Fourier Analysis for Frog Call Classification

Aniketh Iyengar
Stanford University
aniketh@stanford.edu

December 14, 2022

1 Introduction

Frogs are a diverse group of amphibians, with over 6,000 known species. They are found on every continent except Antarctica, and play important roles in the ecosystems in which they live. Despite their importance, many frog species, and more generally amphibian species, are threatened or endangered due to their sensitivity to human activities. There is thus a need for active conservation efforts and biodiversity assessment.

These efforts often include species tracking and identification to evaluate regional differences in fauna or determine the temporal changes in certain populations. Such studies involve large amounts manual labor, including physical field research and the processing of large amounts of visual and audio data.

Fraught with the potential of human error and the urgent need to improve efficiency of these efforts, modern day conservation must to incorporate the use of technology. Thus, one of the target applications would be to use machine learning to process the large amounts of field recordings taken by biologists and classify any recognizable sounds. However, this can have its difficulties with the large amounts of background noise in the wild. Nevertheless, in this project I attempt to use linear algebra techniques to improve the effectiveness of machine learning in classifying frog calls from raw field research audio data.

2 Data Set and Sampling

AmphibiaWeb is an online database created and curated by UC Berkeley, with a mission to connect ”people around the world by synthesizing and sharing information about amphibians to enable research, education, and conservation.” The data base contains information about 7000+ species, including 832 sound recordings. The audio files come from several contributors, often taken in natural/noisy conditions, of variable length, and saved in different file formats with inconsistent sample and bit rates.

To account for imbalances and variety in the data available, I curate a sample of 8 frog species, each with an audio recording containing instances of relatively clear calls. From these recordings, I extract 15, arbitrary 1 second segments from different parts of the call patterns per frog. These recordings are then saved in 22,050 Hz 32-bit mono WAV formats, resulting in approximately 22,050 samples per recording.

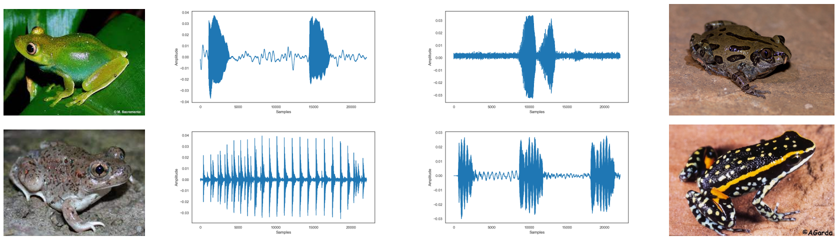

Figure 1: Recordings from curated sample set with image of frog: Aplastodiscus Leucopygius, Kassina Senegalensis, Spea Multiplicata, Ameerega Flavopicta (Top to Bottom, Left to Right)

3 Model

The model for this task was split into two components. Using the Short-Time Fourier Transform to convert from the time-domain into frequency-domain where we can analyze the Power Spectrum values of the recording, which provide information about the most prominent frequencies in the given signal over time. Then a machine learning classifier, in this case K-Nearest Neighbors, can be trained and tested on these new features.

3.1 Fourier Analysis

3.1.1 Theory

The Fourier Transform is one of the most important function decompositions. It is established by the Fourier Basis that the functions $\frac{1}{\sqrt{2}}$, $cos(\frac{2\pi kx}{L})$, and $sin(\frac{2\pi kx}{L})$ are a basis for any smooth (differentiable) function $f(x)$ on $[0, L]$. Euler's formula also establishes that $e^{ix} = \cos x + i \sin x$. Thus, the Fourier Transform uses these principles to decompose a signal/function into it's component frequencies. It is defined as:

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \omega} dx.$$

The output is a complex coefficient, whose magnitude represents the strength of that frequency in the original signal and whose angle represents the phase of the frequency. Often times $|\hat{f}(\omega)|^2$ is used instead to indicate the relative importance of the frequency, known as the Power Spectrum Density. This can be quite useful in denoising or compressing a a signal by reconstructing it with only the frequencies with a PSD above a certain threshold.

Now, as the Fourier Transform is defined continuously, most signals are discrete, and sampled over time. We can say that that given N samples, we can decompose it into N discrete component frequencies. Thus, the Discrete Fourier Transform calculates these complex frequency coefficients when defined by:

$$F(k) = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi i n k}{N}} \text{ for } k = 0, 1, ..., N-1.$$

However, notice that for every one of the N discrete frequencies, a computation of $O(N)$ must occur, and thus result in a total $O(N^2)$ computation for the the DFT. In solution to this, in 1995, James Cooley and John Tukey established the Fast Fourier Transform algorithm for calculating the Discrete Fourier Transform in $O(NlogN)$ complexity, exploiting redundancies in the sinusoidal functions. To visualize this, define the following Vandermonde matrix to represent the DFT where $w = e^{\frac{-2\pi i}{N}}$:

$$F_{nxn} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w & w^2 & \dots & w^{N-1} \\ 1 & w^2 & w^3 & \dots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \dots & w^{(N-1)(N-1)} \end{bmatrix}.$$

Given any vector f with N samples such that N is a power of two, the Fast Fourier Transform allows us to split f into its even indexed and odd indexed values such that:

$$F_{n \times n} f = \begin{bmatrix} I_{\frac{n}{2}} & D_{\frac{n}{2}} \\ I_{\frac{n}{2}} & -D_{\frac{n}{2}} \end{bmatrix} \begin{bmatrix} F_{\frac{n}{2} \times \frac{n}{2}} & 0 \\ 0 & F_{\frac{n}{2} \times \frac{n}{2}} \end{bmatrix} \begin{bmatrix} f_{even} \\ f_{odd} \end{bmatrix} \text{ where } D = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & w & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w^{\frac{n}{2}-1} \end{bmatrix}.$$

We can see that this decomposition is recursive in nature and that the first block matrix contains diagonal matrices. Through this, we arrive at an overall $O(NlogN)$ complexity.

3.1.2 Application to Frog Call Classification

To process the recordings, I can apply the Short-Time Fourier Transform to it. This technique involves applying the Fast Fourier Transform to small localized sections of the overall sound sample using small, overlapping windows. For a window size of s samples, we get $1 + \frac{s}{2}$ complex coefficients due to a symmetry in the coefficients outputed by the DFT around the Nyquist frequency ($k = \frac{N}{2}$). Then, for a size of n for our non-overlapping segments and T for total number of samples in our recording, we can calculate that we will have about $\frac{T}{n} - 1$ time bin values. Thus, our output matrix will be will be a $\mathbb{C}^{(1+\frac{s}{2})\times(\frac{T}{n}-1)}$ feature matrix. We can then convert this matrix into the Power Spectrum of the frequencies over time in terms of a logarithmic decibel scale, becoming $\mathbb{R}^{(1+\frac{s}{2})\times(\frac{T}{n}-1)}$, and then viewed as a spectrogram.

3.2 K-Nearest Neighbors

Once these feature matrices / dB-scaled spectrograms are calculated per sample recording, I built a K-Nearest Neighbors classifier to identify the frog based on the decibel values for the component frequencies over time. I flatten each sample's spectrogram matrix in column major order to create feauture vectors for training/testing and used a default $k = 5$.

3.3 Results

3.3.1 Feature Extraction

As each recording sample I curated from the data base was 1 second long and sampled at 22,050 Hz, each recording contained 22,050 samples, thus in $\mathbb{R}^{22050}$. For my Short-Time Fourier Transform, I used a window length of 512 samples, or 0.23 milliseconds per my sample rate of 22,050 Hz, and a hop length of 128 samples. This resulted in $\mathbb{R}^{257\times 173}$ spectrogram matrices. The following images are the four recording samples shown earlier converted into their respective dB-scaled spectrograms.

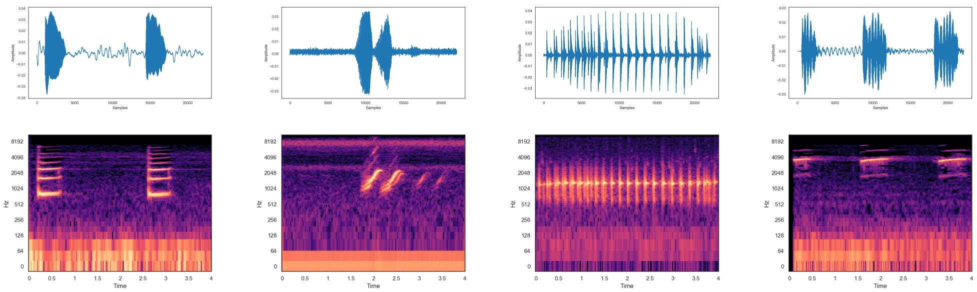

Figure 2: Pairwise (Vertical) Comparison between initial time-domain waveform and frequency domain dB-scaled spectrogram: Aplastodiscus Leucopygius, Kassina Senegalensis, Spea Multiplicata, Ameerega Flavopicta (Left to Right)

3.3.2 Classifier Evaluation

I used Leave-One-Out Cross Validation to evaluate the K-Nearest Neighbors classifier on the 120 sample items. The classifier attained a 97.5% accuracy score on the data.

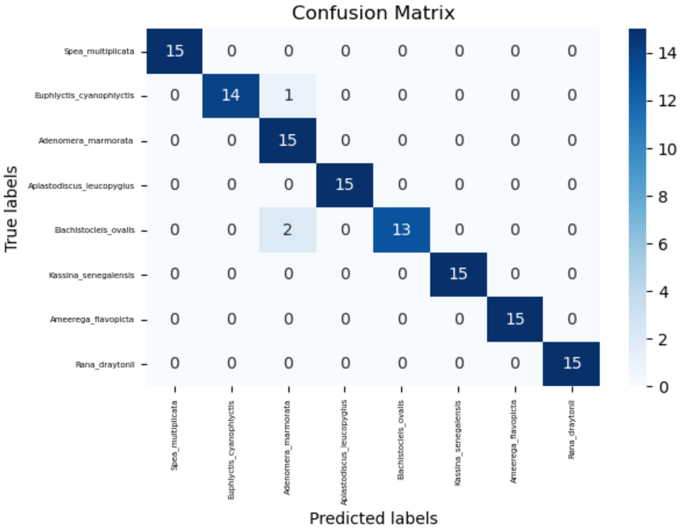

Figure 3: Confusion Matrix of K-Nearest Neighbor Classifier for Frog Calls

3.3.3 Discussion

I was pretty happy with my classifier's performance. The conversion of time-domain wave-forms to a frequency-domain dB-scaled spectrogram achieved a quite significant accuracy score given its simple nature. However, there are a few aspects to keep in mind: Most the frogs selected had distinctly different sounds and my hand curation of samples may have introduced human bias. Further studies with more robust sampling would need to occur for better validation of the classifier.

3.4 Conclusions and Future Steps

In this project I sought to create a classifier for Frogs based on their calls. I was able to create a K-Nearest Neighbors classifier with 97.5% accuracy with a sample data set of 120, 1 second call segments, containing 15 samples for each of 8 frog species.

In reading literature about audio classification, there are many more frequency feature extraction methods for machine learning such as Mel-Filter Spectrograms, Mel-Frequency Cepstrum Coefficients, Shannon Entropy, etc. In future steps, I would explore the implementation of these. Moreover, I would also search for a larger breadth of training data and look into the potential use of algorithmic syllable segmenters to decrease human bias in choosing samples.

Regarding the scalability of my model, large field recording could be segmented into semi overlapping fragments, where each fragment would be run in an ML classifier like mine. Each frame could then have a vote on the identity/identities of the frogs in the recording. Additionally, with audio classification being a large predictive task, the application of Deep Learning would also prove to be quite useful.

References

[1] AMPHIBIAWEB: Information on amphibian biology and conservation. AmphibiaWeb, Berkeley, California, http://amphibiaweb.org/

[2] FFT - math.uconn.edu. (n.d.). Retrieved December 13, 2022, from https://www2.math.uconn.edu/ olshevsky/classes/2018$spring/math$3511$/FFT.pdf$

[3] Han, N. C., Muniandy, S. V., amp; Dayou, J. (2011). Acoustic classification of Australian anurans based on hybrid spectral-entropy approach. Applied Acoustics, 72(9), 639–645. https://doi.org/10.1016/j.apacoust.2011.02.002

[4] Huang, C.-J., Yang, Y.-J., Yang, D.-X., amp; Chen, Y.-J. (2009). Frog classification using Machine Learning Techniques. Expert Systems with Applications, 36(2), 3737–3743. https://doi.org/10.1016/j.eswa.2008.02.059