Rangsit Journal of Arts and Sciences, July-December 2011
Copyright © 2011, Rangsit University

RJAS Vol. 1 No. 2, pp. 139-148

An improved note segmentation and normalization for Query-by-Humming

Nattha Phiwma[1*] and Parinya Sanguansat[2]

[1]Faculty of Information Technology, Rangsit University, Patumthani 12000, Thailand
E-mail: phewma@hotmail.com
[2]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi 11120, Thailand
E-mail: sanguansat@yahoo.com

*Corresponding author

Submitted 5 August 2011; accepted in final form 27 November 2011

Abstract

To improve Query-by-Humming, in this paper, we propose a note segmentation by humming sound method, melody contour extraction technique, and new normalization methods. The noise interference from both the environment and acquisition instrument is the critical issues in humming sound. The query problems about variation of pitch and timing are because most users are not professional singers. The advantage of the note segmentation by humming method is it can separate the sound and silent parts from each other through the process. The melody contour extraction can reduce noise resulting in pitch smoothing. Our approach starts from pre-processing by using features for note segmentation by a humming sound. The process consists of three steps as follows: Firstly, the pitch is extracted from the humming sound by Subharmonic-to-Harmonic Ratio (SHR). Afterwards, we used various new normalization methods, including melody contour extraction, for scaling and noise robust. Finally, Dynamic Time Warping (DTW) is applied to the melody contour, for similarity of measurement between the humming sound and the melody sequence. Comparing our proposed technique and the traditional method, the results show that our proposed techniques can perform more effectively.

Keywords: *Query-by-Humming, melody contour, Dynamic Time Warping, pitch, Subharmonic-to-Harmonic Ratio, note segmentation*

1. Introduction

At present, there is one system that results in easier searching called the Query-by-Humming (QBH) system. This system proposes to search a desired music part by humming its tune, which is very useful when a person wants to find a song from a music library, but might forget its title or artist. However users can only type different keywords (titles, singers, etc.), which is inconvenient for them.

Normally, a user always remembers the melody or rhythm and can hum the melody to retrieve the song. QBH increases the usability of a music retrieval system while the user receives a more convenient result. Many researchers have focused on how to improve QBH for measuring the similarity of a humming sound. First of all, a humming sound must be extracted to pitch by using one of many methods such as autocorrelation, maximum likelihood cepstrum analysis or Subharmonic-to-Harmonic Ratio (SHR) (Ghias, Logan, Chamberlin, & Smith, 1995; Sun, 2002). There are two frameworks for QBH, based on feature types: (1) the technique based on string matching (Ghias et al., 1995; McNab, Smith, Witten, Henderson, & Cunningham, 1996; Uitdenbogerd & Zobel, 1999) and (2) the technique based on continuous pitch contour matching (Nishimura, Zhang, & Hashiguchi, 2001; Zhu, Kankanhalli, & Xu, 2001; Zhu, Kankanhalli, & Tian, 2002).

Most previous string matching methods were focused on matching part of a song to a retrieval systems. The technique retrieves a melody and song from a music database. However, features which we use have variable dimensions, thus Dynamic Time Warping (DTW) is the appropriate method to measure the distance. It is the technique to use for non-linear time normalization that can preserve important features. DTW is robust distance measure for time series and it is the best solution known for time series problems in a variety of domains (Jang & Lee, 2001; Keogh, 2002; Vega-L´opez & Moon, 2006). Probably the most prevalent method (Ghias et al., 1995; McNab et al., 1996; Uitdenbogerd & Zobel, 1999) of melodic representation in QBH systems, the three alphabets are used to display whether a note in sequence is up (U), down (D), or the same (S) as the previous note. Then a melodic representation will be analyzed by the above technique.

String matching based on statistical models includes Hidden Markov Models (HMMs) (Dannenberg, Birmingham, Pardo, Hu, Meek, & Tzanetakis, 2007; Shih, Narayanan, & Kuo, 2003a). This approach uses a combination of HMMs for sequence estimation and DTW for hierarchical clustering (Hu, Ray, & Han, 2006). HMM is used to segment a note in the humming waveform (Shih, Narayanan, & Kuo, 2003b; Raphael, 1999). Gao and Wu (2006) proposed a method, using energy difference to accomplish rough note segmentation, two phases of cepstrum with zero-process to track pitch from non-whistle inputs, pitch contour to do note segmentation and cepstrum peak value curve to deal with note insertion.

Subsequent to this technique is continuous pitch contour. From the above techniques, the discriminant information may be 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, which is a time series of pitch values, represents melody content without using explicit music notes, is used (Jang & Lee, 2001; Nishimura, Zhang, & Hashiguchi, 2001; Zhu, Kankanhalli, & Xu, 2001; Zhu, Kankanhalli, & Tian, 2002).

Pitch and fundamental frequency are important features, 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 describes the amplitude ratio between subharmonics and harmonics (Sun, 2002). For our system, we have implemented pitch tracking using SHR.

The Me1 Frequency Cepstral Coefficients (MFCC) was adopted in many speech analysis applications. This type of feature extraction is being widely used in robust speech recognition systems inspired by human auditory perception and focusing on effective signal processing in the ear using cochlear filterbanks (Behroozmand & Almasganj, 2005). MFCCs were also used as features (Kim & Sikora, 2004; Wu, Cai, & Meng, 2006; Liu, Xu, Wei, & Tian, 2007). From these experiments it shows that using MFCC with the dimension 13 and audio recognition will give better results than other dimensions. MFCC is used in our pre-processing.

Median filter is one of the order-statistics filters and well known for being able to remove impulse noise and the smoothing of signals (Astola, Haavisto, & Neuvo, 1990). Astola, Haavisto, and Neuvo (1990) 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 shows how some types of noise can be removed by the median filter and how other types cannot be removed. Median filter is used for smoothing pitch in the QBH system (Gallagher & Wise, 1981). For our system we decided to reduce noise in a part of a pitch by using this method.

Due to the variation of frequency rank, normalization is needed for reducing these influences. Nguyen, Nocera, Castelli, and Van Loan (2008) proposed that fundamental frequency (F0) normalization methods are presented by a statistical approach (min, max, mean, standard derivation, etc.). Furthermore, we proposed two new normalization techniques and compared this with other normalization methods.

This paper is divided into two parts, which are pre-processing and processing. Segmentation without humming is used for pre-processing. We found that appropriate process is as follow: Firstly, pitch tracking by SHR and then our proposed technique for feature extraction and normalization. Finally, DTW is used for signal alignment.

This paper is organized as follows: Objectives is presented in Section 2, describing the concept. Materials and Methods is proposed in Section 3. In Section 4, experimental results are presented. Discussion is in Section 5. Finally, conclusions are contained in Section 6.

2. Objectives

Our techniques are proposed in order to improve the accuracy rate for a QBH system. Our techniques are note segmentation by humming sound, melody contour extraction, and new normalization techniques.

3. Materials and Methods

3.1 Pitch tracking

In this subsection, 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 notes we hear (Ghias et al., 1995). Notes can begin and end when pitches have been identified. The pitch detector decides based on the statistical information of pitch models. The detailing of each component of the pitch detector is given below.

Four pitch tracking methods were used: Autocorrelation, Maximum Likelihood, Cepstrum Analysis and SHR (Ghias et al., 1995; Sun, 2002). The greatest pitch detection autocorrelation is chosen for implementation of pitch tracking (Ghias et al., 1995). In addition, a PDA based on SHR is developed in the frequency domain and describes the amplitude ratio between subharmonics and harmonics (Sun, 2002). 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 (SH) amplitude is defined as

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

where N is the maximum number of harmonics contained in the spectrum, and $A(f)=0$ if $f > f_{\max}$. If they confine the pitch search range in $[F0_{\min}F0_{\max}]$, then $N = \text{floor}\ (f_{\max} / F0_{\min})$.

Assuming the lowest subharmonic frequency is one half of f_0 the Sum of Subharmonic (SS) amplitude is defined as

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

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

$$SH = \sum_{n=1}^{N} LOGA(\log(nf_0)) = \sum_{n=1}^{N} LOGA(\log(n)+\log(f_0)) \qquad (3)$$

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

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

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

Similarly, by shifting the spectrum leftward at log(1), log(3), log(5), …, log(4*N*-1), they have

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

Next, they define a difference function as

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

In searching for the maximum value, they first locate the position of the global maximum 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(1.96375 f_1), log(2.0625 f_1)]. SHR equation is defined as

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

If SHR is less than a certain threshold value, it indicates that subharmonics are weak and they should favor the harmonics. Thus, f_2 is selected and the final pitch value is $2f_2$. Otherwise, f1 is selected and the pitch is $2f_1$. SHR can be effectively used for pitch tracking (Sun, 2002).

3.2 Melody contour extraction (MCE)

In this subsection, our proposed technique for feature extraction in the QBH system is presented. The following algorithm describes how to extract pitch from a humming sound to obtain the melody contour.

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

Algorithm 1 Melody Contour Extraction Algorithm

Require: p, g, T, s

Ensure: m

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

The first step of this technique is to take a pitch to pass through the noise filtering process which uses the median filter in order to make the signal go smoothly. Then, find the different value of *p* by comparing with the defined *g* value by selecting only the exceed value. The value of *s* is determined in order to apply to find the range of signal that changes a little at that period of time. In other words, it discards the signal that changes rapidly in a short time comparing with this interval. There is the spread around the signal and it only needs the group of significant signals. Hence, it finds the range of signal which has a small value of the spread when comparing with the threshold of standard deviation (*T*).

This algorithm was designed for feature extraction. The humming sound consists of a pitch in several values and also has noise fused in the pitch as shown in Figure 1 (pitch). Normally, the humming sound is usually reduced noise by a median filtering method which makes the signal go smoothly as shown in the Figure 1 (noise reduction by median filtering). However, it usually makes the discriminant information of the signal lost at the same time. It is also applied for filtering as part of the signals prior to further processing with a small window. From our method, the noise can be reduced while the information of the signal is still reserved.

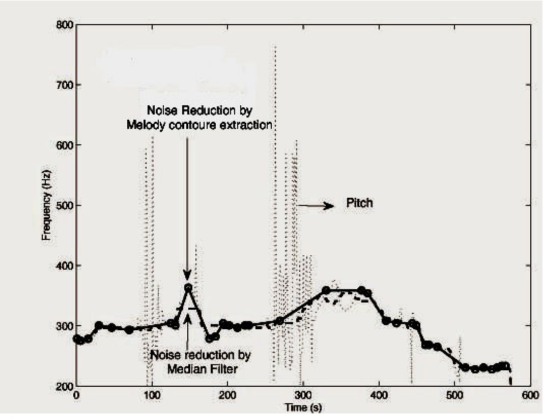

Figure 1 Example of pitch sequence which is obtained by our technique, compare to median filter

From Figure 1, the graph shows the noise reduction by Melody contour extraction, it can be seen that the pitch goes smoothly. Pitch of humming sound is normalized, which is based on logarithmic value and standard variation. The output of our algorithm contains significant pitch.

3.3 Pitch normalization methods

In continuous speech, the pitch contour of the humming sound is affected by many factors. Therefore, pitch normalization is necessary. Let $x(t)$ be the pitch and $\sigma_{\log\ x(t)}$ represent the standard deviation of the logarithm of pitch. In this paper we propose two new techniques for pitch normalization. A1 and A2 are the new proposed methods, while the others are the old methods used for comparison. For these methods, the logarithm of the standard variation are used instead of the standard variation of the logarithm as shown in Eq. (9). Besides in Eq. (10), logarithm of mean is used instead of the mean of the logarithms. The following pitch normalization methods are presented:

A1. Using mean and standard deviation value of pitch and normalizing this new value by logarithmic of each sequence.

$$y(x_t) = \frac{\log\ x_t - \log\ \overline{x_t}}{\log \sigma_{x_t}} \tag{9}$$

A2. Using mean of pitch value and normalizing this logarithmic value of pitch by logarithm of each sequence.

$$y(x_t) = \frac{\log\ x_t}{\log \overline{x_t}} \tag{10}$$

A3. Using logarithm of pitch value and normalizing this logarithmic value of pitch by min and max of each sequence.

$$y(x_t) = \frac{\log x(t) - \min \log x(t)}{\max \log x(t) - \min \log x(t)} \tag{11}$$

A4. Pitch normalization by pitch mean of each sequence.

$$y(x_t) = \frac{x(t)}{\overline{x(t)}} \tag{12}$$

A5. Pitch normalization by min pitch and max pitch of each sequence.

$$y(x_t) = \frac{x(t) - \min x(t)}{\max x(t) - \min x(t)} \tag{13}$$

A6. Pitch normalization by mean and standard deviation of the pitch of each sequence.

$$y(x_t) = \frac{x(t) - \overline{x(t)}}{\sigma_{x(t)}} \tag{14}$$

A7. Using logarithmic value of pitch and normalizing this new value by mean and standard deviation of each sequence.

$$y(x_t) = \frac{\log x(t) - \overline{\log x(t)}}{\sigma_{\log x(t)}} \tag{15}$$

A8. Using logarithm of pitch value and normalizing this logarithmic value of pitch by mean of each sequence.

$$y(x_t) = \frac{\log x(t)}{\overline{\log x(t)}} \tag{16}$$

3.4 Note segmentation by humming sound

For this paper, we propose the method of note segmentation by humming sound to differentiate the sounds part from the silence parts in order to choose the most important part, which is the sound part, to use in the next process. From the sound wave in Figure 2, the silence interval is removed manually as preprocessing before being fed to the HMM.

As shown in Figure 3, the HMM contain 3 states with left-to-right topology using 2 Gaussian mixture distributions. Both the note and the silence are used to train these HMMs.

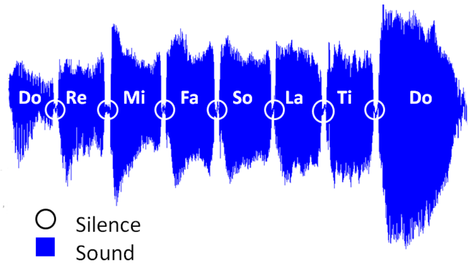

Figure 2 Sound wave from humming standard note in C major scale (do, re, me,...,do)

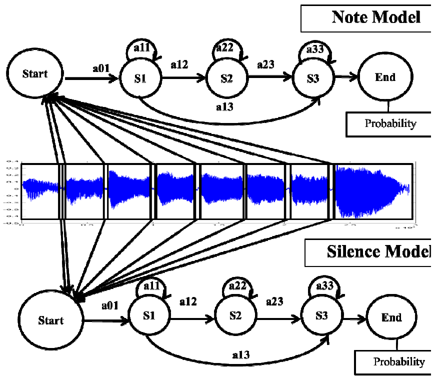

Figure 3 Note Model and Silence Model

We propose three methods in our framework as shown in Figure 4 which are note segmentation, Melody Contour Extraction (MCE) and new normalization methods. Our framework starts from pre-processing by using a feature to facilitate note segmentation by a humming sound. The process consists of three steps as follows: firstly, the pitch is extracted from the humming sound by SHR. Consequently, the feature is extracted by melody contour extraction through our new normalization methods. Finally, DTW is applied to melody contour, for a similarity of measurement between the humming sound and the melody sequence.

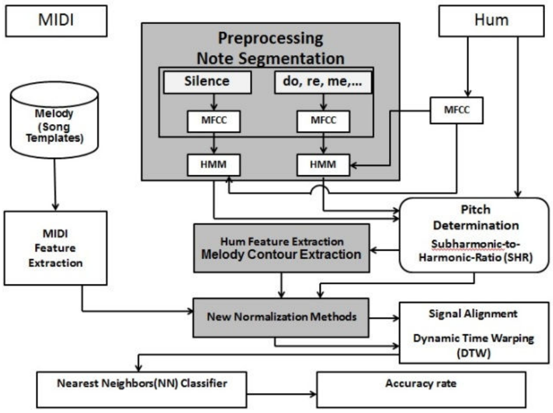

Figure 4 Block diagram of our framework

3.5 Dynamic time warping

Due to the tempo variation of length of sequence, we cannot measure the similarity by any traditional distances. 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 the 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:

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

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

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

The best path is the one with the least global distance, which is the sum of cells along the path. This method exhibits good performance for word speech recognition and QBH (Jang, & Lee, 2001).

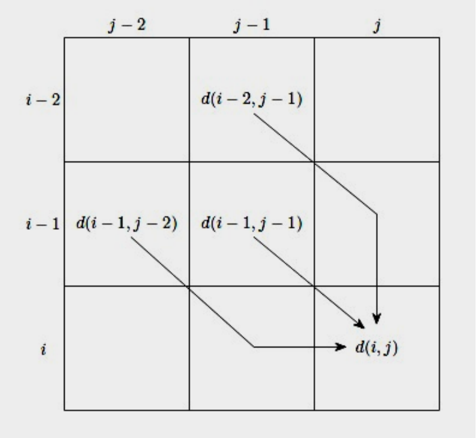

Figure 5 The calculation pattern for the dynamic time warping in the Melody Contour

4. Results

Experiments have shown the effectiveness of the system according to the various conditions. For the effectiveness of this system, the measures were setup to explore such as the variation of the number of songs in database, normalization methods, and note segmentation by humming sound.

4.1 Dataset

In our system, there are 100, 300 and 500 MIDI format songs in the database. The test query is a 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 a time duration of 10 seconds, starting at the beginning of the song. The "10 seconds" data come from the "vocal track" inside the MIDI data.

4.2 Variation of normalization

The pitch of the humming sounds are normalized by our new normalization methods in Eq. (9) and Eq. (10), compared with the normalized pitch by other methods i.e. A3-A8 normalization. The experimental results show that normalized pitch of each sequence by logarithm, mean and standard derivation gave better results than other methods. From Figures 6-8 show that the retrieval accuracies of normalized pitch by A1 and A2 normalization, obtain a higher accuracy rate compared with other normalization methods presented in (Nguyen et al., 2008).

4.3 Variation of feature extraction and denoising

In these experiments, the method used was median filtering, the baseline noise reduction is described in detail (Wang et al., 2008) for comparing with our proposed methods. In our experiments, we set the values of variables such as s, g, and T to 5, 2, and 5 respectively. For a median filter, we found that the optimal size of a window is 53 to achieve the highest performance. Our proposed methods used DTW for alignment and normalized and with our new normalization methods a 76% accuracy rate was achieved, as shown in Table 1. Tables 1-3 and Figures 6-8 show the retrieval accuracies that retrieved 100 humming sounds from 100, 300, and 500 MIDI songs. In order to show the advantage of our proposed technique, the accuracy is better than using only a median filter to reduce noise. Our new normalization methods have a higher accuracy rate when compare to other normalization methods. 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 conventional one by around ten times.

4.4 Variation of note segmentation by humming sound

From the experiment, we added pre-processing into the system after feature extraction as mentioned in 4.3 (Variation of feature extraction and denoising). Note segmentation is used for pre-processing. With this feature, we get through note segmentation by humming sound as we proposed earlier and it will compare between extracted feature that goes through pre-processing and the other group of extracted feature which doesn't go through pre-processing, which will lead to our normalization method as A1 and A2 and other methods.

From Tables 1-3 and Figures 6-8, it shows that the extracted features through pre-processing and normalization methods as we propose on A1 and A2, results in 77% and 74% accuracy respectively. For MIDI 100 songs, A1 gets higher accuracy than other normalization methods as we use it as a baseline. However, A1 and A2 gets a higher accuracy than features that goes through normalization and without pre-processing. While a feature is extracted by melody contour extraction through normalization as A1 and A2 and traditional methods without pre-processing, which gives higher accuracy than the feature by melody contour extraction through pre-processing and also more than the feature from reducing the noise by median filtering only.

Table 1 Test results of experiments with 100 test queries and 100 MIDI songs

Normalization Method	Pitch with Note Segmentation (%)	Pitch with Median Filter (%)	Pitch with Melody contour Extraction (%)	Pitch with Melody Contour Extraction and Note Segmentation (%)
N1	77	37	76	58
N2	74	41	78	62
N3	67	29	60	58
N4	72	40	71	56
N5	71	18	61	46
N6	77	29	77	52
N7	76	30	70	52
N8	74	41	80	62

Table 2 Test results of experiments with 100 test queries and 300 MIDI songs

Normalization Method	Pitch with Note Segmentation (%)	Pitch with Median Filter (%)	Pitch with Melody contour Extraction (%)	Pitch with Melody Contour Extraction and Note Segmentation (%)
N1	74	21	69	53
N2	72	26	73	56
N3	62	16	54	48
N4	71	23	66	52
N5	66	6	48	41
N6	74	9	56	42
N7	74	8	57	39
N8	72	22	71	56

Table 3 Test results of experiments with 100 test queries and 500 MIDI songs

Normalization Method	Pitch with Note Segmentation (%)	Pitch with Median Filter (%)	Pitch with Melody contour Extraction (%)	Pitch with Melody Contour Extraction and Note Segmentation (%)
N1	73	16	68	47
N2	70	18	70	49
N3	62	9	45	43
N4	69	20	66	46
N5	64	5	45	35
N6	72	9	51	35
N7	72	7	52	32
N8	70	19	70	49

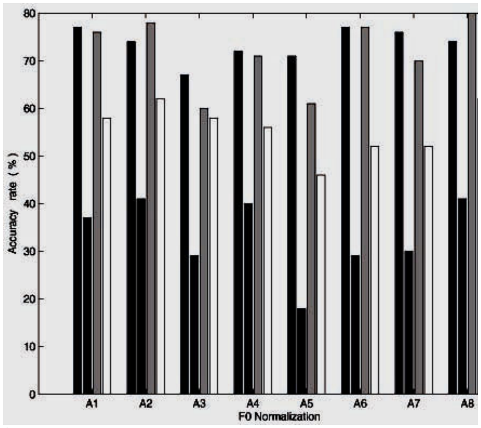

Figure 6 100 Test queries and 100 MIDI songs

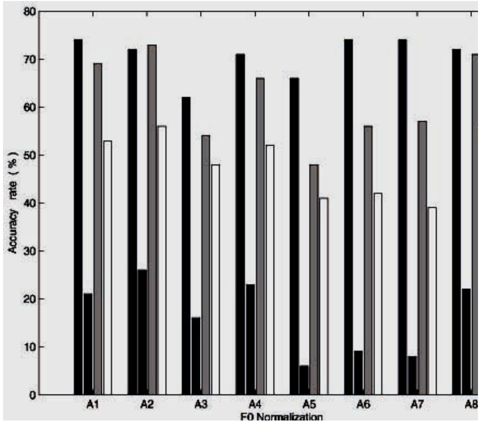

Figure 7 100 Test queries and 300 MIDI songs

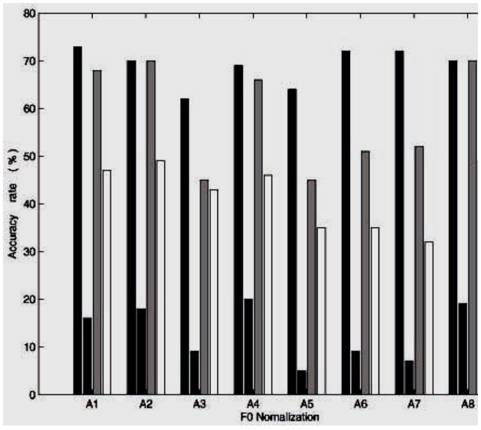

Figure 8 100 Test queries and 500 MIDI songs

■ Pitch with Note Segmentation
▩ Pitch with Median Filter
▨ Pitch with Melody contour Extraction
□ Pitch with Melody Contour Extraction and Note Segmentatioı

5. Discussion

We discussed novel techniques to improve QBH. Obviously, to having features that go through note segmentation by humming sound gives better results than using features without going through this process. The reason for this is that note segmentation can separate the sound and silent part quite well and the humming sound will be used for training before feeding to HMM, in order to build the sound and silent model. This means our proposal which consists of segmentation notes and two normalization methods can improve accuracy for QBH.

The effectiveness of the technique proposed in this paper has been an accuracy rate confirmed as shown in Tables 1-3. Regarding the disadvantage in this work, one of the factors that gives the highest accuracy of only 77% is because the feature we used might not have enough information. However, we believe it can be improved and it will receive more accuracy, which we are going to work toward on feature extraction.

The feature from melody contour extraction technique, when it goes through note segmentation by humming sound might result in less accuracy. The reason is because the feature has significant features already and when it goes through note segmentation by humming sound, some part of the features might be lost.

Therefore, it is desirable that much more attention should be provided in the development on how to increase the effectiveness and accuracy by improving more methods of the feature extraction and then use them together in order to have a variety of information.

6. Conclusion

In this paper, we propose a note segmentation by humming sound method, a new melody retrieval method by the similarity matching of continuous melody contours which is a melody contour extraction technique, and new normalization methods. QBH was improved by our method, a segmentation note by humming sound through pre-processing was improved by the process of feature extraction from various humming inputs. Furthermore, we used a melody contour extraction technique for feature extraction and normalized pitch with our new normalization methods. The experimental results show that the performance of using these features through pre-processing is better than without. Yet, if the feature that comes from melody contour extraction through pre-processing, will turn out less accuracy than not going through pre-processing. However, this feature extraction will give more accuracy than reducing the noise by median filter. The melody contour extraction method offers several advantages: higher accuracy and low complexity. First of all, it can reduce noise while the discriminant information is extracted. That improves the accuracy as shown in our experimental results. Secondly, the query process is faster and consumes lower memory because the feature vector dimension is smaller than a traditional one. The advantage and disadvantage is the feature that goes through the pre-processing segmentation note by humming sound might give more accuracy but it will increase the complexity into the system, while the feature from the melody contour extraction will give less accuracy but it doesn't make the system complicated.

7. Acknowledgements

This study is supported by Rangsit University, and Suan Dusit Rajabhat University Foundation. Further more, we would like to thank students of Suan Dusit Rajabhat University for their great help and also all people who hummed a lot of tunes for us. Additionally, the invaluable recommendations and supervision from the anonymous reviewers are much appreciated.

8. References

Astola, J., Haavisto, P., & Neuvo, Y. (1990). Vector median filters. In *Proceedings of the IEEE*, 78, 678-689.

Behroozmand, R., & Almasganj, F. (2005, December). Comparison of neural networks and support vector machines applied to otimized features extracted from patients' speech signal for classification of vocal fold inflammation. In *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology*, 844-849.

Dannenberg, R. B., Birmingham, W. P., Pardo, B., Hu, N., Meek, C., &Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the MUSART testbed. *Journal of the American Society for Information Science and Technology,* 58(3), 687–701. doi: 10.1002/asi.20532

Gallagher, N. J., & Wise, G. (1981). A theoretical analysis of the properties of median filters. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 9(6),1136-1141.

Gao, L., & Wu, Y. (2006). A system for melody extraction from various humming inputs. In *IEEE International Symposium on Signal Processing and Information Technology*, 680-684. doi: 10.1109/ISSPIT.2006.270886

Ghias, A., Logan, J., Chamberlin, D., & Smith, B. C. (1995, November 5 - 9). Query by humming: Musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*, at San Francisco, CA, USA, 231-236. doi>10.1145/217279.215273

Hu, J., Ray, B., & Han, L. (2006). An Interweaved HMM/DTW approach to robust time series clustering. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, at Washington, DC, USA, 145-148.

Jang, J-S. R., & Lee, H-R. (2001). Hierarchical filtering method for content-based music retrieval via acoustic input. In *Proceedings of the ninth ACM International Conference on Multimedia*, at New York, 401-410.

Keogh, E. (2002, August 20-23). Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB '02)*, at Hong Kong SAR, China, 406--417.

Kim, H-G., & Sikora, T. (2004, September 6 - 10). Audio spectrum projection based on several basis decomposition algorithms applied to general sound recognition and audio segmentation. In *XII European Signal Processing Conference*, Vienna, Austria, 1047-1050.

Liu, Y.,Xu, J-P., Wei, L., & Tian, Y. (2007). The study of the classification of chinese folk songs by regional style. In *Proceedings of the International Conference on Semantic Computing (ICSC)*, at Washington, DC, USA, 657-662. doi>10.1109/ICSC.2007.99

McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L. & Cunningham, S. J. (1996). Towards the digital music library: Tune retrieval from acoustic input. In

Proceedings of the first ACM international conference on Digital libraries, at Bethesda,11-18.

Nishimura, T., Zhang, J. X., & Hashiguchi, H. (2001). Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming. In *Proceeding of the third International Symposium on Music Information Retrieval Continuous Dynamic Programming*, 211-218.

Nguyen, H. Q., Nocera, P., Castelli, E., & Van Loan, T. (2008, June 4 - 6). Tone recognition of Vietnamese continuous speech using hidden Markov model. *Second International Conference on Communications and Electronics*, at Hoi an, Viatnam, 235-239.

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 360-370.

Shih, H. H., Narayanan, S. S., & Kuo, C. C. J. (2003a). An hmm-based approach to humming transcription. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, 337-340.

Shih, H. H., Narayanan, S. S., & Kuo, C. C. J. (2003b). Multidimensional humming transcription using a statistical approach for query by humming systems. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5, 541-544.

Sun, X. (2002, May 13 - 17). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Proceedings of the International Conference on Acoustics, Speech, and Signal,* 2002, at Orlondo, Florida, USA, 1, 333-336.

Uitdenbogerd, A. L., & Zobel, J. (1999). Melodic matching techniques for large music databases. *Proceedings of the seventh ACM International Conference on Multimedia (Part 1)*, at Orlando, Florida, USA, 57-66.

Vega-L´opez, I. F., & Moon, B. (2006, January 23 - 25). Quantizing time series for efficient similarity search under time warping. In *Proceedings of the 2nd IASTED International Conference on Advances in Computer Science and Technology*, at Puerto Vallarta, Mexico, 334-339.

Wang, Lei, et al. (2008). An Effective and Efficient Method for Query by Humming System Based on Multi-Similarity Measurement Fusion. *International Conference on Audio, Language and Image Processing*, at Shanghai, Chaina, 471-475.

Wu, Z., Cai, L., & Meng, H. (2006). Multi-level fusion of audio and visual features for speaker identification. In *In: Proc. Int. Conf. Biometrics LNCS 3832*.

Zhu, Y., Kankanhalli, M. S., & Xu, C. (2001). Pitch tracking and melody slope matching for song retrieval. In *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia*, at London, UK, 530-537.

Zhu, Y., Kankanhalli, M., & Tian, Q. (2002, December 9 - 12). Similarity matching of continuous melody contours for humming querying of melody databases. In *proceedings of IEEE Workshop on Multimedia Signal Processing*, at St. Thomas, Virgin Islands, USA, 249-252.