COVER SONG IDENTIFICATION WITH IF-F0 PITCH CLASS PROFILES

Alexey Egorov
CBMS Networks, Inc.
alexey@cbmsnetworks.com

Gene Linetsky
CBMS Networks, Inc.
gene@cbmsnetworks.com

ABSTRACT

The extraction of time-dependent Pitch Class Profiles (PCP) is widely used in machine listening. PCP describes effective note amplitudes at a given point in time.

Our PCP extraction methodology ([1]) was originally developed for rhythm and harmony detection. It has proven very effective and is used in our free Winamp plug-in, Easy Chords.

This paper describes the results of applying our PCP extraction approach to cover song identification.

Keywords: Pitch Class Profile, Audio, Cover, Identification, Derivative Works, Machine Listening

1. INTRODUCTION

Traditionally, PCP extraction involves windowed Fourier transform with subsequent collapsing of the spectrum into a single octave. We extract PCP by applying a bank of gammatone-like filters, time integration of the resulting signal amplitudes of each channel, extraction of spectrum peaks, suppression of higher harmonics, and mapping the resulting F0-spectrum into a single octave. The resulting PCP is particularly well suited for rhythm and harmony detection.

It is intuitively clear that similarity of higher-level musical characteristics such as intra-measure melodic patterns, rhythmic structures, and harmonic sequences should help identify different renditions of the same musical material, even if they vary in structure, key, instrumentation, tempo, or genre. Indeed, our results show that PCP extraction approach tuned to rhythm and harmony detection performs well in identifying cover songs.

2. DISCUSSION

In our PCP extraction scheme we use F0 spectrum instead of Fourier spectrum ([1]). M. Goto's classical algorithm for F0 extraction is presented in [2]. Its main drawback is high computational cost. Our F0 extraction algorithm is based on Instantaneous Frequency (IF). It achieves significant speed improvement over Goto's algorithm without any loss of precision. We have discovered that the signal reconstructed from IF-spectrum has no perceptual difference from the original with respect to its basic musical content (rhythm, melody, harmony). Consequently, our PCP performs well in extracting beats, bars (measures), and harmonic sequences, as well as such higher-level musical features as segments (large sections of the work such as verses, choruses, bridges of a pop song, etc.) and genres. To apply our PCP to cover song identification, we use the post-processing method described in [3]. Instead of 36-bin PCP we use 12-bin PCP. This is accomplished by adding the detuning detection module to account for tunings that deviate from the standard A at 440Hz. For our MIREX submission we have experimented with three different implementations of Optimal Transposition Index (OTI), each of which differ from the method suggested in [3]. Only the best-performing OTI is described here.

3. IMPLEMENTATION

3.1 Band-Pass Filter

The first step consists of passing the signal $x[n]$ through the set of f -frequency filters, each of which yields the signal $y[n]$ in the following recursive fashion:

$$u[n] = a[1]u[n-1] + a[2]u[n-2] + x[n] - x[n-1];$$
$$w[n] = r \cdot w[n-1] + |u[n]|;$$
$$y[n] = \pi\xi \exp(-\xi 2\pi f / Fs)\sin(2\pi f / Fs)(1-r)w[n];$$
$$a[1] = 2\exp(-\xi 2\pi f / Fs)\cos(2\pi f / Fs);$$

$a[2] = -\exp(-2\xi 2\pi f / Fs)$;
$r = \exp(-\eta 2\pi f / [(\eta 2\pi f T + 1) Fs])$;
$T = 0.02,\ \ \xi = 0.02,\ \ \eta = 0.04$;

Each of these filters introduces the effective delay (measured in samples) of

$$delay = 1/(\exp(\xi 2\pi f / Fs)/r - 1); \quad (1)$$

3.2 Filter Bank

The bank consists of the abovementioned filters

$$f(k) = 440 \exp((k - 95)\ln 2/24); \quad (2)$$

where $k = 0,...,200$. According to (1) the filters have different delays. These delays are aligned across the channels with additional delaying filters. The cumulative delay of the channel is achieved by padding the end of the input signal with zeroes and skipping the corresponding number of samples in the beginning of the output signal. Each channel's output signal is downsampled to 50Hz.

3.3 IF-Spectrum Extraction

In the second step, we extract peaks from the profile comprising the output values of channels $y[k]$ at the same moment in time. A peak is characterized by amplitude A_{PEAK} and Instantaneous Frequency (IF) f_{PEAK}. The extraction is performed as follows:

1) $y[k-1] < y[k] > y[k+1]$:

$$f_{PEAK} = f\left(k + \frac{1}{2}\frac{y[k-1] - y[k+1]}{y[k-1] + y[k+1] - 2y[k]}\right);$$

2) $y[k-1] < y[k] = y[k+1] > y[k+2]$:

$$f_{PEAK} = f(k + 1/2);$$

3) $y[k-1] < y[k] = y[k+1] = y[k+2] > y[k+3]$:

$$f_{PEAK} = f(k+1);$$

etc. where $f(k)$ is defined in (2);

$A_{PEAK} = y[k]$.

3.4 F0-Spectrum Extraction

In the next step, the suppression of higher harmonics in the IF spectrum is done as follows. Let i, j be peak indices, $h = 0,...,15$ - harmonic indices. The harmonic correlation coefficients are computed from peak frequencies $f[i]$:

$$Q[j,i,h] = \exp\left\{-\frac{h^2}{2U^2} - \frac{1}{2}\left(\frac{12}{W}\log_2\frac{(1+h)f[i]}{f[j]}\right)^2\right\};$$

The new amplitudes $A0[i]$ are computed from peak amplitudes $A[i]$:

1. $w[i,h] \leftarrow 0, B[j] \leftarrow A[j]$;

Steps 2-12 are repeated M times:

2. $R[j,i,h] \leftarrow B[j]Q[j,i,h]$;
3. $\Delta w^{(0)}[i,h] \leftarrow \sum_j R[j,i,h]$;
4. $\lambda^{(0)}[i] \leftarrow \sum_h \Delta w^{(0)}[i,h]$;
5. $\Delta B^{(0)}[j] \leftarrow \sum_i \lambda^{(0)}[i] \sum_h R[j,i,h]$
6. Exit cycle if $\max_j \Delta B^{(0)}[j] = 0$;
7. $\kappa \leftarrow \left(\max_j A[j]/M\right) / \max_j \Delta B^{(0)}[j]$;
8. $\Delta B[j] \leftarrow \kappa \cdot \Delta B^{(0)}[j]$;
9. $\lambda[i] \leftarrow \kappa \cdot \lambda^{(0)}[i]$;
10. $\Delta w[i,h] \leftarrow \lambda[i] \cdot \Delta w^{(0)}[i,h]$;
11. $B[j] \leftarrow \max(B[j] - \Delta B[j], 0)$;
12. $w[i,h] \leftarrow w[i,h] + \Delta w[i,h]$;
13. $A0[i] \leftarrow \sum_h w[i,h]$;

$U = 5.5, W = 0.45, M = 100$;

Note that the coefficients $w[i,h]/A0[i]$ describe the signal's timbral profile. These coefficients are used for timbre, instrumentation, and genre identification in various other applications of our PCP.

3.5 Detuning Detection

For further use, peak frequencies $f[i]$ are converted into semitones $s[i]$. We use the logarithmic scale in accordance to the MIDI standard (A at 440Hz is note number 69):

$$s(f) = 12\log_2(f/440) + 69; \quad (3)$$

The degree of detuning in a particular piece of music is computed as follows:

$$\Delta s = \arg\min_{\Delta = -0.5, -0.4, ..., 0.4} \sum_{n,i} \left|\text{round}(s[n,i] - \Delta) - s[n,i] + \Delta\right| A0^2[n,i];$$

where n - time index and i - peak index at that moment in time.

3.6 Pitch Class Profile (PCP)

Before mapping of the F0 spectrum into a single octave, a window is applied:

$$w(s) = \left(\tanh\{(s-a)/c+1\}+1\right)\left(\tanh\{(b-s)/c+1\}+1\right);$$

The mapping yields a 12-bin PCP vector, computed (after initialization with zeros for each time slice) by the following process for each peak i:

$$PCP[k[i] \bmod 12] \Leftarrow I[i]\left(1+\cos\{\pi(s[i]-\Delta s-k[i])\}\right)$$

$$PCP[(k[i]+1) \bmod 12] \Leftarrow I[i]\left(1-\cos\{\pi(s[i]-\Delta s-k[i])\}\right)$$

$$I[i] = \left(w(s[i])A0[i]\right)^V,\ k[i] = \text{floor}(s[i]-\Delta s)$$

where $x \Leftarrow y$ is a shortcut for $x \leftarrow x+y$.

We compute two types of PCP: type 1 for use in Optimal Transposition Index (OTI) and type 2 for use in Binary Similarity Matrix, with the following parameters:

Type 1: $a = s(20), b = s(5000), c = 5, V = 1$;

Type 2: $a = s(100), b = s(5000), c = 5, V = 1$;

where $s(f)$ is defined by (3).

3.7 Low-Pass Filtering and Downsampling

Next, the PCP is downsampled, which dramatically reduces computational cost. The new time slice Δt is applied as follows:

$$v^{(0)}[n,k] = \int PCP[n\cdot\Delta t + t, k]w(t)dt;$$

where $w(t)$ is Blackman window with the width $2.1\Delta t$, and n is time index, $k = 0,...,11$. The final vector $v[n,k]$ is obtained by normalizing the downsampled PCP:

$$v[n,k] = v^{(0)}[n,k] \Big/ \sqrt{\sum_{p=0}^{11}\left(v^{(0)}[n,p]\right)^2};$$

3.8 Optimal Transposition Index (OTI)

Let A and B be two pieces of music to be compared. The OTI is computed as follows. For each $OTI = 0,...,11$, a special $Score$ is computed and then the OTI that maximizes the $Score$ is selected. Let v_A be the vector of type 1 from section **3.7** for A, and v_B for B. Let the vector dimensions of v_A and v_B be, respectively, n and m. To compute the $Score$ for the given OTI, we build the similarity matrix

$$S_{i,j} = \begin{cases} 0, \text{ if } |i-j|\Delta t > D \\ \sum_{k=0}^{11} v_A[i,(k+OTI) \bmod 12]v_B[j,k], \text{ otherwise} \end{cases};$$

Based on this matrix, we compute matrices $H[i,j]$ and $L[i,j]$ recursively as follows:

$i = 0,...,n-1$:

$H[i,0] \leftarrow L[i,0] \leftarrow 0,\ P_A[i,0] \leftarrow P_B[i,0] \leftarrow -1$;

$j = 0,...,m-1$:

$H[0,j] \leftarrow L[0,j] \leftarrow 0,\ P_A[0,j] \leftarrow P_B[0,j] \leftarrow -1$;

$i = 1,...,n-1;\quad j = 1,...,m-1$:

1. $H_1 \leftarrow H[i-1,j-1] + 2\left(S_{i-1,j-1}+S_{i,j}\right) + S_{i,j-1} + S_{i-1,j}$;

2. $L_1 \leftarrow L[i-1,j-1] + 6$;

3. $S_2 \leftarrow \left(S_{i-1,j}+S_{i,j}\right)/2,\ S_3 \leftarrow \left(S_{i,j-1}+S_{i,j}\right)/2$;

4. $h_2 \leftarrow \begin{cases} 6\theta/\sqrt{\theta^2+1}, & \text{if } L[i-1,j]S_2 > H[i-1,j] \\ 6/\sqrt{\theta^2+1}, & \text{otherwise} \end{cases}$;

5. $h_3 \leftarrow \begin{cases} 6\theta/\sqrt{\theta^2+1}, & \text{if } L[i,j-1]S_3 > H[i,j-1] \\ 6/\sqrt{\theta^2+1}, & \text{otherwise} \end{cases}$;

6. $H_2 \leftarrow \begin{cases} -\infty, & \text{if } P_A[i-1,j] = i-2 \wedge P_B[i-1,j] = j \\ H[i-1,j]+S_2h_2, & \text{otherwise} \end{cases}$;

7. $H_3 \leftarrow \begin{cases} -\infty, & \text{if } P_A[i,j-1] = i \wedge P_B[i,j-1] = j-2 \\ H[i,j-1]+S_3h_3, & \text{otherwise} \end{cases}$;

8. $L_2 \leftarrow L[i-1,j]+h_2,\ L_3 \leftarrow L[i,j-1]+h_3$;

10. $Z_{MAX} \leftarrow \arg\max_{Z=1,2,3} H_Z/L_Z$;

11. $\{H,L,P_A,P_B\}[i,j] \leftarrow \begin{cases} \{H_1,L_1,i-1,j-1\}, \text{if } Z_{MAX}=1 \\ \{H_2,L_2,i-1,j\}, \text{if } Z_{MAX}=2 \\ \{H_3,L_3,i,j-1\}, \text{if } Z_{MAX}=3 \end{cases}$;

Then the *Score* is computed:

$$Score = \max H[i,j]/L[i,j] \text{ on } (i,j) \in P \wedge L[i,j] > gL_{MAX};$$

where $L_{MAX} = \max_{(i,j)\in P} L[i,j]$;

$$P = \{(1,m-1),...,(n-1,m-1)\} \cup \{(n-1,1),...,(n-1,m-1)\};$$

The parameters are selected as follows:

$D = 180\,\text{sec.}, \theta = 2, g = 0.8$.

3.9 Binary Similarity Matrix

Let v_A be the type 2 vector (as described in **3.7**), for A, v_B, for B. The cyclical mapping is represented by the formula

$$c[i, j, p] = \sum_{k=0}^{11} v_A[i, (k + p) \bmod 12] v_B[j, k], \quad p = 0,...,11;$$

Based on this, the Binary Similarity Matrix is computed:

$$S_{i,j} = \begin{cases} 1, \text{ if } |i - j| \Delta t \le D \wedge \arg\max_{p=0,...,11} c[i, j, p] = OTI \\ 0, \text{ otherwise} \end{cases};$$

3.10 Dynamic Programming Local Alignment

Further processing of the Binary Similarity Matrix is performed as described in [3].

4. RESULTS

We have emulated several MIREX 2008 test environments using a ground truth database of over 2300 songs in classical, rock, pop, hip hop, metal, country, and electronica genres, including over 40 cover sets (with between 11 and 28 songs in each). **Table 1** lists average results from ten runs on ten different test sets.

Measure	Range	Result
Total number of covers identified in top 10	0-3300	2300.7
Mean number of covers identified in top 10 (average performance)	0-10	6.9718
Mean (arithmetic) of Avg. Precisions	0-1	0.7177
Mean rank of first correctly identified cover	1-990	5.1358

Table 1**. Average results of 10 test runs**

5. FURTHER WORK

Our analysis of the test results shows that when a cover song follows the original song's segment structure, its identification precision is very close to 1. This points to segmentation (and segment-wise similarity matrix) as the next obvious source of dramatic gains in precision. Another direction is fine-tuning the parameters, especially Δt и V. Lower Δt will increase precision but also increase computational cost. The parameter V defines relative weights of harmonic and melodic component and should be tuned on large ground truth collections of cover songs.

6. A NOTE ON THE MIREX EVALUATION SETUP

The current evaluation setup at MIREX is vulnerable to test results inflation without any increase in actual cover identification precision. The mere knowledge of the way the Cover Song Database is structured (330 cover songs in 30 buckets, 670 control songs that are not covers of any others) is sufficient to effectively eliminate the non-covers from consideration. The precision can be further "improved" by multidimensional scaling analysis (via the k-means procedure, for example) of the remaining 330 songs, placing them in 30 clusters with 11 songs each. **Table 2** lists the results of our algorithm after applying these "improvements".

Measure	Range	Result
Total number of covers identified in top 10	0-3300	2615
Mean number of covers identified in top 10 (average performance)	0-10	7.924
Mean (arithmetic) of Avg. Precisions	0-1	0.8317
Mean rank of first correctly identified cover	1-990	2.4355

Table 2.**Average results of 10 test runs after taking advantage of vulnerabilities in evaluation setup**

7. REFERENCES

[1] Egorov, A. *IF-F0 based Pitch Class Profile extraction for Harmony Detection* (unpublished), Moscow, 2007.

[2] Goto, M. *A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals*, Speech Communication, Vol. 43, No. 4. (September 2004), pp. 311-329.

[3] Serrà, J. *Music similarity based on sequences of descriptors: tonal features applied to audio cover song identification*. Master's thesis, MTG, Universitat Pompeu Fabra, Barcelona, Spain, 2007.