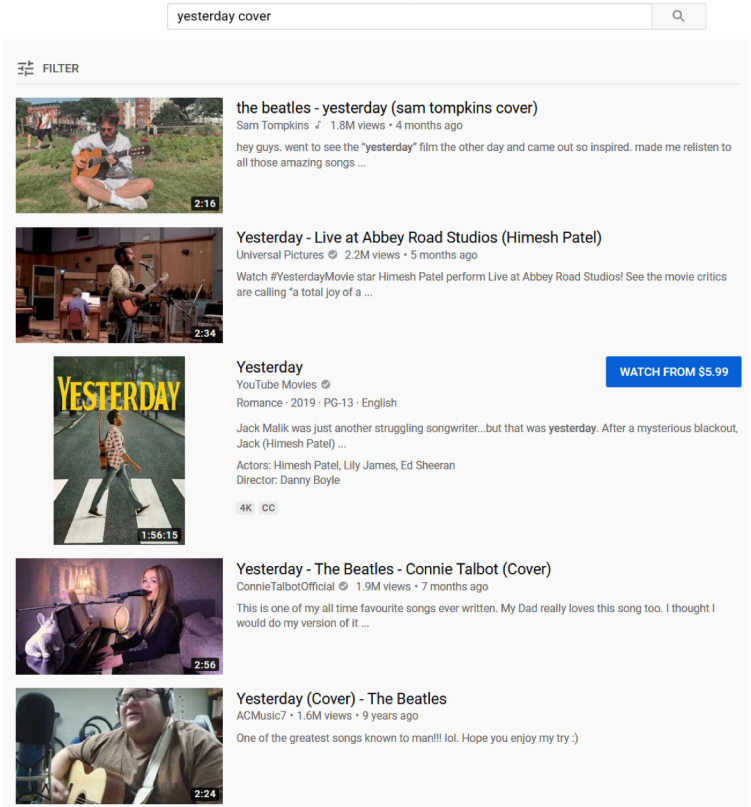
Cover Song Identification

Lukas Dillingham and Varun Khatri

What is Cover Song Identification?

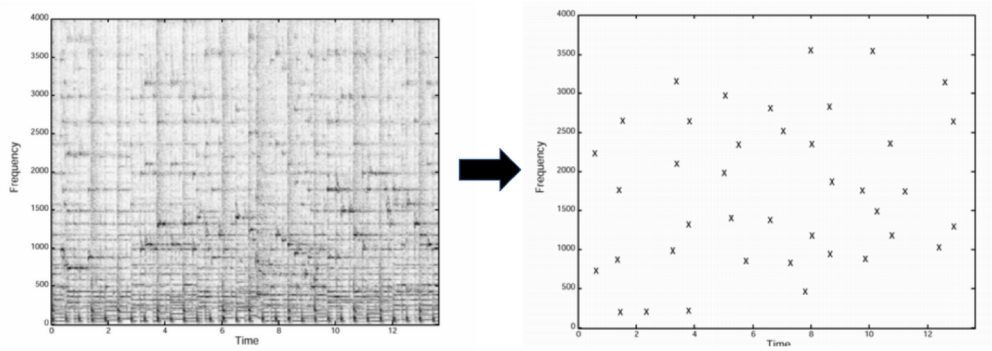
- YouTube has tons of cover songs and in a lot of cases these are not properly annotated.
- Purpose is to identify if song being played is a cover of an existing song so that they can be tagged or be recognized live setting
- We can use Shazam to identify the song if the same version is being played somewhere.
- It is based on a technique called Audio Fingerprinting
- Can the Shazam algorithm be used for this purpose?



Audio Fingerprinting Limitations

[1]

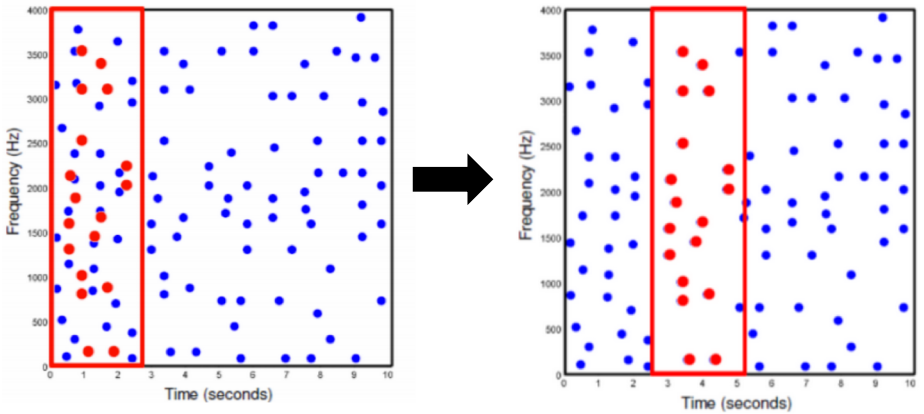
- Find peaks in spectrogram
- Compare peaks of query to fingerprints in database



[2]

Problem:

- Sensitive to changes in pitch and tempo



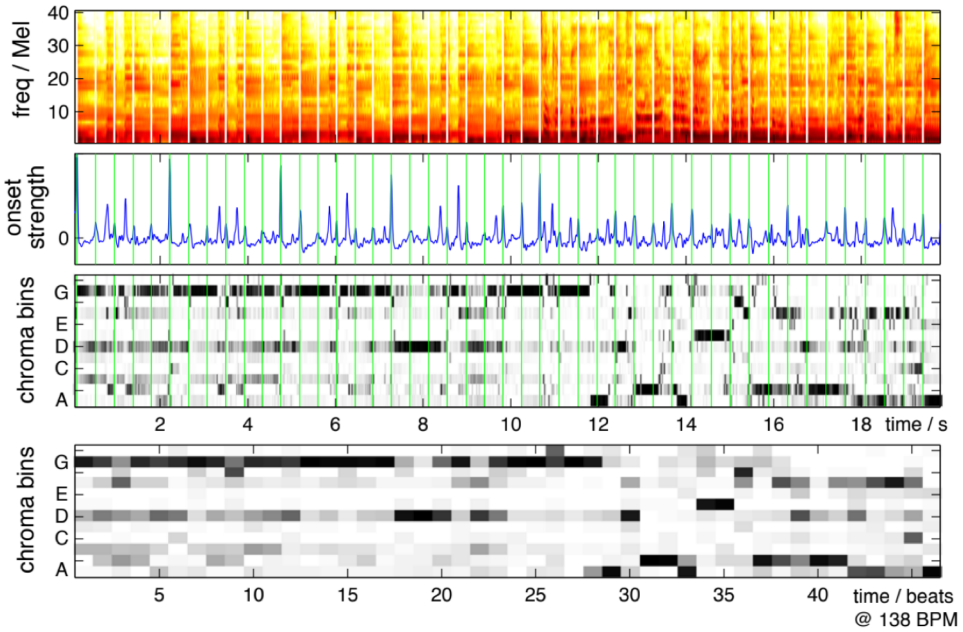
Cover Song Identification - Basic Idea

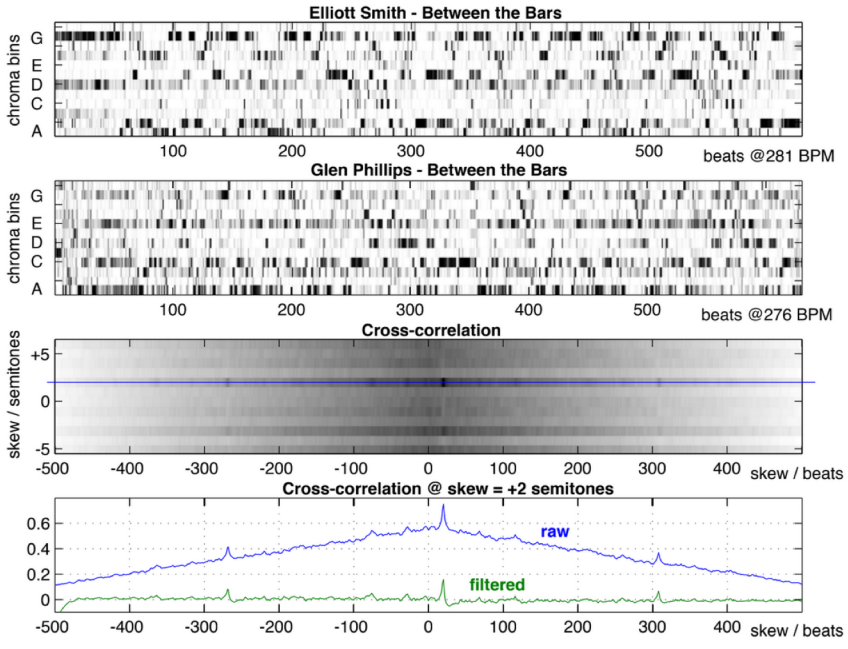
- Select features which represent melody and harmony and are robust to
  - Tempo changes
  - Key changes
  - Change in instrumentation
- Distance computation between original (database) and cover (query)
  - Cross-correlation
  - Dynamic Time Warping
  - Similarity Matrix
- Additionally can also recognize lyrics to identify the song

Methods

Title	First Author	Year	Main Idea
Large-Scale Cover Song Recognition Using The 2D Fourier Transform Magnitude	Bertin-Mahieux	2013	Uses 2DFT of Beat-sync chroma followed by PCA for reducing dimensionality
Audio Cover Song Identification using Convolutional Neural Network	Chang	2017	Obtains likelihood from cross similarity between chromas using CNN
Large-Scale Cover Song Detection In Digital Music Libraries Using Metadata, Lyrics And Audio Features	Correya	2018	Uses text-based features along audio features
Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification	Serra	2008	Uses DP to locally align Binary cross similarity of HPCP features
Cross recurrence quantification for cover song identification	Serra	2009	Introduces measurement techniques in CRP generated from HPCP
Cover Song Identification With 2D Fourier Transform Sequences	Seetharaman	2017	Creates fingerprint using 2DFT of Q transforms
Data Driven And Discriminative Projections For Large-Scale Cover Song Identification	Humphrey	2013	Creates embeddings using sparse coding on 2DFMs of beat-sync chromas
Identifying ‘Cover Songs’ With Chroma Features And Dynamic Programming Beat Tracking	Ellis	2007	Performs cross-correlation between beat-sync chromas
Cover Song Identification With Metric Learning Using Distance As A Feature	Heo	2017	Uses metric learning to learn song embeddings of similar songs
Music Fingerprint Extraction For Classical Music Cover Song Identification	Kim	2008	Creates fingerprint using covariance matrix of chroma and delta chroma
Cover Song Detection: From High Scores To General Classification	Ravuri	2010	Extracts features from chromas and classifies using SVM/MLP
The Intervalgram: An Audio Feature for Large-scale Melody Recognition	Walters	2012	Introduces new feature using SAI-based chroma
Early MFCC And HPCP Fusion for Robust Cover Song Identification	Tralie	2017	Combines beat-synchronous blocks containing HPCP and MFCC features

Approach 1: Using beat-synchronous chroma [Ellis et. al, 2007]

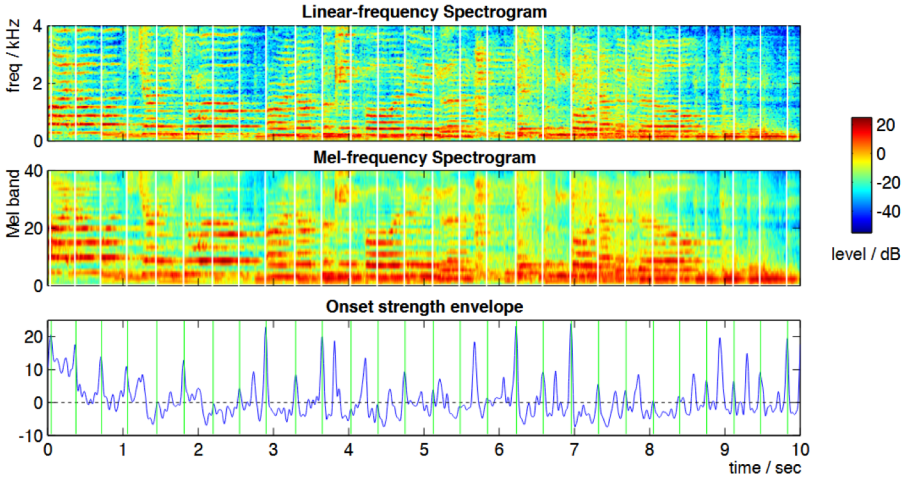




Beat Detection

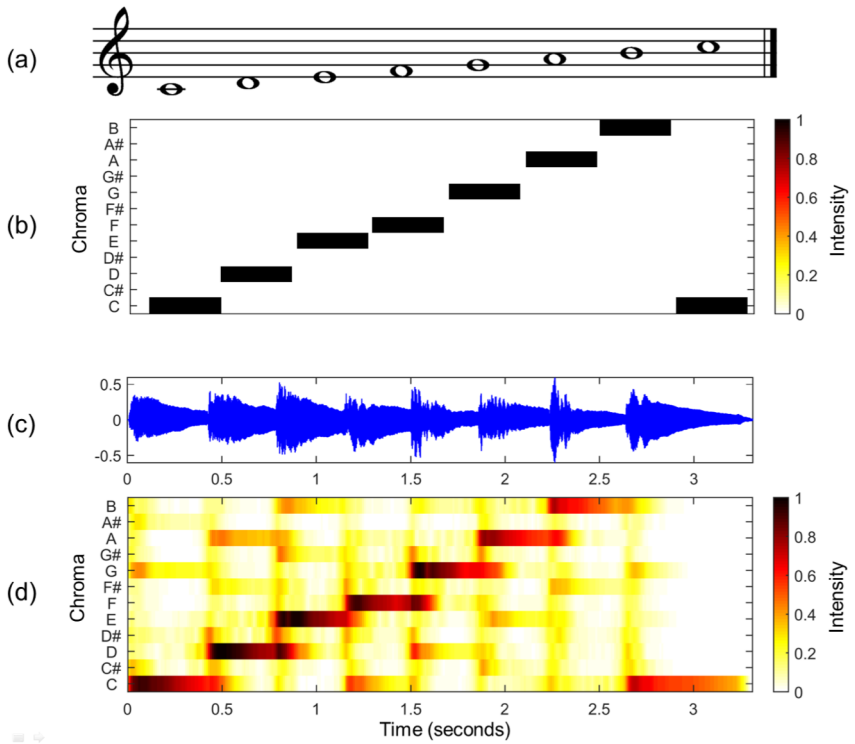
[Ellis et. al, 2007]

- Same process as (Ellis 2006) which we did as a homework
- Take 1st order difference in time in the Mel-frequency spectrogram
- Remove slow varying onsets by applying a high pass filter
- Global tempo is calculated using auto-correlation
- Use dynamic programming to select beats that optimize onset strength and time between beats



Chromagram

- Records the intensity associated with each of the 12 semitones within one octave
- Captures dominant note (typically the melody)
- Use the phase derivative (instantaneous phase) to get higher resolution estimate of note
- Best results using only frequencies up to 1000 Hz
- Adjusted each peak ±0.5 semitones to align with chroma bin center to compensate for the cover song being out of tune

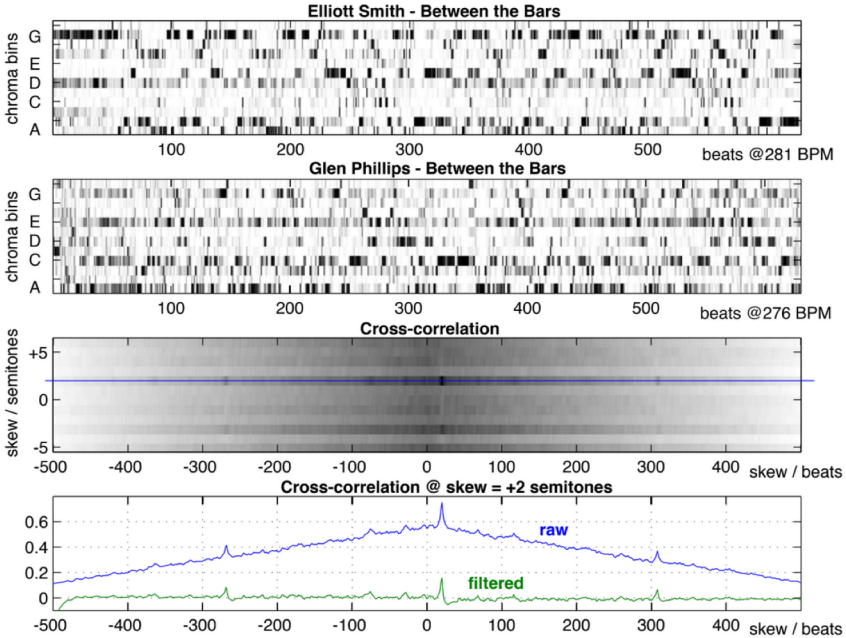


[3]

Matching

[Ellis et. al, 2007]

- Expect cover versions to have long stretches (verses, choruses, etc.) that match reasonably well, but not perfectly
- Perform cross-correlation between original and cover song to calculate similarity
- Perform cross-correlation for ±500 beats and 12 chroma skews
- Take skew with highest cross-correlation and and filter it to remove triangular baseline correlation
- A peak in the filtered cross-correlation represents a match



Evaluation

[Ellis et. al, 2007]

- MIREX (Music Information Retrieval Evaluation eXchange)
- Accuracy is the proportion of time that the correct cover version was returned as most similar to a query
- Mean Reciprocal Rank (MRR) reflects the rank of the first correct response in ordered returns

System	Accuracy	MRR
Standard	59%	0.63
P-Chroma	29%	0.40
Two-Tempos	35%	0.50
120 BPM	51%	0.53

Method Advantages and Disadvantages

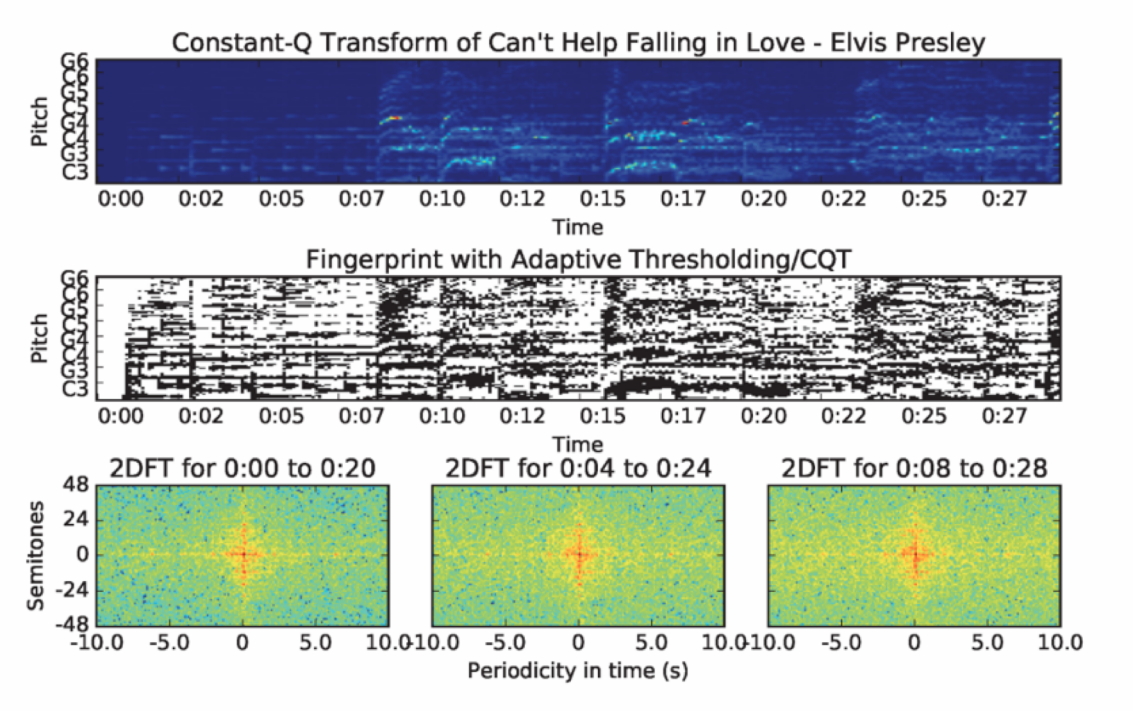
Advantages

- Shifting chromagram is resilient to different tuning and timbre changes
- Dependence on beats gives resilience to tempo changes

Disadvantages

- Different rhythm in cover song would change beats resulting in less matching
- Shifting in chromagram and beats makes algorithm very computationally expensive
- Cannot handle noise degradations

Approach 2: Using 2D Fourier Transform Sequences [Seetharaman et al, 2017]



Constant Q Transform

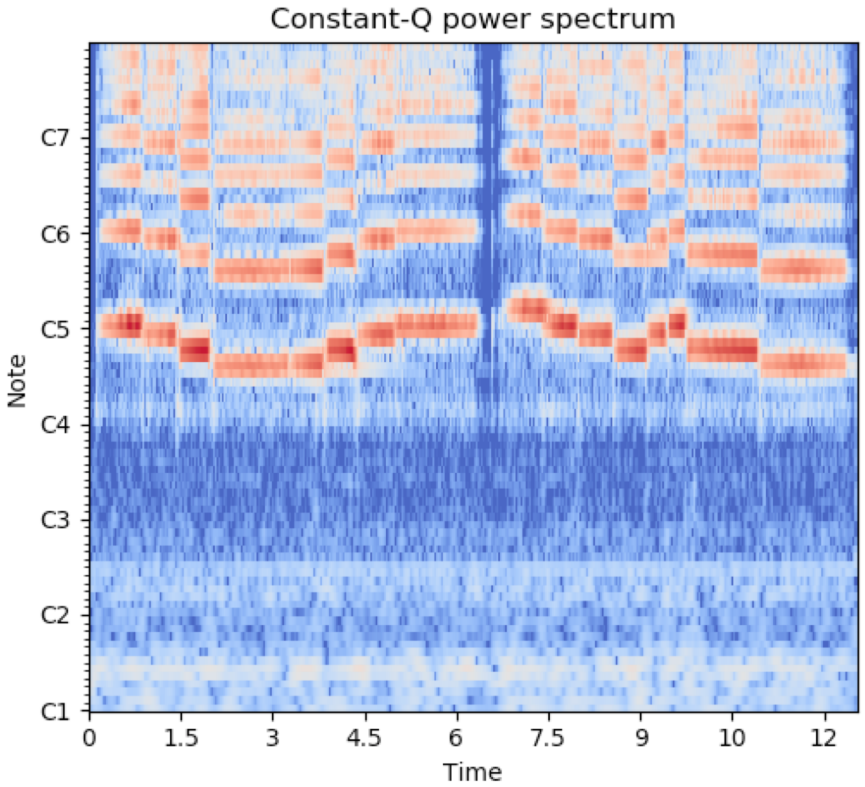
- Transform with logarithmically spaced bins (like Mel or Bark scales) based on Western music scale

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{\frac{-j2\pi Qn}{N[k]}}$$

$$Q = \frac{f_k}{\delta f_k},$$

$$N[k] = \frac{f_s}{f_k} Q$$

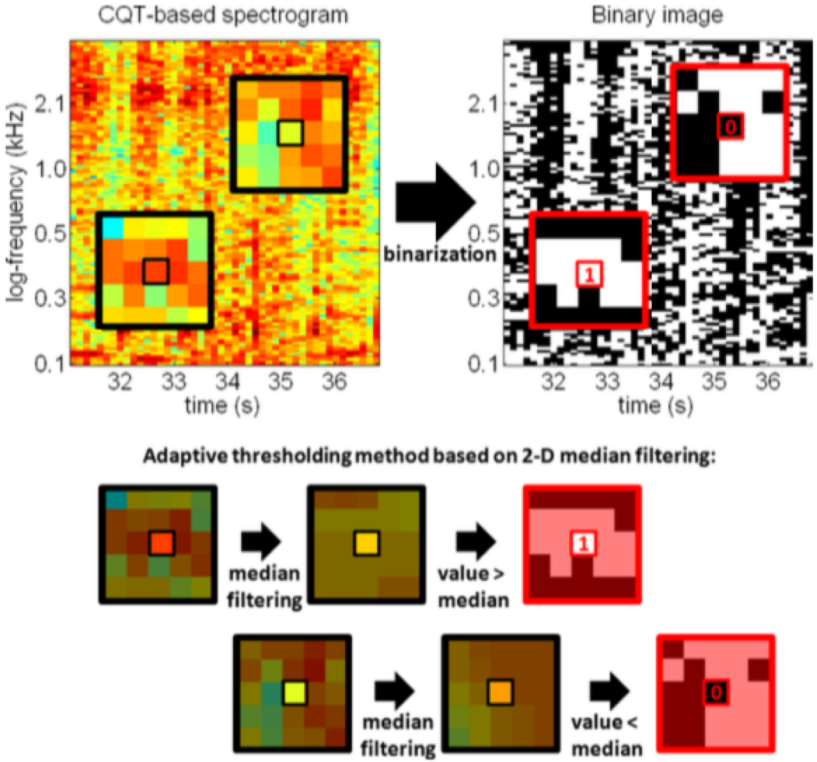
- Can easily handle key variations



Adaptive Thresholding

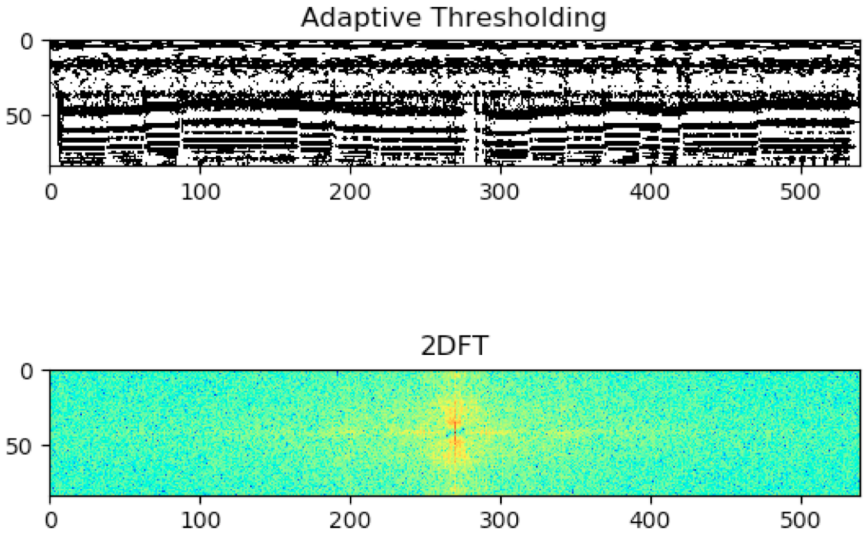
[Rafii et. al, 2014]

- Discard finer timbral information and accentuate the main patterns by thresholding locally
- Set local high energy bins/frames to 1 and low energy bins/frames to 0
- Median value as threshold



2D Discrete Fourier Transform

- 2DFT of 20 seconds long segments at 4 seconds intervals
- Helpful in capturing periodic patterns
- 2DFT is invariant to key changes
- Gaussian blurring to avoid small tempo deviations
- Large tempo deviations are addressed by creating fingerprints using multiple sampling rates



Search

[Seetharaman et al, 2017]

- Similarity matrix constructed using Euclidean distance between 2DFT sequences
- Matrix is normalized and convolved with checkerboard kernel to obtain diagonals

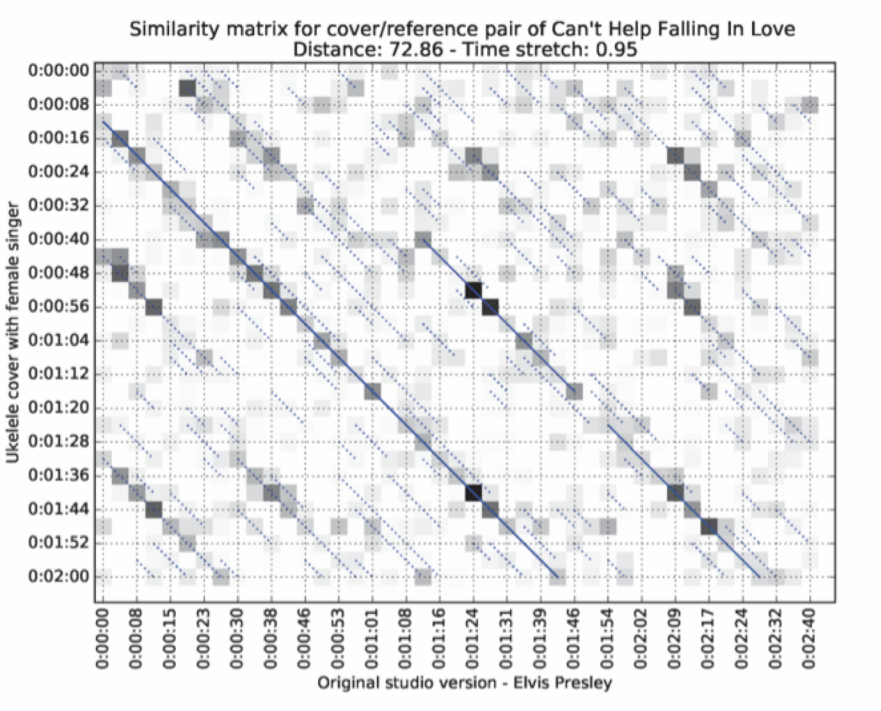
$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

- 3 longest diagonals are identified, and the following distance is then computed as:

$$d(q,r) = \frac{E}{\sum_{i=1}^{3} w_i l_i}$$

E – Unnormalized SM

w, l – weight and length of diagonals



Evaluation

- Train: Covers80 dataset - 80 songs, 2 versions of each
- Test set: YouTube covers dataset - 50 songs, 7 versions of each
- Metrics: Mean average precision, precision at ten, mean rank of correctly identified song
- Overall 164 out of 250 covers correctly identified

Algorithm	**MAP**	**P@10**	**MR1**
DTW [5]	0.425	0.114	11.69
Silva et al. [5]	0.478	0.126	8.49
Serra et al. [2]	0.525	0.132	9.43
Silva et al. [6]	0.591	0.140	**7.91**
Proposed (on CQT)	0.521	0.122	9.75
Proposed (on fingerprint [19])	**0.648**	**0.145**	8.27

Method Advantages and Disadvantages

Advantages

- Robust to instrumentation and other timbral changes (Adaptive Thresholding)
- Resilient to noise degradations (CQT and AT)
- 2DFT accommodates for pitch shifts without having to create multiple fingerprints

Disadvantages

- Still computationally expensive
- Still not good enough performance

Other methods

- Fusion of MFCC and HPCP [Tralie, 2017]
  - Computes blocks synchronized with beats
  - Combines MFCC distance matrices, MFCC SSM distance matrices, and HPCP cosine similarity matrices into a single matrix
  - Performs SNF (similarity network fusion) to obtain a score from this matrix

- Using Convolutional Neural Networks [Chang et al, 2017]
  - Uses chroma features and computes cross-similarity matrix
  - This is input to 10 layer CNN and we obtain likelihood for that version
  - The likelihoods are then ranked and we obtain top@N possible songs

References

[1] A. Wang, "An Industrial-Strength Audio Search Algorithm," ISMIR, 2003.Meinard Müller and Joan Serrà, “Audio Content- Based Music Retrieval (tutorial),” 12th International Society for Music Information Retrieval, Miami, FL, USA, October 24-28, 2011.
[2] D. Ellis. Beat tracking with dynamic programming. In MIREX 2006 Audio Beat Tracking Contest system description, 2006.
[3] https://en.wikipedia.org/wiki/Chroma_feature
[4] Meinard Müller and Joan Serrà, “Audio Content- Based Music Retrieval (tutorial),” 12th International Society for Music Information Retrieval, Miami, FL, USA, October 24-28, 2011.
[5] Z. Rafii, B. Coover, and J. Han, "An audio fingerprinting system for live version identification using image processing techniques," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.
[6] Seetharaman, P., & Rafii, Z. (2017). Cover Song Identification with 2D Fourier Transform Sequences. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
[7] Christopher J Tralie. “MFCC And HPCP Fusion for Robust Cover Song Identification”. In: 18th International Society for Music Information Retrieval (ISMIR). 2017
[8] Sungkyun Chang, Juheon Lee, Sang Keun Choe, and Kyogu Lee. “Audio cover song identification using convolutional neural network”. In Workshop Machine Learning for Audio Signal Processing at NIPS, 2017.