Cover song identification with 2D Fourier Transform Sequences

Prem Seetharaman (*prem@u.northwestern.edu*) and Zafar Rafii
Northwestern University, Gracenote

Overview

Cover song identification is the act of identifying when two musical recordings are derived from the same music composition (e.g., Jimi Hendrix's *All Along The Watchtower* is a cover of the original by Bob Dylan). Many things can change in a cover song, including:

- Key
- Tempo
- Instrumentation
- Music structure

However, some things will stay the same between the cover and the original, such as:

- Melody
- Harmony
- Signature patterns (hooks)

Successful automatic cover song identification requires representations that are invariant to changes while keeping the aspects of music that are transferred between the two.

In this work, we present a time-series representation of audio based on the 2D Fourier transform.

Results

To test our approach, we use the YouTube Covers dataset. YouTube Covers consists of 50 compositions, with 7 recordings of each composition. Of these 7, 5 are covers, 1 is a live recording by the original artist, and 1 is the original studio recording.

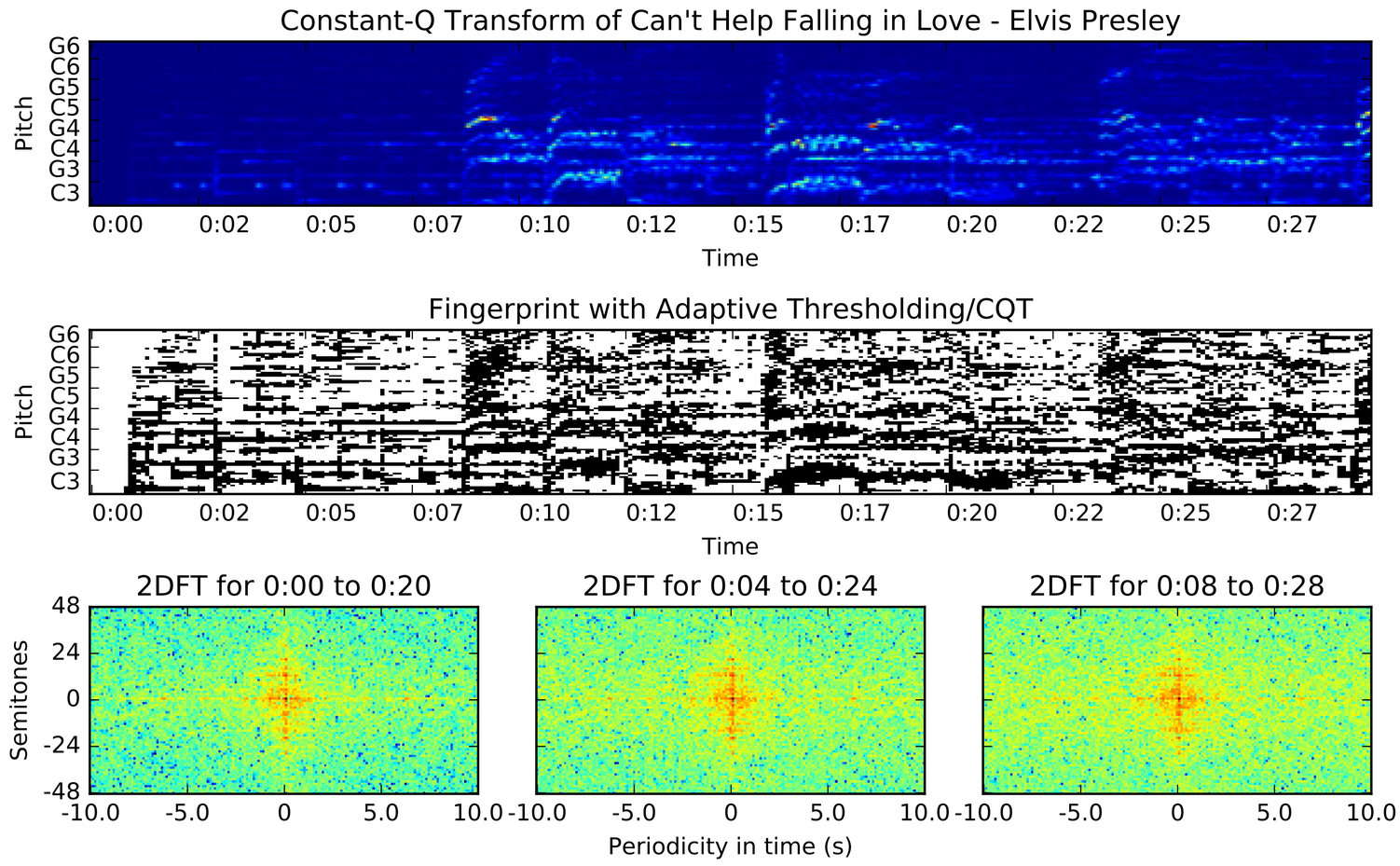
Algorithm	**MAP**	**P@10**	**MR1**
DTW [1]	0.425	0.114	11.69
Silva et al. [1]	0.478	0.126	8.49
Serra et al. [2]	0.525	0.132	9.43
Silva et al. [3]	0.591	0.140	**7.91**
Proposed (on CQT)	0.521	0.122	9.75
Proposed (on fingerprint [4])	**0.648**	**0.145**	8.27

In mean average precision and precision at 10, the proposed approach surpasses current state-of-the-art methods on this dataset. The proposed approach finds 164 covers out of 250 correctly at top one. The impact of the adaptive thresholding step is significant, causing a jump in mean average precision of 0.127, making it the best performing approach in terms of P@10.

Acknowledgements

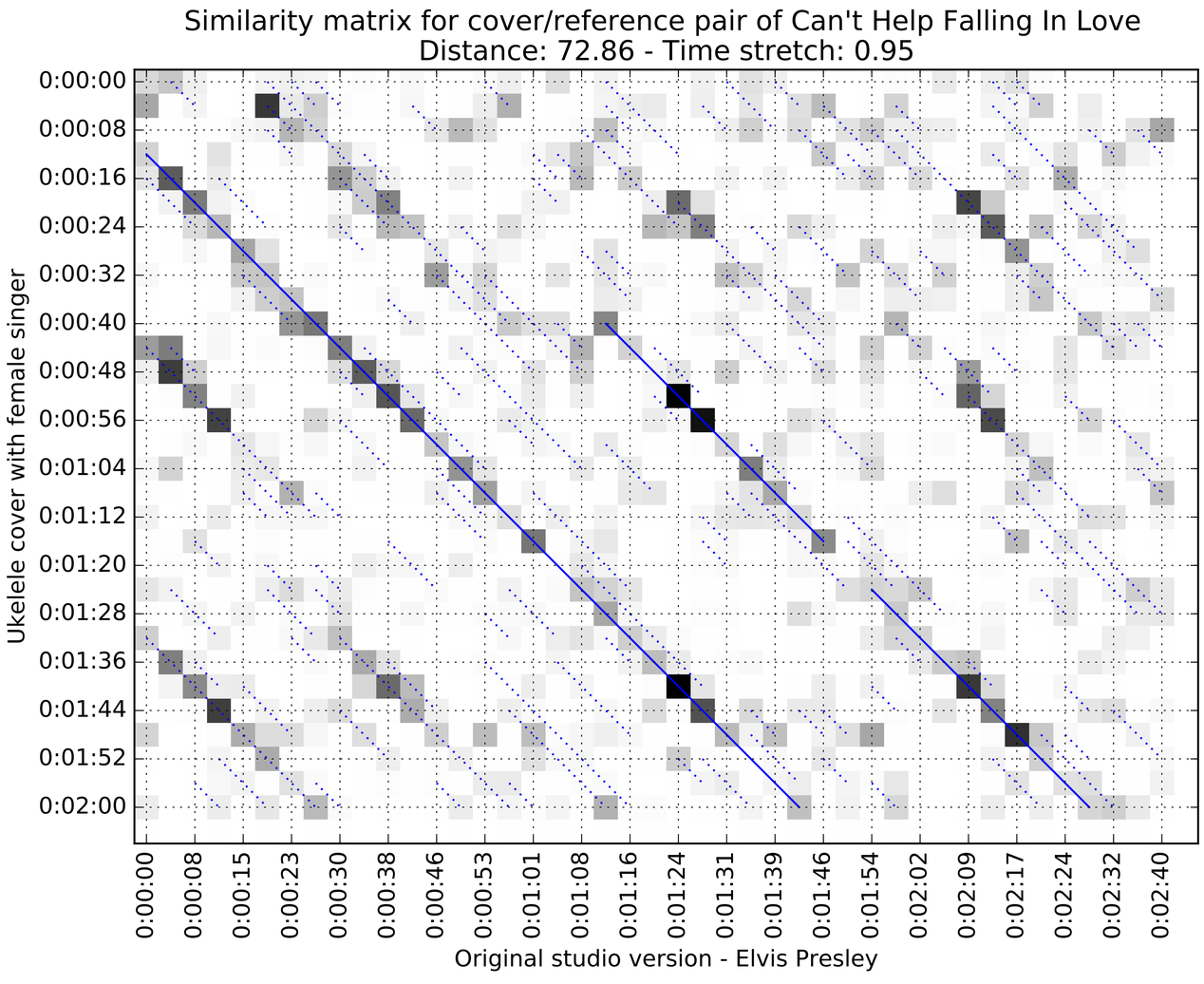
This work was funded by a Gracenote internship.

Proposed approach



1. The audio is first converted into a Constant-Q transform. In the CQT, key changes are linear shifts along the frequency axis. These linear shifts are phase differences in the 2DFT, giving us key invariance.
2. We then use the method from [4] to convert the CQT into a binary matrix using adaptive thresholding. For every time-frequency point in the CQT, it is replaced with a 1 if it is above the median of its surrounding neighborhood and a 0 otherwise. This makes the representation invariant to instrument (timbre) changes.
3. We take 20 second segments with 16 second overlap from the fingerprint. For each segment, we compute the magnitude 2DFT. We apply a small amount of Gaussian blurring to the 2DFT. The sequence of magnitude 2DFTs is then used to represent the audio.

We extract this representation at many different sampling rates, to account for tempo changes between the original and the cover.



We compute the pairwise Euclidean distance between the sequences of magnitude 2DFTs. We then filter the similarity matrix to emphasize diagonal matches. We then use the length and weight of contiguous diagonals to compute distance between the cover and the original.

[1] D. F. Silva, V. M. A. d. Souza, G. E. d. A. P. A. Batista, et al., "Music shapelets for fast cover song recognition," in International Society for Music Information Retrieval Conference, 2015.
[2] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, 2008.
[3] D. F. Silva, C.-C. M. Yeh, G. E. Batista, and E. Keogh, "SIMPle: Assessing music similarity using subsequences joins," in International Society for Music Information Retrieval Conference, 2016.
[4] 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.