Landon Buell

Kevin Short

PHYS 799.01

18 April 2020

**Possible Features for Musical Instrument Classification**

“The ﬁrst step in a classiﬁcation problem is typically data reduction. The data reduction stage which is also called feature extraction, consists of discovering a few important facts about each class. The choice of features is critical as it greatly affects the accuracy of audio classiﬁcation. The selected features must reﬂect the signiﬁcant characteristics of each class of audio signals. In order to better discriminate different classes of audio, we consider features that are related to the temporal and spectral domains.” [1]

**Time Basis Features**

10% - 90% Rise time

This measures the time difference between the first time an amplitude crosses 10% of its maximum value to the first time the same waveform crosses 90% of its maximum amplitude. This time difference gives a rough measurement of the initial attack of the waveform. In some cases, like brass or percussive instruments, this value would be extremely small. In the case of bowed string instruments or sustained woodwind instruments, the maximum amplitude may occur towards the end of the waveform, thus be a large part of its whole. To avoid inconsistency between signals of varying length, this metrics is taken as a fraction or ratio of the length of the signal.

90% - 10% Decay time

This measures the time difference between the last time an amplitude crosses 90% of it’s maximum value to the last time the same waveform crosses 10% of its maximum amplitude. This time difference gives a rough measurement of the decay and release time of a waveform. In the case of percussive instruments, this value is likely to be very large and make of up of the waveform. To avoid inconsistency between signals of varying length, this metrics is taken as a fraction or ratio of the length of the signal.

Percentage of “Low Energy Frames” [1]

This measures the number of *frames* in a waveform that have an RMS power of less than 50% of the RMS power of the rest of the signal. A frame is a subset of audio data, usually around 10 – 40 ms. A frame with “low” energy in indicative of silence or lower energy waveforms. If a waveform has a generally low amplitude, then the number of frames below the mean RMS would relatively large – this a “left skewed”.

Spectral Flux [1]

Spectra flux is (also the delta-spectrum magnitude) a measurement of frame-to-frame spectral difference. It is analogous to take a discrete – first derivative of information in a time or frequency spectrum. In most cases, this is used again with respect to *frames* (waveform subset) rather than each individual audio sample.

Spectral Flux =

Waveforms that experience more dynamic amplitude shifts would have a higher time- spectrum flux. Similarly, a signal that changed frequency a great deal would have a higher frequency – spectrum flux.

Linear Predictive Coefficients (LPC) [1]

We use an algorithm to predict the value of the *n+1* point based on the previous *n* points. We compare this predicted result to the actual waveform sample. The general idea is that each prediction is a weighted sum of the previous *n* samples. Thus the *n*-th predicted sample of a waveform of *N* samples, is given by: ­

Where is a prediction coefficient. These can be computed by minimizing an RSS score or MSE score for previous samples. By taking an element-wise difference of the actual and predicted signal functions, we can compute and the value of any standard error metric and use that as a feature in audio classification.

Zero Crossings

This feature simply counts the number of times that a waveform crosses the horizontal axis. In a time- domain setting, this creates a very crude frequency measurement. To avoid inconsistency between signals of varying length, this metrics would be used and zero-crossing per unit time.

Range of Zero Crossings (R-ZC) [1]

**Frequency Basis Features**

Isolated Peaks

Certain instruments may on average, show much stronger overtones in their frequency spectrum. I propose an algorithm that computes the number of modes in the spectrum that have spikes above a certain value. By calculating the number of peaks above a certain threshold, within a certain frequency range could indicate the rough harmonic range of the instrument, and the properties of the overtones of that instrument.

Power of Successive Overtones

This feature would allow for the extraction of the ratio between the strengths of successive overtones. Instruments with strong fundamentals and weak overtones would return a high value for this metric. Instruments with roughly balanced of similar strength overtones would present a low value for this metric. We can also extend this to include the average power, standard deviation and variance of the heights of each peak.

Spectrogram Density

In a time vs. frequency vs. power plot, it can become apparent that certain overtones have only exist in certain subset of the time-frequency domain. We can isolate the total number of overtones in the entirety of a signal, and then compute what fraction of them are present in what fraction of time. Again we can also expand this to include the mean, standard deviation and variance.

Complex Spectral Phase Evolution [2]

This method could be used as an alternative to the standard Fast-Fourier Transform to produce a significantly more accurate frequency-space representation of a signal. This improve accuracy would also help improve the certainty and validity of features that are further drawn from the frequency spectrum. I am producing a python wrapper that can call the CSPE.m MATLAB script. Details to follow.

**References**

[1] Khan, M. Kashif Saeed, and Wasfi G. Al-Khatib. “Machine-Learning Based Classification of Speech and Music.” Multimedia Systems, vol. 12, no. 1, 2006, pp. 55–67., doi:10.1007/s00530-006-0034-0.

[2] Short, Garcia, “Signal Analysis Using the Complex Spectral Phase Evolution (CSPE) Method”. Journal of the Audio Engineering Society.