THE 2007 LABROSA COVER SONG DETECTION SYSTEM

Daniel P.W. Ellis and Courtenay V. Cotton
LabROSA, Columbia University
New York NY USA
dpwe@ee.columbia.edu, cvcotton@ee.columbia.edu

ABSTRACT

We describe our cover song detection system, as submitted to the MIREX 2007 Cover Song Detection evaluation. The system is developed from our 2006 MIREX system, which was the best-performing entry in last year's evaluation. Using the new "covers80" dataset of 80 pairs of songs and covers, we improve the overall detection from 42.5% to 67.5% through a collection of minor modifications relating to correlation normalization, tempo tracking, and temporal filtering of the beat-synchronous chroma representation.

1 INTRODUCTION

The cover song detection task involves identifying which music audio pieces in a collection represent different versions of a particular piece – typically performed by different artists, with different instrumentation, style, tempo etc. The task is of interest because it relies on identifying the deeper underlying musical information in music audio, rather than the surface timbral/instrument features.

The first MIREX Audio Cover Song Detection task was run in 2006 [2]. Our submission performed best out of the 4 algorithms tested that year [3, 7]. This year we are submitting essentially the same approach, but with several relatively minor that modifications that have, however, substantially improved the performance on our development data.

2 THE "COVERS80" DATASET

The MIREX evaluations are based on a set of eleven versions of each of thirty different songs. This collection of 330 tracks is kept secret by the organizers of the evaluation to prevent over-tuning of approaches. However, this left the community with no common dataset with which to experiment.

For 2006 we developed our algorithm on a very small set of 15 pairs of cover songs, manually identified from within the 8764 pop music tracks of USPOP2002 [6]. Subsequent to the evaluation, we made a more thorough search for cover versions within USPOP, and augmented these with some albums consisting entirely of covers ("Medusa" by Annie Lennox, and "Strange Little Girls" by Tori Amos) for which most of the pairs were found. In the end, we amassed a fairly diverse collection of 80 pieces with two versions of each (160 tracks total). We have dubbed this collection "covers80" and made it available to the research community [5]. We distribute pre-calculated features as well as low bitrate/bandwidth audio; the limited audio quality has been verified to have negligible impact on our cover song detection system, which only considers spectral components up to about 2 kHz.

In our evaluations, we define two sets of 80 songs each, with one version of each cover song in each list. Then each of the 80 items from the first (query) list is compared with all of the 80 items from the second (reference) list, and the most similar is chosen as the matching cover version. Note that this assumes that there is exactly one cover version to be found, but does not attempt to prevent a single reference track being matched as the cover to several queries.

3 BASELINE MIREX06 SYSTEM

The basic cover song system we submitted in 2006 is illustrated in figure 1 and described in [3, 7]. Each song is represented as as single beat-synchronous chroma feature matrix with 12 rows and typically around 2 to 4 columns per second of the song (i.e. a tracked tempo in the range 120-240 BPM). Chroma bin intensity is extracted from a Fourier transform from overlapping 93 ms windows, weighted to emphasize energy from a couple of octaves above and below 400 Hz, and filtered to select tonal concentrations by looking for consistent instantaneous frequency (phase advance) across adjacent Fourier transform bins [1].

Since the chroma representation mainly represents the melodic and harmonic information without much influence of the instrumentation, and since the representation is on a time base defined by the tempo of each piece, cover versions of the same song are likely to have similar beat-chroma matrices. Because the songs may have had structural alterations (different numbers of verses etc.) and also as a result of any local beat-tracking errors, we do not expect an exact match end-to-end. However, rather than trying to identify the largest matching subset between two pieces, we have found it expedient to simply cross-correlate the entire beat-chroma representations of the songs being compared. Any long stretch of correlated features

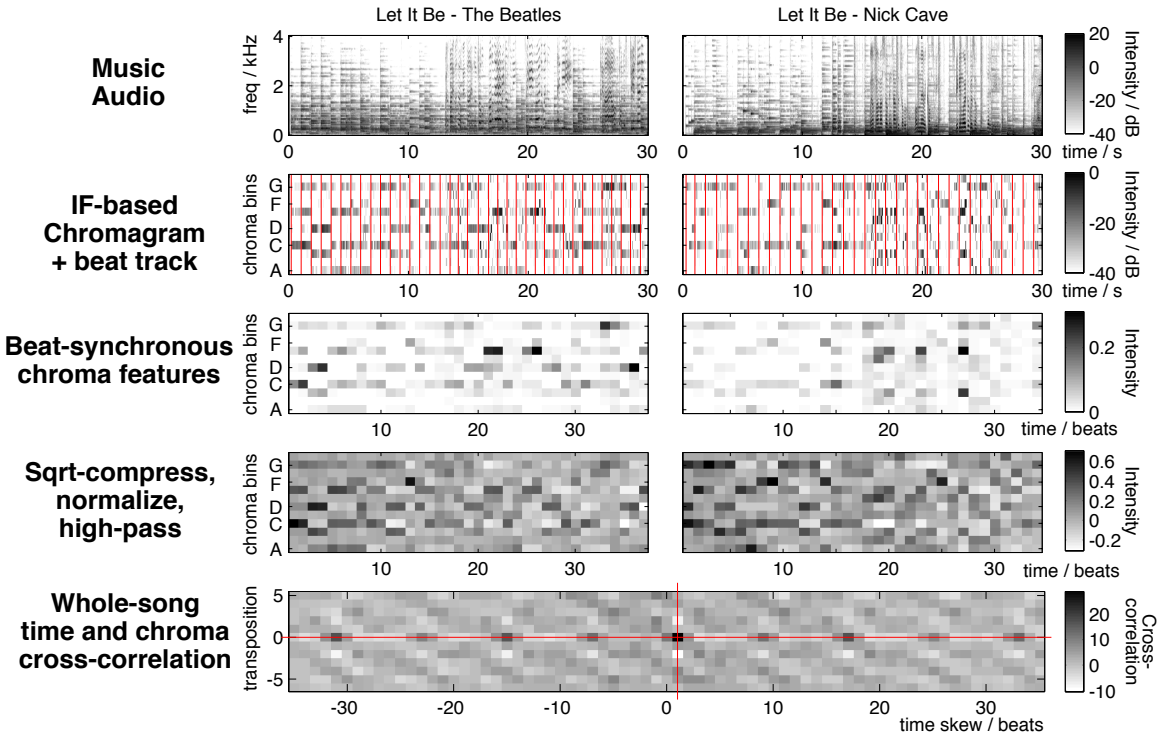

Figure 1. Illustration of the basic cover song detection system. Songs to be compared are first analyzed for 12-bin chroma content via instantaneous frequency, and also have the beat times estimated (shown overlaid in the second row). Then chroma energy is averaged within each beat to create the beat-synchronous chroma representation (beat-chroma). These beat-chroma matrices are compressed by square-root, normalized to make each column sum to unity, then high-pass filtered along time to de-emphasize sustained note. The similarity between two tracks is then simply the largest value of the full cross-correlation between the two entire songs, where all 12 chroma rotations are tested.

will result in a large value at the appropriate lag; we also use circular cross-correlation along the chroma dimension to search all possible transpositions (chroma shifts).

In 2006, we observed improved performance by first compressing the magnitude of the beat-chroma features with a square root, then normalizing the total energy at each time frame by scaling each 12-bin vector to have unit norm. Further, we observed that true matches were typically very specific to a precise lag – shifting the cross-correlation one beat earlier or later gave a much lower correlation score. That led us to high-pass filter the cross-correlations on the lag axis to emphasize such rapid variations. We now interpret this step slightly differently: a lot of spurious matches arise from sustained sequences of a single chroma bin with a large value. Cross-correlation in both dimensions can lead to a large score if sustained blocks in both pieces line up, but there is not really any evidence of the same musical structure being revealed. High-pass filtering the output of the cross-correlation is equivalent to cross-correlating appropriately high-pass filtered versions of the original matrices, and this operation can be understood to de-emphasize sustained (slowly-changing) structure and instead to put weight on the changes in the chroma signature, which are more informative of the particular musical piece.

In the bottom pane of figure 1, we see a large cross-correlation value of around 25 at a relative beat timing of +1 beat, and a chroma shift of zero, indicating the correct match between these two cover songs despite visibly rather different initial beat-chroma representations (the one-beat skew comes from the beat tracker missing the initial beat in one version). We also notice strong negative correlation at ± 2 semitones transposition, and a weaker, recurrent correlation at multiples of 16 beats, the length of the basic chord progression in this piece at this tempo level.

The basic 2006 system found 34 correct covers out of the 80 trials in the 'covers80' set, for 42.5% correct. The entire cover song system (in Matlab) is available for download at `http://labrosa.ee.columbia.edu/projects/coversongs/`.

4 IMPROVEMENTS FOR 2007

In this section we describe the relatively minor changes applied in this year's system. The successive improvements on the 'covers80' dataset are reported in table 1.

4.1 Correlation (un)normalization

The peak cross-correlation value tends to grow with the length of the beat-chroma matrices, since this in a sense dictates the amount of 'opportunity' for correlation there

Table 1. Performance of system variants on the "covers80" [5] dataset.

System	Correct
MIREX 06 baseline	34/80 = 42.5%
Without cross-corr norm	41/80 = 51.3%
Improved high-pass filtering	46/80 = 57.5%
Tempo bias = 120 BPM	48/80 = 60.0%
Dual tempo levels	54/80 = 67.5%

is. In 2006, we normalized every cross-correlation by dividing by the column count of the shorter matrix. This, combined with the normalization of each column, guaranteed that all correlation scores lay between 0 and 1. Unfortunately, it also introduced a variable scaling of the scores of some reference pieces against each target, and could therefore alter which piece is chosen as most similar. In further experimentation, we found the system to be more successful when this normalization is removed; we now use the raw peak cross-correlation value (of the partially-normalized input matrices) as the similarity measure. This improved performance to 41 out of 80, a 20.6% relative accuracy increase.

4.2 High-pass filtering

As described above, high-pass filtering the cross-correlation results along lag was used to highlight local maxima that were sensitive to precise temporal alignment. We further tuned this filter (e.g. cutoff frequency) and moved it earlier in the processing stream, specifically to be applied to just the query beat-chroma matrix prior to cross-correlation. This improved performance to 46 correct (12.2% improvement relative to the previous modification) as well as making the comparisons faster.

4.3 Target tempo bias

Our beat tracker operates by first identifying a global tempo from the autocorrelation of an "onset strength envelope" derived from a Mel-frequency spectrogram [4]. This tempo estimation includes a window that weights the 'preferred' range of tempos, similar to the known human bias towards 'tapping' at 120 BPM. For the 2006 system, we biased our beat tracker to prefer tempos of 240 BPM, which is likely to make it find a faster metrical level; we felt that a more rapid sampling of the chroma structure would lead to a more accurate model. This year we experimented with using a slower, 120 BPM bias and found not only more compact descriptions and hence faster comparisons, but a slight improvement to 48 correct (4.3% relative to the previous step), although this is not statistically significant.

4.4 Multiple tempo levels

In looking at the errors made in the development data, we noticed a couple of occasions when the two cover versions ended up with beat tracks were at different metrical levels e.g. one version might have 16 beats (chroma vectors) per line of the verse, but the second only has 8. Clearly, such a radical transformation will prevent a match from being found.

To accommodate this, we run the beat tracker on each piece twice, differing only in the initial BPM bias value, which was set to both 120 and 240 BPM (although in some cases, the same tempo was found despite these settings. Then, all four correlations between the two versions of the query beat-chroma matrix, and the two version of the test item beat-chroma matrix, are cross-correlated, and the single largest value taken as the score. This improved accuracy to 54/80, a 12.5% improvement over the best single-tempo system.

We had briefly tried an approach of this kind last year, but seen no benefit. However, the beat-chroma matrices obtained at different metrical levels will differ greatly in length (e.g. by a factor of 2), and it may be that the misguided length-based normalization had messed it up at that time.

5 DISCUSSION AND CONCLUSIONS

We have inspected the overall performance of the system through a web interface that shows the top 10 matches for each piece. Once the problem of tempo mismatch has been resolved, the remaining error pieces showed no clear trend, and indeed it was rare for any true match to appear in ranks 2-10; the true match was either in first place, or did not match at all. Some of the cover versions are very different in style.

Cover song detection is not in itself a tremendously compelling application, but the MIREX cover song evaluation is important and noteworthy because it encourages the development of techniques to recover the deeper, musical information from music audio recordings – the kind of processing that used to be the exclusive domain of symbolic music representations. Our underlying goal it the investigation of how this kind of musical core can be efficiently mined for, and managed within, large pop-music databases.

6 ACKNOWLEDGMENTS

Thanks to Jesper Højvang Jensen for the 2D FFT trick that greatly sped up the core correlation operation. Thanks also to Suman Ravuri for various supporting investigations.

This work was supported by the Columbia Academic Quality Fund, and by the National Science Foundation (NSF) under Grant No. IIS-0238301. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

7 REFERENCES

[1] F. J. Charpentier. Pitch detection using the short-term phase spectrum. In *Proc. ICASSP-86*, pages 113–116, Tokyo, 1986.

[2] J. S. Downie, K. West, E. Pampalk, and P. Lamere. Mirex2006 audio cover song evaluation, 2006. http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song_Identification_Results.

[3] D. P. W. Ellis. Identifying 'cover songs' with beat-synchronous chroma features. In *MIREX-06 Abstracts*, 2006.

[4] D. P. W. Ellis. Beat tracking by dynamic programming. *J. New Music Research*, 2007. Special Issue on Tempo and Beat Extraction, to appear.

[5] D. P. W. Ellis. The "covers80" cover song data set, 2007. Web resource, available: http://labrosa.ee.columbia.edu/projects/coversongs/covers80/.

[6] D. P. W. Ellis, A. Berenzweig, and B. Whitman. The "uspop2002" pop music data set, 2003. http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html.

[7] D. P. W. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*, pages IV–1429–1432, Hawai'i, 2007.