E. Machine Learning Audio Analysis - Traditional Algorithms

With little to no experience in the field of machine learning (ML), the first task was to attain a fundamental level of knowledge in using machine learning algorithms. We decided to start by generating our own simple machine-learning code. However, one of the most important parts of machine learning is determining which algorithm you are going to use. For machine learning on sound, it is especially important to determine which features are intended to be extracted from the audio files to use for identification characteristics in the audio model. So, several different features were tested as identification systems before moving on to more advanced code including Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff, Mel-Scale Spectrogram, Chroma Frequencies, and Mel-Frequency Cepstral Coefficients.

Spectral Centroid identifies the perceived brightness or center of mass (COM) within a spectrum, often utilized in music genre classification and sound analysis tasks due to its straightforward computation, offering a relatively fast processing speed (Giannakopoulos & Pikrakis, 2014). Spectral Bandwidth measures the frequency range width within a spectrum, aiding in distinguishing timbral characteristics in music analysis and sound classification, with a moderately fast processing speed suitable for real-time applications. Spectral Contrast captures spectral differences among adjacent frequency bands, enhancing audio feature discrimination in tasks like music genre classification, and processing at a moderate speed due to its detailed spectral analysis (Sable, 2021). Spectral Rolloff determines the concentration of spectral energy below a specific frequency, useful for distinguishing between harmonic and noisy components in audio signals, processing at a moderately fast speed suited for real-time applications (Tjoa, n.d.). Chroma features measure the presence of musical notes within an audio signal, crucial for tasks like chord recognition and melody extraction, with moderate processing speed suitable for music analysis applications.

Lastly, Mel-Frequency Cepstral Coefficients (MFCCs), known for capturing spectral characteristics akin to the human auditory system, are widely employed in speech and audio processing tasks such as speech recognition and speaker identification (Deruty, 2022). Although slightly more computationally complex, MFCCs offer efficient processing, making them suitable for real-time applications and preferred due to their superior performance in various audio analysis tasks. These algorithms were incorporated in machine learning code to determine which

were most effective at identifying the sounds of semi-aquatic sounds, specifically anurans, when trained with an online database of frog and toad sounds.

• • •

IV. Results & Discussion

A. Accuracy of Machine Learning Algorithms

Before performing any testing, it was assumed that the Mel-Frequency Cepstral Coefficients method of identification would likely be the most accurate. Since the algorithm is designed to process speech patterns, it would perform the most meticulous identification of the varying audio patterns found in wildlife vocalizations. Spectral Contrast should also perform fairly well in classification, as it may differentiate the tempo, pitch, and melody of the different anuran species. Due to their simplicity, Spectral Bandwidth, Centroid, and Rolloff were not expected to perform well, though Spectral Rolloff was expected to perform better than Spectral Centroid or Spectral Bandwidth since it is less susceptible to noise.

Algorithm	**Parameter Values**	**Percent Accuracy (n=5)**
MFCC	n/a	97.14%
Spectral Centroid	n/a	14.01%
Spectral Bandwidth	n/a	12.93%
Spectral Contrast	n/a	88.58%
Chroma Frequencies	n/a	72.20%
Spectral Rolloff	95% cutoff	13.36%
	90% cutoff	14.01%
	85% cutoff	15.52%
	75% cutoff	14.44%
	50% cutoff	12.93%
	35% cutoff	11.42%

Table 2 *Results of Each ML Sound Identification Algorithm for Anuran Sounds*

To test this hypothesis, machine learning code was created to test the algorithm of each method, located in the attached "Frog ML Code" Google Drive folder as "frogConfusionML.py". The audio data used to train and test the system was a Kaggle data set uploaded by Turker Tuncer (in F23 Audio Folder in Frog Simple ML Code), consisting of 50 audio files for each species (Tuncer, 2023). The accuracy results for each method, and parameter values for each algorithm, are listed in Table 2.

As predicted, MFCCs were the most accurate method by far. As we mentioned, this makes sense since MFCC is designed to distinguish particular speech patterns. Although originally designed for human speech, frog "speech" seems to be similar enough for the method to still work. Spectral Contrast also did well as expected, providing an accurate prediction around 9 of 10 times. Classifying frog ribbits as though they were different genres provides a fairly accurate model. Similarly, Chroma Frequencies, which measure the presence of musical notes, were fairly accurate as well, providing an accurate prediction in almost 3 out of 4 of the cases. The methods that performed poorly were Spectral Bandwidth, Spectral Centroid, and Spectral Runoff. All of these methods have very simple feature extractions, trading complexity for speed. Spectral Bandwidth and Centroid were not expected to have done well since they essentially provide a range and COM of the frog sounds, it would have been surprising if this was enough to differentiate 43 frog species. These two methods are also very susceptible to noisy systems. Repeated runs showed Spectral Centroid consistently performing slightly better than Spectral Bandwidth; nonetheless, both still lacked significant accuracy. What was surprising, however, was how Spectral Centroid performed about as well as Spectral Runoff, even better in some cases, depending on the cutoff point. Upon further thought, we believe this may be due to the fact Spectral Runoff only cuts off noise *above* a certain frequency (i.e. the *cutoff point*). This means, for instance, when the Spectral Runoff has a *cutoff percent* of 85%, it will return the highest frequency for which 85% of the noise is below it. Since frog ribbits are notoriously low frequency, this may be negatively affecting this prediction. Therefore, the effect of the ambient noise, which is likely of a higher frequency than the frog sound, is even more pronounced in this method.

As a new hypothesis and the results of the previous trials, it was predicted that using cutoff points that cutoff sounds *below* a certain frequency would more accurately capture the

notoriously low frequency of frog sounds. This new method is called Reverse Spectral Rolloff and is located in the attached “Frog ML Code" Google Drive folder as “reversingSpectralRolloff.py”. This new algorithm was run on the same data with cut-off points at 90%, 85%, 50%, and 25% which produced mean accuracy results of 5.17%, 6.47%, 5.60%, and 6.03% respectively over three trials for each parameter value. While a bit better than just guessing (2.33%), Reverse Spectral Rolloff performed the worst of any algorithm. This suggests, contrary to what we originally believed, that frog sounds are more concentrated/differentiated in the higher relative frequencies. Upon further research, these results make sense: against common wisdom, frog croaks are often high frequency. It is only a few notable exceptions, in particular the bullfrog, that have given them this reputation. Smaller frogs, which make up the majority of frog species, have a higher frequency (Narins, 1995). Moreover, tree frogs, due to strong ambient noise in the rainforests, often have even higher frequencies to differentiate their calls (Gridi-Papp, 2014). In fact, frogs have the highest auditory processing capabilities of any amphibian (University of California, Los Angeles, 2009).

B. Analysis of Data & Algorithm Results

After performing this simple machine learning analysis, we decided to take a deep look into the data we are using. The first analysis follows the discussion on the high frequency of frog sounds after the surprising failure of the Reverse Spectral Rolloff feature extraction method (Figure 12).

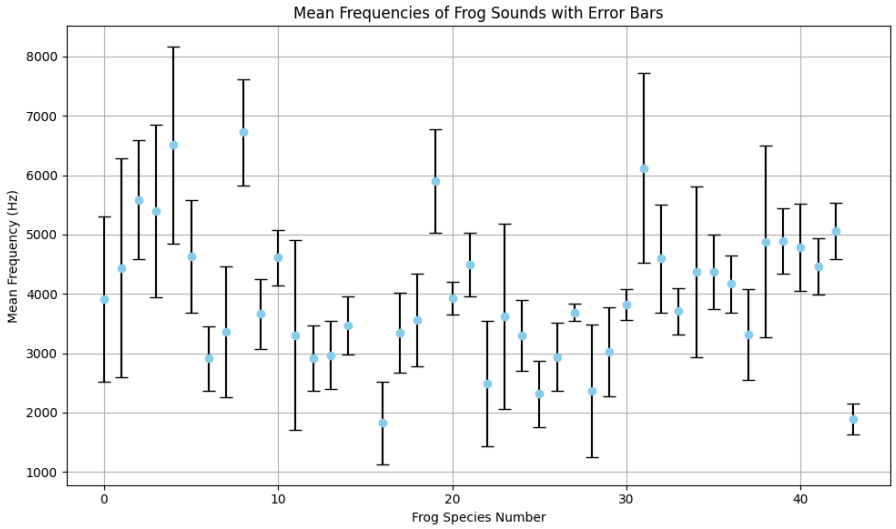

Figure 12. *Graph of mean frequencies for each recording in the used dataset, grouped by frog species number (top), and the average frequency across all sounds particular to a frog species, plotted with error bars at ±1 standard deviation (bottom).*

Interestingly, there is a large variance among the sound frequencies of each frog species. Frogs, 20, 27, 30, and 43 are the only notable exceptions, with 27 having an especially low variance. Additionally, all frequencies averaged over 1500 Hz for each sound recording with some as high as 9 kHz. For reference, the frequency of human speech ranges from 75 to 300 Hz (Lofft 2020). The record for the highest frequency sung by a male is D#8, around 5 kHz (Songstuff 2023). So, as we predicted in our conclusions from the Reverse Spectral Rolloff method, these frog sounds are high frequency. While we do not have the exact names of the frog species, we can conclude that these frog species are of a smaller variety, likely different types of tree frogs.

Confusion matrices are like a scorecard that helps us see how well our algorithm method understands the sounds. Along the diagonal, it shows how many frog sounds the computer correctly guessed. Off the diagonal, it displays the wrong guesses along with which frog species it thought the sound came from. In essence, it helps us understand where the algorithms made mistakes and how accurate they were for each sound it guessed. Therefore, confusion matrices were created for the outcome of each machine learning characteristic algorithm explored in this project, as displayed in Figure 13.

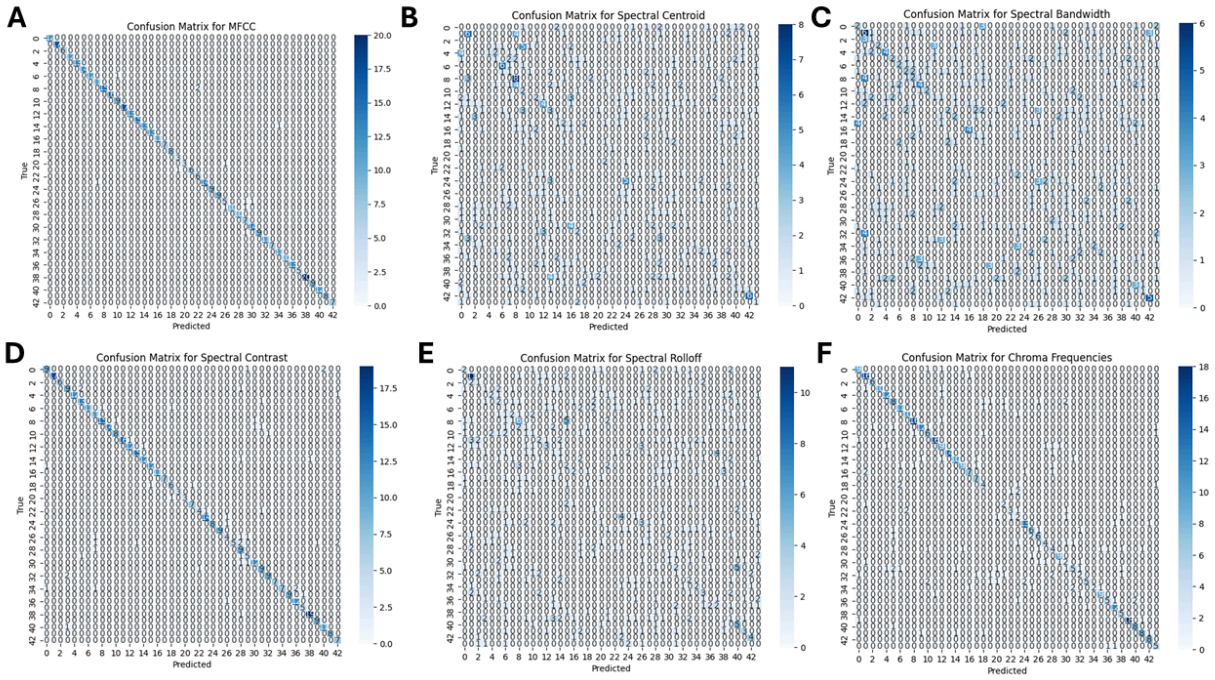

Figure 13. *Confusion matrices representing the results of A) MFCC, B) Spectral Centroid, C) Spectral Bandwidth, D) Spectral Contrast, E) Spectral Rolloff, F) Chroma Frequency algorithms.*

The accuracy for MFCC is 97.41%, so apart from a few scattered mistakes, the model was very accurate. Moreover, all mistakes were singular apart from 2 Frog 8 sounds being predicted to be Frog 22, as shown in Figure 13a. Other notes are that Frog 29 was incorrectly predicted twice, and Frog 26 was mistaken twice. Ultimately, it is a sign of high accuracy that we can see a clear diagonal line in the data. Lastly, while MFCC had the highest accuracy of any method, it was also the least efficient in processing speed. In contrast to MFCC, the Spectral Centroid Confusion Matrix does not have a clear diagonal band, as shown in Figure 13b. As a reminder, the accuracy for Spectral Centroid is 14.01%. So, the Spectral Centroid is inaccurate for this dataset. The spectral centroid measures the Center of Mass of the sound. Despite the inadequacy of this model, a few sounds, Frogs 1, 6, 8 12, 24, and 42, have several correct glances. However, when you account for the different sizes in each frog species sample set, you see that only Frog 42 was predicted accurately more than 50% of the time. This suggests Frog 42's center of mass is fairly distinct. Accuracy for Spectral Bandwidth is even worse than Spectral Centroid, averaging at 12.93%. The analysis of this data set is similar to Spectral Centroid, as shown in Figure 13c. There were a few sounds with a fair number of correct predictions, but Frog 42 seemed to stand out as the only one with over 50% accuracy. As a

reminder, the accuracy for Spectral Contrast is 88.58%. The Spectral Contrast method is fairly reliable, second in accuracy only to MFCC, as shown in Figure 13d. It is also the best of the methods with average processing efficiency (Spectral Contrast and Chroma Frequencies). Similar to MFCC and in contrast with Spectral Centroid and Bandwidth, there is a clearly defined diagonal line. However, unlike MFCC, there are a few instances of the same incorrect prediction for a particular Frog sound. To note on Frog 42, which seems to be an outlier in previous methods, it was predicted accurately 100% of the time. As a reminder, the accuracy for the Spectral Rolloffs (with a cutoff of 85%) is 15.52% which is very poor, but is the best of our time-efficient methods (Spectral Centroid, Bandwidth, and Rolloffs), as shown in Figure 13e. Interestingly, the accuracy of Frog 42 fell below 50% for the first time, suggesting while its center of mass and bandwidth may be distinct, its rolloff is not. In this case, only Frog 1 has an accuracy above 50%, predicting correctly 11 of 17 times. However, it should be noted that Frog 1 is likely overfitted since it was overpredicted and wrong 13 of 24 times. As a reminder, the accuracy for Chroma Frequencies is 72.20%. This was the third best, though the worst of the algorithms with medium-to-high processing cost (Spectral Contrast, MFCC, and Chroma Frequencies). Interestingly, there seems to be a divide in accuracy across species. For instance, some sounds were highly accurate, with Frogs 2, 17, 38, and 42 even having 100% accuracy (again, Frog 42 seems to have a unique sound), as shown in Figure 13f. Yet, others, like Frogs 19, 30, 32, and 34 had very low accuracy. Frog 19 was never accurately predicted—Frog 20 was also never correctly predicted, but in this particular random sample a sound from it was never selected for testing. Since Chroma Frequencies are used to extract musical notes from sounds, perhaps some Frog species have a more musical quality, with an emphasis on notes, than others.

Overall, we found that the more complex feature extraction methods resulted in higher accuracies. The MFCC method, which is used to model human speech, performed best. Second to MFCC, the methods model musical rhythm and notes did moderately well. The method of measuring notes was the most divisive, performing accurately for some Frog sounds, but not others. The simple methods of center of mass, range, and, and measuring lower/upper points of the range, did very poorly. Lastly, we wanted to investigate Frog 42, which seemed to have a unique sound. Upon listening to the audio files, we discovered that it was in fact a recording of bird sounds, not frogs. This is promising for applications of the MFCC software on species other than anurans since MFCC was able to classify Frog 42 just as well as the other sounds.