Feature Extraction

Shan Jiang, Thiago Roque, & Matt McMullen

Computational Music and Audio Analysis

Professor Alexander Lerch

10/17/2021

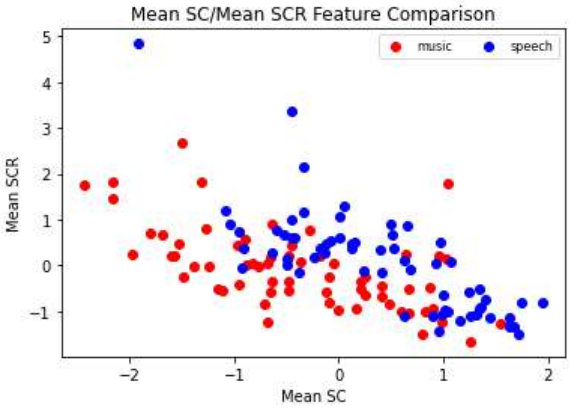

Figure 1. Scatter plot comparing the mean spectral centroid to the mean spectral crest factor for each audio file.

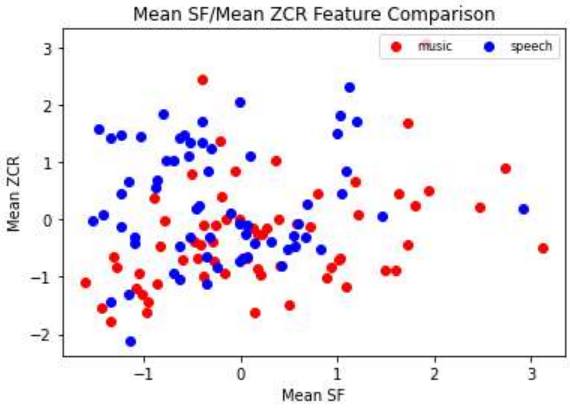

Figure 2. Scatter plot comparing the mean spectral flux to the mean zero-crossing rate for each audio file.

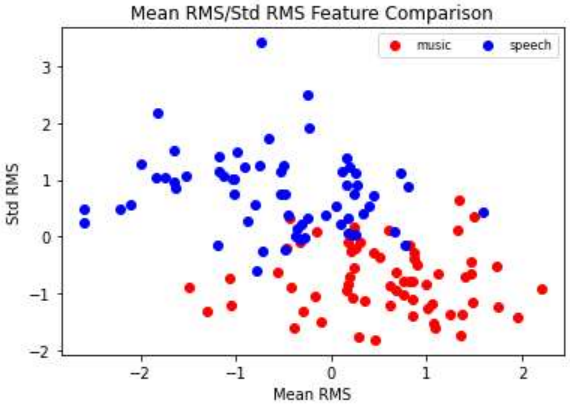

Figure 3. Scatter plot comparing the mean RMS to the RMS' standard deviation for each audio file.

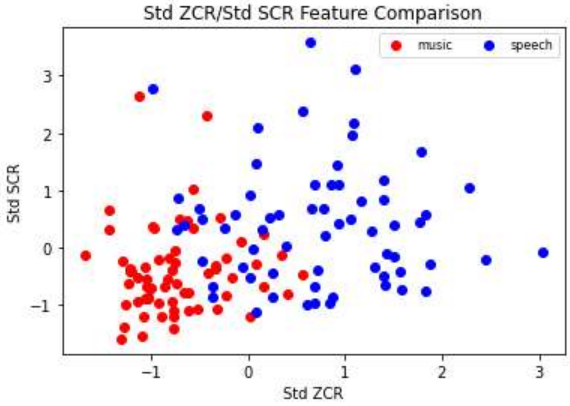

Figure 4. Scatter plot comparing the zero crossing rate's standard deviation to that of the spectral crest factor for each audio file.

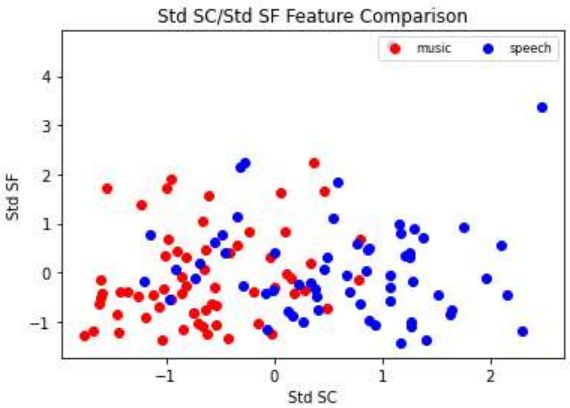

Figure 5. Scatter plot comparing the spectral centroid's standard deviation to that of the spectral flux for each audio file.

By viewing Figures 1-5, several inferences can be made about both the audio files within the dataset that were analyzed and the features extracted from them. Firstly, Figures 3-5 appear to be potentially useful for the purpose of categorizing an audio file as either music or speech. All three scatter plots show general clustering with respect to those two categories; and interestingly, they all visualize the standard deviations of the features of focus. This highlights a key observation that for most features, namely the spectral centroid (SC), zero-crossing rate (ZCR), spectral crest (SCR), and root mean squared (RMS), the standard deviation is noticeably higher for speech audio than for music audio, indicating that speech is generally more variable with respect to these features. The one exception is that the standard deviations are comparable between the spectral fluxes of music and speech (Figure 5).

In contrast to the seemingly clearly distinct nature of the standard deviations, the mean values for all of the features show a lot of overlap between music and speech (Figures 1-3). While the mean values of these features may not serve as reliable metrics for categorization, some further observations can be made. In Figure 2, music and speech both show mean ZCR values across a similar wide range. However, the music files also had a wide range of possible mean SF values while speech did not. This pattern makes sense as SF is a measure of spectral change from frame to frame and one could reasonably assume that there would be more spectral change in even monophonic music than in standard speech. Additionally, the mean RMS values for music were generally larger than those of speech. This is consistent with another reasonable assumption that the dynamic range of music would generally be larger than that of speech, especially when the music is polyphonic or contains changes in number or power of voices.

While some of the plots of mean feature values seem to follow a preconceived expectation, it's important to note that the features computationally extracted from audio files my not always be perceptually relevant. With this in mind, it's possible that the observations previously mentioned in-fact have nothing to do with the "reasonable assumptions" and interpretations provided along side them. Moreover, results that don't make immediate intuitive sense, such as the higher average spectral centroid values seen in speech compared to music (Figure 1), may have a perfectly logical explanation that has nothing to do with how either speech or music sound. The most valuable information obtained through the extraction, normalization, and visualization of these features is what and how features can be used for a given task. In the case of this dataset, it is immediately obvious that certain features seem to be especially relevant if the given task is to categorize an audio file into either music or speech.