MIREX 2015 QBSH Task: SOGOU's Solution

Wenqi Tang[1,2], Guangchao Yao[2], Wei Chen[2] and Limin Xiao[1]
1. Beihang University
{twq}@buaa.edu.cn
2. Sogou Technology (Beijing) Co., Ltd.
{tangwenqi, yaoguangchao, chenweibj8871}@sogou-inc.com

ABSTRACT

This short document mainly describes our submission to the QBSH (Query by Singing/Humming) task of MIREX (Music Information Retrieval Evaluation eXchange) 2015. Our system uses a four-stage QBSH matcher to achieve both low latency and high accuracy. In our system, FFT (Fast Fourier Transform) and PAA (Piecewise Aggregate Approximation) are used for quick filtering. EMD (Earth Mover's Distance) and DTW (Dynamic Time Warping) are used for accurate matching. In addition, we take advantage of a re-matching method based on the idea of self-learning for further improvement.

1. INTRODUCTION

A QBSH system can tell its user which song he is singing or humming. A typical QBSH system mainly consists of three parts which are model builder, pitch tracker and matcher. Model builder is an offline module. It tries to build a database based on MIDI files. Pitch tracker and matcher are online modules. Pitch tracker transforms a user's query into pitch and note sequences. And matcher finds the song which is closest to user's query from the MIDI database.

There are two sub-tasks in the QBSH task of MIREX, Classic QBSH task and Variants QBSH task. In Classic QBSH task, there are four test sets. Two of them are from Roger Jang's corpus and the other two are from ThinkIT's corpus. In each test set, 2000 Essen MIDIs are added to the ground truth MIDIs as noise. The result is measured by MRR (Mean Reciprocal Rank).

Test Set	Hidden Jang	Jang	ThinkIT	IOACAS
Number of Queries	1790	4431	355	404
Number of Ground Truth MIDIs	48	48	106	106

Table 1. Summary of the test sets in Classic QBSH task.

In Variants QBSH task, queries are treated as variants of "ground-truth" midis. In this task, the test set is from Roger Jang's corpus. 2039 of the 4431 queries are used as queries and others are used as "ground-truth" midis. The result is measured by top-10 simple count.

2. SYSTEM DESCRIPTION

Our submission includes three modules. They are model builder, pitch tracker and QBSH matcher. The overall architecture is shown in Fig. 1.

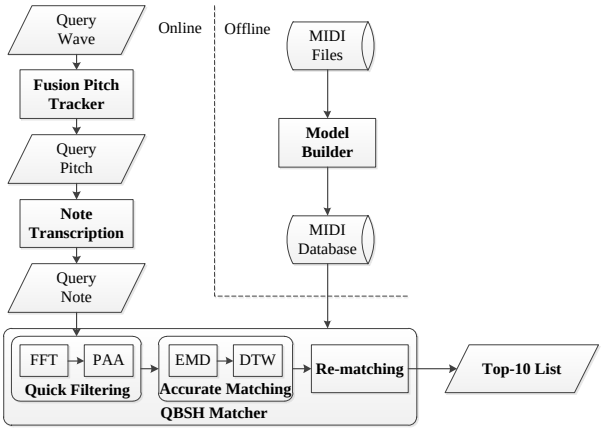

Figure 1. Overall architecture of our system

2.1 Model Builder and Pitch Tracker

2.1.1 Note Duplication Removal and Phrase Segmentation

While building the database, firstly duplicated melodies are removed from each MIDI file based on Max Repeated Substring (MRS) algorithm. Then each MIDI file is spilt into several phrases. This is based on the idea that most people hum from the beginning of a music phrase [1]. At last FFT is performed to each phrase to generate the vectors needed in the matcher module.

2.1.2 Fusion Pitch Tracker and Note Transcription

When the user sings or hums a song, the fundamental frequency (F0) sequence is extracted. Here we use a fusion of three methods, autocorrelation [2], swipe [3] and PYIN [4]. After the F0 sequence is got, query note sequences used in EMD are also generated.

2.2 QBSH Matcher and Re-matching

2.2.1 Quick Filter of FFT and PAA

In order to filter out dissimilar sequences quickly, FFT and PAA are used. Other than measuring the distance between sequences in the time domain, FFT performs distance computation in the frequency domain [5]. And it has been proved to be efficient for filtering.

PAA is a dimensionality reduction technique [6]. It's useful for high-dimensional or large-scale data. Though PAA is slightly slower than FFT, it can achieve higher accuracy. Through the combination of these two methods, more than 80% of the dissimilar sequences are filtered out in a short period.

2.2.2 Accurate Matching of EMD and DTW

After quick filtering, EMD and DTW are used to get more accurate results. This is based on Wang's idea that EMD has robust properties to errors brought from the front end and a weighted combination of EMD and DTW can bring better results [7]. In our system, we carefully tuned the parameters to get higher accuracy.

2.2.3 Re-matching Method

Even after accurate matching, the confidence of some results is still not high enough. In this occasion, an additional post-process stage of re-matching is added. In this additional stage, history query information is used for system self-learning to improve the performance further.

3. CONCLUSION AND DICUSSION

In this document we describe our submission to the QBSH task of MIREX 2015. Our four-stage QBSH system can find out the proper result in most cases. However sometimes the top-1 result is not what you want. Such case usually does not mean that the system is working improperly. If you compare the curves of query pitch and the incorrect top-1, you will find that they are indeed similar to each other. Such problems are still needed to be solved in the future.

4. REFERENCES

[1] Huang S, Wang L, Hu S, Jiang H and Xu B: "Query by humming via multiscale transportation distance in random query occurrence context," Multimedia and Expo, 2008 IEEE International Conference on, pp. 1225-1228, 2008.

[2] Boersma P: "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proceedings of the institute of phonetic sciences, Vol. 17, No. 1193, pp. 97-110, 1993.

[3] Camacho A: "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music", University of Florida, 2007.

[4] Mauch M and Dixon S: "pYIN: A fundamental frequency estimator using probabilistic threshold distributions", Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 659-663, 2014.

[5] Tsai W H, Tu Y M and Ma C H: "An FFT-based fast melody comparison method for query-by-singing/humming systems", Pattern Recognition Letters, Vol. 33, No. 16, pp. 2285-2291, 2012.

[6] Keogh E, Chakrabarti K, Pazzani M, et al: "Dimensionality reduction for fast similarity search in large time series databases", Knowledge and information Systems, Vol. 3, No. 3, pp. 263-286, 2001.

[7] Wang L, Huang S, Hu S, et al: "Improving searching speed and accuracy of query by humming system based on three methods: Feature fusion, candidates set reduction and multiple similarity measurement rescoring", Ninth Annual Conference of the International Speech Communication Association, 2008.