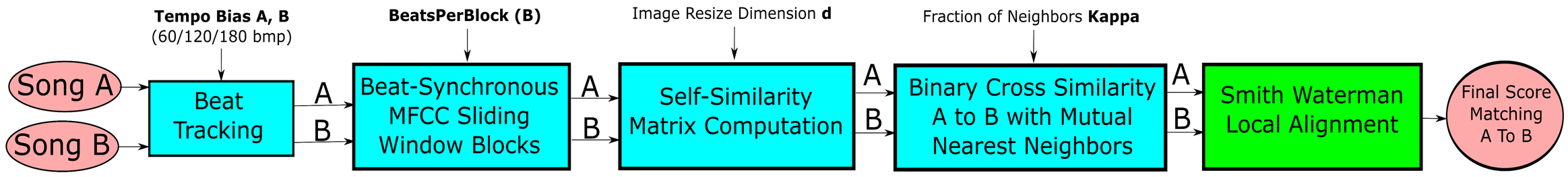
Cover Song Identification with Timbral Shape Sequences

Chris Tralie. chris.tralie@gmail.com
Ph.D. Candidate, NSF Graduate Fellow, Electrical and Computer Engineering, Duke University
Paul Bendich, bendich@math.duke.edu
Assistant Research Professor, Mathematics, Duke University

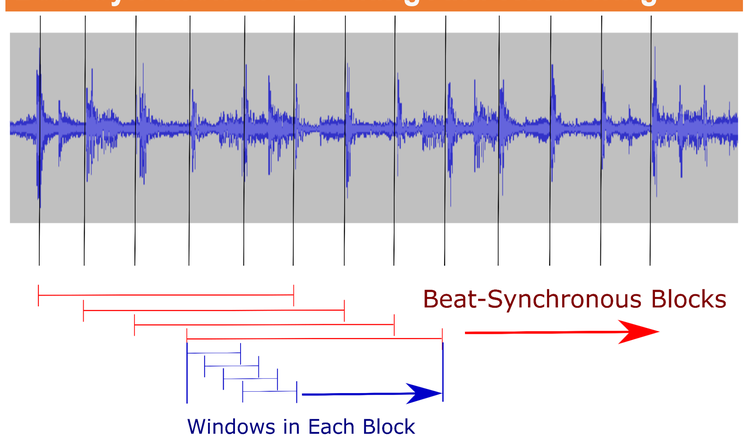
Pipeline



Abstract

We introduce a novel low level feature for identifying cover songs which quantifies the relative changes in the smoothed frequency spectrum of a song. Our key insight is that a sliding window representation of a chunk of audio can be viewed as a time-ordered point cloud in high dimensions. For corresponding chunks of audio between different versions of the same song, these point clouds are approximately rotated, translated, and scaled copies of each other. If we treat MFCC embeddings as point clouds and cast the problem as a relative shape sequence, we are able to correctly identify 42/80 cover songs in the ``Covers 80'' dataset. By contrast, all other work to date on cover songs exclusively relies on matching note sequences from Chroma derived features.

Beat-Synchronous Blocking And Windowing



- **B** beats per block. Take all such blocks in the song
- Take MFCC sliding window features to summarize each block
- MFCC Window size average beat interval

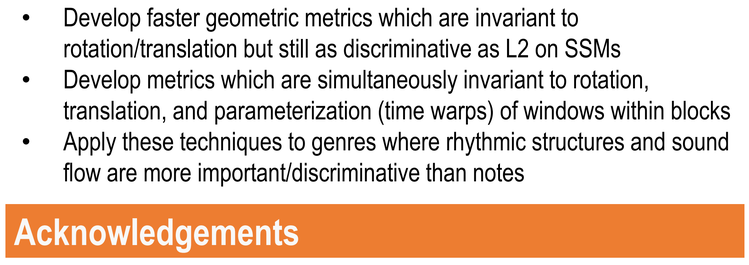
LoopDitty: Music As A Shape

http://www.loopditty.net

Interactive App for Viewing PCA of shapes synchronized to music

- Each MFCC window is a point in 20-dimensional space
- Longer MFCC window size helps smooth path
- Similar relative shapes for cover songs

"**Addicted To Love**" hook, by Robert Palmer



PCA Window Size 0.05 sec

PCA Window Size 0.5 sec

References/Code

[1] Daniel PW Ellis. The "covers80" cover song data set. URL: http://labrosa.ee.columbia.edu/projects/coversongs/covers80, 2007.
[2] Joan Serra, Emilia Gomez, Perfecto Herrera, and Xavier Serra. Chroma binary similarity and local alignment applied to cover song identification. Audio, Speech, and Language Processing, IEEE Transactions on, 16(6):1138–1151, 2008.
[3] Jose A Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of omputational Mathematics, pages 1–40, 2013.
[4] Suman Ravuri and Daniel PW Ellis. Cover song detection: from high scores to general classification. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pages 65–68. IEEE, 2010

Please see our paper for a more complete list of references

Code

https://github.com/ctralie/PublicationsCode/tree/master/ISMIR2015_CoverSongsShape

Future Work

- Develop faster geometric metrics which are invariant to rotation/translation but still as discriminative as L2 on SSMs
- Develop metrics which are simultaneously invariant to rotation, translation, and parameterization (time warps) of windows within blocks
- Apply these techniques to genres where rhythmic structures and sound flow are more important/discriminative than notes

Acknowledgements

Chris Tralie was supported under NSF-DMS 1045133 and an NSF Graduate Fellowship. Paul Bendich was supported by NSF 144749. John Harer and Guillermo Sapiro are thanked for valuable feedback. The authors would also like to thank the Information Initative at Duke (iiD) for stimulating this collaboration.

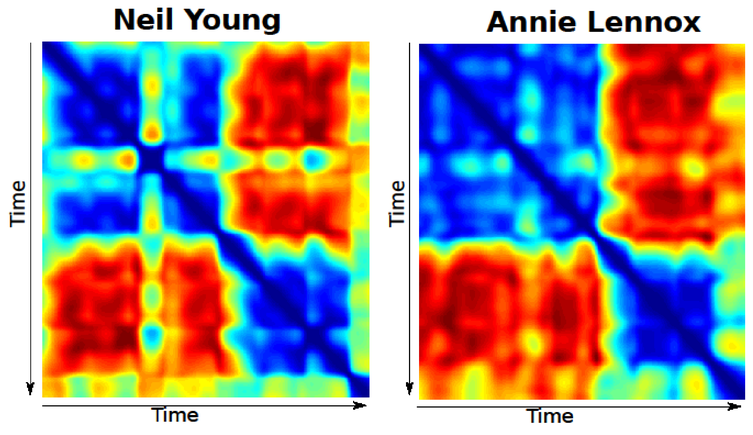
Self-Similarity Matrices

$$SSM_{ij}^{l} = ||X_l[i] - X_l[j]||_2$$

- Computed for each block **l** of **B** contiguous beats for each song
- Invariant to rotation/translation
- Point-center and sphere-normalize windows within each block to help make invariant to scale

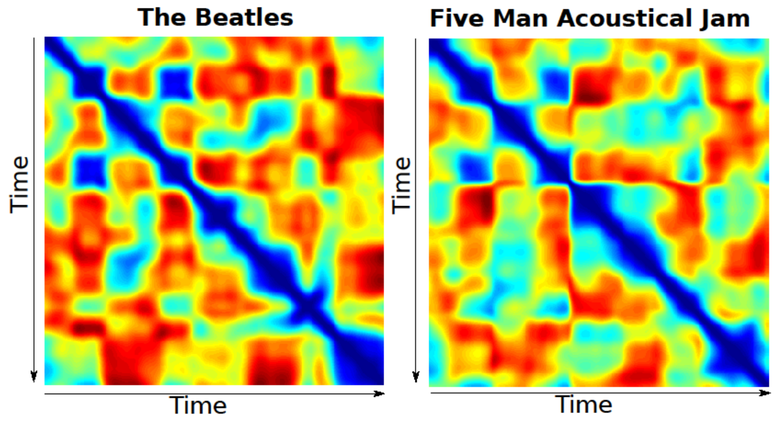
"Don't Let It Bring You Down"

Different gender singer, different instruments, different vocal/instrument balance



"We Can Work It Out"

Different band, live versus studio



Cross-Similarity Matrices

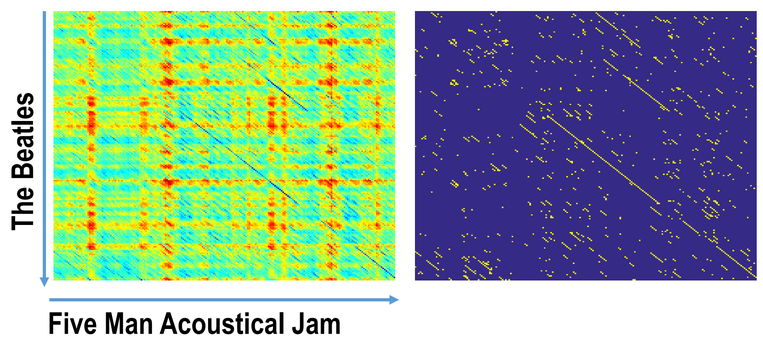
$$CSM_{ij} = ||SSMA^i - SSMB^j||_2$$

- Resize all self-similarity images to common dimension **d x d**
- Comparing SSM from each beat-synchronous block in song A to SSM from each block in song B.
- Long diagonals indicate good matches
- Converting to binary matrix makes more robust. A pixel (i, j) is one if it is within the **kappa** fraction of nearest neighbors of block i in A to all blocks in B and likewise for block j in B to blocks in A
- Exploit Matlab's fast matrix multiplication to compare all images simultaneously

```
CSM = bsxfun(@plus, dot(Ds1,Ds1,2),dot(Ds2,Ds2,2)') …
                - 2Ds1*Ds2';
```

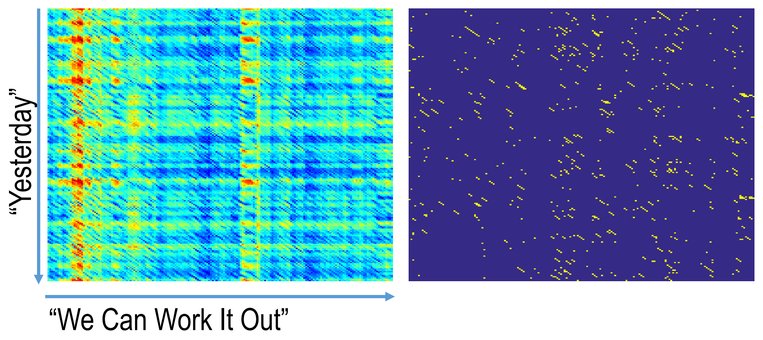
True Cover: "We Can Work It Out"

Long diagonals visible for many well-matching blocks in sequence



False Cover: "We Can Work It Out" vs"Yesterday"

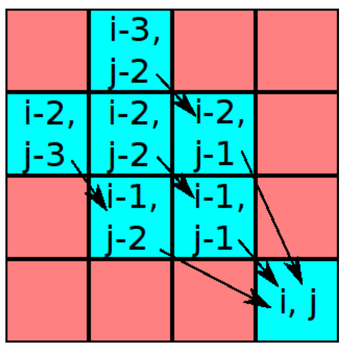
No long diagonals; binary matrix is noisy



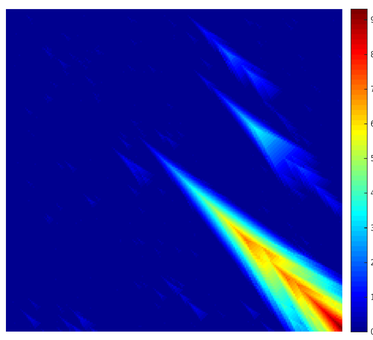
Smith Waterman Alignment with Diagonal Constraints

- Allow gaps as in Smith Waterman for extra beats, but promote near-diagonal paths
- Score of 1 for matching SSMs
- Affine gap penalty
  `-0.5 - 0.7(g-1)`
  For gap of length **g**
- Similar to approach in [2]

Warping Paths Considered

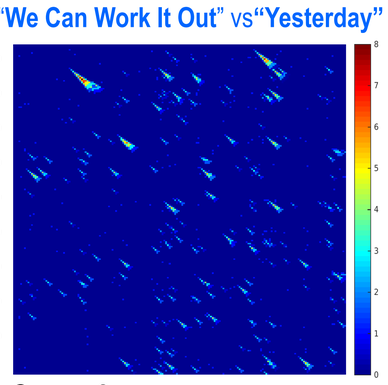


True Cover
"We Can Work It Out"



Score: 93.1

False Cover
"We Can Work It Out" vs"Yesterday"



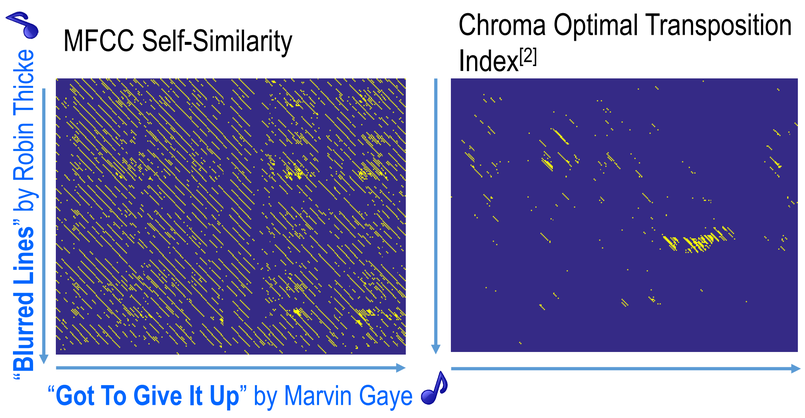
Score: 8

Results: Covers 80

- 80 pop song benchmark [1]. Best results reported in [4]
- Most correct top rank with our scheme: **44/80**
- Results below for 50 x 50 SSMs with 10 windows per beat (mean/median rank of correct song shown in parentheses)

Kappa	B = 6	B = 8	B = 10	B = 12	B = 14	B = 16	B = 18	B = 20	B = 22	B = 24
0.05	24 (8.5 / 19.6)	29 (6 / 17.3)	30 (5.5 / 16.0)	34 (3 / 14.0)	39 (2 / 12.45)	42 (1 / 13.5)	42 (1 / 12.1)	41 (1 / 11.7)	41 (1 / 11.3)	**44 (1 / 1.6)**
0.1	27 (11 / 21.1)	32 (6 / 18.0)	39 (2 / 11.8)	40 (1.5 / 13.0)	39 (2 / 12.2)	43 (1 / 11.5)	43 (1 / 12.6)	42 (1 / 13.2)	42 (1 / 14.1)	41 (1 / 3.7)
0.15	27 (7 / 17.5)	34 (2 / 14.1)	39 (2 / 14.2)	42 (1 / 12.5)	42 (1 / 13.5)	**44 (1 / 12.5)**	40 (1.5 / 12.9)	42 (1 / 13.2)	43 (1 / 12.4)	**44 (1 /13.2)**
0.2	26 (6 / 17.8)	29 (5.5 / 16.7)	34 (2.5 / 15.9)	38 (2.5 / 15.2)	42 (1 / 14.9)	41 (1 / 13.9)	40 (1.5 / 13.2)	41 (1 / 13.5)	40 (1.5 / 13.9)	40 (1.5/13.8)

Results: "Blurred Lines" Cross-Similarity Matrices



- Every 4 beats rhythmic pattern repeats itself (many diagonals)
- Note sequences are different, so traditional chroma-based approaches fail to recognize similarities