Individual Audio Detection of Mexican Fishing Bats Using a Triplet Network

Joshua R. Lo, Abhishek Mishra, Jessica Qin, Hongyu Tu, and Kavya Dagli

Abstract— **The objective of this paper is to create a model which will effectively create a model to detect individual bats within a specific species and give insight into the metrics that differentiate the individuals of that species from one another. Past research indicates that while advancements have been made in methods of audio detection and metric extraction, there are few studies on animal audio detection specifically that takes into account individuals of that species. Our paper aims to fill this gap in audio detection research by combining elements of audio detection and image recognition. By passing pre-processed spectrograms into our model we are able to use a triplet network specific to images in combination with a residual network for increased accuracy. Using this model we hope to work towards answering if it is possible to distinguish individual bats of the Mexican Fishing Bat species. Our findings indicate that a triplet loss network is both an efficient and feasible method of calculating calculating the loss between input data. The functionality of the data gives us the opportunity to find exactly what features were used to distinguish one individuals calls from the others. It is also apparent that the residual network required data with a dimensionality that did not correspond to the inputted spectrograms, thus we remedied this passing the data through fewer layers. Based on our findings, research has the potential to aid the revival of endangered species such as the Mexican Fishing Bats by providing a greater understanding of the behavior and speech patterns of the individuals that make up a species.**

I. INTRODUCTION

Mexican Fishing bats (Myotis vivesi) is an endangered bat species that is found on islands in the Gulf of California. As Mexican Fishing bats are nocturnal, they fly across the surface of the sea at night to catch fish. Mexican Fishing bats use echolocation sequences to communicate. It is known that these calls are varied among the activity that the bat is performing (i.e. catching prey), but little is known about whether there are distinguishable features tied to an individual's call. Previous research on the calls fishing bats of different species generally reveal that those species of fishing bats generally modify their pulse intervals and lower their frequency when attacking fish.

Similarly to the fact that humans have distinguishable features of their voice, despite the various situations when they may modify their own voice, we were curious as to whether it was possible to detect the difference between the bat calls of two distinct individuals. Since Mexican Fishing bats are endangered, their behavior is especially important knowledge to having the potential to aid the species in their continuation in the California gulf region.

In this research, we explore whether it is possible to map the features of a bat call into embeddings that can be later used to determine the probabilities of a bat call being from various individuals based on measurable metrics such as distance between embedding mappings or characteristics of sound such as frequency and amplitude. We also take into account that the bat calls may only be distinguishable at various points at the call such as the frequency/ amplitude at the start, highest point, or end of a call. In breaking down the aspects of sound and time we were able to gather from audio recordings of bat calls, we isolated the bat call in order to perform necessary analysis on the resulting data.

In order to guide our methodology behind answering our researching question, we saw it fit to turn to metric learning problems, which have been applied in state-of-the-art voice recognition research projects. Using Deep Ranking: Triplet Matchnet for Metro Learning we were able to determine a possible method of identifying calls would be to utilize a triplet network (similar to the matchnet described in the paper) along with a residual network. While the problem being solved in this paper concerned metric learning for music and the purpose of music recommendations and classifications, it was useful in identifying technologies that could be modified to aid us in answering our research question. Also looking at Deep Speaker: an End-to-End Neural Speaker Embedding System and VoxCeleb2: Deep Speaker Recognition, which are both works on speaker recognition, we not only reaffirmed to use of triplet loss and ResNet as main structure of model, but also discovered that a 2-stage training strategy, pre-training with softmax and then triplet loss on hard negatives, could potentially increase model performance.

II. METHODS

A. Data

The data used in this project are audio files of bats collected by Edward Hurme, a PhD candidate in the University of Maryland. The data given to us include bat calls from 30 individual bats recorded from 2015 to 2018. The audio files are in .wav format and are unprocessed raw data recorded with microphones attached to the back of the bats. All files are categorized into folders labeled with the year and the bat number. The lengths of the audio files range approximately from 0.5 to 3 seconds, and each of them contains numerous chunks of periodic bat calls no more than 20 milliseconds long.

B. Pre-processing

The raw data require a lot of preprocessing work since they have to be transformed into spectrograms, which are easier to input into the machine learning model. The data also contains

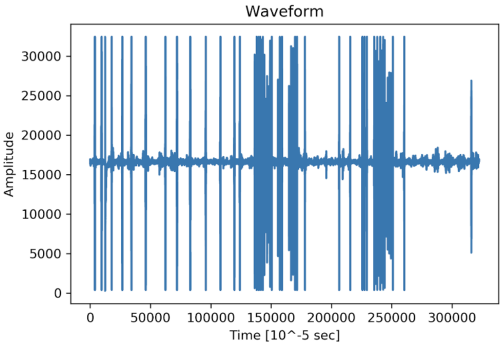

Fig. 1. Example of bat call audio data extracted from .wav file

background noise, incomplete bat calls, false detections, and suppressed waveforms due to microphone hardware limitations. First, we performed Fast Fourier Transform (FFT) to convert raw audio into spectrogram, as spectrogram can take into account audio recorded at different sampling rates.

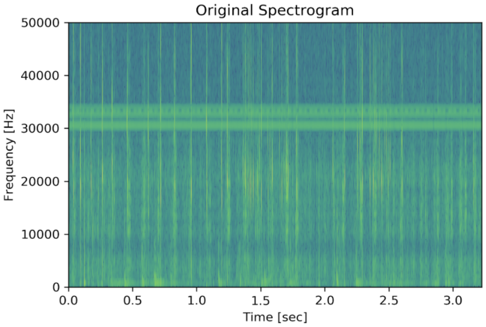

Fig. 2. A sample spectrogram before FFT was applied to extract background noise.

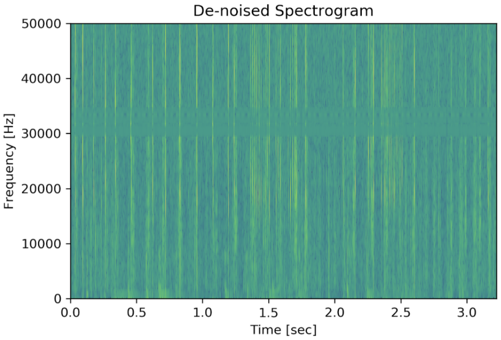

Fig. 3. A sample spectrogram after denoising measures have been applied.

Also, spectrograms are stored as 2D NumPy arrays, so we can later input the spectrograms into networks similar to the way images can be dealt with. Then we eliminate the background noise from the spectrograms. De-noising is performed by removing the mean amplitude in each frequency band. After de-noising the spectrograms, we detected individual bat calls within each file by comparing the mean frequency strength at each timestamp and comparing it to the overall mean frequency strength. If the frequency strength at the current timestamp was higher than average, we marked the timestamp as a bat call. Once the timestamps are acquired, fix-sized 15 millisecond slices of the spectrogram corresponding to individual bat calls were cropped, and they are stored as input data for our model. This threshold for cropping was determined by discussing the typical length of bat calls with professionals in the field who had analyzed the collected the data and assured us that calls rarely exceeded 15 milliseconds. An alternative method for clipping bat calls from the spectrogram would be locating the starting and ending timestamps of each bat call, and crop the enclosed slice of spectrogram. Then the cropped clips would be padded to the same size.

C. Model

Since we are matching bat calls according to their similarity, we are going to compare embeddings of bat call spectrograms. For the model, we are using triplet loss with a ResNet as the convolutional neural network backbone. Triplet loss is defined over triplets of embeddings of an anchor, a positive (object of the same class as the anchor), and a negative (object of a different class). The goal of triplet loss is to minimize the dissimilarity between the anchor and the positive, while maximizing the dissimilarity between the anchor and the negative. By training the triplet network, we will get good embeddings of the spectrograms, as spectrograms of the same class would have very similar embeddings. Selecting training examples for the triplets is very important; if we train with hard triplets, which are triplets where the negative is closer to the anchor than the positive is, training would be more efficient, and the model would be more accurate when predicting harder cases.

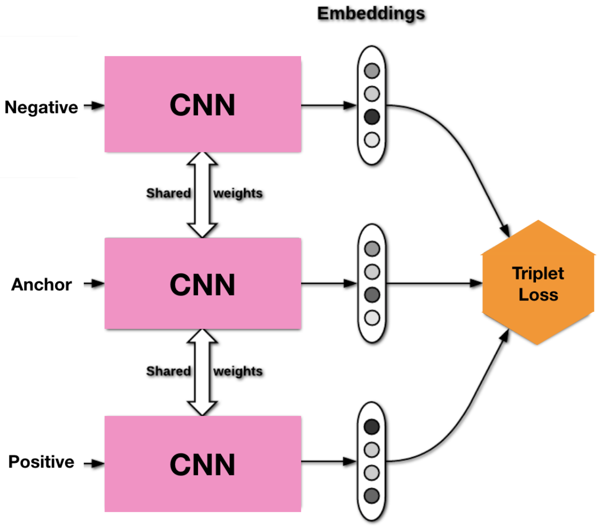

Fig. 4. A depiction of the proposed Convolutional Neural Network in combination with a Triplet Loss function.

D. Backbone Architecturee

The model has a deep convolutional neural network (CNN) as the backbone, which is used to extract features from the input spectrograms. Extracted features of an anchor, a positive, and a negative are then grouped into a triplet that is used for calculating triplet loss. We plan to eventually use ResNet as the CNN backbone, since it has various advantages over many other CNNs, such as faster speed, more layers, fewer parameters, and increased accuracy by solving issues such as vanishing gradients. The final fully connected layer of a regular ResNet has to be left out since we would want the features of the spectrogram, not classifications. The model is designed such that any CNNs can be used as the backbone, so for actual training and demonstration, several different CNN structures can be tested.

E. Implementation

We have tried different methods for implementation, referring to open-source codes of different libraries. Currently we are working on a Keras implementation of the model, with certain parts, such as the triplet loss, done with Tensorflow. The Keras.applications library comes with many models commonly used for image classification, such as ResNet50, VGG16, MobileNet, etc. For efficiency reason, we chose to build our scratch model with MobileNet due to its significantly faster and easier training process. Since our model is designed to work with different kinds of CNN, we can later experiment with different models. For any CNN chosen, we would set argument $include_top = False$, since we wouldnt need the fully connected layer for feature extraction. For the scratch model, we manually selected bat signal data from two bats, preprocessed them, and loaded them into input format for the CNN (3 channel x W x H) by stacking copies of the original spectrograms, which were in (1 channel x W x H) format. When implementing the triplet loss function, margin was set to 0.2, and it was the only loss method used in the scratch model. We didnt implement the 2-stage training strategy, which pre-trains with softmax and cross-entropy loss then fine-tunes with triplet loss. However, we do plan to implement such strategy in future works. We trained the scratch model for 30 epochs on 350 triplet sets created from 2 bats, with batch size of 32. Implementation code can be referred to on our GitHub code repository (github.com/umd-fire-coml/Bat-Calls)

III. Results

In summary, our results concluded that with an effective pre-processing technique that consisted of denoising and clipping, it is possible to isolate bat calls and convert them to data that can be passed into our triplet network. Once the training process is complete, our model is able to show that loss converges which indicates a successful functionality. Additionally, the test triplet achieved a fairly low loss compared to our previous implementations so we can conclude that it is performing relatively well.

This output will allow us to later calculate the probability that will determine which previously recorded bat the inputted sound data is most likely to be from. Since we received a relatively low loss, this probability is more likely to be accurate and consistent among inputs.

This effectively answers our research question: Is it possible to to determine the probabilities of a bat call being from various individuals based on measurable metrics of sound such as frequency and amplitude? Since our model functions according to our predictions, it is apparent that identifying individual bats is feasible using deep learning and sound analysis practices.

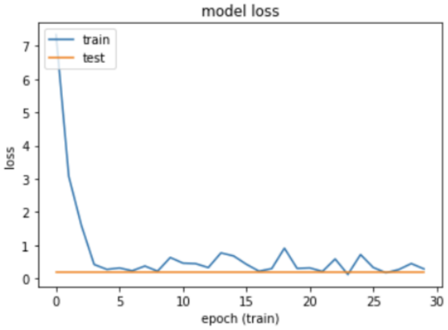

Fig. 5. Training loss graph showing conversing loss.

IV. Conclusion

This paper analyzes the differences between individual Mexican Fishing bats in hope of understanding their calls. From a biological standpoint, this contributes to a better understanding of how the Mexican Fishing bat lives and communicates with others for there have not been many studies on this endangered species. From a technological standpoint, we are attempting to understand which features the triplet network uses to analyze the differences between the calls.

Our data was gathered with various types of microphones over the 4 year collection period. This may have caused variations in the data resulting from calls as well as the level of background noise that was recording. Additionally, the earlier model of microphones collecting data recorded extraneous clipping noises periodically which may have been interpreted as actual calls by our pre-processing models. Another issue with this microphone was that it could only detect a limited range of frequency so high frequencies would back-fold. These hardware limitations have the potential to affect the accuracy of our pre-processing and therefore the accuracy of our model.

Our future work will consist of working to refine our model so it makes more accurate classifications and once that is implemented, we will create a module that analyzes what features the triplet network uses to learn and classify the bats. Additionally, we hope to focus on types of calls an individual bat makes (such as catching prey, mating, etc.) in order to determine further metrics of individual calls.

References

[1] Chung, Joon Son, et al. VoxCeleb2: Deep Speaker Recognition. Interspeech 2018, 2018, doi:10.21437/interspeech.2018-1929.

[2] Clemins, Patrick J., et al. Automatic Classification and Speaker Identification of African Elephant (Loxodonta Africana) Vocalizations. The Journal of the Acoustical Society of America, vol. 117, no. 2, 2005, pp. 956963., doi:10.1121/1.1847850.

[3] Lu, Rui, et al. Deep Ranking: Triplet MatchNet for Music Metric Learning. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, doi:10.1109/icassp.2017.7952130.

[4] R. Lu, K. Wu, Z. Duan and C. Zhang, ”Deep ranking: Triplet MatchNet for music metric learning,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 121-125.

[5] Snyder, David, et al. Deep Neural Network-Based Speaker Embeddings for End-to-End Speaker Verification. 2016 IEEE Spoken Language Technology Workshop (SLT), 2016, doi:10.1109/slt.2016.7846260.

[6] Wang, Huiqin, et al. New Audio Embedding Technique Based on Neural Network. First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC’06), doi:10.1109/icicic.2006.479.