SDSC2102 Group Project

Recognize Animals from Sound

Python libraries: librosa, noisereduce, IPython.display, pandas

Data Preparation:

Audio data background knowledge: https://youtu.be/vrXGaFV1AmE

Raw data: the animals audio from the ESC-50 dataset. 320 segment audios with each 5 seconds, 40 for each animal ('dog', 'cat', 'rooster', 'hen', 'pig', 'frog', 'cow', 'crow').

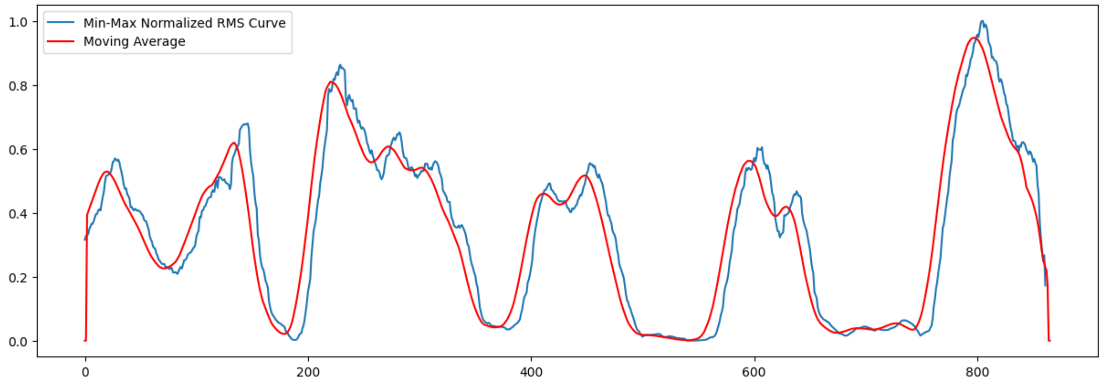
Data processing: the audio dataset is not normalized and contain a lot of noise.

I. Noisereduce library is used for noise reduction. Even it kind of like too easy to bypass the part of noise reduction. Although we know noise reduction is also the job of data processing, it requires a lot of techniques and maths that far away from the course covering, also our project is not focused on this part. No details.

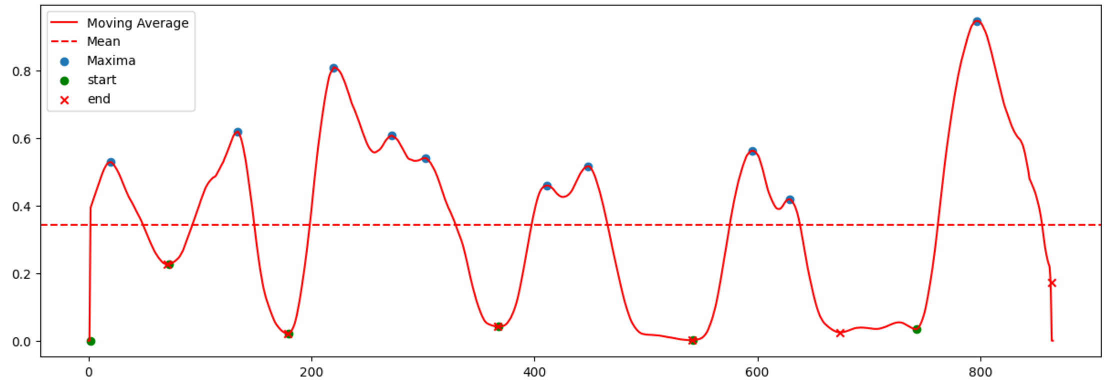
II. Segmentation: We believe that segmentation of the audio into smaller clips that contain only the period when the animal was producing sound is helpful for improving the model. Thur, we decided some segmentation.
Step:

1. Conduct a root-mean-square (RMS) curve for getting a smoot signal envelope. However, you can see the RMS curve still contain some small peak that are unprofitable to our detection methods.
RMS curve parameter: FRAME_SIZE=2048, HOP_LENGTH=128.

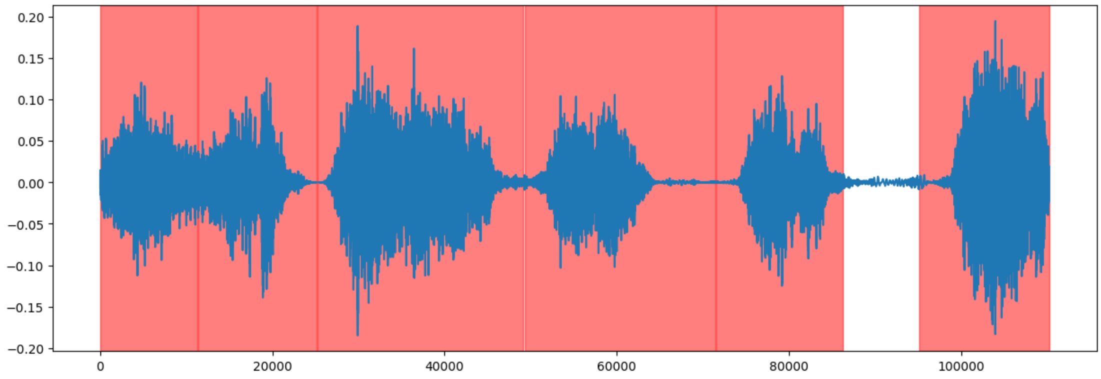
2. Conduct a moving average of the RMS curve with k=10. That smoother the curve.
Moving Average curve parameter: MOVING_SIZE=20



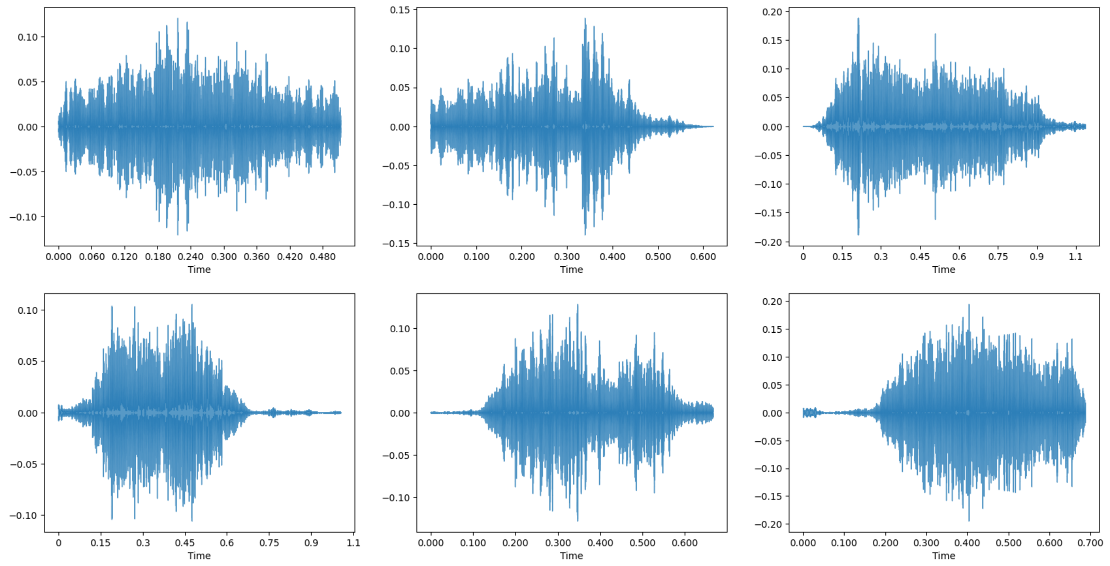
3. Find maximas above mean. Then find the first minimas before and after each maximum below mean as the start and end of each sound produced by the animals.



Length of RMS curve: 862



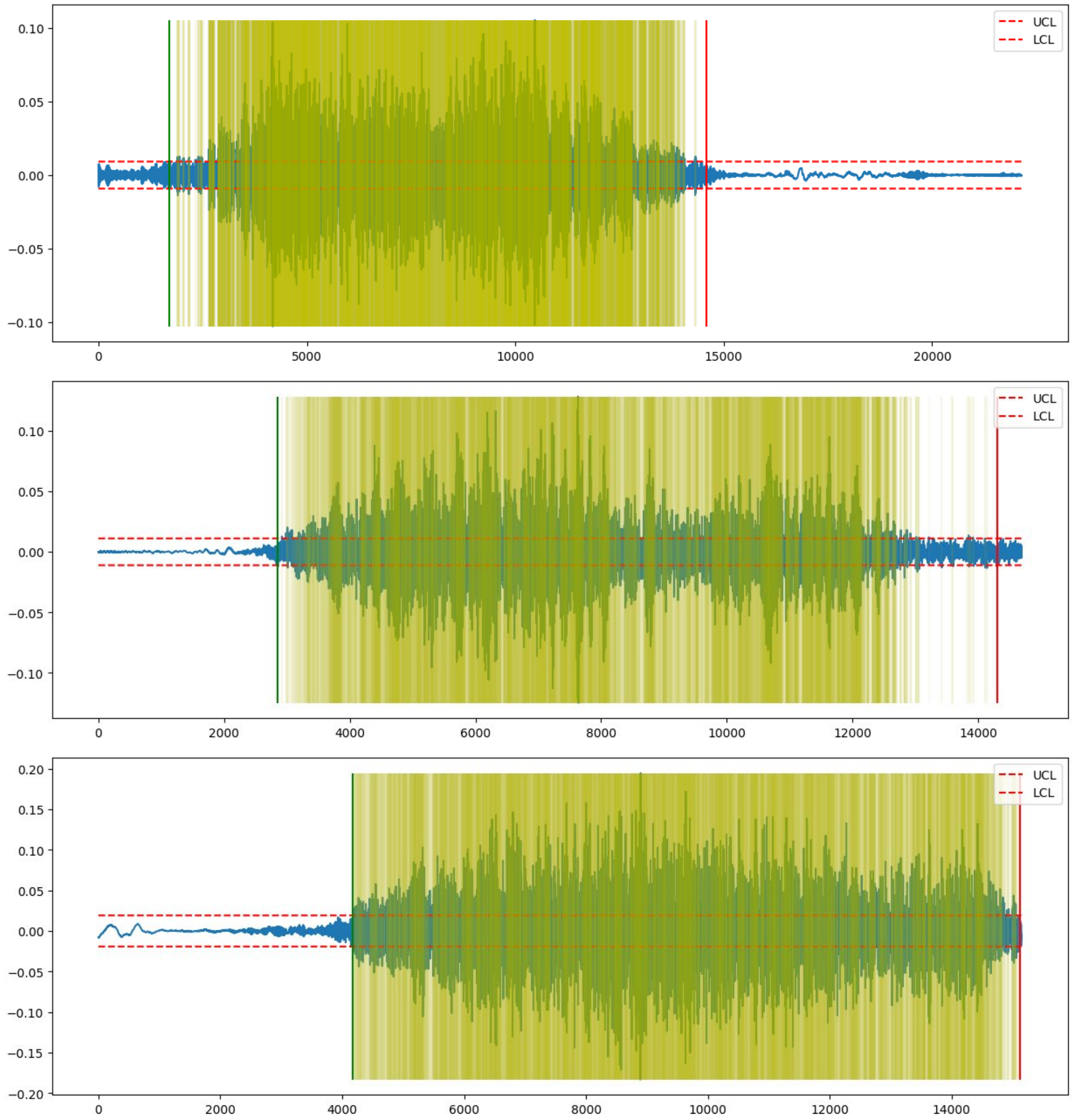
Start Sample Point: [0, 11421, 25245, 49309, 71581, 95104]
End Sample Point: [11293, 25117, 49181, 71453, 86272, 110250]



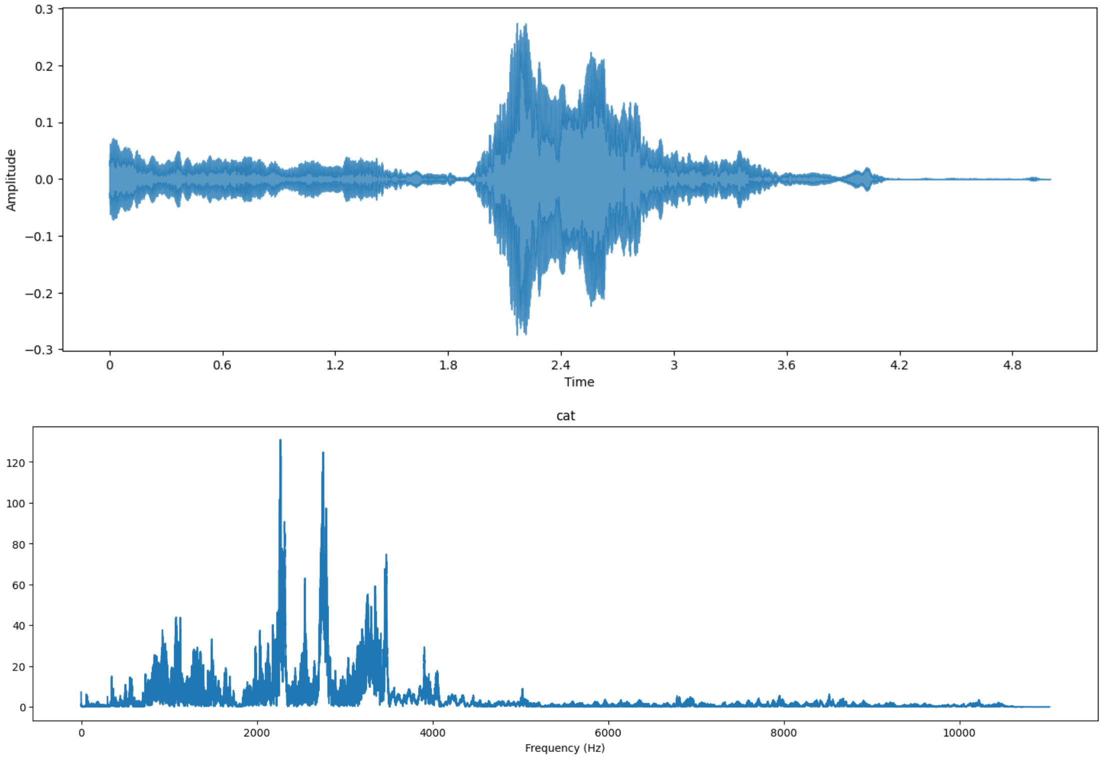
This method may not accurate, and we can’t test its accuracy as the periods of the sound from the present are not labeled in the metadata for the audio. (Although we can find it by head)

III. Second segmentation: To further cut out the off-set part in the clips. We apply second segmentation procedure. standard deviations
Step:

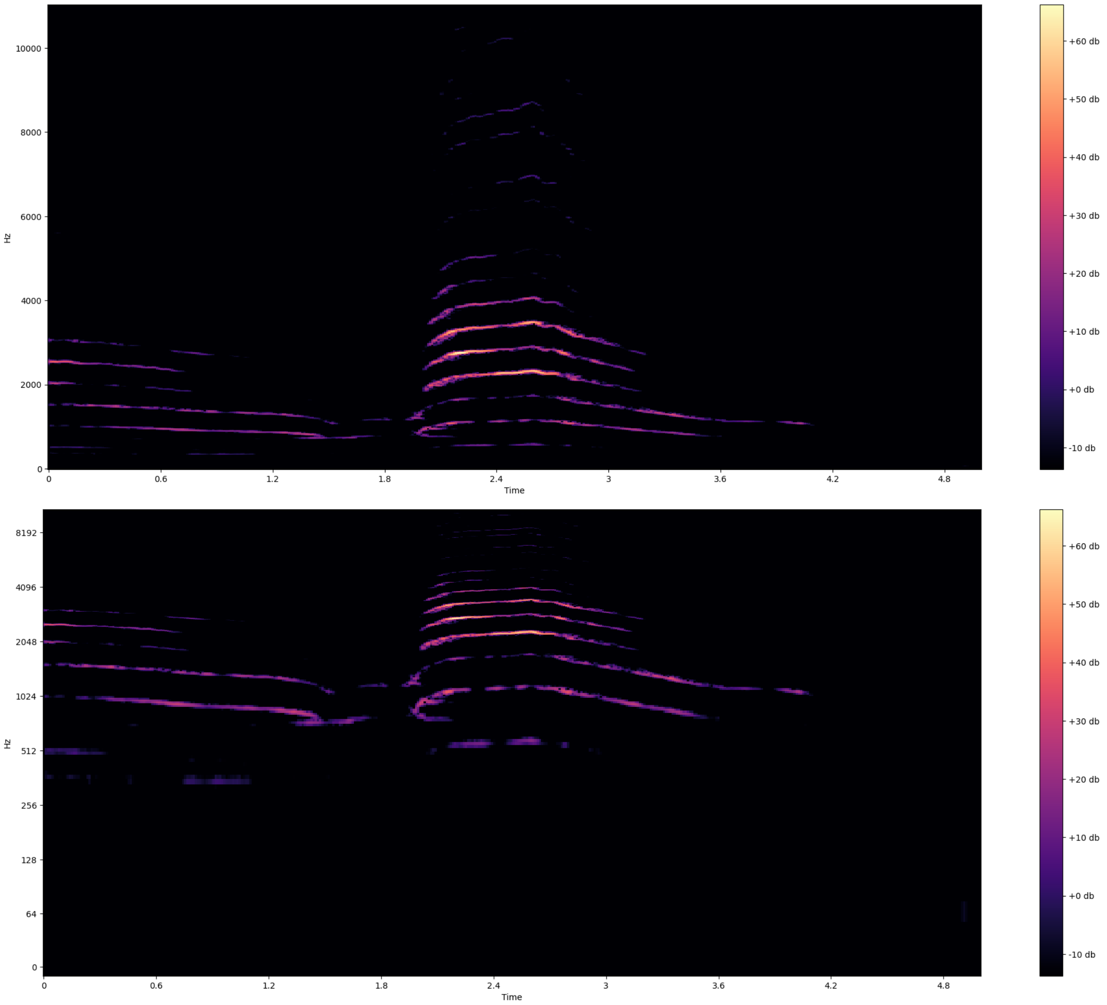
1. Perform second segmentation by applying change detection techniques discussed in class.



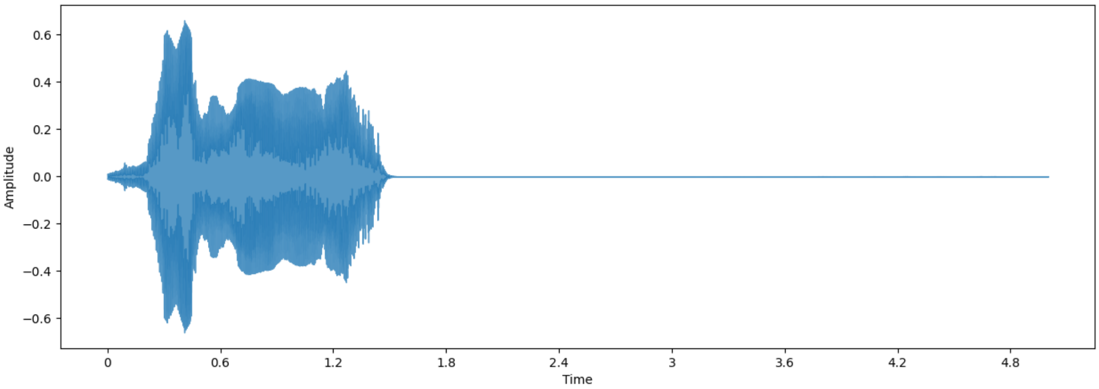
Cat raw wav: Time domain and frequency domain plot (frequency domain etracted by FFT)

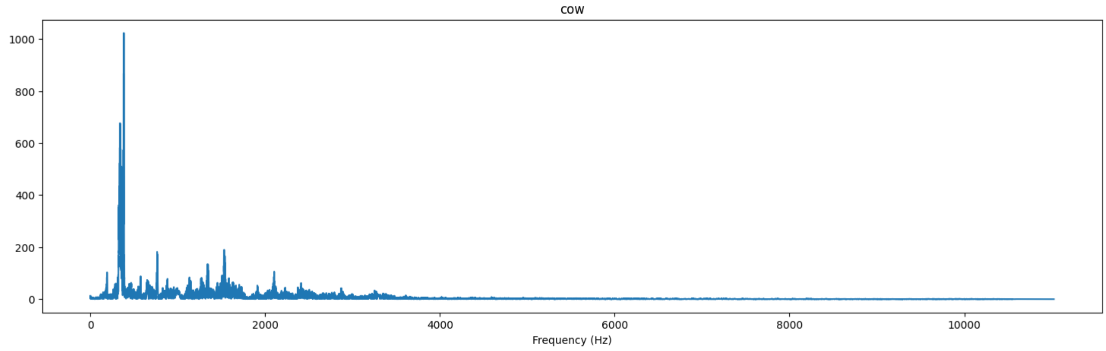


Spectrogram: Linear and log persentation

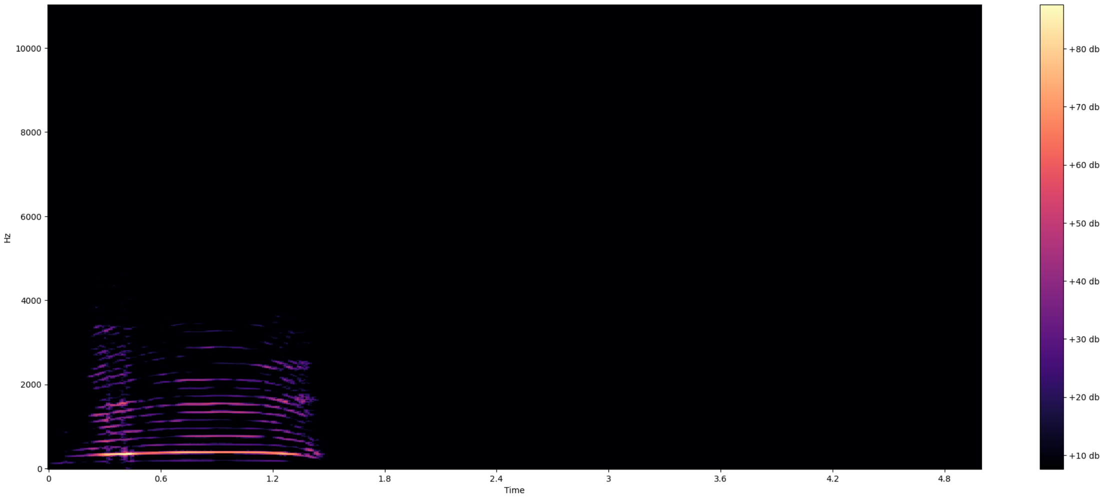


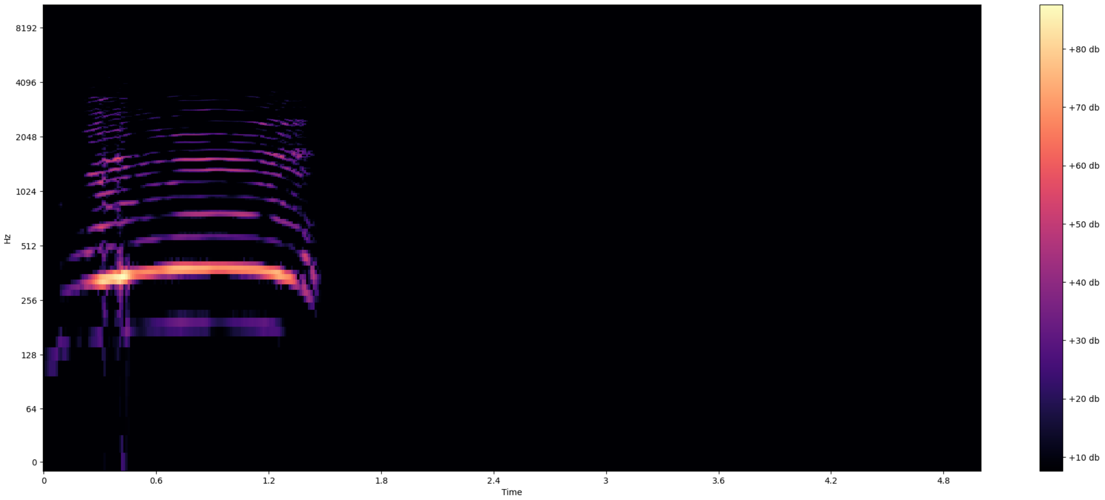
Cow raw wav: Time domain and frequency domain plot (frequency domain etracted by FFT)





Spectrogram: Linear and log persentation





Features extraction:

We extract audio features mainly from its time domain and frequency domain. There are lot of features already researched or discovered by the field of research in audio analysis. One power future, Mel-Frequency Cepstral Coefficients (MFCCs), are commonly used as features in speech recognition systems. Extracting some features contain rich information of the audio often require complex procedures. Still, there are some features come from statistical concept that are worth mention.

Spectral centroid: its nature is ***weight mean*** of the frequencies

$$SC = \frac{\sum_{n=1}^{N} m(n) \cdot n}{\sum_{n=1}^{N} m(n)}$$

$$SC = \frac{\vec{n} \cdot \vec{m}}{\| \vec{m} \|_1}$$

Bandwidth: its nature is Weighted mean of the distances of frequency bands from SC ***(standard deviations)***

$$BW = \frac{\sum_{n=1}^{N} | n - SC | \cdot m(n)}{\sum_{n=1}^{N} m(n)}$$

$$BW = \frac{\overrightarrow{| \vec{n} - SC |} \cdot \vec{m}}{\| \vec{m} \|_1}$$

Band energy ratio: its nature is comparison of energy in the lower/higher frequency bands ***(ratio)***

$$BER = \frac{\sum_{n=1}^{F-1} m(n)^2}{\sum_{n=F}^{N} m(n)^2}$$

Some others descriptive statistical value like frequency with maximum energy magnitude can also be the feature for recognition models.