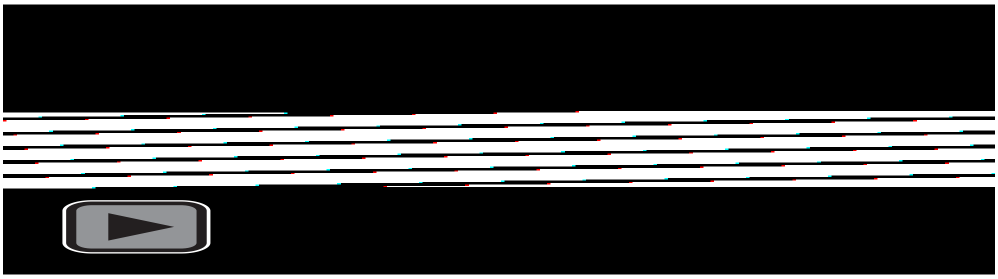
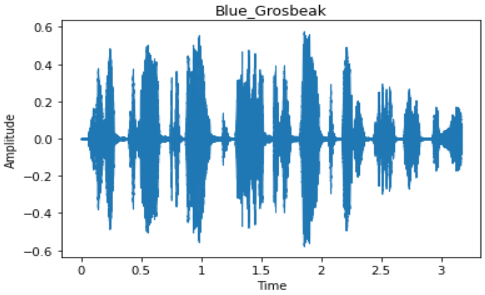
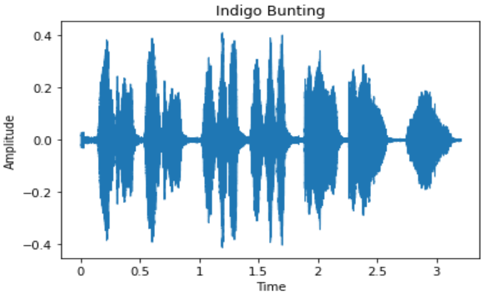
Animal sound recognizer:

Can you tell the difference?







Animal welfare 動物福祉

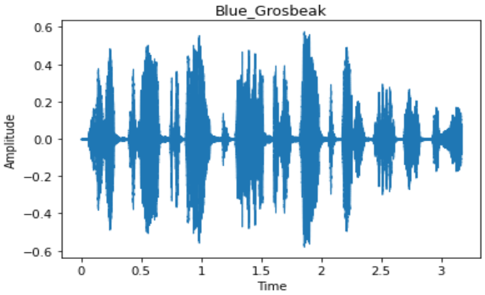
Respecting and caring for animals, especially livestock, has been an essential movement for the past few years. Providing chickens and cows with a better environment does not only have a morally meaning but also can result in better and higher protein production.

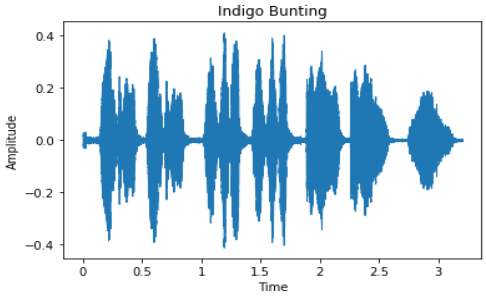
Smart farming is an excellent example of how AI technology is applied to boost the efficiency and effectiveness of the labor-intensive business, potentially contributing to the entire animal welfare.

One example is using animal vocal data to predict the livestock's health conditions. One of the challenges is that the sound from the same animal species can vary largely depending on the various conditions such as animal age, its environment, etc.

Here, I’m using two birds' sound file (source: https://www.audubon.org) to demonstrate one of the machine learning methods to recognize different sounds by using the deep learning CNN model.

Data Analysis



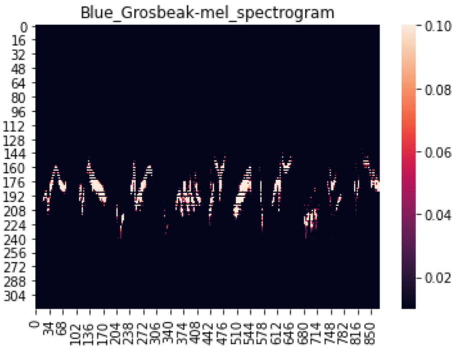


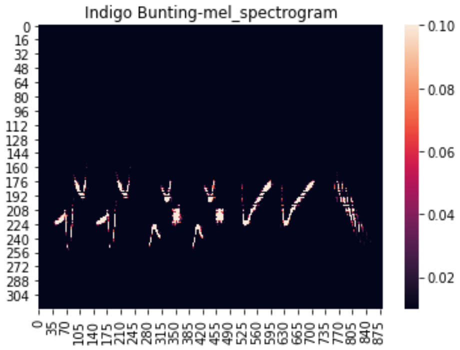
Obviously, the above 2 simple waveform charts cannot provide us enough insight for us to tell how different the 2 bird sounds are. So let dive deeper and analyze the 2 sounds from other different perspectives. I cited the below definitions from Wikipedia.

Over, we see a richer sound, a wider band from the Blue Grosbeak than Indigo Bunting. This may be a special pattern in the Blue Grosbeak that we can use to separate it from Indigo Bunting.

Mel-spectrogram

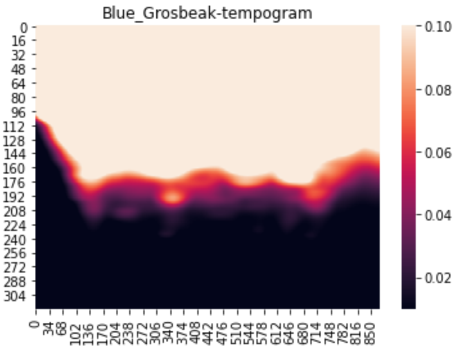
Mel spectrogram is a spectrogram converted from the frequency spectrum of signals and then converted to a Mel scale.

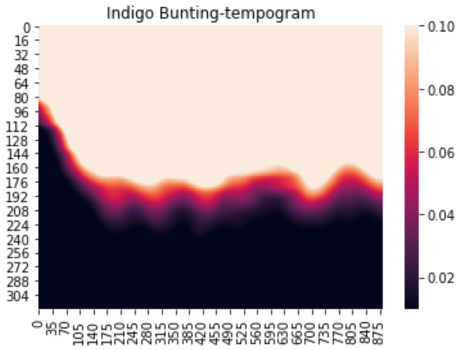




Tempogram

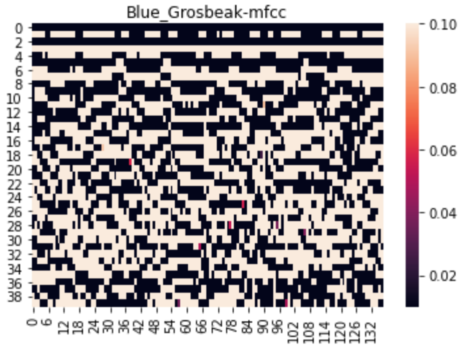
Tempogram, a mid-level representation of tempo information, is constructed to characterize tempo variation and local pulse in the audio signal

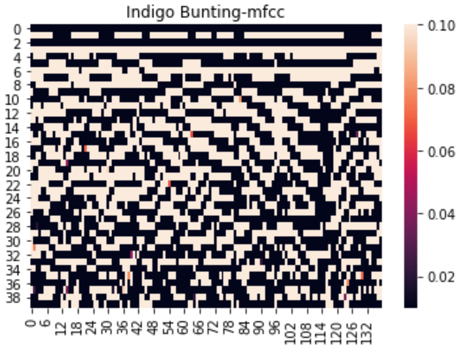




Mel Frequency Cepstral Coefficient(MFCC)

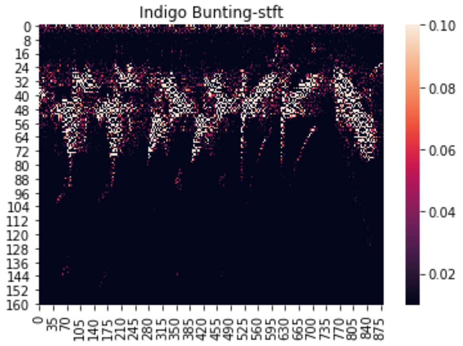
MFCC are coefficients that collectively make up an MFC which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency

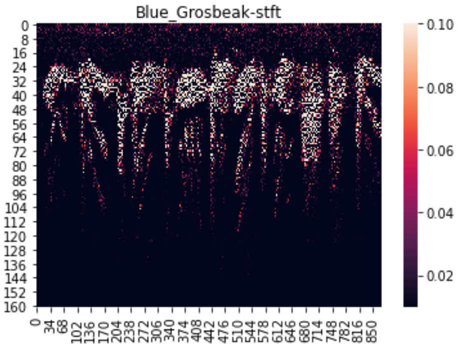

Blue_Grosbeak-mfcc


Indigo Bunting-mfcc

Short-time Fourier Transform(STFT)

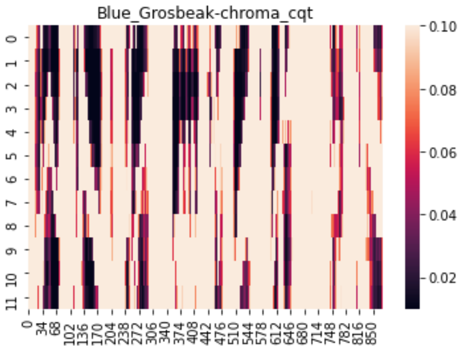
STFT is made by dividing a longer time signal into shorter segments of equal length and then computing the Fourier transform separately on each shorter segment.

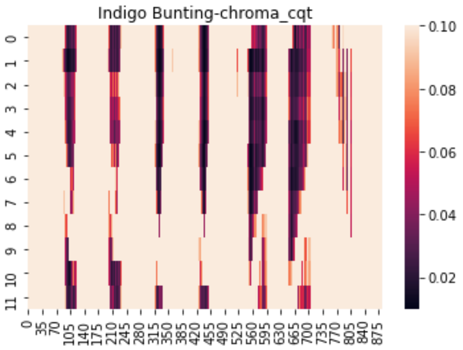

Indigo Bunting-stft


Blue_Grosbeak-stft

Chroma CQT

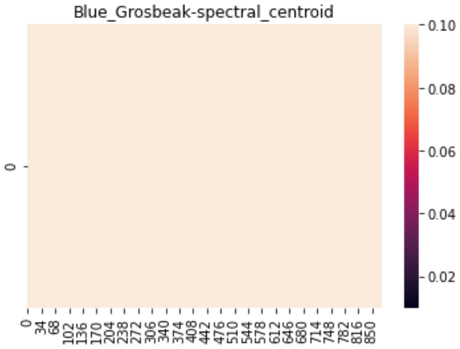
It is made by transforming the chroma vector into the frequency domain.

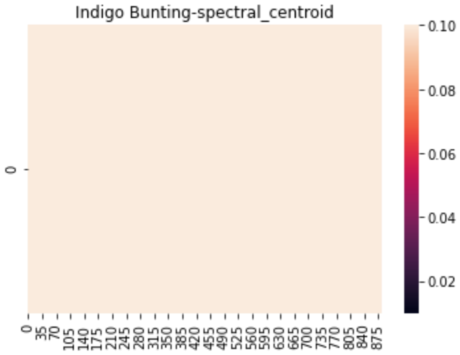

Blue_Grosbeak-chroma_cqt


Indigo Bunting-chroma_cqt

Spectral centroid

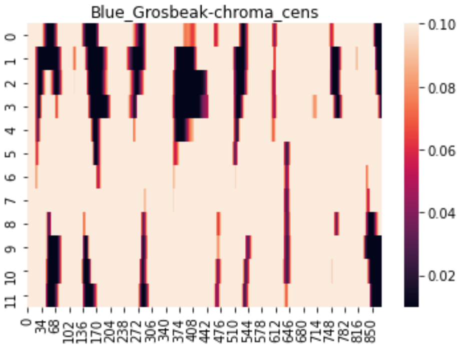
The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located. Perceptually, it has a robust connection with the impression of the brightness of a sound.

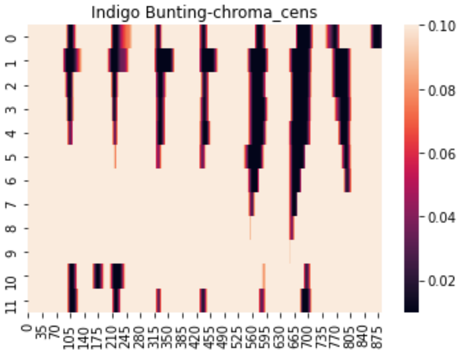

Blue_Grosbeak-spectral_centroid


Indigo Bunting-spectral_centroid

Chroma CENS

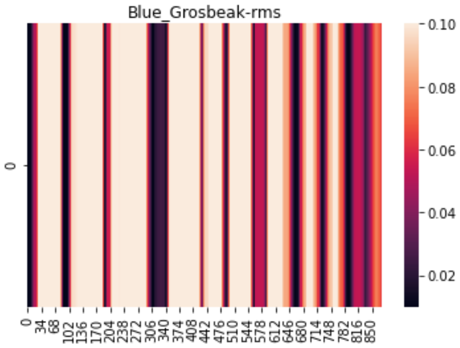
Adding a further degree of abstraction by considering short-time statistics over energy distributions within the chroma.

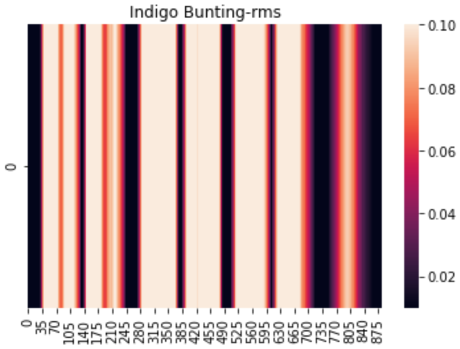

Blue_Grosbeak-chroma_cens


Indigo Bunting-chroma_cens

Root mean square(RMS)

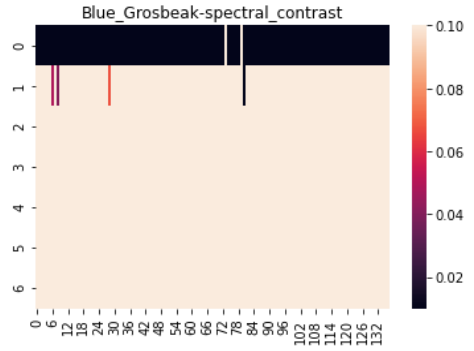
A tool to measure the average loudness of a sound.


Blue_Grosbeak-rms

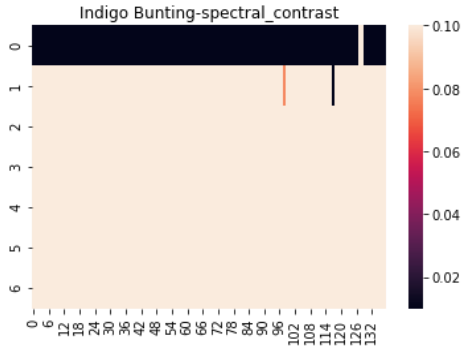

Indigo Bunting-rms

Spectral contrast

The difference between the spectral peak and bottom in each frequency sub-band.



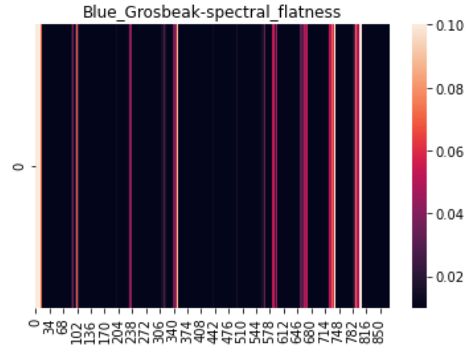
Blue_Grosbeak-spectral_contrast



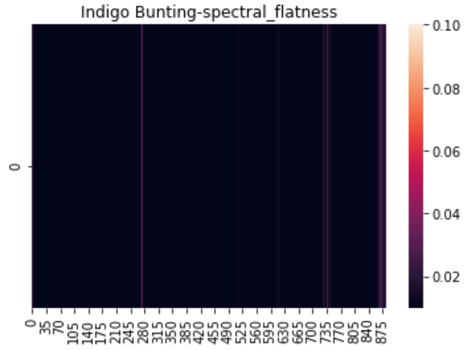
Indigo Bunting-spectral_contrast

Spectral flatness

This measures the white noise, which is the similarity of the amount of energy in each spectrum.



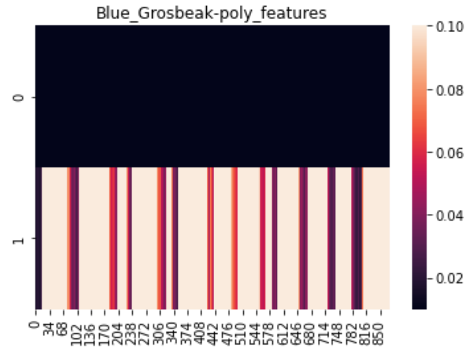
Blue_Grosbeak-spectral_flatness



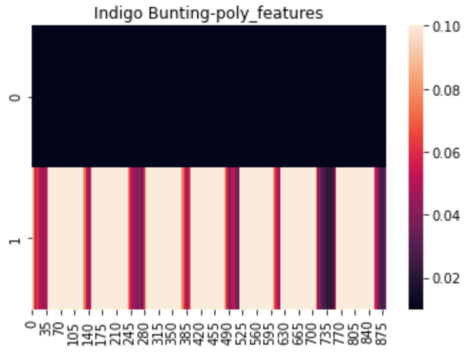
Indigo Bunting-spectral_flatness

Polynomial feature

This method applies polynomial regression to each band to extract the robustness against the background.



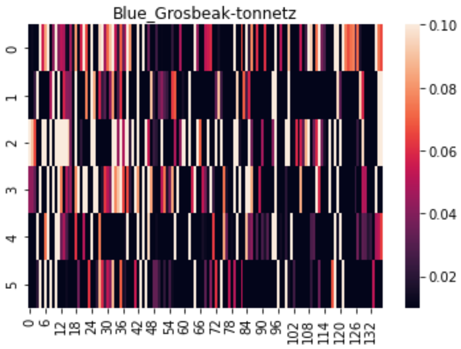
Blue_Grosbeak-poly_features

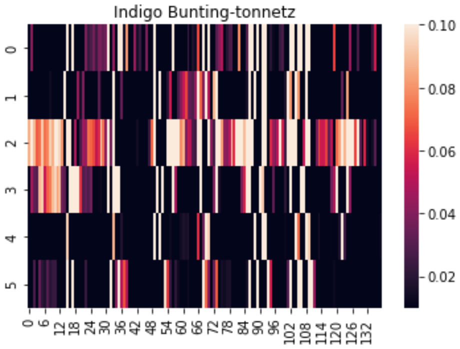


Indigo Bunting-poly_features

Tonnetz

2-dimensional mesh, which represents the tonal space.



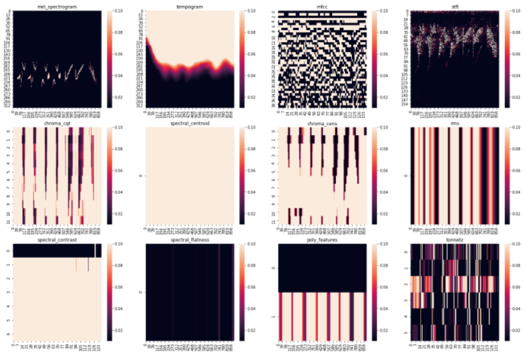


CNN model

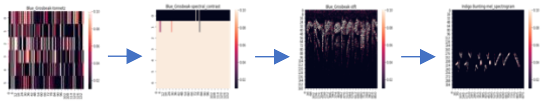
While I'm going to use a relatively simple CNN, some attention should be paid to the data adjustment.

Data layout

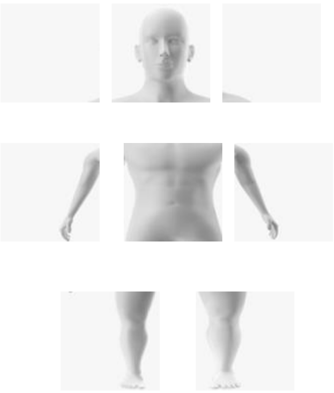
CNN is an image recognition model that can extract and remember patterns within an image. To improve the accuracy and make learning easier for the model, we suggest you combine the 12 charts into one big image instead of asking the CNN model to learn each chart separately.



vs



If you ask why? Let's put it simply: which one do you think is easy to recognize as a human?



vs

We want to leave some white space between the images to tell the model that the two images are not fully connected. The width of the space should be wider than the window size in the convolutional layer. If your largest window size is 8x8, then the white space width should be around 10 with considering any slide you have.

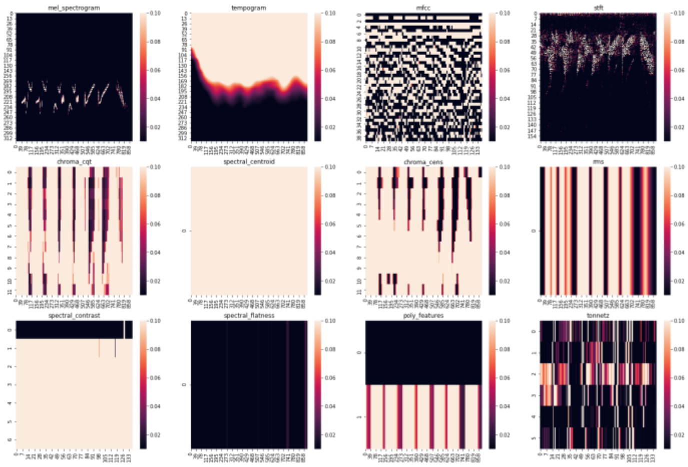
No standardization

It became almost a 'Must step" to standardize the data before you send it to the model. However, for this sound recognition process, I recommend you to be very careful if you want to standardize the data.

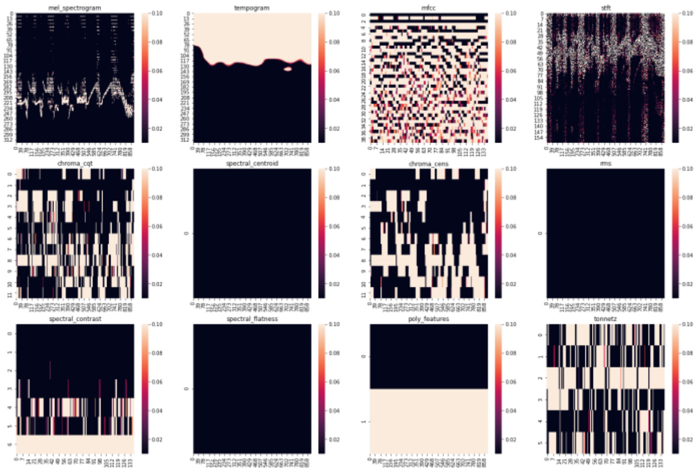
Those 12 features above have already been transformed once, and some of them have already been compressed or "standardized" in certain ways. If you standardize it again, you might lose some important information by twisting the feature in an unpredicted way.

As you can compare the below two charts, the standardized one has been twisted further.

Not standardized



Standardized



Collecting sound after separating the livestock by age, day/night, or any other factors

One of the burdensome works which takes much more cost and time is recording sound by groups. For example, separating children from their mothers can cause stress to the animal, which is against the concept of animal welfare. Also, separating by age may sometimes need more cost (build more houses for different livestock ages) and sometimes are impossible to do.

How about a universal model that can detect all the sounds by all the groups? I think it will be demanding as you can easily imagine that we still haven’t even successfully made a voice translator for all ages and all human countries.

Are voice recognizers for animals doomed to be specialized for each individual farm?

Yes, data shows that even for chickens, farms in the EU have different patterns from farms in the USA.

However, we can start small by only focusing on some key sounds such as ‘cough,’ ‘pain,’ and ‘hungry”.