Segment-based Melody Matching for Query By Singing/Humming

Wen-Hsing Lai[a], Chi-Yong Lee

Department of Computer and Communication Engineering, National Kaohsiung First University of Science and Technology, No. 1, University Rd., Yanchao Dist., Kaohsiung City 824, Taiwan

Abstract. Segment-based melody matching approaches are proposed to solve the problems of pitch instability, tempo inconsistency, and puff noise of singing/humming in Query By Singing/Humming system. Instead of comparing frame by frame from the beginning and trying different scaling ratio, our segment-based approaches reduce the computation complexity by jumping to the possible segment boundary and using the segment ratio as scaling ratio. Our methods predict segment boundary candidates in query singing/humming and songs in database by detecting the abrupt change of pitch curve, match segment pair, scale each segment linearly and calculate the distance. The top 10 recognition rate of our method can reach 70.6%. Our method reduce computational load and performs better in situations of puff noise and inconsistent tempo than popular global Linear Scaling method.

1 Introduction

Due to the vigorous development of the Internet, a large amount of information flow on the web, which makes information more easily accessed. However, in this flood of information, how to find the information needed becomes a major issue and challenge. Keyword search engine is created, for quickly finding the information we want. Similar situation happens on the large amount of music on the Internet. When we think of a piece of music but do not know the song title or artist, how can we find it? Query By Singing/Humming, (QBSH) [1-7] is the technology developed for this demand.

QBSH is a melody recognition system by using a singing or humming query. The system will compare the singing or humming query with the songs stored in database, and report the most similar song to the user. The mostly used feature for recognition is frame-level pitch or note. Pitch information is generally gotten from pitch tracking algorithm [6]. The song with the smallest feature distance with the query is the most possible answer.

There are currently several techniques of QBSH. The most common ones are Linear Scaling (LS), Dynamic Time Warping (DTW) or their combinations [3][5]. Although the recognition rate of DTW, searching for the path of alignment with the minimum pitch distance, is higher, the computational complexity is also higher and the recognition time is longer. LS, which simply stretches or compresses the query pitch contour globally and matches it frame-by-frame (one pitch point per frame) with the target pitch contour, is simpler. It uses more intuitive linear stretch. Its computational complexity is lower and recognition speed is faster.

[a] Corresponding author: lwh@nkfust.edu.tw

But when LS encounter the situation of short time puffs, which usually appears at the beginning of humming or singing, or unstable tempo, which is also very common especially for amateurs, the recognition rate will decline because of its simple global scaling. Besides, we do not know the singing or humming starting position in songs and we do not know the exact humming tempo variation, thus we generally need to compare from the beginning frame by frame with different scaling factor, which increases computational load. We try to solve the puff and unstable tempo problems and speed the recognition time by introducing segment-based approaches.

The segment-based methods we propose determine segment boundaries of query singing/humming and songs in database firstly, and then choose the possible segment matching pair, and stretch each segment by linear scaling. We hope the segment-based methods can eliminate the recognition inaccuracy caused by puff noise and tempo inconsistency. Since the query singing/humming may not start right from the very beginning frame, jumping to the possible segment boundary directly instead of comparing frame by frame from beginning can also reduce computational load.

This paper is organized as follows. The next section will introduce the background of LS, which our methods are based on. Then, the segment-based methods we propose are presented in the third section. Experimental results will be shown in the following section. Finally, conclusions and future works are discussed.

2 Linear Scaling

In this paper, we propose segment-based Linear Scaling approaches. Segment boundaries are predicted and the query and songs in database are split into segments for matching. Inside each matching segment, LS matching is used. To understand it, the traditional LS is introduced in this section.

LS [6][7] is also called uniform scaling or global scaling. It is a frame-based method for melody recognition. Typical LS working on frame-level pitch uses interpolation to uniformly expand or compress the pitch curve of query singing/humming, and compare the scaled pitch curve with the songs in database. Different scaling ratio are tried and compared. Distances to all songs in database with different scaling ratio are calculated. The song in database with the minimum distance is the most likely song for the query singing/humming.

3 Segment-based approaches

In the segment-based methods we proposed, firstly, for query singing/humming and songs in database, we select the points with abrupt change of pitch as the segment boundary candidates. Secondly, from the segment boundary candidates in query singing/humming, four segment boundaries, the first, the last, the highest, and the second highest boundaries, are chosen. Between the segment boundaries, there are three segments. Because the query singing/humming does not necessarily correspond to the very beginning of midi in database, the corresponding matching segment boundary in midi is searched from the midi segment boundary candidates with the same rising or decreasing pitch trend. For all possible matching pairs, we calculate the distance of three segments. Inside segment, linear scaling is applied to stretch or compress the length. The scaling ratio is simply the ratio of segment length, so the query segment is scaled to the same length with the midi segment. Since the expansion or compression of segment and key shifting bring bias in calculating distance, distance normalization, which normalizes the distance by the length of the segment, and key transposition, which shifts the pitch to the same mean as that of midi song in database, are used. The segment pair with the shortest distance is the best match. After compare all the songs in database, the song with the shortest distance is the answer to the query singing/humming.

Our four approaches of the first step, selecting segment boundary candidates, are introduced in the following subsections. We call them "Cross *n* Semitones", "Moving Average", "Composite Moving Average", and "Combination of Cross *n* Semitones and Composite Moving Average".

3.1. Cross n Semitones

"Cross *n* Semitones" selects the points with pitch variation over *n* semitones within *m* frames as possible segment boundary candidates. If *n* is large, we get fewer candidate points. If *n* is small, we get more candidate points, but too small variation means they may be just caused by pitch drift phenomenon instead of song trend, and such selection of *n* may cause mistake. For convenience, when *n* equals 3, it is called "Cross 3 Semitones".

3.2 Moving Average

There are drifts in the pitch of query singing/humming. The technique of moving average on pitch curve can smooth out short-term fluctuations, reflecting the long-term trend, so we can reduce the interference in the process of selecting the segment boundary candidates. When the window length of moving average is *w*, we call it *w*MA. For example, 20MA means moving average with window length 20. The greater the *w* is, the smoother the pitch curve is. Over smooth curve reduces the influence of pitch drift, but loses the original pitch features, resulting a lot of songs look similar. On the other hand, if the *w* is small, it is less smooth, and keeps more original pitch features, but more vulnerable to the influence of pitch drift.

For convenience, we mark the segment boundary with continuously upward pitch frame as a positive boundary, and the segment boundary with consecutive decline pitch frame as a negative boundary. When we try to find the corresponding matching segment boundaries in song database, it is convenient that we only search from the segment boundary candidates in song database with the same rising (positive) or decreasing (negative) pitch trend.

If the pitch trend (upward or decline) value exceeds a threshold *t* we set, we mark it as a possible segment boundary candidates. The same procedure is done on songs in database to find the possible segment boundary candidates. Note that the moving average only applied in finding possible segment boundary candidates. When calculating the distance, the original pitch is used.

3.3 Composite Moving Average

Moving average curves with different *w* overlap at area with large dynamic pitch variation. Therefore, this overlapped point can be used as a possible segment boundary candidate. We call this method "Composite Moving Average". The number of moving average curves considered at one time is fixed as 3 in our experiment. The gap between the *w* is *g*. We express it as *w*MA(*g*). For example, 20MA(5) means composite moving average with window length 20, 15, and 10.

3.4 Combination of Cross n Semitones and Composite Moving Average

This method is a combination of "Cross *n* Semitones" and "Composite Moving Average". If a point is selected by both "Cross *n* Semitones" and "Composite Moving Average", then it is selected as a segment boundary candidate.

4 Experiments

4.1. Database

Two databases are used in our experiment. One is 2009 MIR-QbSH corpus including 35 singers, totally 829 songs. The recording is in wave format with 8kHz sampling rate, 8 bits per sample, and mono channel. Pitch files are also included. The frame size for pitch is 256 samples. The other database is called Mono-Midi including 342 midi songs, which is the corpus we collected from Internet.

4.2. Experimental results

We compare the top 1 and top 10 recognition rate of our four segment-based methods as shown in Table 1. A, B, C, D respectively indicate our segment-based methods using "Cross 3 Semitones with m=5", "20MA with t=10", "20MA(5) with t=35", and "Combination of Cross 3 Semitones with m=5 and 20MA(5) with t=35".

Two examples are also provided and observed. The first example is a query singing with puff noise at the beginning. As shown in Figure 1, the (lower) blue line is the singing pitch curve and the (upper) green line is the corresponding midi pitch curve. It is easy to see the puff noise at the beginning of the blue singing pitch curve. After applying our segment-based methods, the singing and midi pitch curves after segment-based scaling are shown in Figure 2. For comparison, the singing and midi pitch curves after applying traditional LS are also shown in Figure 3. Since our methods are segment-based, they are allowed to jump to the most matching segment boundary and exclude the noisy segment. Therefore, puff noise has less impact on the recognition results of our methods than others. For this example, the correct midi ranks 1, 1, 1, and 1 in experimental results by using our methods A, B, C, and D respectively. It ranks 154 and 139 by using traditional LS and DTW.

Another example is a query singing with inconsistent tempo. As shown in Figure 4, the blue line is the singing pitch curve and the green (stepwise) line is the corresponding midi pitch curve. The singing and midi pitch curves after applying traditional LS are shown in Figure 5. If we align the two curves at the position indicated by black (middle) arrow, we can see the inconsistent tempo costs shifting at the position indicated by (both side) red arrows. The singing and midi pitch curves after segment-based scaling by applying our segment-based methods are shown in Figure 6. The two pitch curves fits better than using traditional LS. For this example, the correct midi ranks 1, 1, 1, and 1 in experimental results by using our methods A, B, C, and D respectively. It ranks 86 and 1 by using LS and DTW.

Based on the above experimental results, the problems of puff noise and inconsistent tempo which usually encountered in singing/humming, can be solved by segment-based methods. DTW can solve tempo inconsistence problem, though, it carries more computational load.

Table 1. Comparion of the Recognition Rate of Our Four Segment-Based Approaches.

	A	B	C	D
Top 1	52.2%	55.4%	52.4%	56.2%
Top 10	60.3%	70.1%	70.6%	67.8%

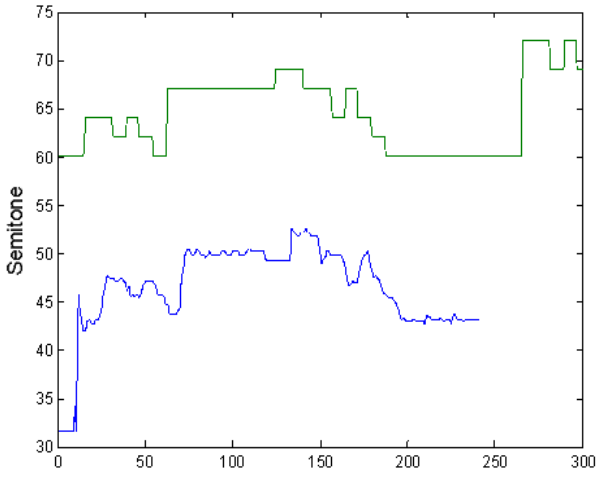

Figure 1. A singing example with puff noise at the beginning. The (lower) blue line is the singing pitch curve and the (upper) green line is the corresponding midi pitch curve.

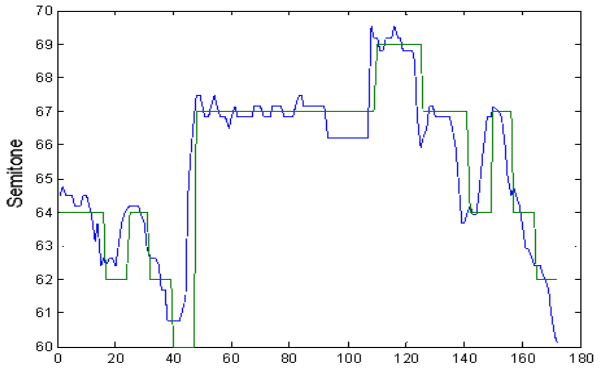

Figure 2. The singing and midi pitch curves after segment-based scaling. The blue line is the scaled singing pitch curve and the green (stepwise) line is the corresponding midi pitch curve.

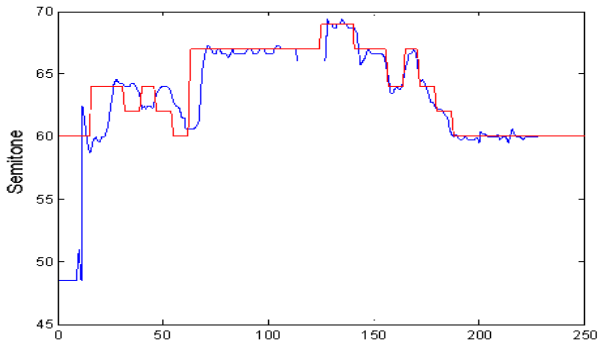

Figure 3. The singing and midi pitch curves after applying traditional LS. The blue line is the scaled singing pitch curve and the red (stepwise) line is the corresponding midi pitch curve.

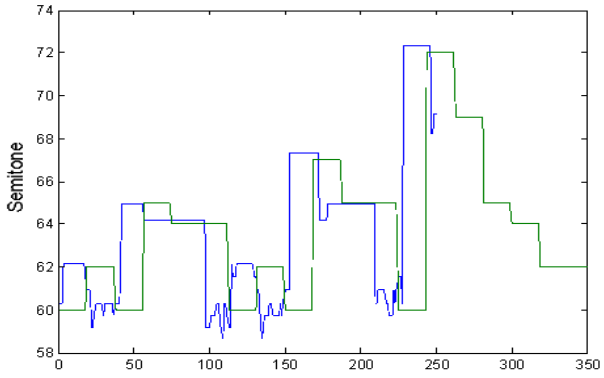

Figure 4. A singing example with inconsistent tempo. The blue line is the singing pitch curve and the green (stepwise) line is the corresponding midi pitch curve.

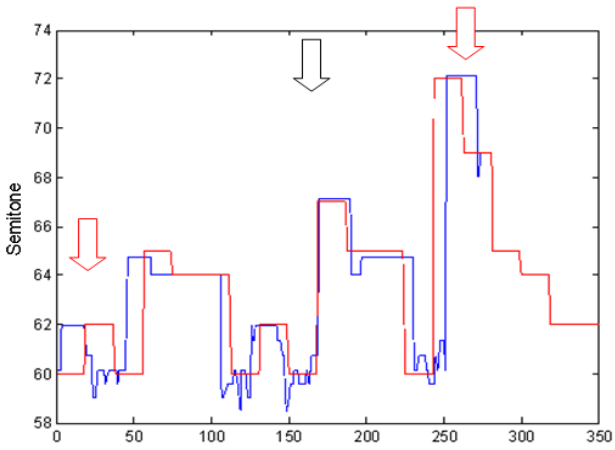

Figure 5. The singing and midi pitch curves after applying traditional LS. The blue line is the scaled singing pitch curve and the red (stepwise) line is the corresponding midi pitch curve.

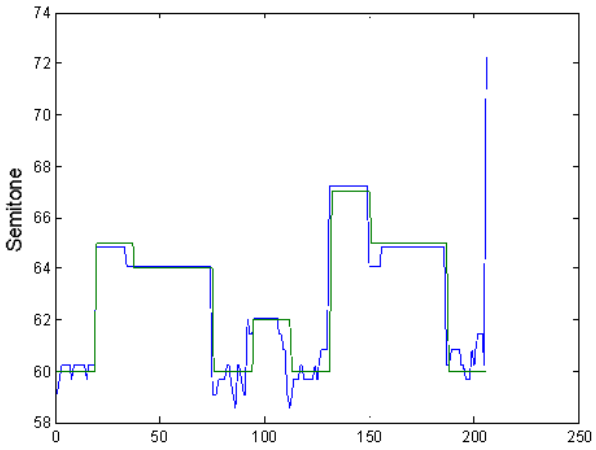

Figure 6. The singing and midi pitch curves after segment-based scaling. The blue line is the scaled singing pitch curve and the green (stepwise) line is the corresponding midi pitch curve.

5 Conclusions and future works

We propose four segment-based approaches to solve the problems of puff noise and inconsistent tempo in Query-By-Singing/Humming system. The segment-based methods also reduce the computation complexity by jumping to the possible segment boundary and using the segment ratio as scaling ratio, instead of comparing frame by frame from the very beginning and trying different scaling ratio in traditional LS method. Experimental results show that the best top 10 recognition rate is 70.6%. In addition, from our examples, it shows that segment-based method can solve the problems of puff noise and inconsistent tempo better than traditional global LS.

Improving the segment accuracy by observing and considering more features besides pitch variation will be the future work that we want to focus on.

References

1. Q. Wang, Z. Guo, G. Liu, C. Li, J. Guo, ICASSP, 3711 (2013)
2. C. C. Wang, C. H. Chen, C. Y. Kuo, L. T. Chiu, J. S. R. Jang, ICASSP, 477 (2012)
3. W. T. Kao, C. C. Wang, K. K. Chang, J. S. R. Jang, W. Liou, APSIPA, 1 (2013)
4. W. Cao, D. Jiang, J. Hou, Y. Qin, T. F. Zheng, Y. Liu, ICME, 942 (2009)
5. G. P. Nam, K. R. Park, INT J DISTRIB SENS N, 1 (2015)
6. S. Y. Cheng, *An Evaluation of Query By Singing/Humming Based on Various Pitch Tracking Methods* (Master Thesis, National Tsing Hua University, Taiwan, 2008)
7. H. Y. Zhuang, *Early Screening of Inappropriate Inputs for QBSH Systems*, (Master Thesis, National Tsing Hua University, Taiwan, 2009)