Two-layer Distance Scheme in Matching Engine for Query by Humming System

Feng Zhang, Yan Song, Lirong Dai, Renhua Wang

University of Science and Technology of China, iFlytek Speech Lab, Hefei
zhangf@ustc.edu, {songy, lrdai, rhw}@ustc.edu.cn

Abstract: In query-by-humming system, minimizing the impact introduced by the pitch tracking errors and humming errors is always a difficult problem. In this paper, we propose a two-layer distance scheme instead of the global DTW distance in traditional QBH system, which is motivated by the people’s perception of humming. The local distance measure tries to find the correct part of humming, which is robust to the hummed error and the noise. Also, the approach to combine the local and global distance measure is proposed. The experiment of our QBH system on the mobile phone channel shows that the performance is greatly improved.

Keywords: Query By Humming, Dynamic Time Warping, Two-layer Distance

1 Introduction

With the development of the multimedia technology, the amount of music data is increasing rapidly everyday. To deal with the overwhelming amount of music data, a Music Information Retrieval (MIR) system is necessary. Query by humming (QBH) is an interesting branch of MIR. By using the QBH system, people can find the song by humming instead of having their name in remembrance.

The earlier QBH system was proposed by Ghias [1], in which the hummed tune was firstly converted into a string in a 3 letter alphabet, and then a string matching algorithm was used to search converted symbols. In later QBH systems, the melody was converted into a series of pitch values. In order to prevent the performance degradation introduced by the humming errors, some methods based on the Dynamic Programming (DP) were proposed to calculate Dynamic Time Warping (DTW) distance (by Jang [2]) or Edit Distance (by Lemstrem [3]) between the hummed tune and the MIDI template in database. The DTW algorithm is widely applied in QBH systems, although there are other methods for matching engine such as Linear Alignment Matching (LAM) proposed by Wu [4] etc.

However, the DTW algorithm in most QBH systems only measures the overall distance of the whole hummed tune, which may not be suitable for the situation that there are a few of pitch tracking errors and humming errors. The performance of such system is still far from satisfactory on the mobile phone channel or noisy environment. But people can well perceive the hummed tune merely according to the
correct part of humming. Motivated by this intuition, in this paper, we propose a novel two-layer distance measure scheme, in which a local DTW distance is calculated to find the corresponding correct hummed part, and the global DTW distance is used to measure the overall similarity. These distance measures are combined into the final criterion for retrieve the right pieces in melody database. Furthermore, a QBH application system using the two-layer distance measure scheme is constructed for querying from the melody database consisted of the 15000 clips extracted from about 1000 MIDI files, and the querying humming is collected from the mobile telephone channel. The experiment results show that the performance of our QBH system is greatly improved (detailed in section 4).

This paper is organized as follows. Section 2 gives the overall description of the QBH system. Section 3 details the two-layer distance scheme. And experiment results are shown in section 4, followed by the conclusion and future work in section 5.

2 Overview of the Proposed QBH System

The proposed QBH system consists of 3 components including Melody Extraction, Matching Engine and Melody Database Construction, as shown in figure 1. The acoustic input of the QBH system is the hummed tunes recorded from mobile phone on the CDMA channel (China Unicom) in a common office environment, and the sample rate is 8000, 16-bit resolution.

2.1 Melody Extraction

Melody extraction primarily consists of 2 steps including Pitch Extraction and Note Segmentation. After extracting the pitch features, the hummed tune clips into notes according the pitch and intensity features.

It is worth noting that the pitch of humming is difficult to be extracted precisely in mobile telephone channel. The pitch errors such as double pitch and half pitch may greatly degrade the system performance. Here, the autocorrelation algorithm proposed by Boersma [5] is used to extract the pitch. Although this algorithm works well in our experiment, there are also some double pitch errors mainly caused by the strong second harmonic components.

2.2 Matching Engine

In matching engine, a two-layer distance scheme is proposed, which will be detailed in Section 3.

2.3 Melody Database Construction

The database is consisted of 1000 MIDI files. These MIDI files are downloaded from Internet including Chinese pop songs and western pop songs. Their main tracks are
picked up manually. Because people are always humming only at several boundaries of the MIDI file, each MIDI file is divided into several clips by the local boundary detection model proposed by Cambouropoulos [6]. In our database, the 1000 MIDI files are divided into about 15000 clips automatically.

3 Matching Engine Using the Two-layer Distance Scheme

In this section, we first briefly introduce the matching engine in classical QBH system. Then, based on the analysis of this scheme, the matching engine using a novel two-layer distance measure is proposed in detail.

3.1 Matching engine in classical QBH system

To account for tempo variation of each individual user, dynamic time warping (DTW) is used to compute the warping distance between the input note sequences and MIDI note sequences in the database. To deal with key variation, relative pitch value is used.
Note durations are also used and measured by the Inter Onset Interval (IOI) ratio, which gives the ratio between the current and the next note duration.

A classical calculation pattern is used to compute the melodic distance \( D(Q, M) \) between the query note sequences \( Q = q_1 q_2 \ldots q_m \) and the MIDI note sequences \( M = m_1 m_2 \ldots m_n \) by filling the matrix \( D = [d_{ij}]_{m \times n} \). Each entry \( d_{i,j} \) denotes the minimal melodic edit distance between two sequences \( q_1 q_2 \ldots q_i \) and \( m_1 m_2 \ldots m_j \), which is calculated as follows,

\[
d_{i,j} = \min \begin{cases} d_{i-1,j} + w(q_i, \emptyset) & \text{(deletion)} \\ d_{i,j-1} + w(\emptyset, m_j) & \text{(insertion)} \\ d_{i-1,j-1} + w(q_i, m_j) & \text{(replacement)} \\ d_{i-k, j-1} + w(q_{i-k+1} \ldots q_i, m_j), 2 \leq k \leq i & \text{(consolidation)} \\ d_{i-1, j-k} + w(q_i, m_j-k+1 \ldots m_j), 2 \leq k \leq i & \text{(fragmentation)} \end{cases} \tag{1}
\]

where \( w() \) is the distance between the query note \( q_i \) and MIDI note \( m_j \) for different situations such as deletion, insertion etc., which can be termed as \( w_{ij} \).

The end point constraints were loosed to cope with different starting or ending points. We choose \( n = a \times m \) (\( a \) is a tunable parameter, for example, \( a = 1.4 \)). Since the query note sequences is same for each MIDI clip in the database, the final distance value \( d_{\text{final}} \) can be calculated from the \( m \)-th row of the matrix \( (d_{0 \ldots m,0 \ldots n}) \) as follows:

\[
D(Q, M) = \min (d_{m,k}), \quad 1 \leq k \leq n \tag{2}
\]

For each MIDI clip \( M_i \) (\( i = 1, 2 \ldots N \)) in database, where \( N \) is the size of database, the melodic distance \( D(Q, M_i) \) is calculated as equation (2). And by sorting \( D(Q, M_i) \), the rank list is obtained.

In practical situation, there are a few of pitch tracking errors and humming errors, which may degrade the querying performance heavily. However people can still discriminate the tune based on the correct hummed part. According to this observation, the performance of the QBH system can be improved by introducing the local distance measure, which will be detailed next.

### 3.2 Matching engine using two-layer distance measure scheme

In this section, we propose a two-layer distance measure scheme. The global distance measure \( D(Q, M) \), termed as \( D_{\text{global}} \), is calculated as aforementioned, and the local distance measure \( D_{\text{local}} \), is calculated as follows.

#### 3.2.1 Local Distance Measure

In calculation of the \( d(i,j) \) as equation (1), the best path \( P_{\text{best}}(i,j) \) from start point \( p(0,0) \) to end point \( p(i,j) \) is obtained. The point sequence of \( P_{\text{best}}(i,j) \) is termed as \( s_1 s_2 \ldots s_{k(i,j)} \)
where \( k(i,j) \) is the number of the point sequence. For further calculating the local distance measure according to current \( P_{\text{best}}(i,j) \), the local path in \( P_{\text{best}}(i,j) \) needs to be determined. We can set the point \( p(i,j) \) as the end point of the local path, and trace back along the \( P_{\text{best}}(i,j) \) to find the appropriate start point, and the corresponding local distance is termed as \( d_{\text{local}}(i,j) \). After obtaining \( d_{\text{local}}(i,j) \) of each \( P_{\text{best}}(i,j) \), the final local distance \( D_{\text{local}} \) is the minimal of the matrix \([d_{\text{local}}]_{m\times n}\). This approach can be shown to be equivalent to find the best local path by greedy method.

In addition, if the length of the path is too short, the corresponding melody may be meaningless. Here, the low-bound threshold of path length is heuristically set to 5 notes in our experiments.

The start point \( p_{\text{start}}(i,j) \) of the path \( P_{\text{best}}(i,j) \) and the \( d_{\text{local}}(i,j) \) are calculated in following procedure.

**End point searching procedure:**

**Initialization:**

\[
p_{\text{start}}(1,1) = p(1,1) \quad (3)
\]

\[
d_{\text{local}}(1,1) = \infty \quad (4)
\]

**Iteration:**

1. Given the end point \( p(i,j) \), the point \( s_{k(i,j)-1} \) is the point just before \( p(i,j) \) in the path \( P_{\text{best}}(i,j) \). The start point of \( P_{\text{best}}(s_{k(i,j)-1}) \) termed as \( s_q \) has already been obtained before.

2. If the length of the path with start point \( s_q \) and end point \( p(i,j) \) in \( P_{\text{best}}(i,j) \) is less than 5 notes, then \( p_{\text{start}}(i,j) = s_q \).

3. If the length of the path with start point \( s_q \) and end point \( p(i,j) \) in \( P_{\text{best}}(i,j) \) is more than 5 notes, \( d_{\text{local}}(i,j) \) is calculated as follows:

\[
d_{\text{local}}(i,j) = \begin{cases} 
  d_{\text{increase}}(s_q,s_q + 1,\ldots,s_{k(i,j)}) & \text{if } p_{\text{start}}(i,j) = s_q \\
  \min \left\{ d_{\text{increase}}(s_q + 1,s_q + 2,\ldots,s_{k(i,j)}) \quad \text{if } p_{\text{start}}(i,j) = s_q + 1 \\
  \min \left\{ d_{\text{increase}}(s_q + 2,s_q + 3,\ldots,s_{k(i,j)}) \quad \text{if } p_{\text{start}}(i,j) = s_q + 2 \right. \right. \\
\right. 
\end{cases} 
\quad (4)
\]

\[
d_{\text{increase}}(s_q,s_q + 1,\ldots,s_{k(i,j)}) = \frac{\text{sum}(w_q,w_q + 1,\ldots,w_{k(i,j)})}{\text{sum}(t_q,t_q + 1,\ldots,t_{k(i,j)})} \times \left(1 + \frac{l}{\text{sum}(t_q,t_q + 1,\ldots,t_{k(i,j)})}\right) 
\quad (5)
\]

(where \( w_q \) is the \( w \) part of the point \( s_q \) in equation (1) and \( t_q \) is the duration of \( s_q \)'s corresponding note. \( l \) is a tunable parameter, which is set to 10.0 in our experiment.

To reduce the computation load, the length of the new local path is around the length of the old local path in equation (4).

**End Procedure**
3.2.2 Matching Criterion

Given a two-layer distance, $D_{\text{global}}$ and $D_{\text{local}}$, it is necessary to determine which one is more reliable. Note sequences’ posterior probability is proposed to solve this problem.

For each MIDI clip $M_i$, note sequences’ posterior probability $P(M_i|T)$ can be calculated as follows based on Bayes Theory:

$$ P(M_i | N) = \frac{P(N | M_i) \times P(M_i)}{P(N)} \quad (6) $$

where $P(N)$ is constant for each MIDI clip, and the prior probability $P(M_i)$ is assumed to be the same. The class conditional probability $P(N|M_i)$ can be used instead of posterior probability $P(M_i | T)$. For each MIDI clip, $P(N|M_i)$ can be calculated from distance $D_i$ ($D_{\text{global}}$ or $D_{\text{local}}$) as shown in equation (6),

$$ P(N | M_i) = \frac{1}{\text{SUM}(1 / D_k)} \quad (1 \leq K \leq \text{number of MIDI clips}) \quad (7) $$

Using equation (7), two prior probabilities: $P(N|M_i)_{\text{global}}$ and $P(N|M_i)_{\text{local}}$ can be obtained. The layer with higher prior probability is assumed to be more reliable and the corresponding rank list is taken as a result.

4 Experiment

The melody database is detailed in section 2.3. In our experiment, 16 people including 9 men and 7 women hum 200 pieces of randomly selected songs to test the two-layer distance scheme. The average searching time for one piece is one half of the humming time on the platform of PC(P4 1.4G). The result is shown in Table 1. The accuracy for rank 1 is 74% compared with 58% for the base DTW method. The mean reciprocal rank (MRR) rises from 0.6563 to 0.7971.

Table 1. Query-by-humming experiment

<table>
<thead>
<tr>
<th></th>
<th>Top1</th>
<th>Top3</th>
<th>Top5</th>
<th>Top10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>the base DTW method</td>
<td>58.04%</td>
<td>69.93%</td>
<td>73.42%</td>
<td>79.72%</td>
<td>0.6563</td>
</tr>
<tr>
<td>Two-layer Distance Scheme</td>
<td><strong>74.12%</strong></td>
<td><strong>82.23%</strong></td>
<td><strong>85.31%</strong></td>
<td><strong>88.81%</strong></td>
<td><strong>0.7971</strong></td>
</tr>
</tbody>
</table>

5 Conclusion and Future Work

In this paper, we propose a two-layer distance scheme based on the DTW algorithm in our QBM system, which follows people’s perception to search the hummed tune. The experiment has shown that our approach improves the performance of the QBM
system. The accuracy of rank 1 is 74%, while for the accuracy of rank 1 in traditional DP method is 58%.

There are still some issues to be investigated as future extension. First, we only use distances of two layers. An approach to use multi-layer distance scheme needs to be exploited to further improve the performance of QBH system. Second, the more reliable local distance measure needs to be found in the future work.

Acknowledgments. The work is supported by iFlytek Co., and this QBH system has been used in the service of China Unicom Co.

References