[4] [4] Irwan Bello, Barret Zoph, Vijay Vasudevan, Quoc V. Le, Neural optimizer search with reinforcement learning, https://dl.acm.org/citation.cfm?id=3305429

[5] [5] Hao Li, Zhien Zhang, Zhijian Liu, Application of Artificial Neural Networks for Catalysis: A Review

[6] Yuanzhi Li, Yang Yuan, Convergence Analysis of Two-layer Neural Networks with ReLU Activation, http://papers.nips.cc/paper/6662-convergence-analysis-of-two-layer-neuralnetworks-with-relu-activation

[7] Elodie F. Briefer, Vocal contagion of emotions in non-human animals,https://royalsocietypublishing.org/doi/10.1098/rspb.2017.2783

[8] Elodie F. Briefer, Vocal expression of emotions in mammals: mechanisms of production and evidence, https://zslpublications.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7998.2012.00920