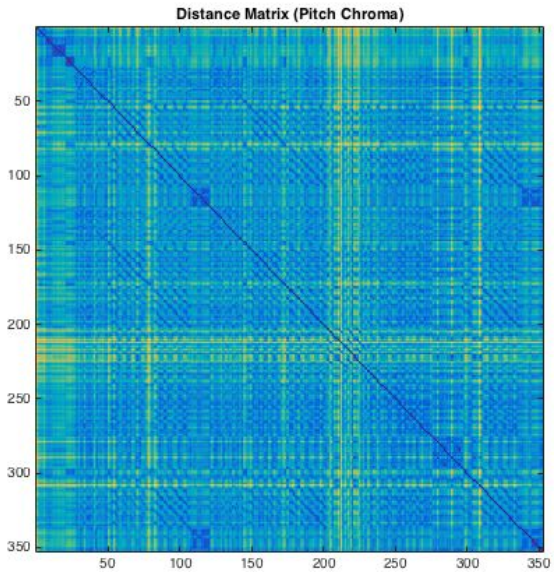
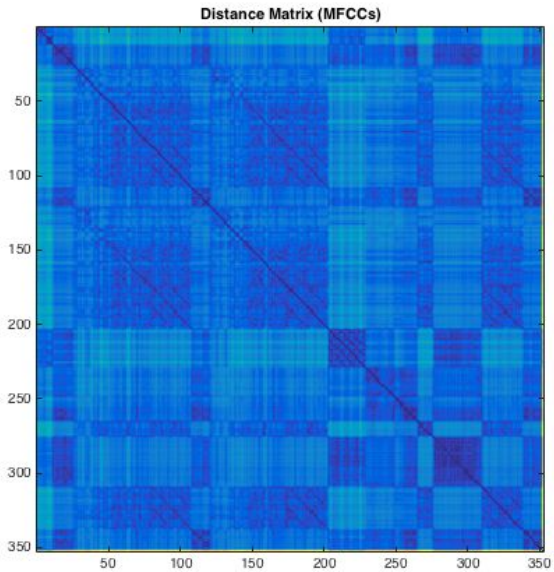
MIR: Assignment 4
Chris Latina and Liang Tang

Question 1: self-distance matrix





What are the similarities and dissimilarities between these two representations? (in terms of geometric shapes and intensity of the colors)

In pitch chroma, the borders of each segment are most noticeable, whereas with MFCCs, the structural segmentation is noticeable by contigious differentiation. Texturally, the blocked patterns visible in the distance matrix for MFCCs are much clearer. It is easy to see the features of one within the other (ie the borders in MFCCs are present but less pronounced, and the blocks in PC).

How would you interpret the horizontal and vertical stripes in the figures?

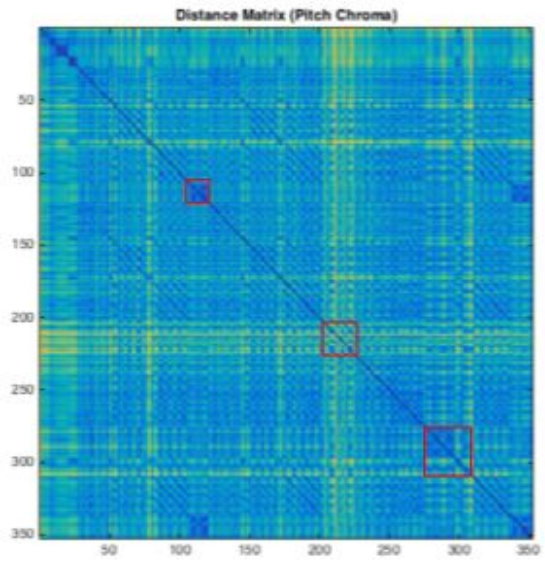
The horizontal and vertical stripes are the portions of change between structural parts, ie intro to verse to chorus to bridge. In PC, the stripes are pronounced, which means low similarity (like the noisey silence at the end of MFCCs). The stripes represent changes in key.

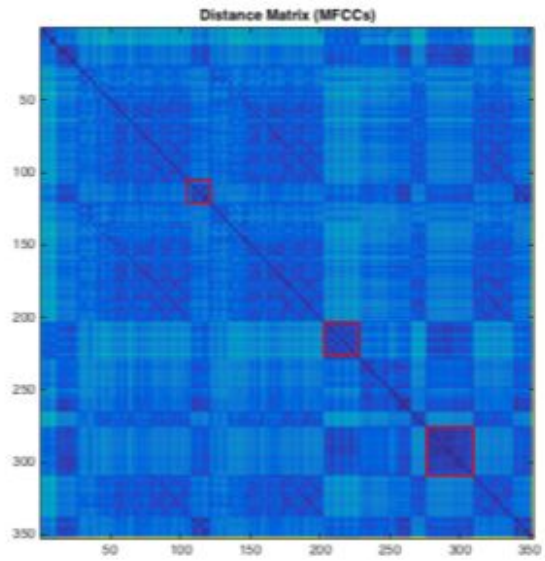
For both of the SDMs, you should be able to observe some stripes that are parallel to the diagonal. How would you interpret these parallel stripes?

These parallel stripes are self similar passages (in regards to both melody/harmony and instrumentation).

For both of the SDMs, you should be able to observe "blocks" along the diagonal. Annotate the time stamps of 2 ~ 3 most distinctive blocks, label them on your figures, listen to their corresponding audio segments and discuss their relationships with the audio content.

There is a distinct block shared between PC and MFCCs around [105,105]. Other significant portions are at [200,200] and [275,275].

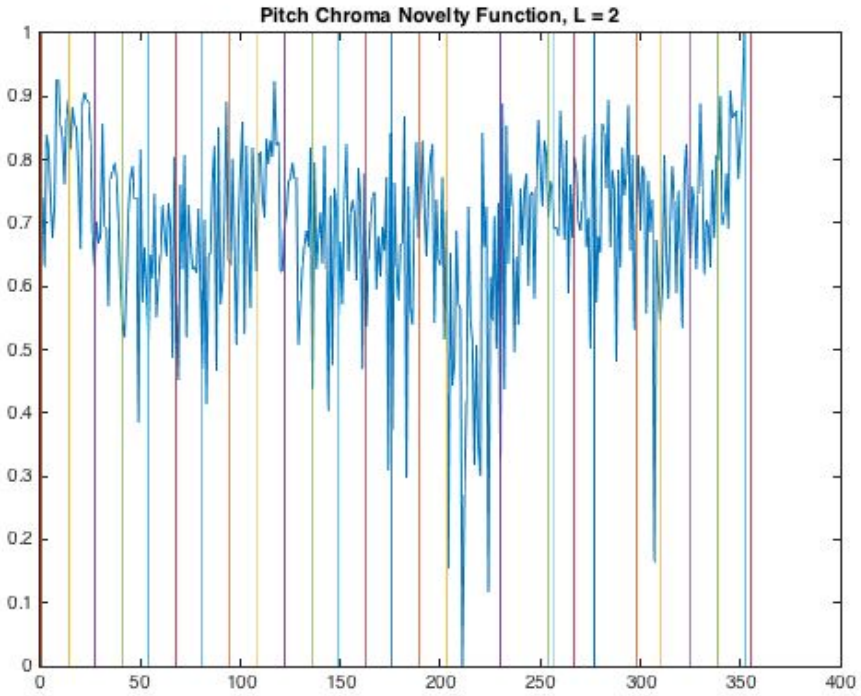


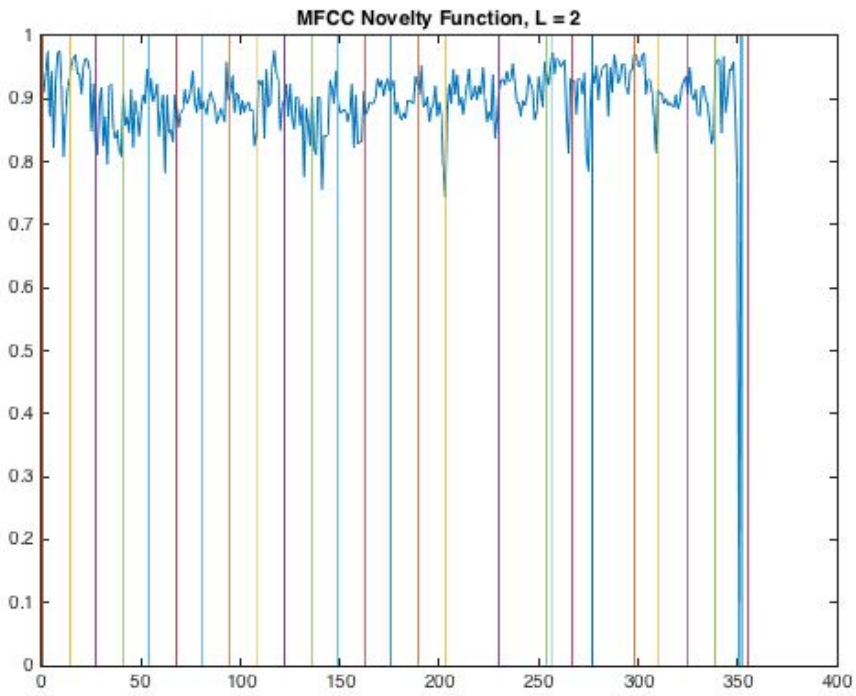


Around 105 seconds there is a bridge with a guitar solo (the drums drop out) that becomes apparent in both PC and MFCC. A similar repetitive guitar lick occurs around 200 seconds. Between 275 and 300s there is a breakdown after a quick vocal passage.

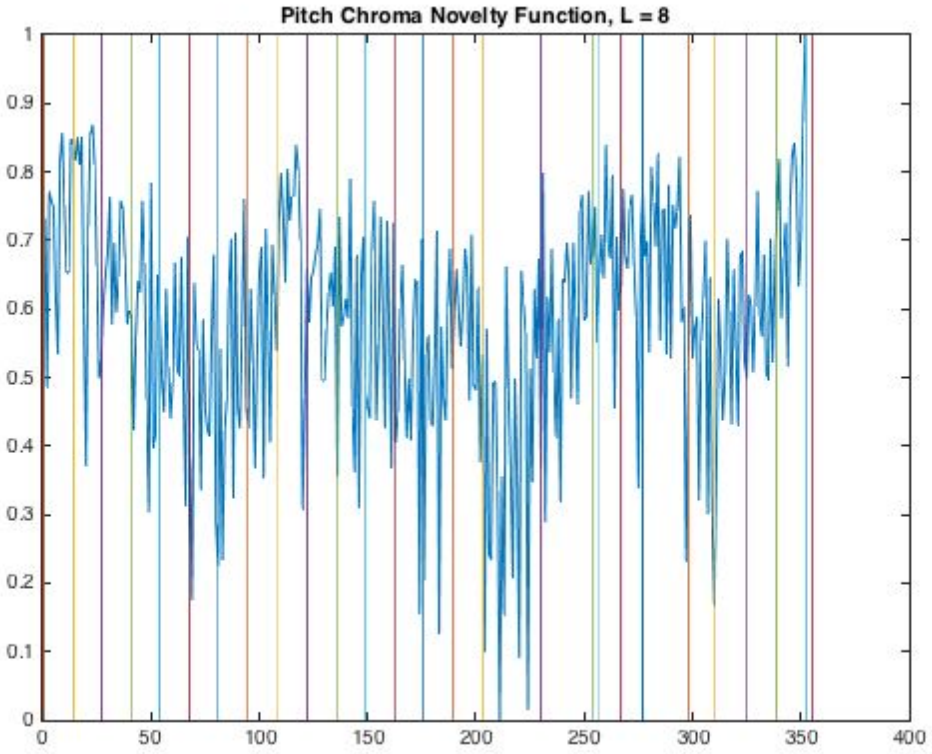
Question 2: Checkerboard Kernel Audio Novelty Function

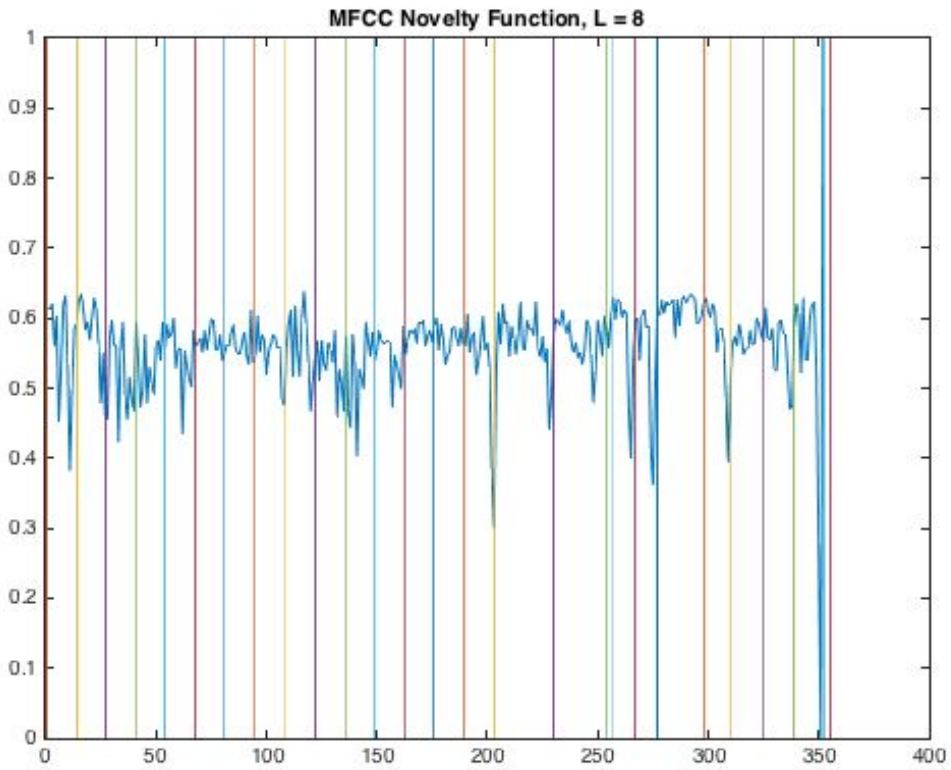
L = 2:



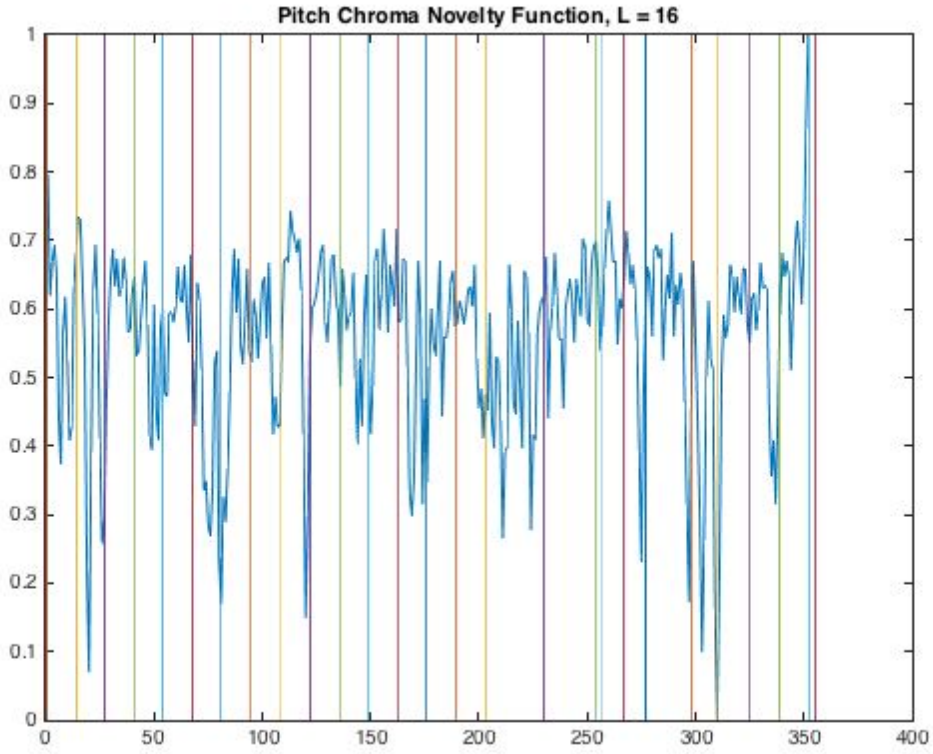


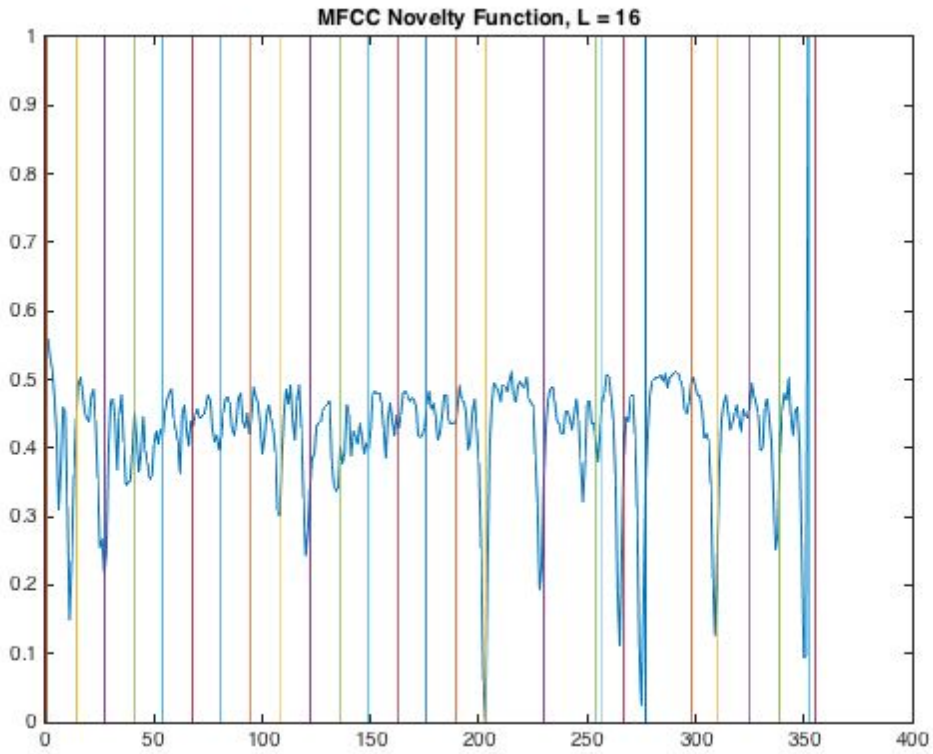
L = 8:


Pitch Chroma Novelty Function, L = 8
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
400


MFCC Novelty Function, L = 8
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
400

L = 16:


Pitch Chroma Novelty Function, L = 16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
400


MFCC Novelty Function, L = 16
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
400

Which L is returning a better result? What is the relationship between the novelty and the ground truth?

The larger kernal, L = 16 shows the clearest result. You can see the drop towards 0 then back up towards 1 in the novelty functions aligning with the ground truth.

Based on the extracted novelty functions, how would you interpret the Checkerboard kernel?

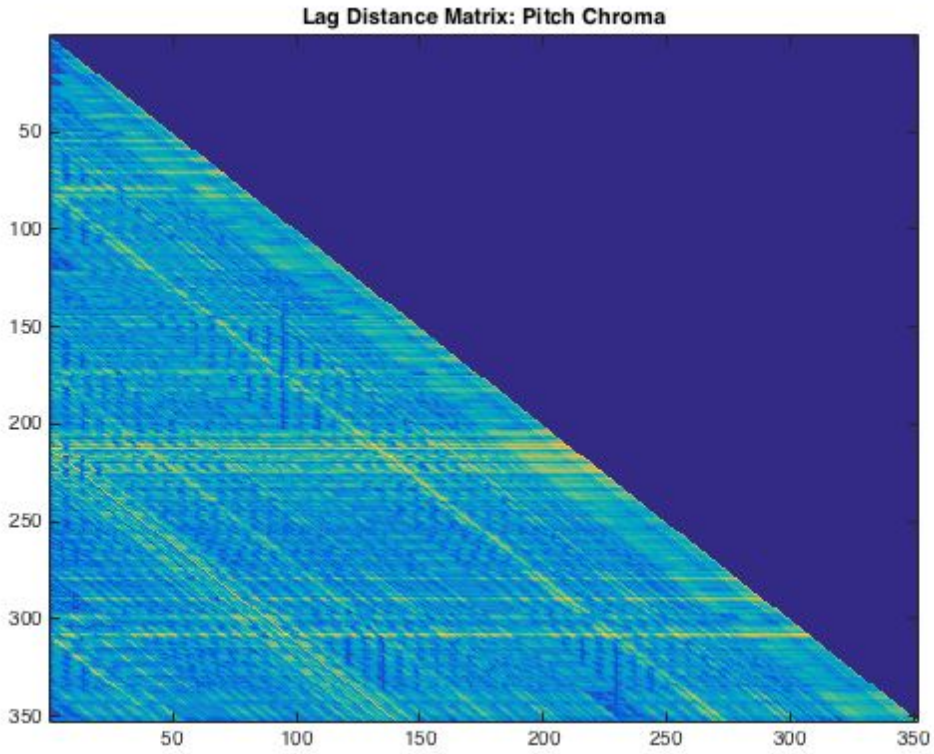
We used a gaussian version of the checkerboard. This kernel is similar to a two dimensional edge detector. It accentuates the normalized novelty function so major portions are easily readable when normalized. The gassuain portion also smooths out the noise. Larger filter values obviously smooth the novelty function significantly more.

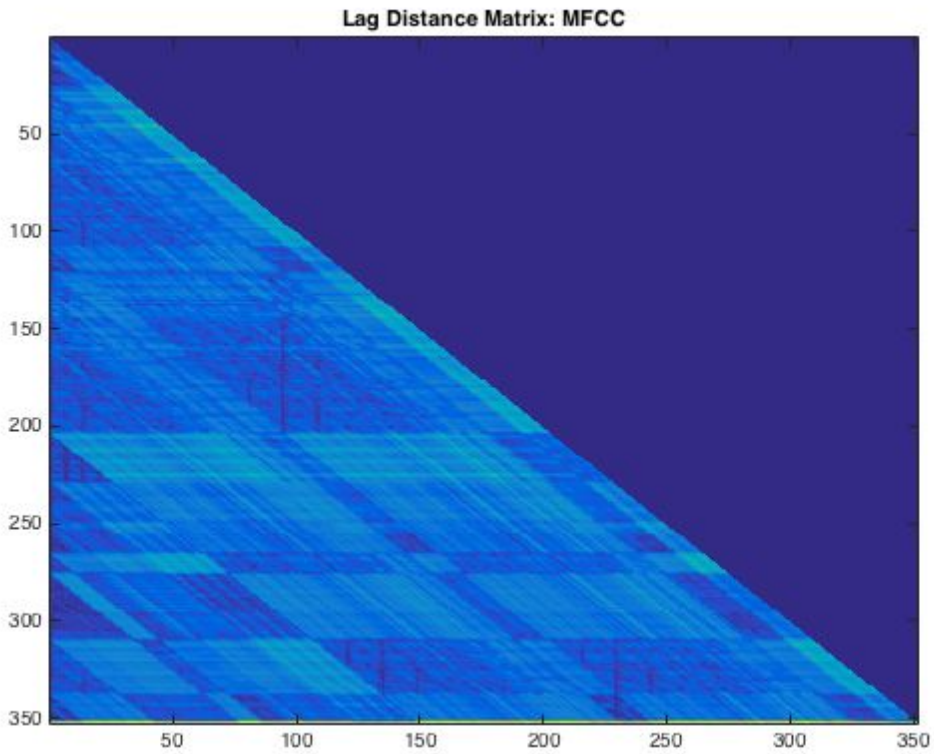
In this audio example, which feature (chroma and mfcc) provides a more meaningful novelty function for the audio segmentation task? Why? (Justify your answer based on your observation through listening)

The MFCCs are clearer both visually and from listening. Certarin passages based on pitch chroma are especially easy to differentiate. Overall, similarities between MFCC and Pitch Chroma indicates a bridge or solo but structurally MFCCs are more appropriate.

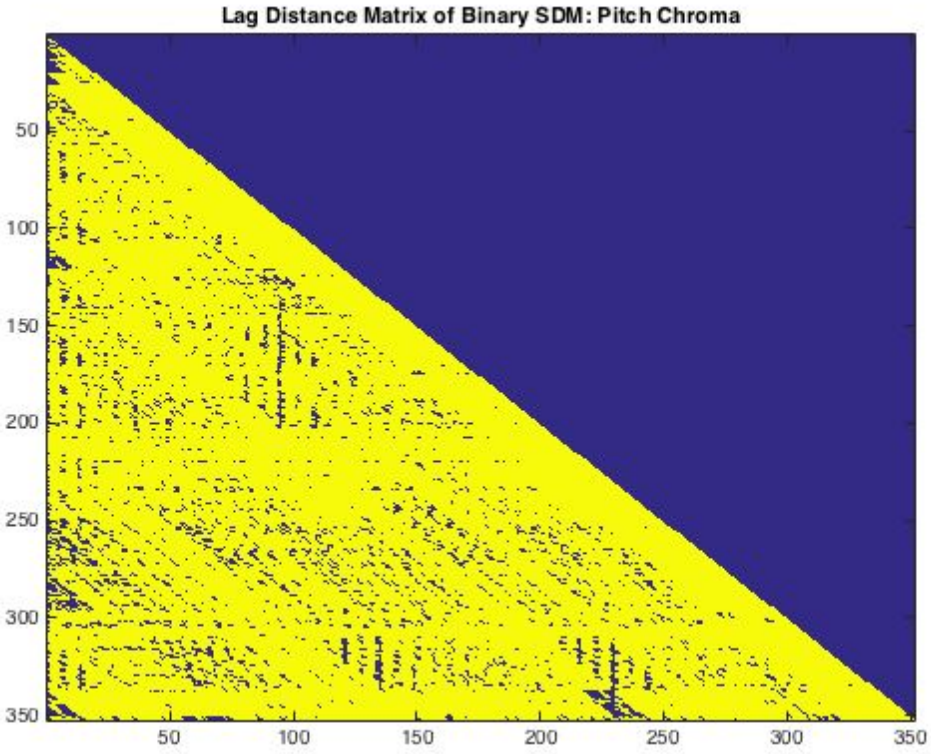
Question 3: Lag-Distance Matrix & repetition detection

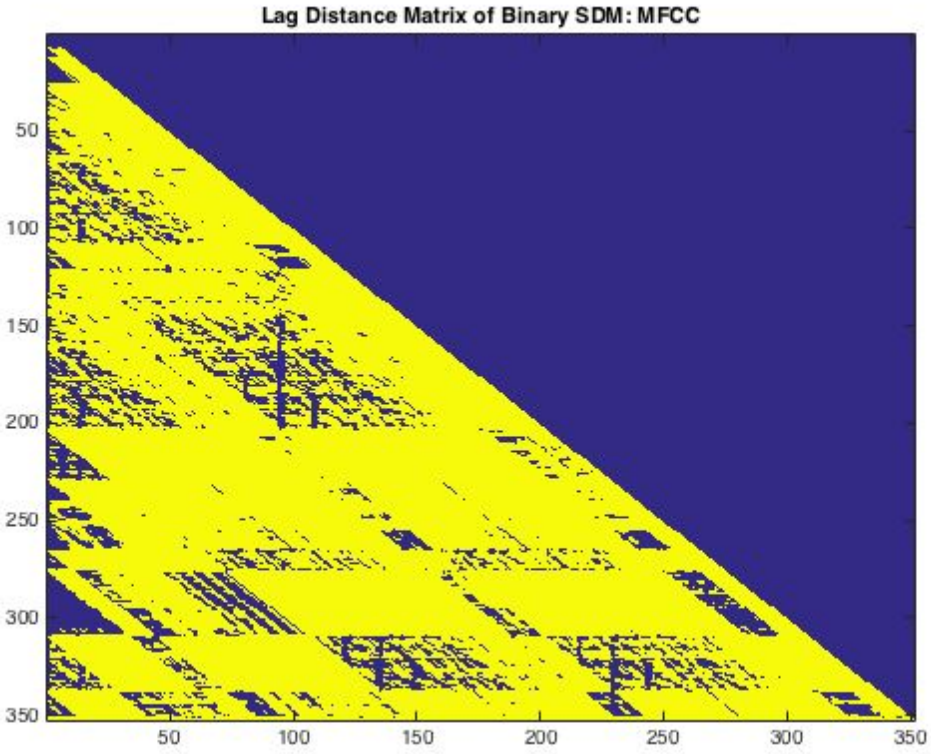
Time Lag Matrix for both SDMs



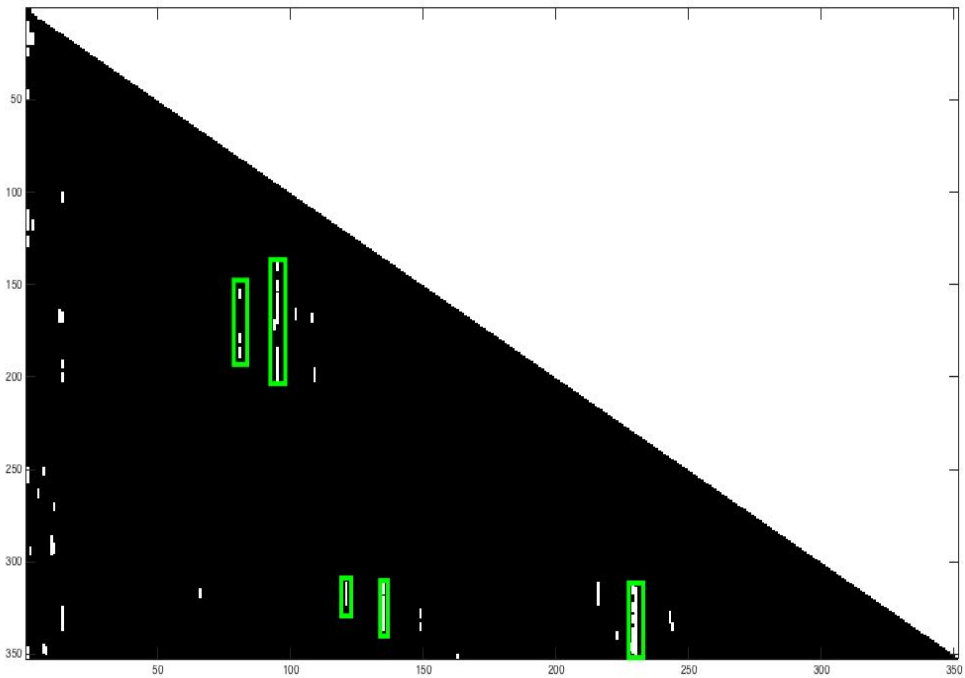


Binary Lag Distance Matrix

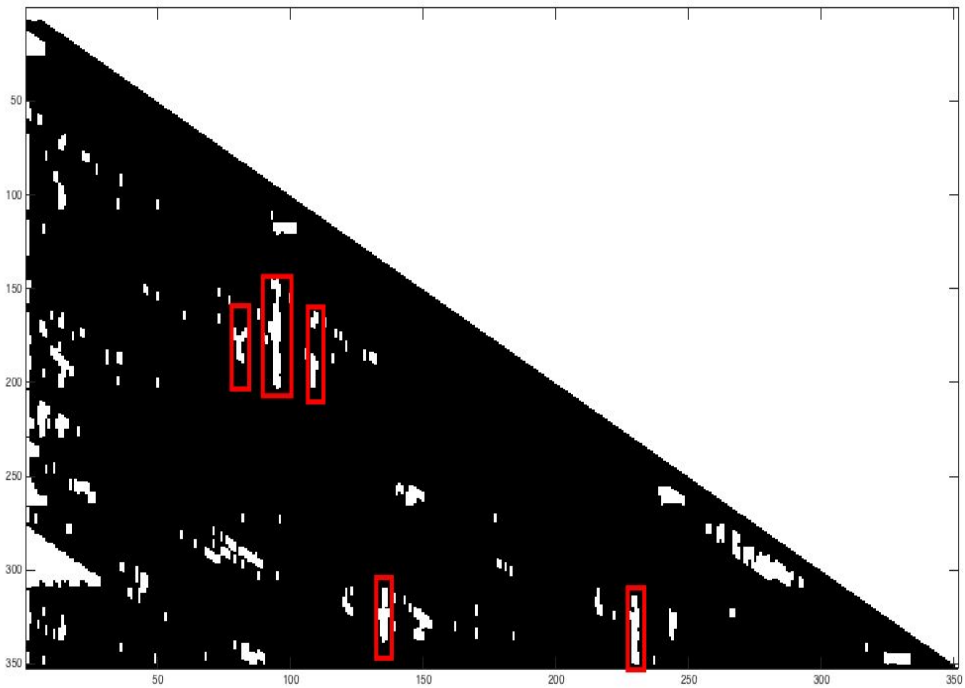




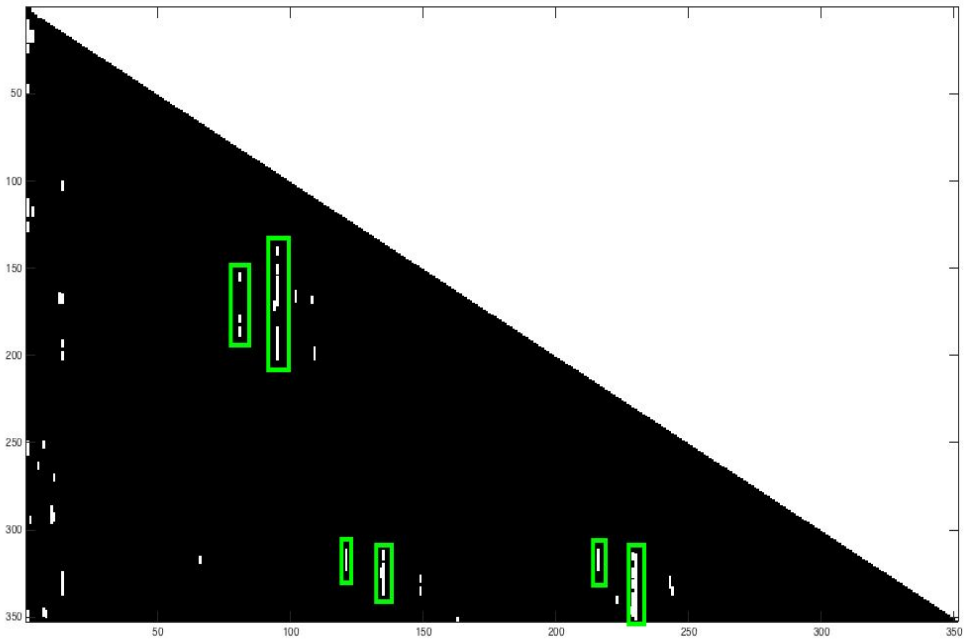
For both features, plot your results and mark the start and end time of these lines.
Pitch Chroma (L = 16)



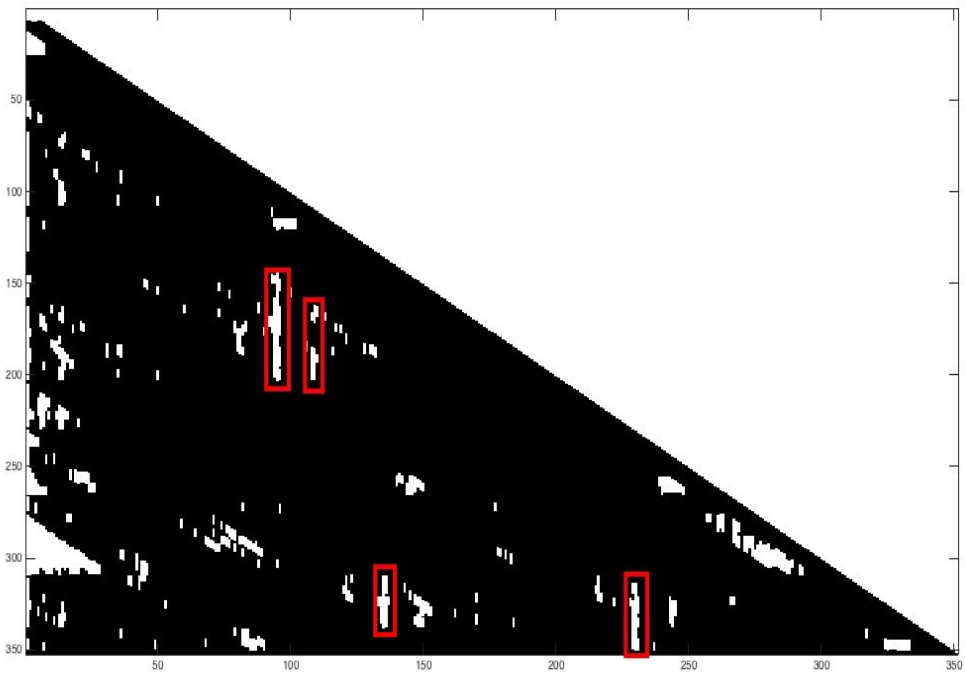
MFCC (L = 16)



Pitch Chroma (L = 8)



MFCC (L = 8)



Listen to the audio of your extracted regions. What do they have in common? Also, discuss the discrepancy between the ground truth and the extracted regions?

The pitch chroma appears to be slightly less noisy than the MFCCs. Overall however, the segmentation results were pretty poor, so the descrepancies are very high. The portions we

mentioned earlier, between 75 and 100s and between 200 and 250s are clearly visible in the eroded versions. The similarity between these part is this sort of polka sounding guitar lick repeated between verses.

Bonus question: If we want to capture the hierarchical structure of a song (for example, a larger part might contain many smaller subparts), how can we do it? (This is an open-ended question, therefore, the more insights you can provide the better)

The heirarchical approach may look at the lengths of repeated sequences and sub-sequences that build up to that length. The song we used was a poor example becasue it was timbrally simplistic, noisy, and repetitive throughout. Taking some sort of recursive approach to approximating and segregating the introduction, outro, and some combination of verse and chorus (or simply just labelling A vs B to be more generic) may work. Then within the subsections, sublabels like A1-A4 and B1-B4 could be used. This would make it relatively easy to assign a “chorus/verse” or “call/response” structure if it exists.