A Novel Query-by-Singing/Humming Method by Estimating Matching Positions Based on Multi-layered Perceptron

Tuyen Danh Pham, Gi Pyo Nam, Kwang Yong Shin, and Kang Ryoung Park*
Division of Electronics and Electrical Engineering, Dongguk University, 26, Pil-dong 3-ga, Jung-gu, Seoul 100-715, Republic of Korea
[e-mail: phamdanhtuyen@gmail.com]
[e-mail: oscar1201@dgu.edu]
[e-mail: skyandla@dgu.edu]
[e-mail: parkgr@dgu.edu]
*Corresponding author : Kang Ryoung Park

Received May 1, 2013; revised July 1, 2013; accepted July 9, 2013; published July 30, 2013

Abstract

The increase in the number of music files in smart phone and MP3 player makes it difficult to find the music files which people want. So, Query-by-Singing/Humming (QbSH) systems have been developed to retrieve music from a user's humming or singing without having to know detailed information about the title or singer of song. Most previous researches on QbSH have been conducted using musical instrument digital interface (MIDI) files as reference songs. However, the production of MIDI files is a time-consuming process. In addition, more and more music files are newly published with the development of music market. Consequently, the method of using the more common MPEG-1 audio layer 3 (MP3) files for reference songs is considered as an alternative. However, there is little previous research on QbSH with MP3 files because an MP3 file has a different waveform due to background music and multiple (polyphonic) melodies compared to the humming/singing query. To overcome these problems, we propose a new QbSH method using MP3 files on mobile device.

This research is novel in four ways. First, this is the first research on QbSH using MP3 files as reference songs. Second, the start and end positions on the MP3 file to be matched are estimated by using multi-layered perceptron (MLP) prior to performing the matching with humming/singing query file. Third, for more accurate results, four MLPs are used, which produce the start and end positions for dynamic time warping (DTW) matching algorithm, and those for chroma-based DTW algorithm, respectively. Fourth, two matching scores by the DTW and chroma-based DTW algorithms are combined by using PRODUCT rule, through which a higher matching accuracy is obtained.

Experimental results with AFA MP3 database show that the accuracy (Top 1 accuracy of 98%, with an MRR of 0.989) of the proposed method is much higher than that of other methods. We also showed the effectiveness of the proposed system on consumer mobile device.

Keywords: QbSH, MP3 Files, Multi-layered Perceptron, Dynamic Time Warping.

This work was supported by the Ministry of Knowledge Economy grant funded by the Korea government (No. 2010-S-001-01), and in part by the Public welfare & Safety research program through the National Research Foundation of Korea (NRF) funded by the MSIP (Ministry of Science, ICT & Future Planning) (No. 2012-0006554).

http://dx.doi.org/10.3837/tiis.2013.07.008

1. Introduction

The development of internet allows users to access an enormous amount of music. In addition, the increase in the number of music files in smart phone and MP3 player makes it difficult to find the music files which people want.

Conventionally, a music file can be browsed based on categories such as its title, singer, or composer's name. An emerging technology known as Query-by-Singing/Humming (QbSH) provides a convenient method for users to search for music by humming/singing the corresponding melodies, without having to know detailed information about the title or singer of song. The QbSH searching process is carried out by the matching algorithms, which are attracting much interest in the field of pattern recognition.

There have been many studies in QbSH since this music information retrieval problem was first presented in the 1990s [1]. The approaches can be classified into note-based and frame-based methods [2][3][4][5]. The note-based approach focuses on the quantization of the continuous pitch values into discrete musical notes values [6][7][8]. This method has the disadvantage of inaccurate note segmentation, which can reduce the matching accuracy. To overcome this drawback, the original pitch values are used in the frame-based method [2][9][10]. The studies in QbSH can also be categorized into top-down and bottom-up methods [2][3][4][5]. The bottom-up approach solves the matching problems in QbSH by locally comparing the humming/singing query data with the reference music data to find the optimal matching result [6][7][9]. Alternatively, the top-down method uses the global shape of the two waveforms for comparison, and the local information of the waveform for the adjustment of the matching result of global shape matching [2][8][11].

QbSH systems based on a dynamic time warping (DTW) algorithm are one of the most common approaches. In the QbSH system proposed by Jang et al. [9][12], DTW was used to calculate the distance between the pitch features extracted from the queries and the reference data. An extension of DTW with three dimensions was proposed by Heo et al. in order to solve the problem with multiple pitch candidates [13]. Nam et al. used the method of score level fusion of two classifiers, which are a quantized binary (QB) code-based LS algorithm and a pitch-based DTW algorithm [4]. In addition, they developed an enhanced version of the QbSH system by combining five classifiers, which are pitch-based linear scaling (LS), pitch-based DTW, QB code-based LS, local maximum and minimum point-based LS, and pitch distribution feature-based LS [5]. In the method proposed by Phiwma et al. [14], DTW was used to create the distance vectors between humming sequences and the reference template, and these feature vectors were applied to the chosen classifier, which is support vector machine (SVM). Lemström et al. proposed the method for a music comparison and retrieval based on the edit distance which is evaluated DTW algorithm [15]. Mongeau et al. proposed the method for comparison of musical sequences and they used the DTW algorithm for calculating the measure of comparison in order to find out the similarity in melodic line [16]. Kotsifakos et al. proposed a subsequence matching framework, and developed a space and time efficient DTW method using this framework for QbH system [17].

In most of the cases using pitch-based DTW matching method, in order to locally compare the humming/singing data and reference data, the QbSH system has to move the starting position of the humming/singing sequence along the reference music file [4][5]. This process can lead to intensive computation and lower matching accuracy. However, few studies have been conducted to find the correct matching position for the DTW algorithm.

In addition, most previous researches on QbSH have been conducted with the reference songs of musical instrument digital interface (MIDI) files [22][23]. However, the production of MIDI files is a time-consuming process. In addition, more and more music files are newly published with the development of music market. Consequently, the method of using the more common MPEG-1 audio layer 3 (MP3) files for reference songs is considered as an alternative. However, there is little previous research on QbSH with MP3 files because an MP3 file has a different waveform due to background music and multiple (polyphonic) melodies compared to the humming/singing query. So, the full matching of the MP3 file with the humming/singing query results in high rates of error and a low matching speed. To overcome these problems, we propose a new QbSH method with MP3 files by four multi-layered perceptrons (MLPs) and score fusion methods on mobile device.

The remainder of this paper is organized as follows. Section 2 describes the details of the proposed method. The experimental results are presented in Section 3. Finally, the conclusions are stated in Section 4.

2. Proposed QbSH System

2.1 Overview of the Proposed Method

Fig. 1 gives an overview of the proposed system. The pitch information is extracted from each humming/singing file by a musical note estimation method using spectro-temporal autocorrelation (STA) [4][5][22][23]. After this step, the extracted pitch features are preprocessed as follows [4][5][22][23]. We eliminate all the zero values from the extracted features. These values, which contain no feature information in the extracted pitch data, arise because of the intervals of silence in the humming/singing sequence. The zero-eliminated humming/singing sequence is subsequently normalized by using a variety of procedures of mean-shifting, median filtering and min-max scaling in order to reduce the differences between the input and reference data, and remove the noise.

After this normalization step, each processed feature is applied to four separated neural networks, which are trained to estimate the four positions (in the MP3 file), namely the start and end positions for matching by the DTW and chroma-based DTW algorithms. Based on these four positions, the two DTW algorithms calculate the distances between the pitch data of the humming/singing file and those of the reference files. Finally, the matching scores of the two algorithms are combined by score level fusion using PRODUCT rule, and the correct reference song is selected based on the combined score.

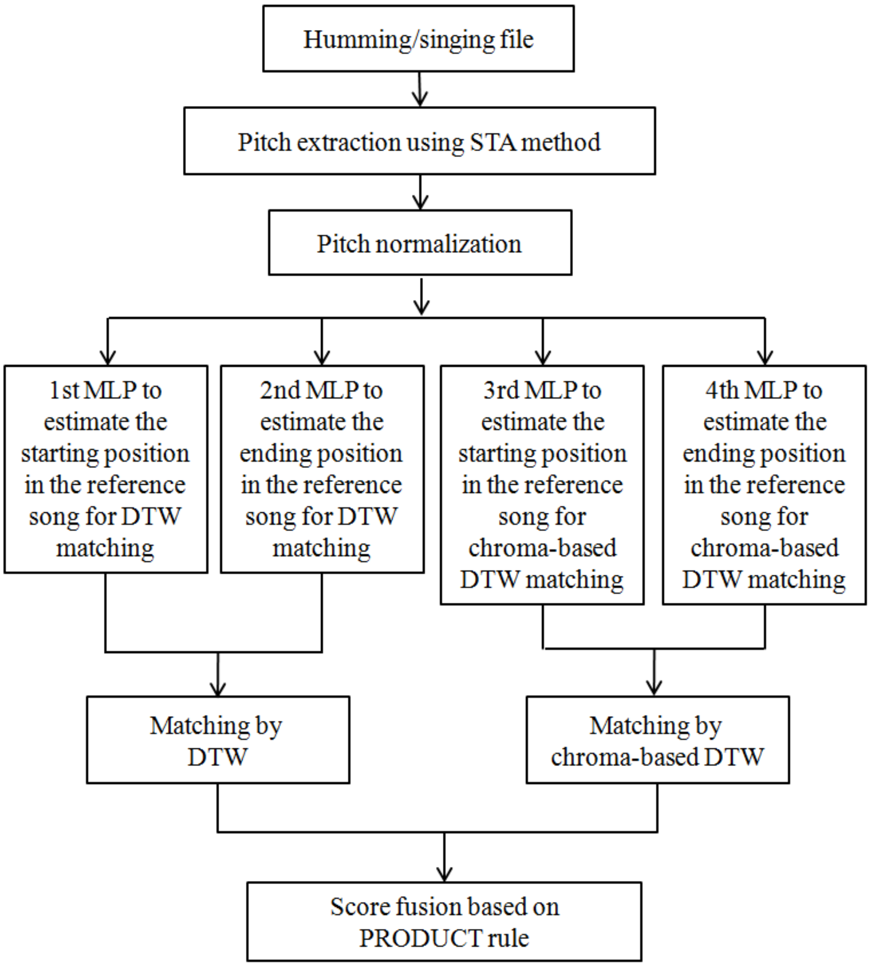

Humming/singing file
Pitch extraction using STA method
Pitch normalization
1st MLP to estimate the starting position in the reference song for DTW matching
2nd MLP to estimate the ending position in the reference song for DTW matching
3rd MLP to estimate the starting position in the reference song for chroma-based DTW matching
4th MLP to estimate the ending position in the reference song for chroma-based DTW matching
Matching by DTW
Matching by chroma-based DTW
Score fusion based on PRODUCT rule

Fig. 1. Flowchart of the proposed system

2.2 Pitch Extraction and Normalization

In our research, the pitch features are extracted from the humming/singing files by STA method [4][5][22][23], in which both spectral and temporal autocorrelations are utilized. The estimated frequency values f_p are converted into pitch values p by (1) [20].

$$p = 12\log_2\left(\frac{f_p}{440}\right) + 69 \tag{1}$$

With the extracted pitch values, the normalization step is performed as follows [4][5][22][23]. In pitch data, there are usually zero values, representing the silent durations in the original humming/singing files. Since these zero values contain no feature information, the zero values are eliminated in the first step of normalization. The zero-eliminated humming/singing sequence is subsequently normalized by procedures of median filtering, mean-shifting, and min-max scaling in order to reduce the differences between the input and the reference data, and remove the noises. Median filtering is used to remove the peak noises caused in the recording process by the surrounding and device noises. Mean shifting and min-max scaling are to set the mean values (DC levels) of two data sequences and adjust the scales of pitch values of these two sequences, respectively. Since the number of input nodes of the MLP network is a fixed number, the normalized humming/singing data sequences should be stretched or shrunk to become equal in length. We stretch or shrink the humming/singing file linearly by (2).

$$input\_data(i) = humming\_data(round(i \times \frac{H}{N})) \tag{2}$$

for i = 1 to N

where $round(\cdot)$ is a round-up operator, N is the number of input nodes of the MLP, H is the size of humming data, and *input_data(i)* is the value for the i^{th} input node of the MLP. The optimal N for the MLP network was experimentally determined as 250 in terms of minimum training error.

2.3 Estimating Start and End Positions in the Reference Song for DTW and Chroma-based DTW Matching

Since a reference song in the MP3 file format usually has a different waveform due to background music compared to the humming/singing query, the full matching in the MP3 file with the humming/singing query results in a high error rate and low matching speed. As such, the start and end positions of the reference song to be matched are estimated by using MLP before performing the matching with the humming/singing file.

An MLP is an artificial neural network with the ability to map sets of input data to a set of desired output values, after being trained by a supervised learning technique known as a back-propagation algorithm [18]. In this method, the errors of the feed-forward network are calculated and propagated back from the output nodes to the inputs in order to appropriately adjust the weights in each layer of the network. There are several popularly used kernel functions, and the performance with the hyperbolic tangent function of (3) is compared to that of the log-sigmoid function in this research (see **Table 1**).

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3}$$

In general, the back-propagation algorithm works well on networks with various numbers of layers, as long as the used kernel function is differentiable. Thus, the 3-layer network made up of input, hidden, and output layers is the chosen model in our experiments. As shown in **Table 1**, the model's performance was compared according to the number of hidden nodes and output nodes of the MLP. Experimental results gave the training performance in the case where four MLPs were used, and where each MLP has one output as shown in **Fig. 1**. For the training of MLP, the desired output should be obtained, and the desired outputs are the start or end positions in the reference song for DTW and chroma-based DTW as shown in **Fig. 1**. In order to obtain the desired output data, the correct start or end positions in the reference song should be obtained, but they are difficult to manually measure because an MP3-based reference song consists of long streams of polyphonic data, making it difficult to discriminate which part of the reference song corresponds to the input humming/singing. The desired output data of the correct starting or ending positions in the reference song are therefore automatically obtained as follows. In general, users can hum or sing any arbitrary part of the song that they remember. Hence, the relative position of the humming/singing file to the reference file has to be moved in a searching procedure [4][5] in order to find the part of the reference song that best matches with the considered humming/singing sequence. By comparing the local DTW distances obtained after each moving step, and from the location at which the calculated distance is minimum, we can estimate the start and end positions in the reference song, which are best matched with the corresponding humming/singing file as shown in **Fig. 2**.

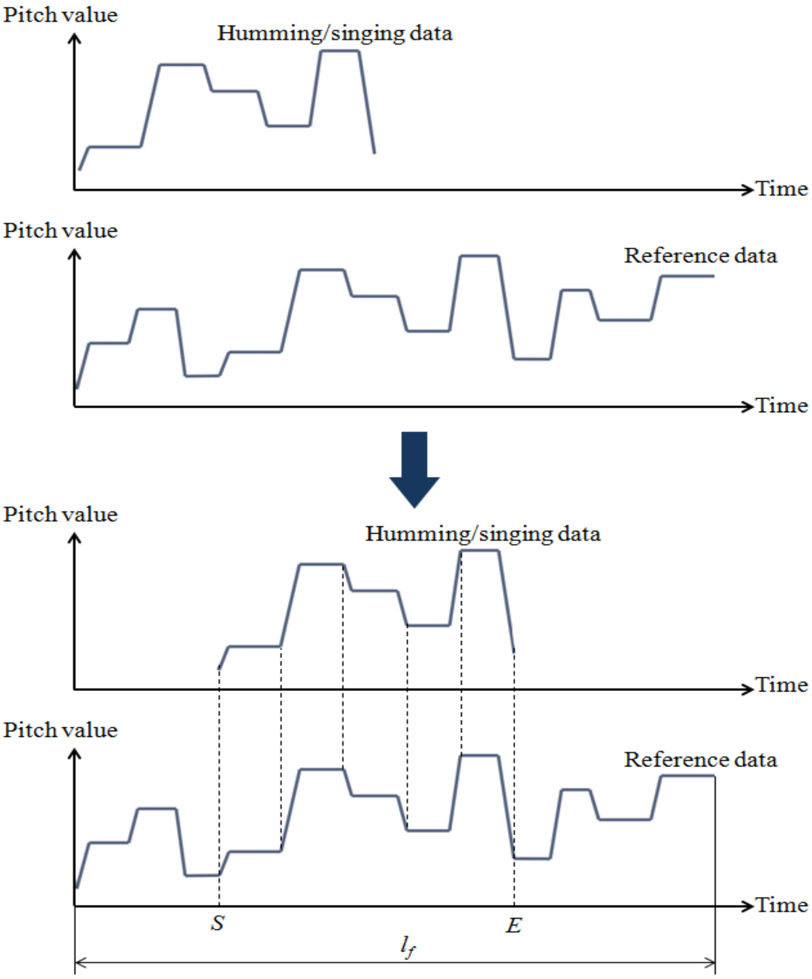


Fig. 2. Moving the humming data relative to reference data to extract the start (*S*) and end (*E*) positions

The extracted position data has to be normalized so that it can be applied to the MLP training procedure as the desired output data. The actual position represent the start position (*S*) and end position (*E*) as shown in **Fig. 2** are normalized to have the values from 0 to 1 by min-max scaling with the length of the reference data l_f. Each pair of input data and desired output data is called a patch. The MLP training process on the patches used for network learning is performed a number of times, known as the number of epochs. The mean squared error (MSE) criterion, which is calculated from n number of target (desired) values t_i and calculated output values z_i as in (4), is ordinarily used to evaluate the mapping accuracy of the back-propagation algorithm.

$$MSE = \frac{\sum_{i=0}^{n-1}(z_i - t_i)^2}{n} \quad (4)$$

Fig. 3 shows the training procedures in the case of a 1-output MLP network with 250 inputs and 100 hidden nodes. The horizontal and vertical axes of **Fig. 3** are training epochs and MSE, respectively.

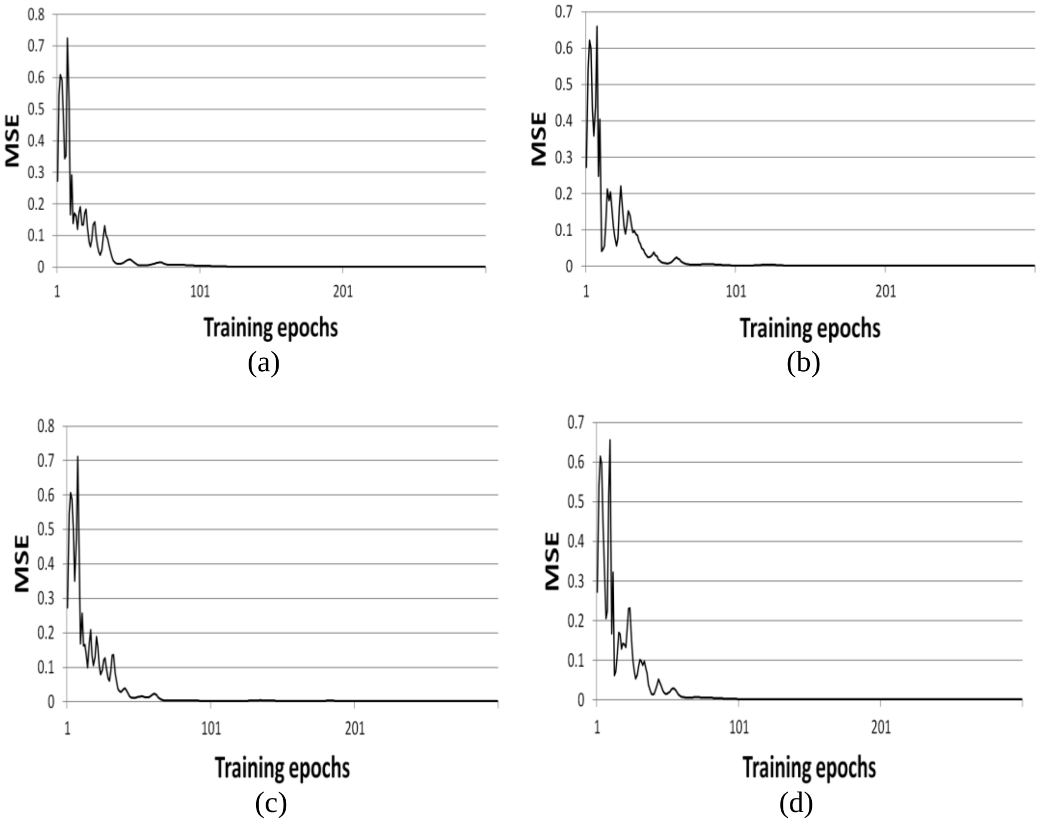


Fig. 3. Training procedure of 3-layer MLP networks with 250 inputs, 100 hidden nodes, and one output in terms of MSE. (a) 1st MLP for estimating the start position for DTW. (b) 2nd MLP for estimating the end position for DTW. (c) 3rd MLP for estimating the start position for chroma-based DTW. (b) 4th MLP for estimating the end position for chroma-based DTW

The functional diagram of the MLP training procedure is shown in **Fig. 3**. The network weights are obtained after the MLP training by using the back-propagation algorithm. With the four trained MLPs, we can obtain the start and end positions in the reference files for DTW and chroma-based DTW matching as shown in **Fig. 3** and **Fig. 4**.

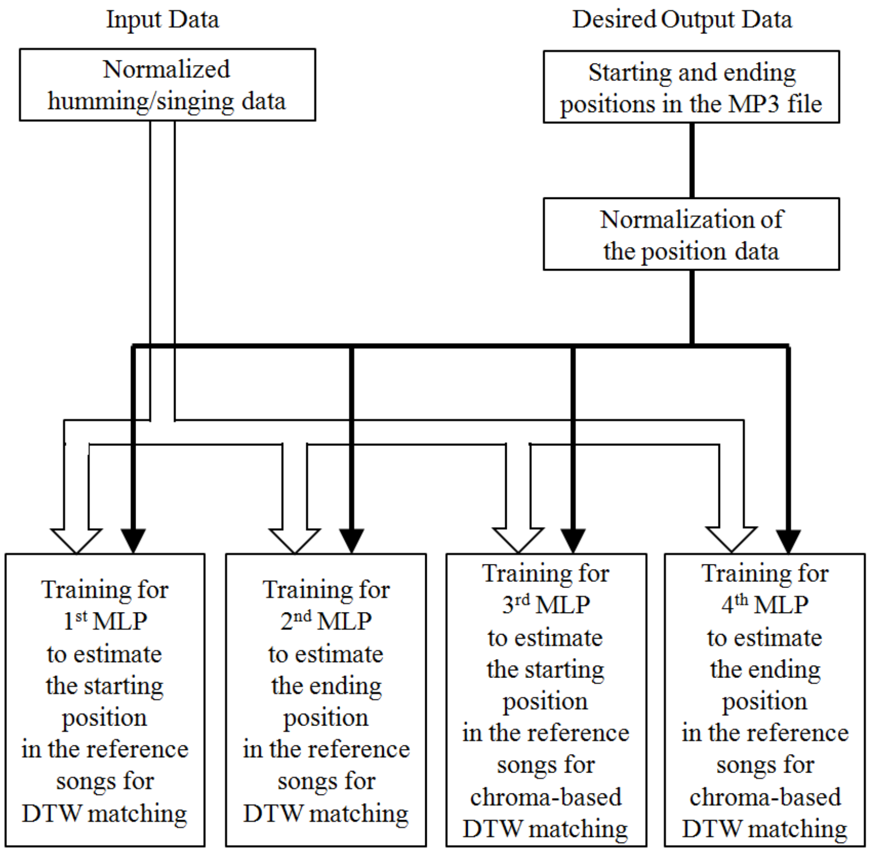

Input Data
Desired Output Data
Normalized humming/singing data
Starting and ending positions in the MP3 file
Normalization of the position data
Training for 1st MLP to estimate the starting position in the reference songs for DTW matching
Training for 2nd MLP to estimate the ending position in the reference songs for DTW matching
Training for 3rd MLP to estimate the starting position in the reference songs for chroma-based DTW matching
Training for 4th MLP to estimate the ending position in the reference songs for chroma-based DTW matching

Fig. 4. Block diagram of MLP training process

2.4 Matching by DTW, Chroma-based DTW Algorithm, and Score Fusion

In conventional QbSH systems, the length of the humming/singing data is different from the reference music data, since a user tends to hum only a part instead of the whole song, and the humming/singing speed can be slower or faster than the original tempo of the melody. For good matching between two data sequences of different length, DTW is used. In previous researches, the DTW-based matching is performed by sliding the input humming/singing data based on the reference song [3][4][5][22], whereas the matching is performed without sliding only at the determined start or end position estimated by the MLPs in this research. For higher matching accuracy, chroma-based DTW is also used [20]. In general, each pitch can be represented by two components, tone height and chroma [19]. For example, the pitch class of the chroma C is the set, and each component of the set is discriminated by an integral number of octaves [21]. Based on this, a DTW based on the chroma feature of the pitch data is used, in order to reduce the pitch doubling and halving errors that make inaccurate feature extraction [20]. In this method, the remainders that are obtained by dividing the pitch values by 12 are used as the pitch values for matching. However, since the original pitch values were substituted by these processed values, it could cause the loss of original query information. For example, the pitch values of 13 and 14 are regarded as being same to those of 1 and 2, respectively, although they can be different from those of 1 and 2. In order to solve this problem, we also used conventional DTW, and combined two scores by the conventional DTW and chroma-based DTW to enhance the matching accuracy.

As shown in **Table 2**, the accuracy when using both conventional DTW and chroma-based DTW is higher than that only using chroma-based DTW or conventional DTW. For the chroma-based DTW, we used the distance measuring method different from that of previous work [20].

Two calculated distances (scores) by DTW and chroma-based DTW were combined by the score level fusion method. Experimental results showed that the accuracy of score fusion through PRODUCT rule (which calculates final score by multiplying two scores) was the highest as shown in **Table 2**. In detail, with one input humming data, two matching distances (scores) by DTW and chroma-based DTW are calculated with one reference MP3 file, respectively. If these two distances are assumed as S_1 and S_2, respectively, we obtain a new distance S_3 (=S_1× S_2), and assign S_3 as the distance between the input humming data and the reference MP3 file. For example, if S_1 and S_2 are 0.1 and 0.2, respectively, the calculated S_3 becomes 0.02. This is named as PRODUCT rule. Like the same method, all the distances between this input humming data and all the reference MP3 files are obtained, and one reference MP3 file whose distance is smallest is determined as the correct reference one to the input humming data.

In case of MIN rule of **Table 2**, the smaller distance among S_1 and S_2 is selected. If S_1 and S_2 are 0.1 and 0.2, respectively, the calculated S_3 becomes 0.1. In case of SUM rule of **Table 2**, the summated distance of S_1 and S_2 is used. If S_1 and S_2 are 0.1 and 0.2, respectively, the calculated S_3 becomes 0.3. All the distances between this input humming data and all the reference MP3 files are obtained by MIN or SUM rule, respectively, and one reference MP3 file whose distance is smallest is determined as the correct reference one to the input humming data. In case of PRODUCT, MIN, and SUM rules, we do not use any weight for calculating the distance.

3. Experimental Results

In our experiments, we used the audio feature analysis (AFA) MP3 database, which includes pop songs and children songs in Korean and English. This database consists of 1,000 input query files corresponding to 100 reference songs as MP3 files [20]. The 1,000 query data, which have 298 humming and 702 singing data, were obtained from 32 volunteers (18 men and 14 women) [20]. The volunteers were required to select the part of the song that they want to sing or hum, and their query data was recorded in the durations from about 12 seconds. The lengths of MP3 database are different from about 34 to 396 seconds [20].

In the training phase, we made various experiments on different MLP network models according to a number of MLPs, input, hidden and output nodes, and kernel functions using the back-propagation algorithm. The training accuracy of each MLP model was evaluated based on the MSE criterion. **Table 1** shows the best training results on various a number of MLPs, input, hidden and output nodes, and kernel functions. In **Table 1**, “SEDC” means “start and end positions for DTW and chroma-based DTW”. “SDC” shows “start positions for DTW and chroma-based DTW”. “EDC” represents “end positions for DTW and chroma-based DTW”. “SD” is “start position for DTW”. “ED” shows “end position for DTW”. “SC” is “start position for chroma-based DTW”, and “EC” represents “end position for chroma-based DTW”, respectively. **Table 1** shows that the experiments with the MLP of single output with the hyperbolic tangent kernel function gave better results than other MLPs in terms of the MSE value. Based on these results, four MLPs (with 250 inputs, 100 hidden nodes, 1 output, and a hyperbolic tangent kernel function) were used for estimating the start and end positions in the reference songs for DTW and chroma-based DTW matching, respectively, as shown in

Fig. 1.

Table 1. Experimental results of MLP training

Number of MLPs	**Number of input nodes**	**Number of hidden nodes**	**Kernel function**	**Number of output nodes**	**MSE**
1	200	250	Log-Sigmoid	4 (SEDC)	2.78×10^{-4}
2	200	250		2 (SDC)	2.75×10^{-4}
				2 (EDC)	2.74×10^{-4}
4 (Proposed Method)	250	100	Hyperbolic tangent	1 (SD)	3×10^{-12}
				1 (ED)	0
				1 (SC)	5×10^{-12}
				1 (EC)	2.5×10^{-11}

With these four MLPs, we evaluated the overall performance of the QbSH system in the testing phase. In this phase, the performance was measured with total 450 MP3 files including the additional 350 files which were not hummed or sung (not used for training) in order to enhance the confidence level of experiment. The commonly used method for the evaluation of QbSH systems is mean reciprocal rank (MRR), which was also widely used for the MIREX contest [3][4][5][20]. The formula of MRR is given in (5), in which k is the number of input query sequences and $rank_i$ is the ranking of the genuine reference file corresponding with the input query data. The ideal result occurs when MRR reaches the maximum value of 1 [4][5].

$$MRR = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{rank_i} \tag{5}$$

Table 2 shows the comparison results in Top 1, Top 10, Top 20 and MRR of the proposed method and other methods. Top 10 shows the probability that the correct MP3 file (corresponding to the input humming/singing file) is included in the 10 highest-ranked candidates among the 450 MP3 files. As shown in **Table 2**, the proposed method showed the best results of MRR of 0.989 and Top 1 accuracy of 98% compared to other methods. By matching only at the start and end positions estimated by MLPs, the matching accuracy of the proposed method was greatly improved compared to conventional DTW and chroma-based DTW matching, which performs matching over the whole reference file. Although the MRR of the proposed method based on PRODUCT rule is the same as that based on the sum rule, the Top 1 accuracy of the proposed method is a little higher than that of the sum rule. Generally, since each song has some parts that are familiar to most audiences, the humming/singing queries sometimes are similar to each other. As such, the trained MLPs can be used to predict the positions of the humming/singing data in the reference files.

Table 2. Comparisons of the matching accuracies of the proposed method and other methods

Algorithm	**Score fusion method**	**Top 1 (%)**	**Top 10 (%)**	**Top 20 (%)**	**MRR**
DTW	N/A	57.1	71.1	75.1	0.621
Chroma-based DTW		62.0	77.8	82.9	0.679
DTW and chroma-based DTW	MIN	71.1	80.1	83.7	0.738
	SUM	64.4	79.0	82.4	0.696
	PRODUCT	64.8	81.2	83.7	0.738

MLPs with DTW and chroma-based DTW	MIN	96.5	99.6	99.9	0.974
	SUM	97.9	100.0	100.0	0.989
	PRODUCT (Proposed method)	**98.0**	**100.0**	**100.0**	**0.989**

In the last experiment, the average matching time of each query was measured, which included the whole matching time with 450 MP3 files. Tests were performed on two platforms. The first one is a desktop computer with CPU of 3.4 GHz (quad core) and RAM of 4 GB. The second one is a mobile computer with CPU of 1.6 GHz (dual core) and RAM of 4 GB.

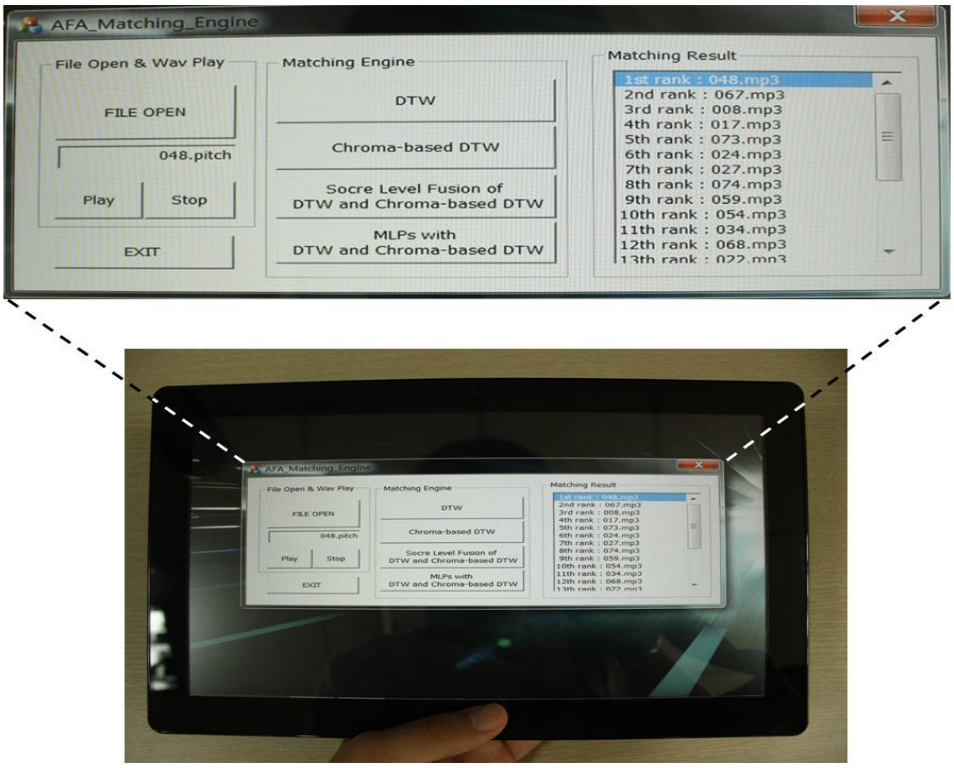
Table 3. Comparisons of the average matching time of the proposed method and other methods

Algorithm	**Score fusion method**	**Average processing time (s)**	
		Desktop computer	**Mobile computer**
DTW	N/A	10.505	57.537
Chroma-based DTW		11.939	63.101
DTW and chroma-based DTW	MIN	22.444	120.638
	SUM		
	PRODUCT		
MLPs with DTW and chroma-based DTW	MIN	**1.329**	**6.010**
	SUM		
	PRODUCT (Proposed method)		

As shown in **Table 3**, the average matching speed of the proposed method was much faster than those of other methods on both platforms, and we confirm that the proposed method can run at real-time speed on various devices even with large number of MP3 reference files. **Fig. 5** shows the example of operating the proposed QbSH system on the mobile computer.

Our method has the shortest processing time because of the following reason: the other methods using DTW or chroma-based DTW have to locally calculate the distance at all the matching positions by sliding the starting position of each humming/singing data relative to the reference files. If not, they should perform the matching with the shorter humming and longer (whole) MP3 file, which include a lot of procedures of insertion or deletion of DTW and chroma-based DTW because they do not know the corresponding range of the humming data to the reference MP3 file. So, it takes much processing time.

In our method, instead of performing this time-consuming procedure, we used the MLP to estimate the starting and ending positions for matching of the query in the reference song. With one input humming and one reference MP3 file, the DTW and chroma-based DTW calculate the distances one time, respectively, only at one matching position (which is determined by the starting and ending positions estimated by the MLP) in the reference song. So, the processing time can be much reduced in our method.



```
1st rank : 048.mp3
2nd rank : 067.mp3
3rd rank : 008.mp3
4th rank : 017.mp3
5th rank : 073.mp3
6th rank : 024.mp3
7th rank : 027.mp3
8th rank : 074.mp3
9th rank : 059.mp3
10th rank : 054.mp3
11th rank : 034.mp3
12th rank : 068.mp3
13th rank : 022.mp3
```

Fig. 5. The example of operating the proposed QbSH system on the mobile computer

4. Conclusion

In this study, we proposed a new method of implementing a QbSH system, using MLP, DTW and chroma-based DTW algorithms. The pitch data of humming/singing for matching was obtained by STA-based pitch extractors and normalization methods. By using the four MLPs, the start and end positions in the MP3 reference song are estimated for DTW and chroma-based DTW matching. Experimental results showed that the proposed method worked effectively on the AFA MP3 database with high accuracy. In future works, we plan to research the methods of enhancing the accuracy by combining MLP and hidden Markov models (HMMs) with various QbSH databases.

References

[1] R. Typke, F. Wiering, and R. C. Veltkamp, “A survey of music information retrieval systems,” in *Proc. of 6th International Conference on Music Information Retrieval*, pp. 153-160, September 11-15, 2005.
http://ismir2005.ismir.net/proceedings/1020.pdf

[2] X. Wu, M. Li, J. Liu, J. Yang, and Y. Yan, “A top-down approach to melody match in pitch contour for query by humming,” in *Proc. of International Symposium on Chinese Spoken Language Processing*, vol. 2, pp. 669-680, December 13-16, 2006.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.110.1802

[3] K. Kim, K. R. Park, S.-J. Park, S.-P. Lee, and M. Y. Kim, “Robust query-by-singing/humming system against background noise environments,” *IEEE Trans. Consumer Electron.*, vol. 57, no. 2,

pp. 720-725, May 2011. Article (CrossRef Link).

[4] G. P. Nam, K. R. Park, S.-J. Park, S.-P. Lee, and M.-Y. Kim, “A new query-by-humming system based on the score level fusion of two classifiers,” *Int. J. Commun. Syst.*, vol. 25, issue 6, pp. 717-733, June 2012. Article (CrossRef Link).

[5] G. P. Nam, T. T. T. Luong, H. H. Nam, K. R. Park, and S.-J. Park, “Intelligent query by humming system based on score level fusion of multiple classifiers,” *EURASIP J. Adv. Signal Process.*, vol. 2011:21, pp. 1-11, July 2011. Article (CrossRef Link).

[6] A. Kornstadt, “Themefinder: a web-based melodic search tool,” *Computing in Musicology*, MIT Press, 1998, vol. 11, pp. 231-236.
http://www.ccarh.org/publications/books/cm/vol/11/contents.html

[7] S. Blackburn and D. DeRoure, “A tool for content based navigation of music,” in *Proc. of ACM International Conference on Multimedia*, pp. 361-368, September 12-16, 1998. Article (CrossRef Link)

[8] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. V. Oostrum, “Using transportation distances for measuring melodic similarity,” in *Proc. of International Conference on Music Information Retrieval*, pp. 107-114, October 26-30, 2003.
http://ismir2003.ismir.net/papers/Typke.PDF

[9] J.-S. R. Jang and M.-Y. Gao, “A query-by-singing system based on dynamic programming,” in *Proc. of International Workshop on Intelligent Systems Resolutions*, pp. 85-89, December 11-12, 2000.
http://ir.lib.nthu.edu.tw/bitstream/987654321/17662/1/2030226030026.pdf

[10] L. Prechelt and R. Typke, “An interface for melody input,” *ACM Trans. Computer-Human Interact.*, vol. 8, no. 2, pp. 133-149, June 2001. Article (CrossRef Link).

[11] M. Ryynanen and A. Klapuri, “Query by humming of MIDI and audio using locality sensitive hashing,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2249-2252, March 31-April 4, 2008. Article (CrossRef Link).

[12] J.-S. R. Jang and H.-R. Lee, “A general framework of progressive filtering and its application to query by singing/humming,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 2, pp. 350-358, Feb. 2008. Article (CrossRef Link).

[13] S.-P. Heo, M. Suzuki, A. Ito, and S. Makino, “An effective music information retrieval method using three-dimensional continuous DP,” *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 633- 639, June 2006. Article (CrossRef Link).

[14] N. Phiwma and P. Sanguansat, "A novel method for query-by-humming using distance space," in *Proc. of International Conference on Pervasive Computing Signal Processing and Applications*, pp. 841-845, September 17-19, 2010. Article (CrossRef Link)

[15] K. Lemström and E. Ukkonen, “Including interval encoding into edit distance based music comparison and retrieval,” in *Proc. of Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, pp. 53-60, April 17-20, 2000.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.6339

[16] M. Mongeau and D. Sankoff, “Comparison of musical sequences,” *Computers and the Humanities*, vol. 24, no. 3, pp. 161-175, June 1990. Article (CrossRef Link).

[17] A. Kotsifakos, P. Papapetrou, J. Hollmén, and D. Gunopulos, “A subsequence matching with gaps-range-tolerances framework: a query-by-humming application,” in *Proc. of the VLDB Endowment*, vol. 4, no. 11, pp. 761-771, 2011.
http://www.vldb.org/pvldb/vol4/p761-kotsifakos.pdf

[18] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*, PWS Publishing Company, 1996.
http://dl.acm.org/citation.cfm?id=249049

[19] M. A. Bartsch and G. H. Wakefield, “Audio thumbnailing of popular music using chroma-based representations,” *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96-104, Feb. 2005. Article (CrossRef Link).

[20] D. Jang, C.-J. Song, S. Shin, S.-J. Park, S.-J. Jang, and S.-P. Lee, "Implementation of a matching engine for a practical query-by-singing/humming system," in *Proc. of IEEE International*

Symposium on Signal Processing and Information Technology, pp. 258-263, December 14-17, 2011. Article (CrossRef Link).

[21] M. Müller, *Information Retrieval for Music and Motion*, Springer, 2007. Article (CrossRef Link).

[22] G. P. Nam and K. R. Park, “Multi-classifier based query-by-singing/humming system on mobile device,” *Multimedia Systems*, in submission.

[23] G. P. Nam and K. R. Park, “Fast Query-by-Singing/Humming System that Combines Linear Scaling and Quantized Dynamic Time Warping Algorithm,” *KSII Transactions on Internet and Information Systems*, in submission.

Tuyen Danh Pham received a B.S degree in Electronics and Telecommunications from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2010. In Feb. 2013, he received Master’s degree and is currently pursuing a Ph.D. degree in Electronics and Electrical Engineering at Dongguk University. His research interests include pattern recognition and digital image processing.

Gi Pyo Nam received a B.S. degree in Digital Media Technology from Sangmyung University, Seoul, South Korea, in 2009. He is currently pursuing a combined Master’s and Ph.D. degree in Electronics and Electrical Engineering at Dongguk University. His research interests include biometrics, pattern recognition, and image processing.

Kwang Yong Shin received a B.S. degree in Electronics Engineering from Dongguk University, Seoul, South Korea, in 2009. He is currently pursuing a combined Master’s and Ph.D. degree in Electronics and Electrical Engineering at Dongguk University. His research interests include biometrics and image processing.

Kang Ryoung Park received B.S. and Master’s degrees in Electronic Engineering from Yonsei University, Seoul, South Korea, in 1994 and 1996. He received a Ph.D. degree in Electrical and Computer Engineering from Yonsei University in 2000. He was an assistant professor in the Division of Digital Media Technology at Sangmyung University until Feb. 2008. He has been a professor in the Division of Electronics and Electrical Engineering at Dongguk University since March 2013. His research interests include image processing and biometrics.