Content-based Music Information Retrieval
using Pitch Histogram of Band Pass Filter Signal

Mansoo Park
School of Engineering, Information
and Communications University, Korea
Information Retrieval & Mining
Research Team, Konan Technology Inc., Korea
mansoo@icu.ac.kr

Chuleui Park, Hoi-Rin Kim
School of Engineering, Information
and Communications University, Korea
{pce337, hrkim}@icu.ac.kr

Sang-Hyuk Im, Dong-Ho Shin, Seung Hyun Yang
Information Retrieval & Mining
Research Team, Konan Technology Inc., Korea
{mdbhyuk, dhshin, yan}@kantechnolog.com

ABSTRACT
Recently, content-based techniques using a signal itself as a query have been required to efficiently identify an unknown music signal. This paper addresses a solution for “query by example” such as the content-based music information retrieval. In this work, only a few seconds’ duration of the music signal is used to recognize the unknown polyphonic music by matching it to pre-indexed audio references. Here, to consider a real-world situation, the query signal is recorded through a general microphone when music sound is played by MP3 player. Thus, we propose a robust and efficient identity matching method using melody pattern so that the distortion in frequency domain is reduced. The experiment shows that the proposed technique is generally effective in the case of retrieval for a query signal recorded in real-world condition. Also, it can be successfully applied to music identification with only a few seconds’ duration of the unknown music signal.

Categories and Subject Descriptors
I.5.4. [Pattern Recognition]: Applications – Signal processing.

General Terms
Algorithms.

Keywords
Non-harmonic signal filtering, Pitch histogram, Melody pattern, IFCR.

1. INTRODUCTION
Lately, the importance of information retrieval has been raised by the use of the web site as a portal site. Most of the information retrieval systems focus on text-based query. However, we can imagine a situation that we want to get the information of a song from a loudspeaker in any place. The problem may be solved by describing the characteristics of the content using the specific features such as the audio features given in the MPEG7 audio descriptor [3].

Traditionally, most of publications related to content-based music information retrieval discuss song identification by matching the stochastic pattern based on histogram. The histogram-based stochastic pattern may be organized by clustering audio feature vectors based on VQ (Vector Quantization) [4][5]. In general, they use SFM (Spectral Flatness Measure), SCF (Spectral Crest Factor), or MFCC (Mel-Frequency Cepstral Coefficient) as the features to describe audio characteristics. In real-world condition, however, recording audio signal for a query is distorted by adverse effects of electrical device (e.g. microphone, sound device, or loudspeaker) and background noise. Thus, spectral features are not efficient to robustly match the similarity relation between original audio signal (e.g. mp3 files or audio CDs) and the distorted one.

In this work, we propose a method that a stochastic model of the melody pattern is used to identify the polyphonic music signal. Here, the pitch histogram is applied to build the melody pattern model since the pitch information is very effective in representing the melody pattern of the music signal. Also, pre-processing by
BPF (Band Pass Filtering) and non-harmonic signal filtering are required to reduce an adverse effect of signal distortion.

The next section will discuss the pre-processing to reduce distortion effect. Section 3 describes the pitch histogram to model the melody pattern. Experiments in real environment are then discussed in Section 4.

(a) The spectrogram of a query signal recorded thru a microphone from a loudspeaker.

(b) The spectrogram of the original music signal.

Figure 1. The spectrogram of a query signal and its original music sound.

2. PRE-PROCESSING

2.1 Stage 1: BPF (Band Pass Filtering)

Recent loudspeaker systems have a woofer to amplify low frequency sound. Moreover, a typical inexpensive microphone cannot perfectly capture both the low frequency signal and very high frequency signal. Background noise can be added much more in low frequency region since noise source is relatively far from the microphone than music signal source. Figure 1 shows the spectrogram distinction between a query signal and an original music signal. Here, a query signal is recorded through a general inexpensive microphone toward not a woofer but 2-channel loudspeakers in a 2.1 channel speaker systems. As shown in the figure 1, a query signal has lower energy than that of the original music signal from the mp3 file in frequency domain.

Pitch information to model the melody pattern suffers from adverse effect by causes mentioned above. Thus, we apply a BPF before pitch detection. Here, the Butterworth filter or the Chebyshev filter could be used to cut off low-frequency signal and high-frequency signal irrelevant to the pitch information. In this work, the Butterworth filter is more appropriate to detect pitch information since the Chebyshev filter has a ripple in the pass band region.

2.2 Stage 2: Non-harmonic signal filtering

To find the melody pattern as pitch histogram, non-harmonic frame should be filtered off because pitch information exists only in quasi-periodic signal and the periodicity of signal can be dwindled by noise. However, accurate detection of periodic signal requires many computations in real implementation.

In this work, we do not focus on fine pitch detection since somewhat coarse pitch detection would be tolerant to organize typical melody pattern. For that reason, harmonic or non-harmonic frame can be simply estimated by using pre-determined pitch interval in this process.

3. PITCH HISTOGRAM

ACF (AutoCorrelation Function) is a well-known pitch detection algorithm picking the peaks of autocorrelation function in time domain. The equation for ACF is given as

$$\phi(\tau) = \frac{1}{K} \sum_{k=0}^{K-1} x(k) x(k-\tau)$$

which computes the correlation between a signal and itself with an offset (\(\tau\)) in time domain. It has the maximum peak at the fundamental period since it would highly correlate with itself at the fundamental period if a signal had periodicity. An alternative is AMDF (Average Magnitude Difference Function) defined as the following equation.

$$\psi(\tau) = \frac{1}{K} \sum_{k=0}^{K-1} |x(k) - x(k-\tau)|$$
This function tends to have the valleys in the peaks of ACF. In this paper, a modified ACF with $c = 1$ which is a combination of ACF and AMDF is used to estimate the pitch period more robust in noise environment [2]. Its function is defined as

$$\varphi(\tau) = \frac{\phi(\tau)}{\psi(\tau) + c} \quad (3)$$

Here, $D$ is the number of histogram bins, $o_n$ is the probability value of the $n^{th}$ histogram bin for the query pattern, and $y_n$ is that of the $n^{th}$ histogram bin for the reference pattern.

4. EXPERIMENTS

4.1 Audio Data

For experiments in real environment, a query signal to identify a song was captured by using an inexpensive microphone such as a general pin or stand microphone which was apart from a 2.1 channel speaker of mp3 player at a distance of 10 ~ 20 cm. The query signal has the 8 seconds’ duration and is converted to standard PCM format which is sampled at 44.1 kHz and quantized with 16 bits in mono channel. Music items for references consist of 1,000 popular songs which have mp3 format converted from audio CDs. They include various genres such as rock/ballad, pop/dance, rap, folk, and so on. Query data were captured from 50 songs which were played at randomly clicked position.

The audio signal was parameterized by texture window of 8 seconds’ duration and hopped with 2 seconds’ rate. And pitch detection was performed every frame with 8 ms rate.

4.2 Performance Evaluation

To evaluate the proposed technique, the query data consist of 4 types of set according to the device and the recording environments.

- Set I: Directly cropping mp3 files.
- Set II: Using a pair of stand microphone and 2.1-channel speakers in very quiet environment.
- Set III: Using a pair of stand/pin microphones and 2.1-channel speakers in noisy environment with TV sound and human voice.
- Set IV: Using a pair of stand/pin microphones and laptop speaker in noisy environment with TV sound, human voice, and sporadic noise. In addition, some cases are overflow into the amplitude range of 16 bits PCM by very loud music sound.

As a histogram distance measure, IFCR (Intra-Feature Component Ratio) is defined to compute the difference between pairs of pitch histogram patterns, since the IFCR is more effective to compare the contours of the pdf of pitch histogram. In the following equation, the closer the IFCR is to 1, the more similar the pairs are.

$$IFCR = \prod_{n=1}^{D} \frac{\min[o_n, y_n]}{\max[o_n, y_n]}$$

Figure 2 shows the process of the pitch histogram modeling with the proposed pre-processing. Absolutely, the pitch histogram should be normalized to organize the stochastic melody pattern. In this paper, band pass filter is cut off from 4 kHz to 12 kHz and non-harmonic frame filter has cut off from 62.5 Hz to 1.5 kHz.

As a histogram distance measure, IFCR (Intra-Feature Component Ratio) is defined to compute the difference between pairs of pitch histogram patterns, since the IFCR is more effective to compare the contours of the pdf of pitch histogram. In the following equation, the closer the IFCR is to 1, the more similar the pairs are.

$$IFCR = \prod_{n=1}^{D} \frac{\min[o_n, y_n]}{\max[o_n, y_n]}$$

To say nothing of, all microphones are omni-directional. Also, all cases of the distance between microphone and speakers are apart...
around 10 ~ 20 cm, and the distance between microphone and noise source is around 2 ~ 5 meters.

### Table 1. Recognition performance with/without pre-processing. (%)

<table>
<thead>
<tr>
<th>Query Set</th>
<th>None</th>
<th>Pre-Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set I</td>
<td>100</td>
<td>99.8</td>
</tr>
<tr>
<td>Set II</td>
<td>23.9</td>
<td>96.0</td>
</tr>
<tr>
<td>Set III</td>
<td>17.8</td>
<td>90.2</td>
</tr>
<tr>
<td>Set IV</td>
<td>10.3</td>
<td>78.0</td>
</tr>
</tbody>
</table>

### Table 2. Recognition performance according to the number of histogram bins. (%)

<table>
<thead>
<tr>
<th>Query Set</th>
<th>145 bins</th>
<th>219 bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set I</td>
<td>99.8</td>
<td>100</td>
</tr>
<tr>
<td>Set II</td>
<td>96.0</td>
<td>98.0</td>
</tr>
<tr>
<td>Set III</td>
<td>90.2</td>
<td>93.2</td>
</tr>
<tr>
<td>Set IV</td>
<td>78.0</td>
<td>82.7</td>
</tr>
</tbody>
</table>

Recognition results for the 1-best match of the proposed pre-processing according to increasing the number of histogram bins are shown in Table 1 and 2. The results shown in the table 1 demonstrate the superiority of the proposed pre-processing when the number of histogram bins is equal to 145. As shown in the above tables, we can observe that non-harmonic noise could be smoothly filtering off because of identity matching by pitch-based pattern, and BPF could be filtering out some background noise relatively far from microphone, which is localized into low-frequency area. In case of the Set IV, the recognition accuracy might be degraded by magnitude clipping caused by overflow because pitch was estimated in time domain. Also, the higher the pitch histogram resolution is, the superior the recognition rate becomes.

### 5. CONCLUSION

From the experiments, we observed that the proposed technique is generally effective in the case of retrieval with a query signal recorded in real-world condition. Also, it can be successfully applied to music identification with only a few seconds’ duration of the unknown music signal. In this work, it is shown that the proposed pre-processing can reduce adverse effects of electrical devices and various noise environments causing linear or non-linear distortions to audio signal.

Even if only the weighted ACF is mentioned to estimate the pitch in this paper, we observed that the performance results according to a well-known pitch detection technique such as ACF, AMDF, or NCC (Normalized Cross Correlation) are very similar or a little lower as compared with Table 2. But it is important that is more proper pitch detection technique for polyphonic music signal.

However, we still remains many challenges, performance comparison in contrast with spectral features-based music identification, how to hybrid spectral features and pitch histogram, and how to verify out-of-set materials.

### 6. REFERENCES


