QUERY BY SINGING/HUMMING (MIREX 2016) THE TUNE FOLLOWER

Bartłomiej Stasiak
Institute of Information Technology
Łódź University of Technology, Poland
bartlomiej.stasiak@p.lodz.pl

ABSTRACT

This extended abstract presents a solution to the QbSH (Query by Singing / Humming) problem, submitted to MIREX 2016 QbSH evaluation contest. The solution is based on the well-known DTW (Dynamic Time Warping) approach, modified by the use of *tune-following* procedure. In this way it is possible to alleviate the problem (resulting i.a. from imprecision in sung queries) of fitting absolute pitch values between a query and a template.

1. INTRODUCTION

Dynamic Time Warping is a standard tool used in QbSH tasks. It is often considered more robust and precise, yet considerably slower, than note-based approaches, with the best results being obtained by combination of the two [6]. One of the problems of direct DTW application is that it is basically not robust to transpositions (both fixed and occurring in the course of a query), while in melody matching only the relative pitch changes between consecutive pitch values are of importance. The solution may be searched for e.g. in mean value subtraction, in using delta sequences or in trying several transpositions of one of the sequences [5] [7]. An alternative to those approaches is presented in the following section.

2. SYSTEM DESCRIPTION

The solution proposed in this extended abstract is to try to follow the melody of the template by gradually decreasing the difference between the query and the template. This is intended to resemble the way in which humans follow the known melody irrespective of pitch inaccuracies and key changes.

The detailed description of the method is presented in [1]. Basically, for any query-template pair of pitch vectors being compared, the optimal warping path is found with the DTW algorithm. Then, the consecutive pitch values q_i of the time-warped query are compared to the corresponding pitch values t_i in the template, in order to produce a modified query sequence $\hat{q}$. The modification aims at making the $\hat{q}$ sequence closer to the template t by gradually "tuning" it to the right pitch values. This is obtained by a simple algorithm depicted in Fig. 1. Finally, the $\mathrm{DTW}(t, \hat{q})$ value is used as a similarity measure instead of $\mathrm{DTW}(t, q)$.

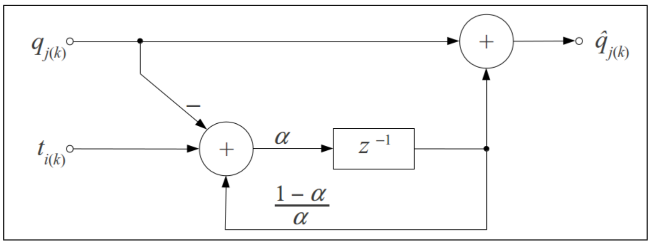

Figure 1. The block diagram of the tune follower [1].

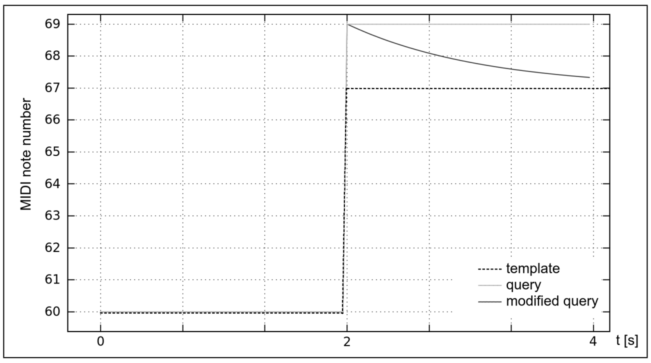

Figure 2. Effect of the tune-following procedure (the last two seconds).

An example, in which a template fragment contains a perfect fifth while the user sings a major sixth instead, is presented in Fig. 2. An example for the whole song is shown in Fig. 3, where the effect of the tune-following procedure, with α parameter set to 0.1, is clearly visible [1].

Several additional pre/post-processing operations are involved to optimize the solution. Median filter is used to remove impulse noise from the pitch vectors. The basic DTW algorithm is based on the original work by Sakoe and Chiba [4] with slope constraint parameter $P = 1/2$. Preliminary alignment of the absolute pitch values between the query and the template (transposition) is performed on the basis of the mean value of the first half of the query, after rejecting some initial values (<500 ms).

Reduction of computation time is achieved be means of

[1] Note, that in practice, much lower values (e.g. $\alpha = 0.03$) seem to yield optimal recognition enhancement [1].

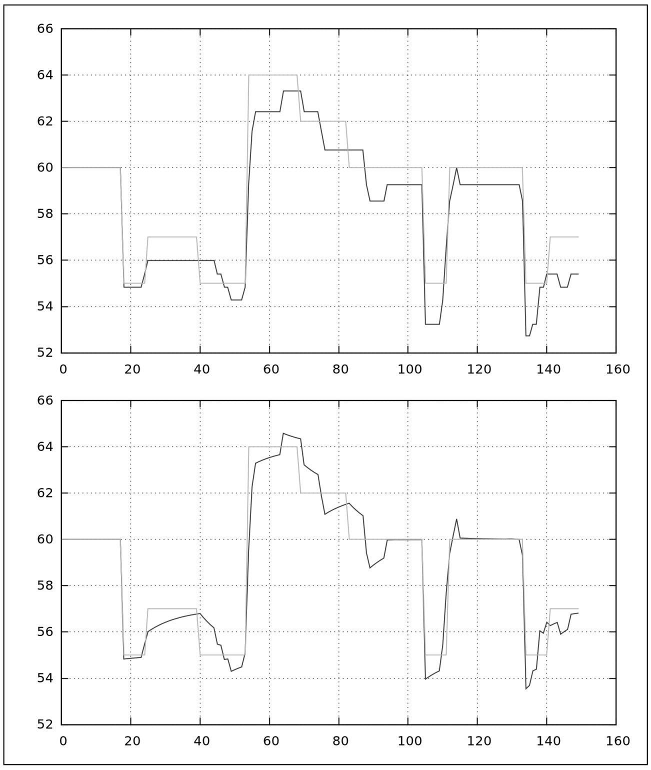

Figure 3. *"Old McDonald had a farm"* - the template (light) and the median-filtered query (dark). Top: original sequences after time-warping. Bottom: the result of application of the tune follower for $\alpha = 0.1$.

indexing the template sequences in R*-tree structure [8]. Every template is represented by 20 points in 16-dimensional search space, where each point is obtained with piecewise aggregate approximation (PAA, [8]) of a different part cut from the template sequence.

A novel cutting scheme based on discrete total variation (DTV) of the melody sequence is applied both for queries and templates, as described in [3]. The one-dimensional DTV may be defined as:

$$\mathrm{DTV}(n) = \sum_{k=1}^{n} |p(k) - p(k-1)| \ , \tag{1}$$

where $p(k)$ denotes the k-th time frame of the pitch vector.

The fundamental property of DTV is that it accumulates pitch changes in the course of the melody, irrespective of the actual direction of the changes (Fig. 4). We may therefore set a threshold T_{DTV} for the accumulated pitch changes and cut all the compared melodies when they reach T_{DTV}, as follows:

$$p_c = [p(0), p(1), ..., p(n_c)] \ , \tag{2}$$

where p_c denotes the pitch vector reduced to the first $n_c + 1$ values and where:

$$n_c = \min\{n \in N; DTV(n) \geq T_{\mathrm{DTV}}\} \ . \tag{3}$$

Setting a fixed threshold value T_{DTV} in Eq. 3 allows to obtain similar melody sections both for the query and for the template, irrespective of any potential tempo discrepancy between the two. For example – considering the first motif of the melody from Fig. 4, shown in Fig. 5 – if we set $T_{DTV} = 5$, both sequences would be cut at the onset of the 4th note. Some further properties of the proposed DTV-based cutting scheme are discussed in [3].

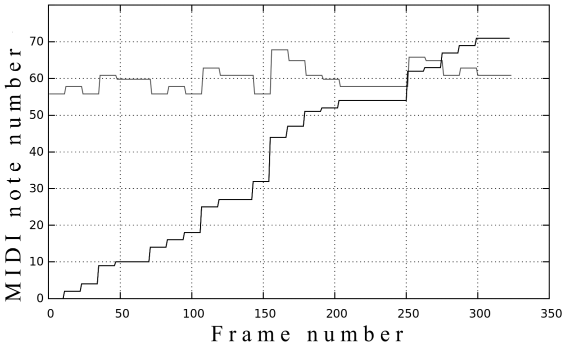

Figure 4. Pitch vector representation of "Happy Birthday" melody (top, light-grey) and the corresponding DTV sequence (bottom, dark-grey).

In order to enhance the results of our QbH solution in the cases where several consecutive notes have the same pitch (repetitions), the elements of rhythm analysis have been introduced. The note onsets are detected both in the templates and in the queries and they are involved in the computation of the DTW cost. The onset detection is straightforward in the case of the MIDI-based templates, while for the queries an approach based on frequency-domain onset detection function (ODF) is applied [2].

3. RATIONALE

It has been shown [1] that the tune-following procedure has a positive influence on the DTW-based melody search. Although it generally makes the matching cost smaller for most of the templates, both matching and non-matching ones, it can be expected that in the first case (the matching templates) this decrease will be more significant.

This may result from the effect of accumulation of the corrections for consecutive notes. For example, when the pitch of a note sung by a user is too high with respect to the matching template, then it is gradually decreased until it reaches the right tune (provided that the note is long enough – otherwise it will get at least *partially* corrected). Generally, if the note was sung too high, then it is probable that the pitch of the next note will also be too high, in which case it will get corrected immediately or – at least – faster. Similar observations may be made if a part of a query is sung too low with respect to the template. In either case, the total cost of matching a query with the corresponding (correct) template will most probably get significantly reduced.

This type of correspondence between the signs of the pitch differences in consecutive notes cannot be generally expected when comparing a query with a non-matching template. In this case, correcting one note may as well result in *increasing* the initial difference between the next

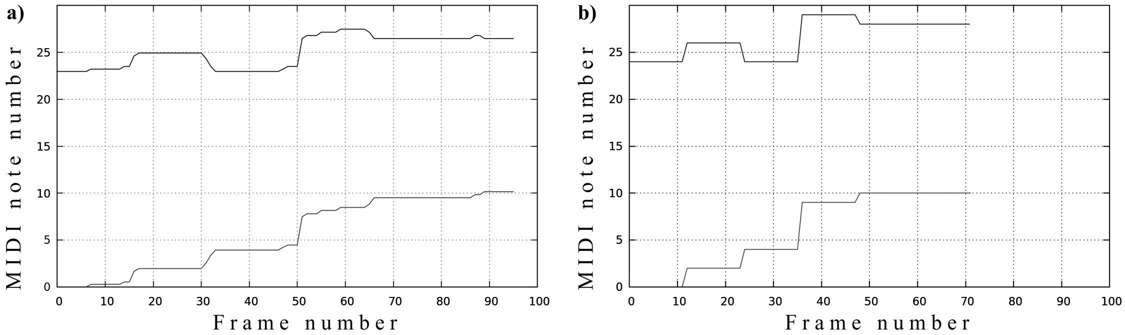

Figure 5. The first motif of the melody from Fig. 4: **a)** query; **b)** template. Top plots present the original pitch vectors (after transposition by 3 octaves down, for visualization purposes); the bottom plots show the DTV sequences.

note and the template. In some cases, this may even lead to increasing the total matching cost for the incorrect templates.

4. REFERENCES

[1] B. Stasiak, *Follow That Tune - Adaptive Approach to DTW-based Query-by-Humming System*, Archives of Acoustics, Vol. 39, No. 4, pp. 467 – 476 (2014)

[2] B. Stasiak, J. Mońko, and A. Niewiadomski, *Note onset detection in musical signals via neural-network-based multi-ODF fusion*, International Journal of Applied Mathematics and Computer Science, Vol. 26 (1), pp. 203 – 213 (2016)

[3] B. Stasiak, *DTV-based melody cutting for DTW-based melody search and indexing in QbH systems*, in Proc. ISMIR 2016

[4] H. Sakoe, and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition*. IEEE Trans. on Acoustics, Speech and Signal Processing, pp. 43–49, 1978

[5] H.-M Yu, W.-H. Tsai, and H.-M. Wang, *A Query-by-Singing System for Retrieving Karaoke Music*. IEEE Trans. on Multimedia, Vol. 10(8), pp. 1626–1637, 2008

[6] L. Wang, S. Huang, S. Hu, J. Laing, and B. Xu, *An effective and efficient method for query by humming system based on multi-similarity measurement fusion*. Int. Conf. on Audio, Language and Image Processing, pp. 471–475, 2008

[7] www.music-ir.org/mirex/abstracts/2011/JSSLP1.pdf

[8] E. Keogh, *Exact indexing of dynamic time warping*. In 28th International Conference on Very Large Data Bases, pp. 406–417, 2002

[9] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Wiley-IEEE Press, 2012

Acknowledgement:
The *Tune Follower* software has been tested on the PLATON project's infrastructure at the Technical University of Lodz Computer Center (http://ck.p.lodz.pl).