Addressing the *Same but different - different but similar* problem in automatic music classification

Unjung Nam
CCRMA
Stanford University
Stanford, CA 94305
1 650 723 4971
unjung@ccrma.stanford.edu

Jonathan Berger
CCRMA
Stanford University
Stanford, CA 94305
1 650 723 4971
brg@ccrma.stanford.edu

ABSTRACT

We present a hybrid method in which we classify music from a raw audio signal according to their spectral features, while maintaining the ability to assess similarities between any two pieces in the set of analyzed works. First we segment the audio file into discrete windows and create a vector of triplets respectively describing the spectral centroid, the short-time energy function, and the short-time average zero-crossing rates of each window. In the training phase these vectors are averaged and charted in three-dimensional space using k-means clustering. In the test phase each vector of the analyzed piece is considered in terms of its proximity to the graphed vectors in the training set using k-Nearest Neighbor method. For the second phase we apply Foote's (1999) similarity matrix to retrieve the similar content of the music structures between two members in the database.

1. ANALYSIS METHODS

1.1 Spectral Centroid

The spectral centroid is commonly associated with the measure of the brightness of a sound. The individual centroid of a spectral frame is defined as (here, $F[k]$ is the amplitude corresponding to bin k in DFT spectrum..)

$$Spectral \quad Centroid \quad = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]}$$

Figure 1 presents the weighted average spectral centroids of the two analyzed sound examples. The lower (magenta) band is an excerpt of the Kremlin Symphony's recording of Mozart's Symphony 25 (K. 183) and the upper (cyan) band is a rock style arrangement of the same musical segment. The high frequency components in the pervasively percussive rock version accounts for its higher placement on the graph.

1.2 Short-Time Energy Function

The short-time energy function of an audio signal is defined as: (where $x(m)$ is the discrete time audio signal, n is time index of the short-time energy, and $w(m)$ is a rectangular window.)

$$En = \frac{1}{N} \sum_{m} [x(m)w(n-m)]^2 \qquad w(x) = \begin{cases} 1, & 0 \le x \le N-1, \\ 0, & \text{otherwise.} \end{cases}$$

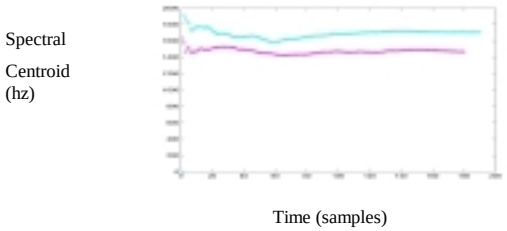

Figure 1.

It provides a convenient representation of amplitude variation over time. Patterns of change over time suggest the rhythmic and periodic nature of the analyzed sound. Figure 2 is the short-time energy change of the same excerpts. The highly fluctuating rock version (cyan) resulting from the persistent drum beats compared to the more subdued but highly contrasting symphonic version suggests one possible determinant for genre classification.

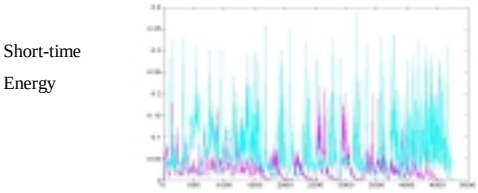

Figure 2.

1.3 Short-Time Average Zero-Crossing Rate

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different signs. The short-time averaged zero-crossing rate (ZCR) is defined as

$$Zn = \frac{1}{2} \sum_{m} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m),$$

$$where \quad \text{sgn}[x(n)] = \begin{cases} 1, & x(n) \ge 0, \\ -1, & x(n) < 0, \end{cases}$$

Figure 3 is the ZCR over time of the same two sound examples, as before, the classical version is magenta and the rock version is cyan. Compared to that of speech signals, the ZCR curve of music has much lower variance and average amplitude and when averaged, shows significantly more stability over time. ZCR curves of music generally have an irregular small range of amplitudes.

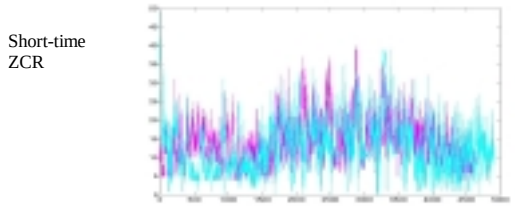

Figure 3.

1.4 Foote's Similarity Method

Foote (1999) represents acoustic similarity between any two instants of an audio recording in a 2D representation, Figure 4 shows the 'similarity matrix' analyzed for the two music samples. The parameterization was done with a Mel-frequency cepstral coefficient function with frame size 30. Both samples are about 16 seconds long and sampled at 11025hz, 16 bits. The analysis visualizes the tripartite segmentation of the phrase in the 16 second excerpt (seconds 1-5, 5-12, and 12-16) in both the classical version (fig 4.1) and the classical version (fig 4.2). Despite the stylistic disparity between the two examples the musical similarity in terms of pitch and rhythmic structure is well represented.

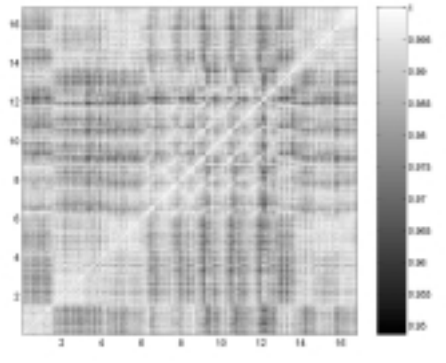

Figure 4.1. orchestral version.

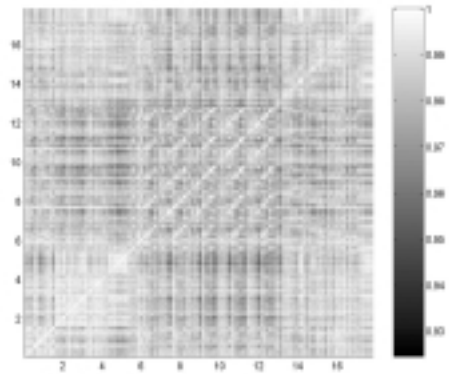

Figure 4.2. rock version

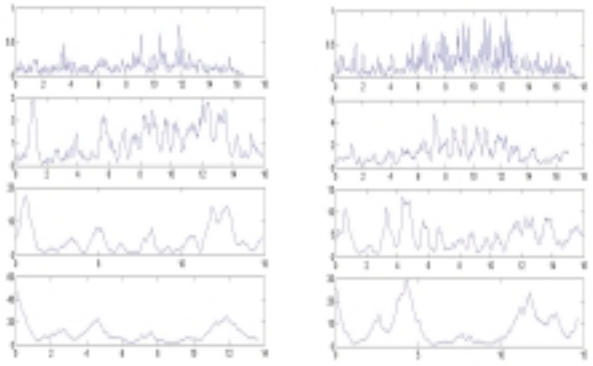

Figure 5. Novelty scores of
orchestral (left) and rock (right) version

Figure 5 presents the novelty scores over time(second) of the two examples. In each figure the outputs with kernel sizes, from top down, 10, 20, 60 and 96. The graph of kernel size 96 displays three high peaks corresponding to the tripartite musical structure. The smaller the kernel size the greater the detail represented. This facilitates detection of discrete musical events. We are currently considering heuristics to find optimal kernel sizes to track appropriate novelty information.

2.CONCLUSION

In this paper we explored a computational model that combines classification and comparison of raw audio signals to explore the perceived similarity between musical recordings. Foote's (1999) similarity matrix retrieved the similar content of the music structures between two music samples even though their spectral components are different. Future research will focus on quantitative measurement of the degree of musical similarity between two works, as well as genre classification by statistical clustering.

3. ACKNOWLEDGMENTS

Our thanks to Professor Julius O. Smith, Jonathan Foote and Malcolm Slaney for their insights and assistance.

4. REFERENCES

Foote, J. (1999) "Visualizing Music and Audio using Self-Similarity." In Proceedings of ACM on Multimedia.

Scheirer, E. and Malcolm Slaney. (1997) "Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator." In Proceedings of IEEE ICASSP.