International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 13 (2017) pp. 3752-3756

Query by Singing/Humming System Based on Deep Learning

Jia-qi Sun * and Seok-Pil Lee**

**Department of Computer Science, Graduate School, Sangmyung University, Seoul, Korea.*

*** Department of Electronic Engineering, Sangmyung University. Korea.*

*(**Correspondence Author)*

Abstract

With the proliferation of digital music, efficient indexing and retrieval tools are required for searching the desired music in a large digital music database (DB). Traditional text-based information retrieval methods (titles, lyrics, singers, etc.) cannot meet people's needs now. Music information retrieval (MIR) has been a matter of interest. Therefore, how to retrieve the music information quickly and effectively becomes the focus of current research. In this paper, we propose a query-by-singing/humming system based on deep learning. The result shows our method is very promising in comparison with the published QbSH system based on monophonic database.

Keywords: Query by Singing/Humming, Deep Learning, MIDI Music Library

INTRODUCTION

With the proliferation of digital music, efficient indexing and retrieval tools are required for searching the desired music in a large digital music database (DB). Query-by-singing/humming is a representatively convenient and intelligent method in the field of content-based music retrieval systems. It can be used for retrieving a music file without singer's name and song title based on the melody of the music hummed/sung by a user.

In previous researches, the various kinds of QBSH systems have been researched [1], [2], [3]. Ghias et al. processed the method of representing the pitch contour features extracted from the humming or whistle data as an up-down-repeat (UDR) string and using them for matching [2]. McNab et al. processed the MELDEX system based on the pitch contour, interval, and duration with string matcher [3], [4]. In the previous research [5], they proposed the Tuneserver representing the pitch contour as the UDR string. Kornstadt et al. developed the Themefinder system which has the capability of searching the theme of music in the Humdrum database of classic music of the 16th century and folk songs on the web [6], [7]. They showed the retrieval method using the changes of melody and the UDR string [8]. Ryynanen and Klapuri proposed the method of extracting the pitch vectors by using a fixed-size time window and matching them by using locality sensitive hashing (LSH) method [9]. In another study [10], they adopted earth mover's distance (EMD) method which could calculate the minimum cost between the features of humming and reference data with the changes of the weight to measure melodic similarity. In the previous research [11], they proposed the method of content-based music retrieval which firstly filters out 80% unlikely candidates by using hierarchical filtering method and compares the input query with the remaining candidates. Salamon and Rohrmeier proposed the two-stage retrieval method for QBSH system [12]. As the first stage, the number of candidates is reduced by the indexing method using n-grams. And detail matching with the remaining candidates is performed with the remaining candidates based on local alignment with modified cost functions. Wang et al. proposed the QBSH system by combining the EMD and dynamic time (DTW) classifiers based on the weighted SUM rule [13].

The rest of this paper is organized as follows. Section 2 presents the general QbSH system. Section 3 explain our QbSH system for monophonic recordings. Section 4 and 5 present experimental results and conclusions, respectively.

GENERAL QbSH SYSTEM

Before introducing our method, we briefly summarize the preparation for QBSH system.

Main research contents of content-based by humming music

retrieval include music database construction, feature extraction, melody matching three parts.

A. Overview of the Proposed Method

For the QBSH system, DTW algorithm has been widely used for matcher.

Dynamic Time Warping(DTW) algorithm was first put forward by the Japanese scholar Itakura, and essentially it is a measure of the length of two different time series similarity method.

In time series, it is necessary to compare similarity of two paragraphs may not be equal to the length of the time series, and in the field of speech recognition performs different people having different speaking speed. Because speech signal has considerable randomness, even the same person at different times to speak the same word, could not have exactly the same time length. As shown in the figure2.1.

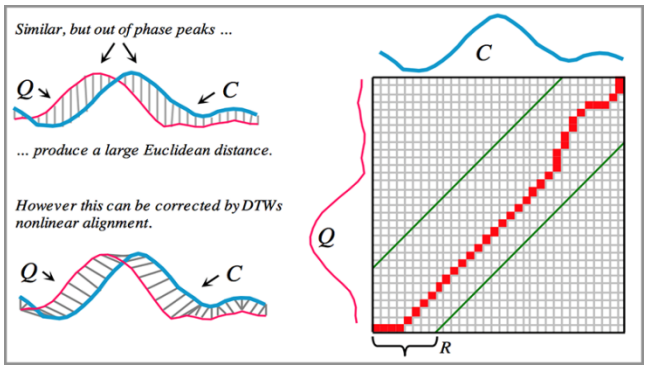

Figure 2.1: Dynamic Time Warping Theory

Dynamic Time Warping is a typical optimization problem. DTW algorithm is to calculate minimum distance of the two template matching.

$$\gamma(i,j) = d(q_i, c_j) + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\}$$

B. Pitch Extraction

Pitch is one of the most important and universal feature of music pieces. Autocorrelation function (ACF) [16] is particularly useful in estimating hidden periodicities in signal. The function is :

$$\mathrm{ACF}(n) = \frac{1}{N-n}\sum_{k=0}^{N-1-n} x(k)x(k+n)$$

Where N is the length of signal x, n is the time lag value. The value of n that maximize ACF(n) over a specified range is selected as the pitch period in sample points. If ACF has highest value at n=K, then K is the chosen time period of signal, and the fundamental frequency is 1/K.

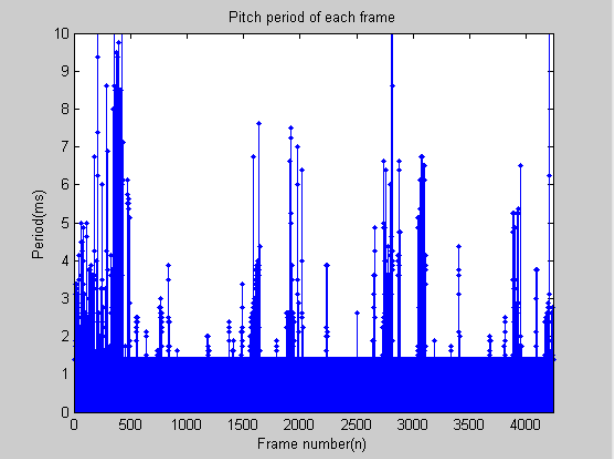

Figure 2.2: An example of autocorrelation function

QBSH SYSTEM CONSTRUCTION

In QBSH system, the most important part is melody matching. Here, we propose to use deep learning method to achieve QBSH system.

A. Deep Belief Networks

Deep belief network is a multi-layer hidden variable probability generation model, in which each layer is to capture the underlying hidden features of a higher order correlated process. At the top of the two layers of deep belief networks formed a undirected bipartite graph, and the bottom two layers formed directed sigmoid belief network, as shown in Figure2.3. Deep belief network is mainly made up of undirected bipartite graph model of Restricted Boltzmann Machine, an exponential model, and it is widely used in the collaborative filtering, information, and image retrieval.[17]

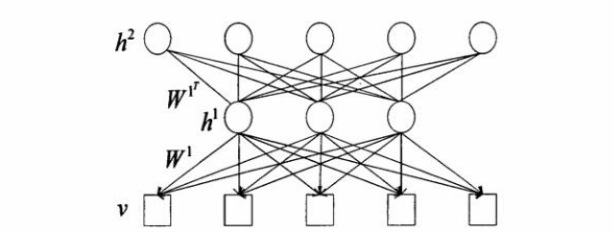

Figure 2.3: Two hidden layers of DBN, $W^2 = W^{1T}$

B. Restricted Boltzmann Machine

Restricted Boltzmann Machine is a kind of special markov random field with two layers structure, one is visible layer with statistical unit $v \in \{0,1\}^D$; Another is hidden layer with statistical with $h \in \{0,1\}^F$, at this point visible layer state is 0,1, named Bernoulli-Bernoulli RBM(BBRBM), when the visible layer state is Gaussian model data, named Gaussian-Bernoulli RBM(GBRBM).

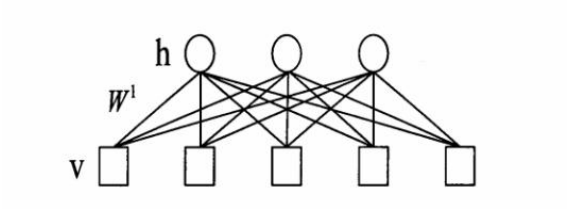

Figure 2.4: Restricted Boltzmann Machine

Restricted Boltzmann Machine energy function: $E(v, h; \theta) = -v^T W h - b^T v - a^T h$

$$= -\sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{F} a_j h_j$$

D is the visible layer unit number; F is hidden layer unit number; $\theta = \{W, a, b\}$ is the model parameters; W_{ij} is weight coefficient between visible layer unit I and hidden layer unit j; a_i and b_j are bias of unit.

C. The Greedy Algorithm

Assuming that deep belief networks have two layers of hidden units $\{h_1, h_2\}$, and the second hidden layer units are same as visible layer units, as shown in Figure2.3. At the top of two layer is composed of undirected binary chart, and at the bottom of two layer is composed of directed sigmoid belief network. In order to present the derivation process of greedy learning algorithm, we will omit offset of visible layer and hidden layer. In this model, the joint probability distribution of v, h^1, h^2 is

$$P(v, h^1, h^2; \theta) = P(v|h^1; W^1)P(h^1, h^2; W^2)$$

$\theta = \{W^1, W^2\}$ is model parameter; $P(v|h^1; W^1)$ is directed sigmoid belief network; $P(h^1, h^2, W^2)$ is the joint probability distribution of second layer Restricted Boltzmann Machine.

$$P(v|h^1; W^1) = \prod_i p(v_i|h^1; W^1)$$

$$p(v_i = 1|h^1; W^1) = g(\sum_j W_{ij}^1 h_j^1$$

$$P(h^1, h^2; W^2) = \frac{1}{Z(W^2)} \exp(h^{1^T} W^2 h^2)$$

Greedy algorithm is mainly composed of the following process. Because a DBN with two hidden layers' weighted parameters meet $W^2 = W^{1^T}$, and the joint distribution $P(v, h^1; \theta) = \sum_{h^2} P(v, h^1, h^2; \theta)$ of the bottom of DBN is same as the joint probability distribution $P(v, h^1, W^1)$ of Restricted Boltzmann Machine. As you can see in Figure2.3 and Figure2.4, marginal distribution $P(h^1; W^1)$ and conditional distribution $P(v|h^1; W^1)$ of the bottom of RBM.

Greedy learning algorithm through constructing multi-layer restricted Boltzmann machine to achieve. As shown in Figure2.5.

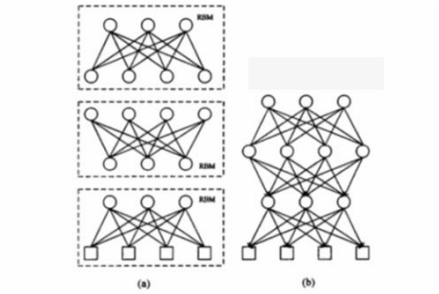

Figure 2.5(a): Greedy learning multi-cascade RBMs
(b) The corresponding three layers of DBN

EXPERIMENT

In this paper, experiments were done in Windows 8 32-bit operating system. Programming environment is Matlab 2010a.

A. Database

We selected 10 people as collection object, including 3 boys and 7 girls. Each person recorded 10 songs. They chose the part they like to hum 10 seconds. Recording environment is relatively quiet.

We used midi database which we download the newest songs from internet, including 100 Chinese songs and 100 Korean songs.

B. Results

For experiments, we used midi database we constructed. To measure the matching accuracy, the mean reciprocal rank (MRR) is used as the criterion of performance, and it has been frequently used for measuring the accuracy of QBSH system[12].

$$\text{MRR} = \frac{1}{K}\sum_{i=1}^{k}\frac{1}{rank_i}$$

Where K is the number of input singing/humming files and $rank_i$ is the ranking of the correct midi file(corresponding to the input file), as calculated by the proposed method.

From that, we can confirm that the deep learning method can show better performance than DTW method.

Table 1. The matching accuracy of DTW method and Deep Learning method

QBSH	Size	Top3(%)	Top10(%)	MRR
DTW	200	73	89	0.79
DL	200	81	93	0.82

In previous researches, with midi 100 database, the result of DTW method is 0.81. As we can see, in our system, we use midi 200 database, and the result of deep learning method showed 0.82.

CONCLUSION

This paper proposes a QbSH system based on deep learning. In previous researches, DTW is typically adopted as a matcher for QbSH systems. But we use a deep learning method as a matcher. The experimental results showed that the proposed method enhanced the matching accuracy.

As a future work, we will focus on the design of more accurate pitch extractor using advanced learning algorithm based on polyphonic music contents like mp3 files.

ACKNOWLEDGMENT

This research was supported by a 2017 Research Grant from Sangmyung University.

REFERENCES

[1]. Typke R., Wiering F., Veltkamp R. C. “A survey of music information retrieval systems” , Proceedings of the International Conference on Music Information Retrieval, September 2005

[2]. Ghias A., Logan J., Chamberlin D., Smith B.C. “Query by humming: musical information retrieval in an audio database ”, Proceedings of ACM International Conference on Multimedia (MULTIMEDIA95) November 1995

[3]. McNab R.J., Smith L.A., Witten I.H., Henderson C.L., Cunningham S.J. “Towards the digital music library: tune retrieval from acoustic input”, Proceedings of the 1st ACM International Conference on Digital Libraries, March 1996

[4]. McNab R. J., Smith L.A., Bainbridge D., Witten I.H. “The New Zealand digital library melody index”, D-lib Magazine 1995

[5]. Prechelt L., Typke R. “An interface for melody input ”, ACM Transactions on Computer-Human Interaction 2001

[6]. Kornstadt A. “Themefinder: a web-based melodic search tool”, Computing in Musicology 1998

[7]. Themefinder ‘http://www.themefiner.org/’

[8]. Blackburn S., DeRoure D. “ A tool for content based navigation of music”, Proceedings of ACM International Conference on Multimedia 1998

[9]. Ryynanen M., Klapuri A. “A query by humming of midi and audio using locality sensitive hashing”, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, April 2008

[10]. Typke R., Giannopoulos P., Veltkamp R.C., Wiering F., Oostrum R.V. “Using transportation distances for measuring melodic similarity”, Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR’ 03), October 2003

[11]. Jang J.-S.R., Lee H.-R. “Hierarchical filtering method for content-based music retrieval via acoustic input”, Proceedings of the ACM International Conference on Multimedia, October 2001

[12]. Salamon J., Rohrmeier M. “A quantitative evaluation of a two stage retrieval approach for a melodic query by example system”, Proceedings of the 10th International Society for Music Information Retrieval

Conference, October 2009

[13]. Wang L., Huang S., Hu S., Liang., Xu B. "An effective and efficient method for query by humming system based on multi-similarity measurement fusion", Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP'08), July 2008

[14]. Wu X., Li M., Liu J., Yang., Yan Y. "A top-down approach to melody match in pitch contour for query by humming", Proceedings of the International Symposium of Chinese Spoken Language Processing 2006

[15]. Jang J.-S.R., Gao M.-Y. "A query-by-singing system based on dynamic programming", Proceedings of the International Workshop on Intelligent Systems Resolutions 2000

[16]. X.-D. Mei, J. Pan, S.-h. Sun, "Efficient algorithms for speech pitch estimation", Intelligent Multimedia, Video and Speech Processing. Proceedings of 2001 International Symposium on. IEEE

[17]. Salakhutdinov R. "Learning deep generative models[D]." Canada: University of Toronto, 2009