2010 International Conference on Signal Acquisition and Processing

A Music Information System Based on Improved Melody Contour Extraction

Nattha Phiwma and Parinya Sanguansat
Department of Information Technology
Rangsit University
Pathumthani, Thailand
phewma@hotmail.com, sanguansat@yahoo.com

***Abstract*—In this paper, we propose a new melody contour extraction technique to improve Query-by-Humming. A critical issue of humming sound is noise interference from both environment and acquisition instrument. Furthermore most users are not professional singers so they cause the other query problems about variation of pitch and timing. Advantage of a proposed technique can reduce noise whereas makes pitch smoothing. Moreover, this method is used to enhance the query accuracy. Our technique consists of three steps as follows: Firstly, the melody contour is extracted from humming sound by Subharmonic-to-Harmonic Ratio (SHR). Subsequently, the melody contour is filtered and smoothed by statistical approach. Finally, Dynamic Time Warping (DTW) is applied to melody contour, for similarity measurement between humming sound and melody sequence. We use a new method to extract melody contour from pitch. Experimental results show that our proposed technique can perform better than traditional method.**

***Keywords*—Query-by-Humming; melody contour; Dynamic Time Warping; pitch; Subharmonic-to-Harmonic Ratio**

I. Introduction

Currently, most people are interested in music both listening and singing, it seems to be an essential part of our lives. People try to find ways to entertain themselves and Karaoke seems to be one of them. The only problem when it comes to this kind of entertainment is that when users can not come up with the names of the songs, however users can only type different keywords (titles, singers, etc.). It is inconvenient for them. Normally, the user always remembers the melody or rhythm and can hum the melody to retrieve the song. At present, there is one system that helps the searching become easier, which is called Query-by-Humming (QbH) system. This system allows users to retrieve a song via simply humming some part of the song and will show the result by different names of songs. The QbH increase the usability of a music retrieval system meanwhile the user receive convenient.

Many researchers have focused on how to improve QbH for measuring similarity of humming sound. First of all, humming sound must be extracted to pitch by using many methods such as autocorrelation, maximum likelihood cepstrum analysis [1] or Subharmonic-to-Harmonic Ratio (SHR) [2]. There are three frameworks of QbH, based on feature types: (1) the technique based on string matching [1], [3], [4]; (2) the technique based on continues pitch contour matching [5], [6], [7]; (3) the technique based on spectral [8], [9], [10], [11]. These techniques can be classified according to feature representations, i.e. string sequence, time and frequency, and spectrogram.

The first framework, most previous methods were focused on matching part of song retrieval systems. The technique based on string matching is used method of melody and song retrieval from a music database. As DTW can be used for measuring sound signals, it allows local flexibility in aligning time series [12], [13]. Probably the most prevalent method [1], [3], [4] of melodic representation in QbH systems, the three alphabets were used to display whether a note in sequence is up (U), down (D), or the same (S) as the previous note. Then melodic representation will be analyzed by above technique. N-grams is another approach, which is widely used in text retrieval and applied to retrieve songs in music system [4], [14], [15], [16]. It is particularly effective for short queries and manual queries not for automatic queries [17]. In addition, string matching based on statistical models including Hidden Markov Models (HMMs) in [14], [18], [19]. This approach uses a combination of HMMs for sequence estimation and DTW for hierarchical clustering [20].

Subsequent to this technique is continuous pitch contour. From the above techniques, the discriminant information may be of lost and the changing of sounds is not different. We can look to probabilistic models being used in speech recognition and production as possible inspiration. Melody contour or pitch contour used in [5], [6], [7], [21] which is a time series of pitch values, represents melody content without using explicit music notes. The method above is based on time and frequency domain analysis which cannot be processed at the same time.

To the best of two domains, there is a technique that both domains possibly work together. According to time and frequency domain analysis, spectral features is the technique that we have classified. In some works, a feature extraction method of the sound recognition framework is used spectrum via spectral basis functions [8], [9], [10], [11]. Spectrogram has been widely used as one of the method for time-varying spectral analysis which is important in many applications such as radar, sonar, speech, geophysics and biological signals [22].

Pitch and fundamental frequency are important feature therefore it must be extracted pitch. A pitch determination algorithm (PDA) based on Subharmonic-to-Harmonic Ratio (SHR) is developed in the frequency domain and describe the amplitude ratio between subharmonics and harmonics [2]. For our system, we have implemented pitch tracking using SHR.

978-0-7695-3960-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSAP.2010.8

Median filter is well known for being able to remove impulse noise and smoothing signal [23], [24]. In [25] described desirable signal properties for signals used in it which if the real signal has added noise, then it may or may not be possible to remove the noise by filtering. It show how some types of noise can be removed the noised by median filtering and how other types cannot be removed. Median filter is used for smoothing pitch in QbH system [26]. Therefore, our system we decided to reduce noise a part of pitch by it.

Among the various methods in music processing, we consider the selection of melody contour to improve QbH. Noises from environment and variation of pitch and time are the problems of humming sound. The melody contour is extracted by our proposed technique. DTW is applied to melody contour, for similarity measurement between humming sound and melody sequence.

This paper is organized as follows: Describing the concept of pitch tracking in Section II and Dynamic Time Warping in Section III. Melody Contour Extraction technique is proposed in Section IV. In Section V, experimental results are presented. Finally, conclusion is in Section VI.

II. Pitch Tracking

In this section, the concept of pitch tracking is described how the system is converted into a sequence of relative pitch transitions. The concept of pitch is the fundamental frequency that matches what note we hear [1]. Notes can begin and end when pitches have been identified. The pitch detector decides based on the statistical information of pitch models. The detailed of each component of the pitch detector is given below.

Four pitch tracking methods: Autocorrelation, Maximum Likelihood, Cepstrum Analysis and SHR [1], [2]. The most of pitch detection autocorrelation is chosen for implementation pitch tracking [1]. In addition, a pitch determination algorithm (PDA) based on Subharmonic-to-Harmonic Ratio (SHR) is developed in the frequency domain and describe the amplitude ratio between subharmonics and harmonics [27], [2]. For our system, we have implemented pitch tracking using SHR. For each short-term signal, let $A(f)$ represents the amplitude spectrum, and let f_0 and $f_{\max}$ be the fundamental frequency and the maximum frequency of $A(f)$, respectively. Then the sum of harmonic amplitude is defined as

$$SH = \sum_{n=1}^{N} A(nf_0), \tag{1}$$

where N is the maximum number of harmonics contained in the spectrum, and $A(f) = 0$ if $f > f_{\max}$. If the pitch search range is defined $[F0_{\min}, F0_{\max}]$, then N=floor$(f_{\max}/f_{\min})$. Assuming the lowest subharmonic frequency is one half of f_0, the sum of subharmonic amplitude is defined as

$$SS = \sum_{n=1}^{N} ((n-1/2) f_0). \tag{2}$$

Let $LOGA(\bullet)$ denote the spectrum with log frequency scale, then we can represent SH and SS as

$$SH = \sum_{n=1}^{N} LOGA(\log(n) + \log(f_0)). \tag{3}$$

$$SS = \sum_{n=1}^{N} LOGA(\log(n-1/2) + \log(f_0)). \tag{4}$$

To obtain SH, the spectrum is shifted leftward along the logarithmic frequency abscissa at even orders, i.e., log(2), log(4),...log($4N$). These shifted spectra are added together and denoted by

$$SUMA(\log f)_{even} = \sum_{n=1}^{2N} LOGA(\log f + \log(2n)). \tag{5}$$

Similarly, by shifting the spectrum leftward at log(1), log(3), log(5),...log($4N$-1), we have

$$SUMA(\log f)_{odd} = \sum_{n=1}^{2N} LOGA(\log f + \log(2n-1)). \tag{6}$$

Next, A difference function defines as

$$DA(\log f) = SUMA(\log f)_{even} - SUMA(\log f)_{odd} \tag{7}$$

In searching for the maximum value, the position of the global maximum is located and denoted as $\log(f_1)$. Then, starting from this point, the position of the next local maximum denoted as $\log(f_2)$ is selected in the range of $[\log(1.9375 f_1), \log(2.0625 f_2)]$. Equation of SHR is defined as

$$SHR = \frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)}. \tag{8}$$

In case of SHR is less than a certain threshold value, it indicates that subharmonics are weak, so that harmonics are preferred. Thus, f_2 is selected and the final pitch value is $2f_2$. Otherwise, f_1 is selected and the pitch is $2f_1$. In [2], SHR can be effectively used to pitch tracking.

III. Dynamic Time Warping

Due to the tempo variation of length of sequence, we cannot measure the similarity by any tradition distances. Dynamic Time Warping (DTW) is adopted to fill the gap caused by tempo variation between two sequences. For our system, we use DTW to compute the warping distance between the input melody contour and that of each song in database. Suppose that the input melody contour vector (or query vector) is represented by $t(i), i = 1, \ldots, m$, and the reference vector by $r(j), j = 1, \ldots, n$. These two vectors are not necessarily of the same size. The distance in DTW is define as the minimum distance starting from the begin of the DTW table to the current position (i, j). According to the dynamic programming algorithm, the DTW table $D(i, j)$ can be calculated by:

$$D(i,j) = d(i,j) + \min \begin{cases} D(i-2, j-1) \\ D(i-1, j-1) \\ D(i-1, j-2) \end{cases}, \tag{9}$$

where $D(i, j)$ is the node cost associated with $t(i)$ and $r(j)$ and can be defined from the L1-norm as

$$d(i,j) = |t(i) - r(j)|. \quad (10)$$

The best path is the one with the least global distance, which is the sum of cells alone the path. This method exhibits good performance for word speech recognition and QbH in [21].

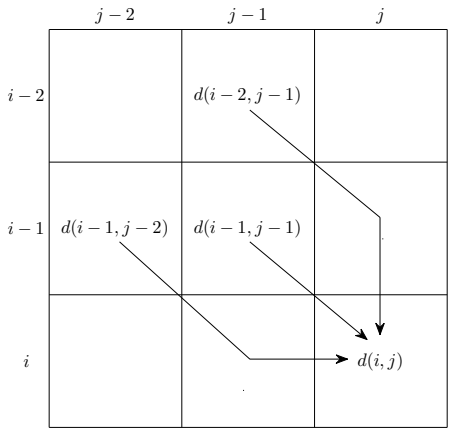

Figure 1. The calculation pattern for the dynamic time warping in the Melody Contour.

IV. Melody Contour Extraction

In this section, our proposed technique for feature extraction in Query-by-Humming (QbH) system is presented. The following algorithm describes how to extract pitch from humming sound to obtain the melody contour.

Algorithm 1 Melody Contour Extraction Algorithm

Require: $\mathbf{p}$, g, T, s
Ensure: $\mathbf{m}$

```
smoothing p by median filter.
initial m_1 ← p_1
N ← length of p
j ← 1
while t ≤ N do
    d = |p_t − p_{t−1}|
    Y ← {y_{t−v}, y_{t−v+1}, ..., y_{t+v−1}, y_{t+v}}
    S_Y ← Standard deviation of Y
    if d > g and S_Y < T then
        m_j ← p_t
    end if
    t ← t + s
    j ← j + 1
end while
return m
```

Let $\mathbf{m}$ represents melody contour and let $\mathbf{p}$ be the pitch. The variables of algorithm are describe as follows: s is the size of window for filtering, g is the gap of pitch difference, T is threshold of standard deviation and v is variance of pitch interval.

This algorithm was designed for feature extraction. The humming sound consists of pitch in several values and also has noise fused in the pitch as shown in Fig. 2(a). Normally, the humming sound is usually reduced noise by median filtering method which makes the signal is better smooth as shown in the Fig. 2(b). However, it usually makes the discriminant information of the signal be lost at the same time. It is also applied for filtering part of signals prior to further processing with small window. We can reduce noise meanwhile the information of the signal is still reserved by our method.

The first step of this method is taking pitch to pass the process of noise filter which uses the median filter in order to make the signal smooth. Then, find the different value of p by comparing with the defined g value by selecting only the value which different value exceed the g value. The value of s is determined in order to apply to find the range of signal that change a little for a while. In other words we discard the signal that change rapidly in short time comparing with this interval. There is the spread around the signal and we need the group of significant signal only. Hence, we find the range of signal which has a little value of the spread when comparing the threshold of standard deviation (T).

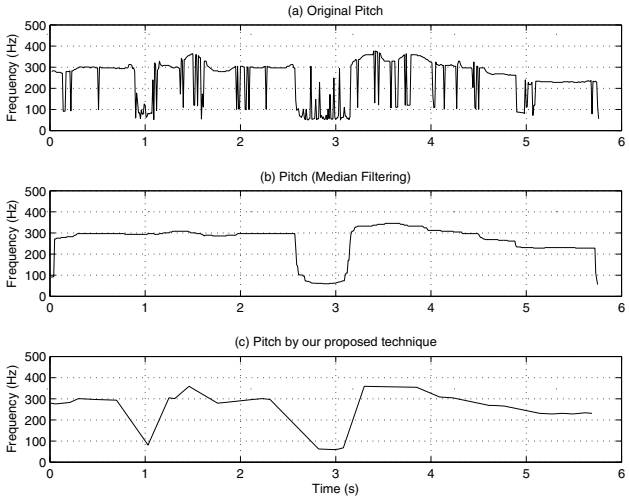

Figure 2. A graph is shown (a) Original Pitch, (b) Pitch (Median Filtering) and (c) Pitch by our proposed technique

From the Fig. 2(c), it can be seen that the pitch which is better smooth. The output of the algorithm melody contour contain significant pitch. Finally, when this technique is applied to retrieval task, it to do retrieval process, the result will be more correct than the traditional method.

V. Experimental Results

Experiments have shown the effectiveness of the system and according to the various conditions, such as number of songs for humming and MIDI in database. The baseline noise reduction is described in detail [26]. In order to show the advantage of our proposed technique, the accuracy is better than use only median filter to reduce noise. The performance

TABLE I
TEST RESULT OF EXPERIMENT WITH 100 TEST QUERIES.

Rank	Method	Number of songs		
		100	*300*	*500*
Top-1	Proposed technique	76	67	66
	Median Filtering	42	24	17
Top-5	Proposed technique	95	89	87
	Median Filtering	80	56	48
Top-10	Proposed technique	96	94	92
	Median Filtering	91	71	61

evaluations include three measurements: top-1 rate, top-5 rate, and top-10 rate.

Our system, there are 100, 300 and 500 MIDI format songs in the database. The test query is humming sound which consists of tunes hummed with *Da Da Da*. We used 100 humming sounds from different people to test our system. The recording was done at 8 kHz sampling rate, mono and time duration 10 seconds, starting at the beginning of song. Pitch of humming sounds are normalized base on logarithmic value and z-score [28]. In our experiment, we set the values of variables such as s, g, and T to 5, 2, and 5 respectively. For median filter, we found that the optimal size of window is 53 to achieve the highest performance.

Experimental results show that top-n rate means the rate of queries that retrieves correct music within top-n rank and can find correct slope-sequence alignment. The comparison of different methods is shown in Table 1 and Fig. 3 that illustrates the performance of accuracy rate of our proposed technique and median filter method. In Table 1, the 100 humming sounds were used to query in 100, 300, and 500 MIDI songs databases. Fig. 3 shows the retrieval accuracies that retrieved 100 humming sounds from 500 MIDI songs database by varying the top-n rank from top-1 to top-25.

Moreover, our technique can reduce the dimension of feature vector, which contains only the significant information. Thus in our experiments, the query time is faster than the conventional one around ten times.

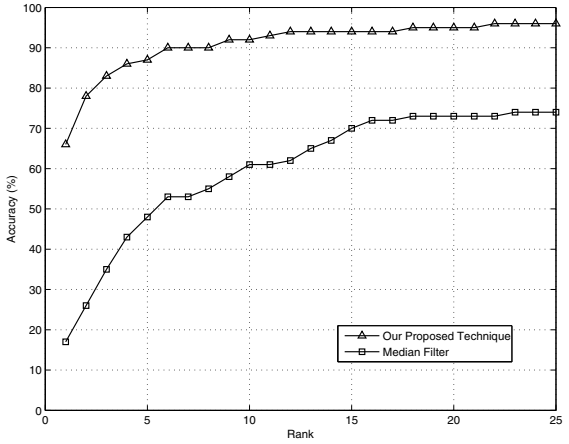

Figure 3. A graph is shown the performance of accuracy rate of our proposed technique and median filter method

VI. CONCLUSION

In this paper, we have proposed a new melody retrieval method by similarity matching of continuous melody contours. We have improved the process of feature extraction from various humming inputs. Our technique offers several advantages: higher accuracy and low complexity. First of all, it can reduce noise meanwhile the discriminant information is extracted. That makes the accuracy improve as shown in our experimental results. Secondly, the query process is faster and consumes lower memory because the dimension of feature vector is smaller than traditional one.

REFERENCES

[1] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith, "Query by humming: musical information retrieval in an audio database," in *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*, New York, NY, USA, 1995, pp. 231–236, ACM.

[2] Xuejing Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proceedings of the IEEE*, 2002, pp. 333–336.

[3] Rodger J. McNab, Lloyd A. Smith, Ian H. Witten, Clare L. Henderson, and Sally Jo Cunningham, "Towards the digital music library: tune retrieval from acoustic input," in *DL '96: Proceedings of the first ACM international conference on Digital libraries*, New York, NY, USA, 1996, pp. 11–18, ACM.

[4] Alexandra Uitdenbogerd and Justin Zobel, "Melodic matching techniques for large music databases," in *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, New York, NY, USA, 1999, pp. 57–66, ACM.

[5] Yongwei Zhu and Mohan Kankanhalli, "Similarity matching of continuous melody contours for humming querying of melody databases," in *of Melody Databases, International Workshop on Multimedia Signal Processing, USVI*, 2002.

[6] Takuichi Nishimura, J. Xin Zhang, and Hiroki Hashiguchi, "Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming," in *Continuous Dynamic Programming, Proc. 3 rd International Symposium on Music Information Retrieval*, 2001, pp. 211–218.

[7] Yongwei Zhu, Mohan S. Kankanhalli, and Changsheng Xu, "Pitch tracking and melody slope matching for song retrieval," in *PCM '01: Proceedings of the Second IEEE Pacific Rim Conference on Multimedia*, London, UK, 2001, pp. 530–537, Springer-Verlag.

[8] Jonathan Foote, Matthew L. Cooper, and Unjung Nam, "Audio retrieval by rhythmic similarity," in *ISMIR*, 2002.

[9] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. International Conference on Multimedia and Expo 2001.*, 2001.

[10] Xiangyang Xue Leon Fu, "A new spectral-based approach to query-by-humming for mp3 songs database," in *World Academy of Science, Engineering and Technology 4 2005.*, 2005.

[11] John N. Gowdyl Sabri Gurbuz and Zekeriyu Tufekci, "Speech spectrogram based model adaptation for speaker identification," in *Proceedings of the IEEE*, 2000, pp. 110–115.

[12] Ada Wai-chee Fu, Eamonn Keogh, Leo Yung Hang Lau, and Chotirat Ann Ratanamahatana, "Scaling and time warping in time series querying," in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*. 2005, pp. 649–660, VLDB Endowment.

[13] Yunyue Zhu and Dennis Shasha, "Warping indexes with envelope transforms for query by humming," in *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 2003, pp. 181–192, ACM.

[14] Roger B. Dannenberg, William P. Birmingham, Bryan Pardo, Ning Hu, Colin Meek, and George Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 5, pp. 687–701, 2007.

[15] Stephen Downie and Michael Nelson, "Evaluation of a simple and effective music information retrieval method," in *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2000, pp. 73–80, ACM.

[16] Yuen-Hsien Tseng, "Content-based retrieval for music collections," in *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1999, pp. 176–182, ACM.

[17] Alexandra Uitdenbogerd and Justin Zobel, "Melodic matching techniques for large music databases," in *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, New York, NY, USA, 1999, pp. 57–66, ACM.

[18] Hsuan-Huei Shih, S.S. Narayanan, and C.-C.J. Kuo, "An hmm-based approach to humming transcription," in *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, 2002, vol. 1, pp. 337–340 vol.1.

[19] Hsuan-Huei Shih, S.S. Narayanan, and C.-C.J. Kuo, "A statistical multidimensional humming transcription using phone level hidden markov models for query by humming systems," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, July 2003, vol. 1, pp. I–61–4 vol.1.

[20] Jianying Hu, Bonnie Ray, and Lanshan Han, "An interweaved hmm/dtw approach to robust time series clustering," *Pattern Recognition, International Conference on*, vol. 3, pp. 145–148, 2006.

[21] Jyh-Shing Roger Jang and Hong-Ru Lee, "Hierarchical filtering method for content-based music retrieval via acoustic input," in *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, New York, NY, USA, 2001, pp. 401–410, ACM.

[22] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, Jul 1989.

[23] J. Astola, P. Haavisto, and Y. Neuvo, "Vector median filters," *Proceedings of the IEEE*, vol. 78, no. 4, pp. 678–689, Apr 1990.

[24] H.-M. Lin and Jr. Willson, A.N., "Median filters with adaptive length," *Circuits and Systems, IEEE Transactions on*, vol. 35, no. 6, pp. 675–690, Jun 1988.

[25] Jr. Gallagher, N. and G. Wise, "A theoretical analysis of the properties of median filters," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 6, pp. 1136–1141, Dec 1981.

[26] Lei Wang, Shen Huang, Sheng Hu, Jiaen Liang, and Bo Xu, "An effective and efficient method for query by humming system based on multi-similarity measurement fusion," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, July 2008, pp. 471–475.

[27] Xuejing Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," in *the 6th International Conference of Spoken Language Processing*, 2000, pp. 676–679.

[28] Hong Quang Nguyen, P. Nocera, E. Castelli, and T. Van Loan, "Tone recognition of vietnamese continuous speech using hidden markov model," in *Communications and Electronics, 2008. ICCE 2008. Second International Conference on*, June 2008, pp. 235–239.