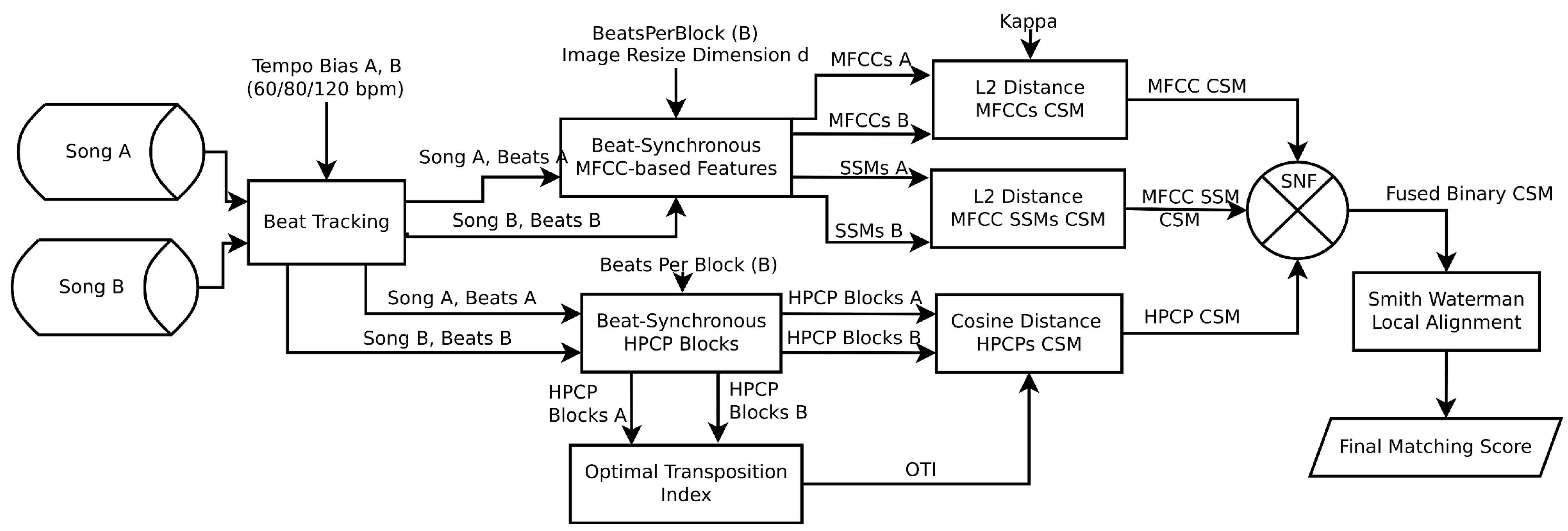
Early MFCC And HPCP Fusion for Robust Cover Song Identification

Christopher J. Tralie. ctralie@alumni.princeton.edu
Department of Electrical and Computer Engineering, Duke University

Abstract

While most schemes for automatic cover song identification have focused on note-based features such as HPCP and chord profiles, a few recent papers surprisingly showed that local self-similarities of MFCC-based features also have classification power for this task. Since MFCC and HPCP capture complementary information, we design an unsupervised algorithm that combines normalized, beat-synchronous blocks of these features using cross-similarity fusion before attempting to locally align a pair of songs. As an added bonus, our scheme naturally incorporates structural information in each song to fill in alignment gaps where both feature sets fail. We show a striking jump in performance over MFCC and HPCP alone, achieving a state of the art mean reciprocal rank of 0.87 on the Covers80 dataset. We also introduce a new medium-sized hand designed benchmark dataset called ``Covers 1000," which consists of 395 cliques of cover songs for a total of 1000 songs, and we show that our algorithm achieves an MRR of 0.9 on this dataset for the first correctly identified song in a clique. We provide the precomputed HPCP and MFCC features, as well as beat intervals, for all songs in the Covers 1000 dataset for use in further research.

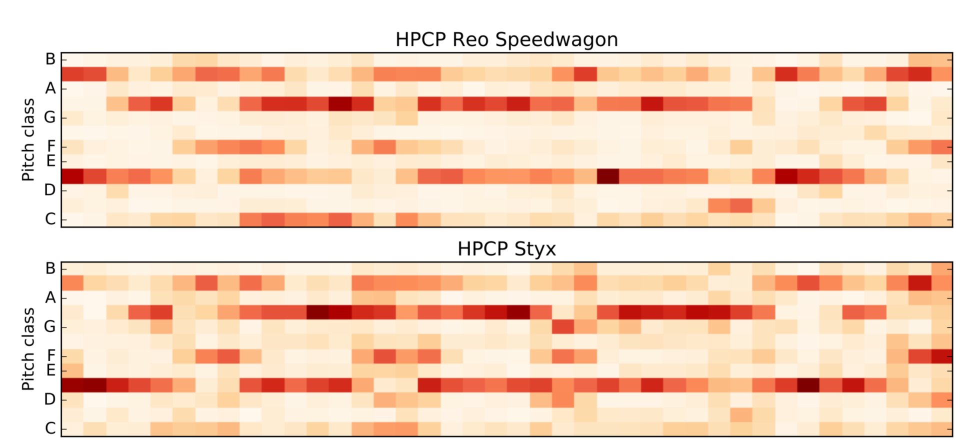
Full Pipeline



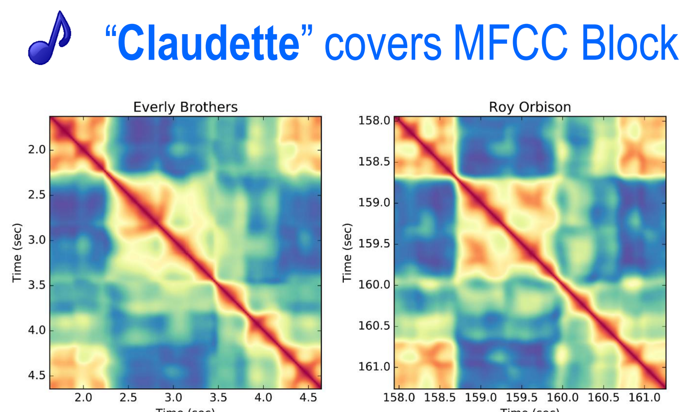
Beat-Synchronous Blocked Features

- Motivation: Chroma-based features good choice but in certain scenarios don't work (e.g. hip hop, drumming)
- We take blocks of 3 types of features synchronized to beats
  1) HPCP Features, 2 windows per beat, OTI for matching
  2) MFCC Features, long window size 0.5 seconds
  3) Local SSMs of MFCC blocks in #2 (as in [3])
- All blocks computed in 20 beat intervals
- Blocks resized to a common number of frames before comparison
- hopSize 23ms for all features

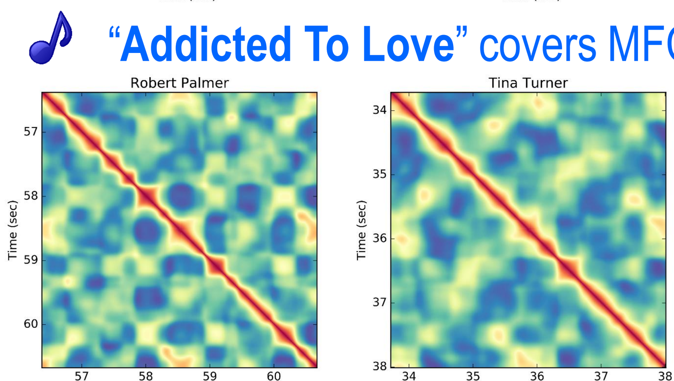
"Grand Illusion" covers HPCP Blocks



"Claudette" covers MFCC Block SSMs



"Addicted To Love" covers MFCC Block SSMs



Improving Beat-Synchronous Cross Similarity Matrices with Early Similarity Network Fusion (SNF)

$$W(i,j) = e^{-\rho^2(i,j)/2(\sigma_{ij})^2}$$

1) Create similarity kernel given distance matrix from a feature type

$$P(i,j) = \begin{cases} \frac{1}{2}\frac{W(i,j)}{\sum_{k\neq i} W(i,k)} & j \neq i \\ 1/2 & \text{otherwise} \end{cases}$$

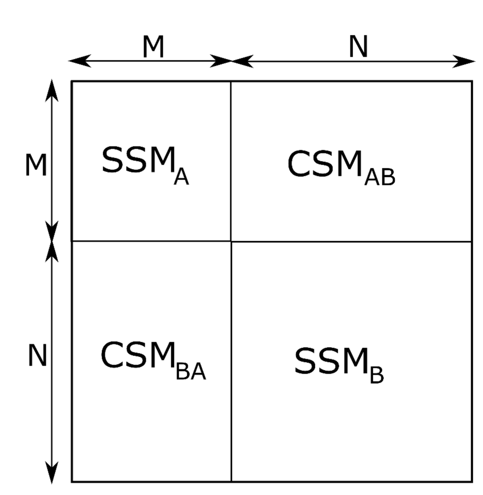
2) Compute self-similarity regularized Markov transition probabilities

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \in N(i)} W(i,k)} & j \in N(i) \\ 0 & \text{otherwise} \end{cases}$$

3) Compute neighborhood truncated Markov transition probabilities

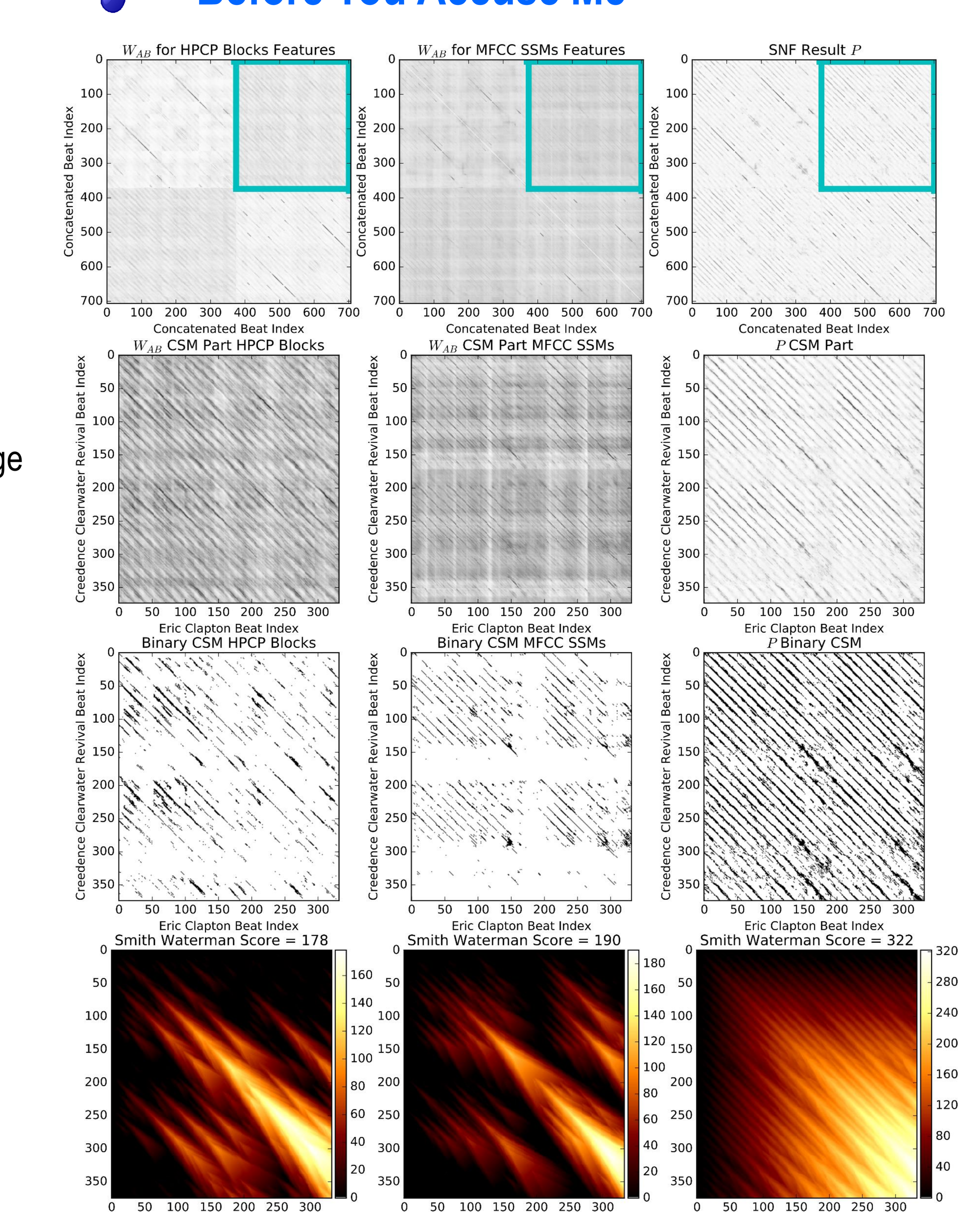
$$P_{t+1}^f = S^f \left(\frac{\sum_{v\neq f} P_t^v}{m-1} \right) (S^f)^T$$

4) Compute random walk probabilities using average probabilities from other features and truncated transition matrix from feature f



- Goal: given different cross-similarity measures, for a pair of songs, fuse into an improved cross-similarity measure
- Technique developed in [5, 6], used by Chen et. al. [1] for cover songs at the level of song similarities, which we call "late fusion" (also similar to [4])
- We focus on local block similarities measured by different features
- For each blocked feature, create an (N+M) x (N+M) "parent SSM" (left) from concatenating song A (M blocks) to song B (N blocks), which captures both self-similarity and cross-similarity, then run algorithm. Example shown on the right

"Before You Accuse Me"



- Similarity kernels are normalized differently for self-similarity and cross-similarity parts
- Apply Smith Waterman on nearest neighbor cross-similarity matrix to score alignment for individual features and fused features
- Can still apply late fusion to similarities from all features (W = 1/Score) **and** results from early fusion to boost classification performance

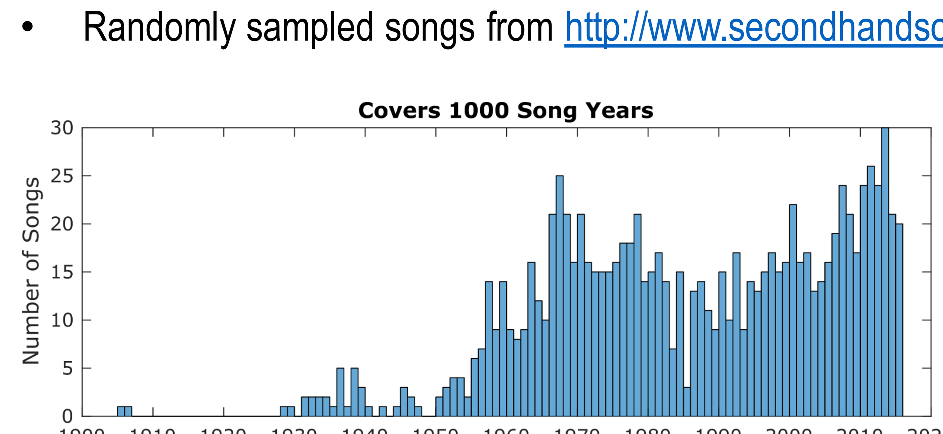
Results: Covers 80

- '80s/'90s pop song benchmark [2]. 80 cliques, 2 songs per clique

	MR	MRR	Top-01	Top-10	../80
MFCCs	29.7	0.538	79	97	42/80
SSMs	15.1	0.615	91	111	48/80
HPCPs	18.2	0.673	102	119	53/80
Late SSMs/MFCCs	14.0	0.7	107	125	55/80
Late All	8.63	0.824	127	141	64/80
Early	7.76	0.846	131	143	68/80
Early + Late	7.59	0.873	136	144	69/80
[1]	?	0.625	?	114	?

Covers 1000 Dataset / Results

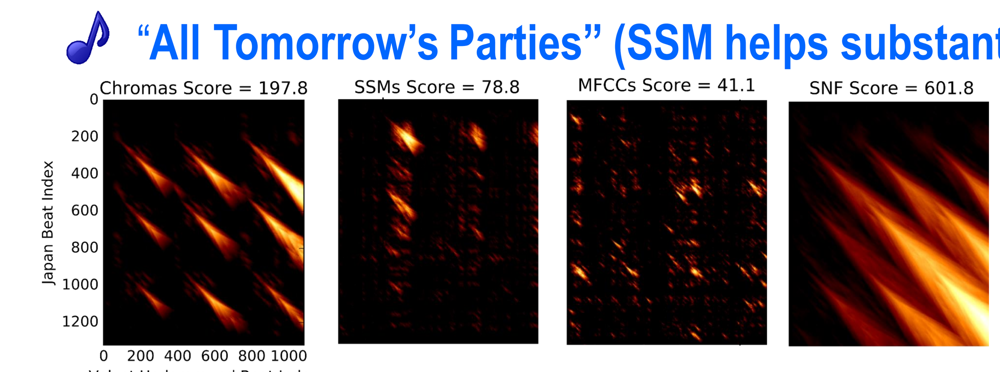
- New dataset curated for the community, features available at http://www.covers1000.net
- Hand designed dataset, 1000 songs (395 cliques total)
- Randomly sampled songs from http://www.secondhandsongs.com



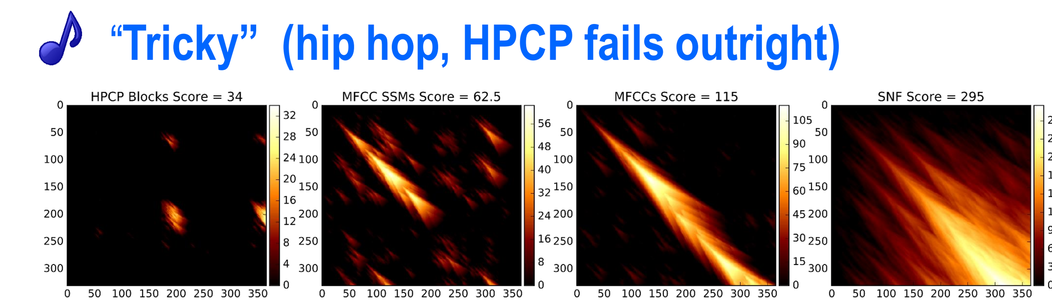
	MR	MRR	Top-01	Top-10
MFCCs	83.3	0.618	583	679
SSMs	72.5	0.623	581	698
HPCPs	44.4	0.757	727	809
Late	19.8	0.875	855	931
Early	22.5	0.829	798	884
Early + Late	14	0.904	884	950

Additional Examples

"All Tomorrow's Parties" (SSM helps substantially)



"Tricky" (hip hop, HPCP fails outright)



Frank Zappa: "The Black Page"

- 8 versions of song with no harmonic content
- Compare to all songs in covers1000 and 8 versions
- Mean Average Precisions:
  - HPCP: 0.014
  - MFCC SSMs: 0.905
  - Raw MFCC: 0.97
  - Early SNF: **0.98**

MIREX

- Use a single tempo level from state of the art beat tracking [7] to reduce computation by a factor of 9.
- Take a slight (but not severe) performance hit

Covers80

	MR	MRR	Top1	Top10	../80
MFCCs	31.5125	0.531269	79	93	41/80
SSMs	24.15	0.578772	87	102	45/80
Chromas	24.9	0.616607	93	107	46/80
SNF	18.3625	0.745443	115	130	59/80
Late	16.1938	0.762246	117	129	61/80

Covers1000

	MR	MRR	Top1	Top10
MFCCs	112.241	0.549329	523	633
SSMs	92.68	0.572648	533	694
Chromas	64.195	0.686407	653	793
SNF	54.588	0.751717	720	835
Late	38.262	0.82227	803	878

References/Code

[1] Ning Chen,Wei Li, and Haidong Xiao. Fusing similarity functions for cover song identification. Multimedia Tools and Applications, pages 1–24, 2017.
[2] Daniel PW Ellis. The "covers80" cover song data set. URL: http://labrosa.ee.columbia.edu/projects/coversongs/covers80, 2007.
[3] Christopher J Tralie and Paul Bendich. Cover song identification with timbral shape sequences. In 16th International Society for Music Information Retrieval (ISMIR), pages 38–44, 2015.
[4] Joan Serr`a, Massimiliano Zanin, Perfecto Herrera, and Xavier Serra. Characterization and exploitation of community structure in cover song networks. Pattern Recognition Letters, 33(9):1032–1041, 2012.
[5] BoWang, Jiayan Jiang,WeiWang, Zhi-Hua Zhou, and Zhuowen Tu. Unsupervised metric fusion by cross diffusion. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2997–3004.
[6] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. Nature methods, 11(3):333–337, 2014.
[7] Sebastian B¨ock, Filip Korzeniowski, Jan Schl¨uter, Florian Krebs, and Gerhard Widmer. madmom: a new python audio and music signal processing library. In Proceedings of the 2016 ACM on Multimedia Conference, pages 1174–1178. ACM, 2016. IEEE, 2012.

Please see our paper for a more complete list of references

Code

https://github.com/ctralie/GeometricCoverSongs

Dataset / Live Demo

http://www.covers1000.net/dataset.html
http://www.covers1000.net/demo.html

Future Work

- Figure out a way around beat tracking
- Address time complexity and scale up to even larger datasets
- Apply block-based SNF to music structure analysis within a song

Acknowledgements

Chris Tralie was supported under an NSF Graduate Fellowship NSF under grant DGF-1106401 and an NSF big data grant DKA-1447491. We would also like to thank Erling Wold for pointing out the 8 covers of "The Black Page" by Frank Zappa, and we would like to thank the community at www.secondhandsongs.com for meticulously annotating songs which helped us to design Covers 1000.