Special Area Exam Part II, April 28, 2001
Unjung Nam

Due to the restricted computer resource and computing time, I have edited the sound files and used for this analysis only 16 seconds portion. They are in the .wav format, 11025 hz sampling rate, and 16 bit quantization.

I'd like to present a few methods I have previously used for the analysis of the digitized music signals. They are Spectral Centroid, Short-Time Energy Function, Short-Time Average Zero-Crossing Rate, and Foote's self-similarity method. These methods are not yet sufficiently experimented with many audio samples and they need to be explored more for figuring out their optimal conditions useful for music genre classification system. The descriptions of each method are based on my previous research reports.

Spectral Centroid

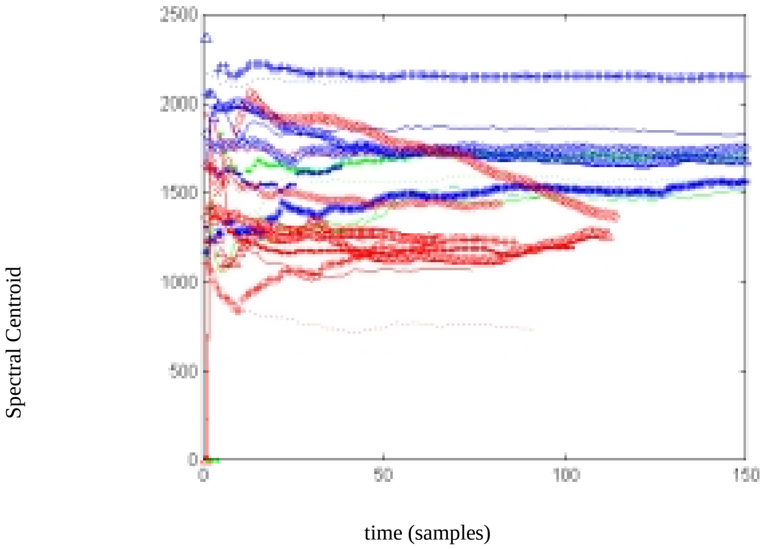
The spectral centroid is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the "center of gravity" using the Fourier transform's frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes, or:

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

Here, *F* [*k*] is the amplitude corresponding to bin *k* in DFT spectrum.

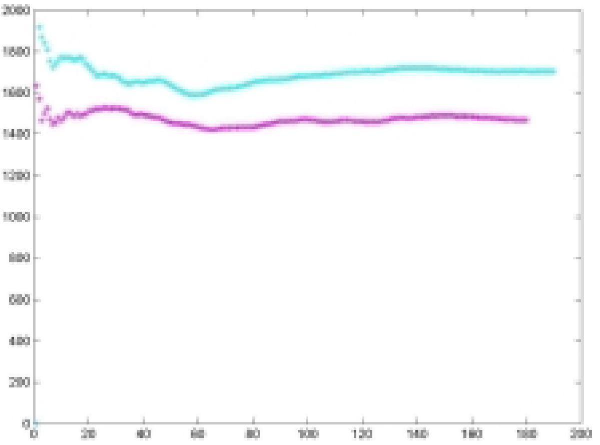
In practice, centroid finds this frequency for a given frame, and then finds the nearest spectral bin for that frequency. The centroid is usually a lot higher than one might intuitively expect, because there is so much more energy above (than below) the fundamental which contributes to the average.

It is not sure if the spectral centroid would be useful for classifying different genres of musics. At least, it will show some spectral components of the music, which are mixed sounds.



The above figure shows the weighted average spectral centroid of each sample of 20 sound files. The red ones are classical music samples while the green ones are jazz. The blue ones are pop/rock music. It shows that the spectral centorid of pop/rock music samples seems higher than that of classical music samples. The spectral centroid of jazz music samples are somewhere between that of the two. Also it seems that the fluctuation of the value over time is somewhat less stable in classical case than the rock music.

Here are the weighted average spectral centroids of the two sound examples that I have analyzed for this exam. The magenta color shows the changes of the averaged spectral centroid over time of the Mozart 25th symphony and the cyan color shows that of its rock version. It is interesting to see that the changes of the averaged spectral centroid of the rock version is placed above the classical version. It seems to imply that the higher spectral centroid of the rock version is due to the high frequency components from the drum playing. However, the fluctuation of the spectral centroid over time looks stable in both cases. The fluctuating value of this doesn't seem to imply anything for music genre classification and it only depends on what kinds of instruments are playing in time.



Short-Time Energy Function

The short-time energy function of an audio signal is defined as

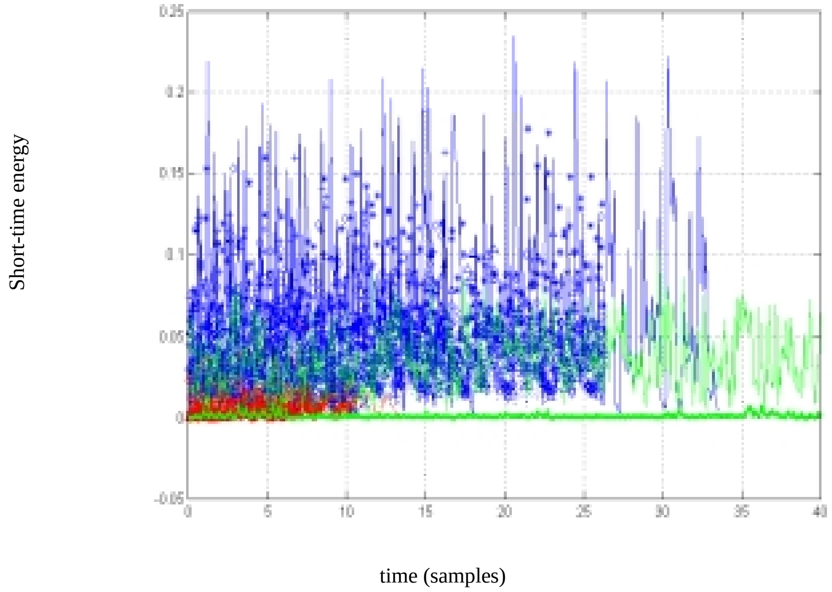
$$En = \tfrac{1}{N} \sum_{m} \left[x(m) w(n-m) \right]^2$$

Where $x(m)$ is the discrete time audio signal, n is time index of the short-time energy, and $w(m)$ is a rectangle window, i.e.

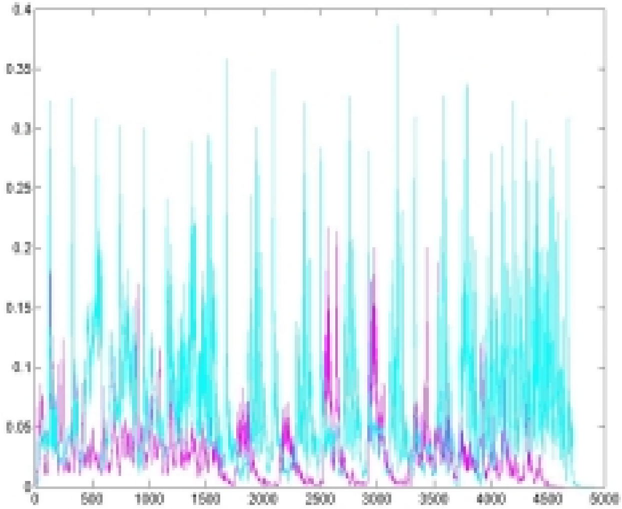
$$w(x) = \begin{cases} 1, & 0 \le x \le N-1, \\ 0, & \text{otherwise.} \end{cases}$$

It provides a convenient representation of the amplitude variation over the time. For the purpose of this experiment, its change pattern over the time may reveal the rhythm and periodicity nature of the underlying sound.

The following figure shows the short-time energy changes over time of each sample of 20 sound files. The red ones are samples of classical music while green ones are jazz.
The blue ones are pop/rock music. The pop/rock music samples show the most fluctuating energy while classical music samples shows stable energy fluctuation. The energy changes of jazz samples seem to show medium fluctuation.



Here are the short-time energy change of the two sound examples that I have analyzed for this exam. The magenta color shows the energy change over time of the Mozart 25th symphony and the cyan color shows that of its rock version. It clearly shows that the rock version is much more highly fluctuation than the classical version. Though the fluctuation of the energy due to the beats produced by drums, it seems that this function can be useful for classifying musics into pieces with or without drum and even into classical and rock/pop genre. This result also confirm my assumption from the previous figure.



Short-Time Average Zero-Crossing Rate

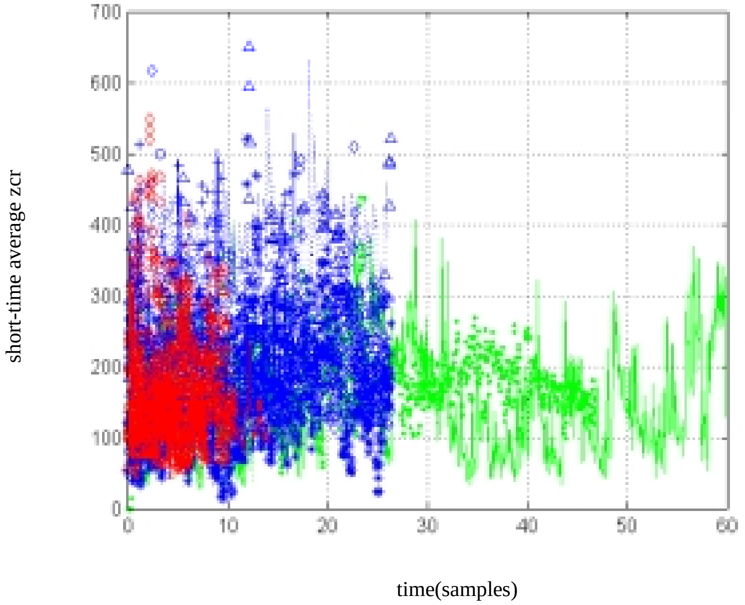
Put simply, Zero-Crossing Rate (ZCR) is a measure of how often the signal crosses zero per unit time. In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different signs. The rate at which zero-crossings occur is a simple measure of the frequency content of a signal. The short-time averaged zero-crossing rate is defined as

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

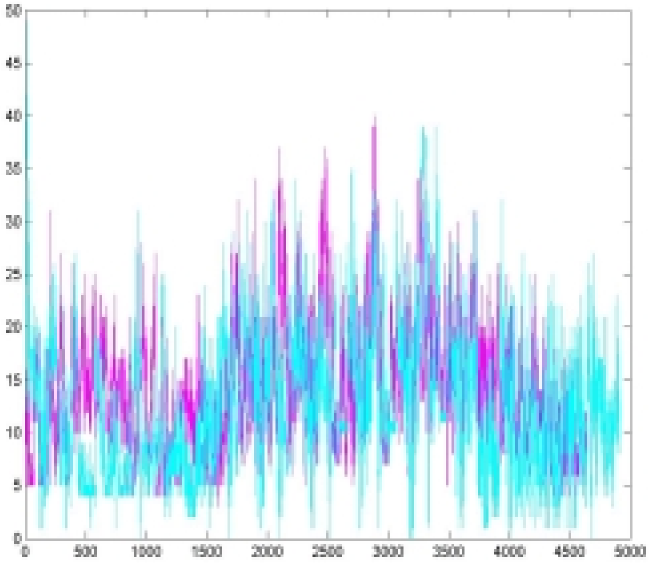
$$where \quad \mathrm{sgn}\left[x(n)\right] = \begin{cases} 1, & x(n) \geq 0, \\ -1, & x(n) < 0, \end{cases}$$

and *w*(*n*) is a rectangle window of length *N*. Compared to that of speech signals,
The ZCR curve of music has a much lower variance and average amplitude. This
Suggests that the averaged zero-crossing rate of music is normally much more stable
During a certain period of time. ZCR curves of music generally have an irregular small range of amplitudes.

The following figure shows the short-time ZCR over time of each sample of 20 sound files. The red ones are samples of classical music while green ones are jazz. The blue ones are pop/rock music. It doesn't seem to show any indication of classifying the three different music genres.



Here are the short-time ZCR over time of the two sound examples that I have analyzed for this exam. The magenta color shows the energy change over time of the Mozart 25th symphony and the cyan color shows that of its rock version. As my assumption in the previous figure, It doesn't seem to show any indication of classifying the two different musical component.



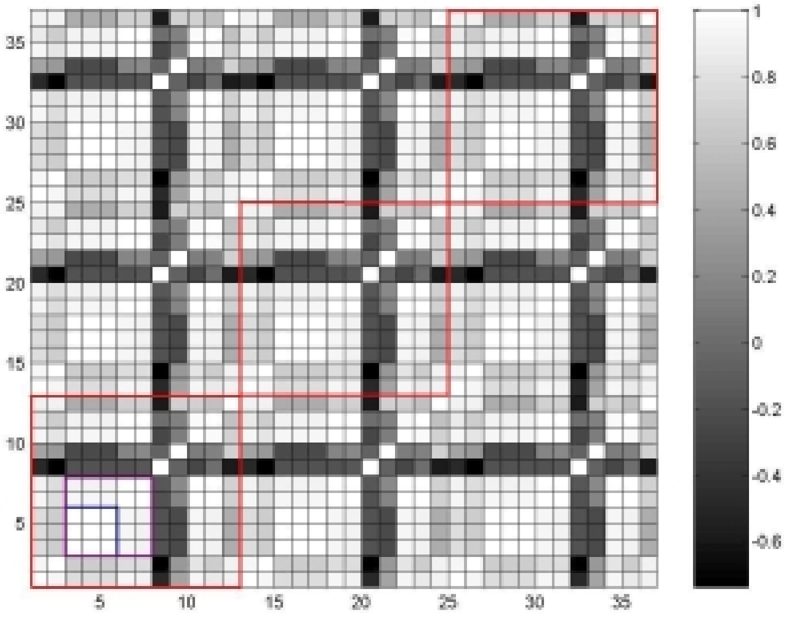
Foote’s Self-Similarity Method

Foote’s self-similarity matrix shows a time series that is proportional to the acoustic novelty of source audio at any instant. Here I present two representations of this method. One is ‘similarity matrix’ and the other is ‘novelty score’. In both representations, high values and high peaks correspond to large audio changes.

Similarity Matrix

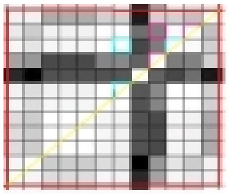
In general similarity matrix will have maximum values on the diagonal (because every window will be maximally similar to itself). Each pixel in the matirx is given a gray scale value proportional to the similarity measure and scaled such that the maximum value is given the maximum brightness. In similarity matrix, the brighter the slanted region is, the less changes it has in the sequential events. Similarity matrix can be visualized as a square image such that regions of high audio similarity, such as silence or long sustained notes, appear as bright squares on the diagonal. Repeated figures, such as themes, phrases, or choruses, will be visible as bright off-diagonal rectangles. If the music has a high degree of repetition, this will be visible as diagonal stripes or checkerboards, offset from the main diagonal by the repetition time.

Here is a matrix ***S*** calculated similarity method.



As I draw the red lines in each pattern, it can be clearly seen that the feature vectors are repeated three times by 13 squares. The repeated pattern (marked as a bracket in the feature vectors and which are the 3rd, 4th, 5th squares in blue box) shows a relatively bright color. The 3,4,5,6,7th squares are not changing much, so they function like a sustained tone or repeated pattern. Thus the purple line portion shows a bright color.

Consider a segment from the matrix above.



White squares on the diagonal correspond to the notes, which have high self-similarity; black squares on the off-diagonals correspond to regions of low cross-similarity. In the figure above, sky-blue lined-square is the value of cross-similarity between two sky-blue lined-rectangles. The same rule is applied to the red lined-square. The sky-blue lined square shows brighter color than the red-lined square which means that the red-lined rectangles have a less similar correlation than the sky-blue lined-rectangles.

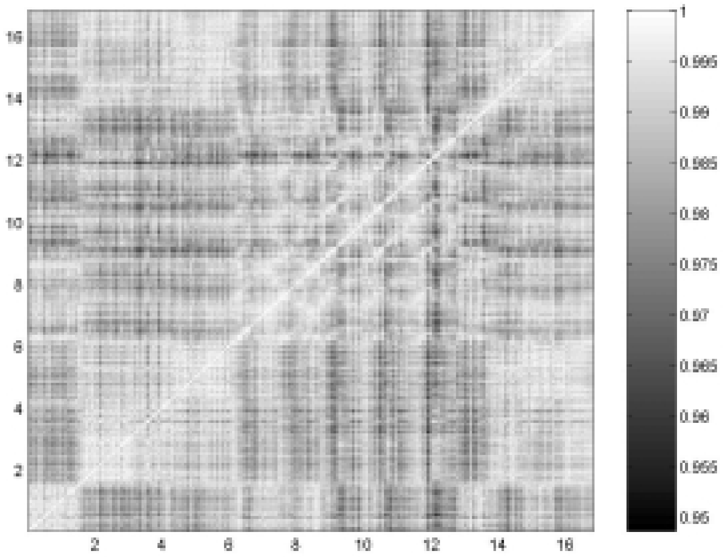
With windowing, you can see the matrix ***S*** smoothed. This windowed matrix ***S*** smoothes the radical change appeared in the original ***S***. It can be useful to track individual notes or other musical events because window rates (or the rates of feature vectors) usually are higher than typical musical events. Window rates cannot be set to same rates as musical events because it must be higher than that and track detailed feature vectors.

Novelty Score

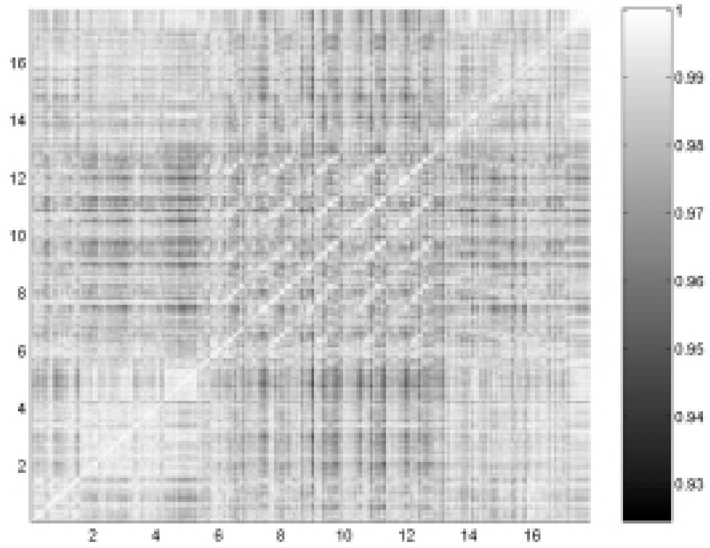
Based on the similarity matrix, we can find the instant when the notes changes by correlating the similarity matrix with some kernel functions. The width of the kernel (kernel size) directly affects the properties of the novelty measure. A small kernel detects novelty on a short time scale, such as beats or notes. Increasing the kernel size decreases the time resolution, and increases the length of novel events that can be detected. Larger kernels average over short-time novelty and detect longer structure, such as musical transitions like that between verse and chorus, key modulations, or symphonic movements. This method has no a-priori knowledge of musical phrases or pitch, but finding perceptually and musically significant points.

Here are the ‘similarity matrix’ and the ‘novelty score’ analyzed for the two music samples. The parameterization was done with Mel-frequency cepstral coefficient function with frame size 30. Both samples are about 16 seconds long and sampled at 11025hz, 16 bits. I could not include the score of this music, but by listening to them, the portion of the first 16 seconds consists of three musical sections: exposition (1-5 secs.), development (5-12 secs.), closing (12-16 secs.). It is interesting to note that the similarity matrices in the next page clearly shows this division in both cases 1. and 2. Also, it should be noted that these two pieces have the same musical ‘content’, but their ‘styles’ are in quite different format: one in classical style, the other is in rock style. Even though their styles are different, this similarity matrix only represents their musical content. I think that this is almost like a symbolic representation without indication of its stylistic information (especially, timbral information). Though this method may not be effective in classifying different genre of music in terms of its timbral/other musical content, it may be useful for grouping musics according to some overall structural features. Obviously, this is a very useful method to cluster the music with same content in different genres like crossover genre.

1. Mozart 25th Sysmphony orchestra version.



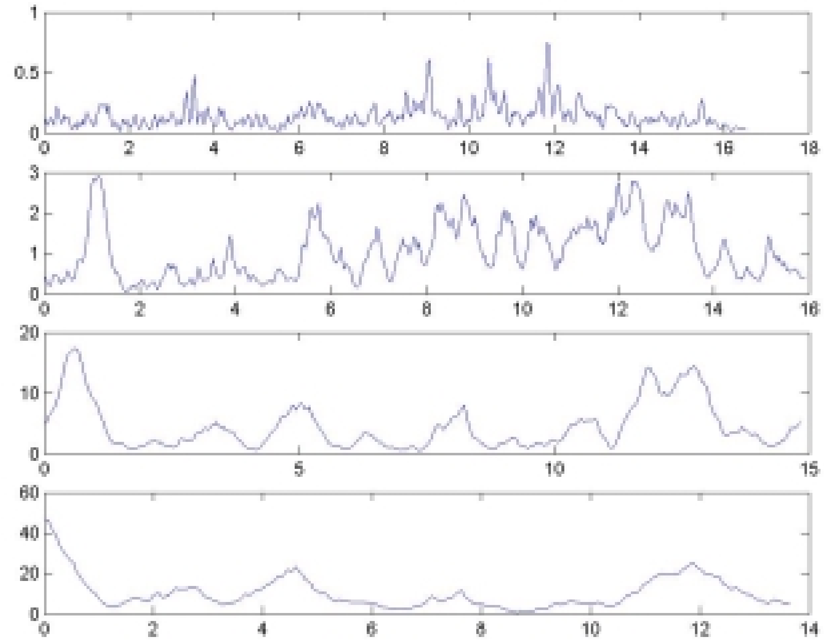
2. Its rock version



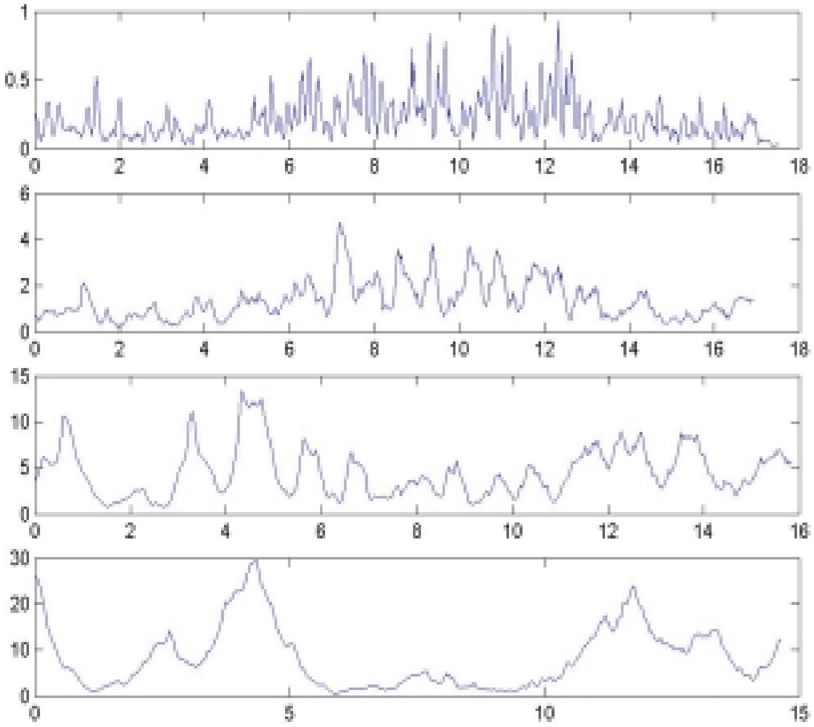
Novelty score

The following are the novelty score of the two versions of the music. In each firgure, the first one has a kernel size 10, the second 20, the third 60, the fourth 96. The figure whose kernel size is 96 shows the three high peaks which indicate the three sections (exposition, development, closing) of this music. As the kernel size gets less, the novelty score becomes in detail and try to detect the smaller novelty like notes, or beats. In this figure, the first three don't seem to show any useful novelty information due to non optimal kernel sizes. The kernel sizes for tracking appropriate novelty information should be optimized by some heuristical experiments.

1. Mozart 25th Symphony orchestra version



2. Its rock version


1
0.5
0
0
2
4
6
8
10
12
14
16
18
6
4
2
0
0
2
4
6
8
10
12
14
16
18
15
10
5
0
0
2
4
6
8
10
12
14
16
30
20
10
0
0
5
10
15