REAL-TIME AUDIO TO SCORE ALIGNMENT USING LOCALLY-CONSTRAINED DYNAMIC TIME WARPING OF CHROMAGRAMS

Kosuke Suzuki, Yushi Ueda, Stanisław A. Raczyński, Nobutaka Ono, and Shigeki Sagayama
Graduate School of Information Science and Technology, The University of Tokyo
{suzuki, ueda, raczynski, onono, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This extended abstract describes our score follower submitted to the MIREX 2010 Real-time Audio to Score Alignment (a.k.a. Score Following) task. The score alignment is obtained by on-line Dynamic Time Warping (DTW). Feature extraction used in our system is sum of chroma and delta chroma vectors.

1. SYSTEM OVERVIEW

Score alignment problem is time alignment of a musical score and the audio signal of its performance. In this MIREX task, the score and the audio signals are given in the formats of MIDI and WAV.

Here we describe an overview of our score follower. First we convert MIDI data into a reference audio signal using a sequencer. By generating the reference audio signal, the audio-to-MIDI matching problem is simplified as an audio-to-audio matching problem. Next, both the performance and reference audio signals are transformed to chromagrams. Finally, the alignment is obtained by Dynamic Time Warping (DTW) between the performance and reference chromagrams. We show an overview of the score following system and an example of the alignment in Figure 1.

2. FEATURE EXTRACTION

Although difference of octaves is useful information for score following, difference of octave is not always very clear in spectrum. Thus for the robustness, we used chroma feature extraction which does not distinguish octaves. Chroma vector is a 12-dimensional vector, each of whose element is total power of the frequencies corresponding to one note in all octaves [1]. To take dynamics into account, we used sum of chroma and delta chroma vectors as feature vectors, refered to as $\mathrm{chroma} + \Delta\mathrm{chroma}$ in the following. First chroma vectors are normalized by the sum of the elements. Second, $\Delta\mathrm{chroma}$ is obtained by differentiating the normalized chroma vectors of the current and previous frames. Finally, we obtain the $\mathrm{chroma} + \Delta\mathrm{chroma}$ vector

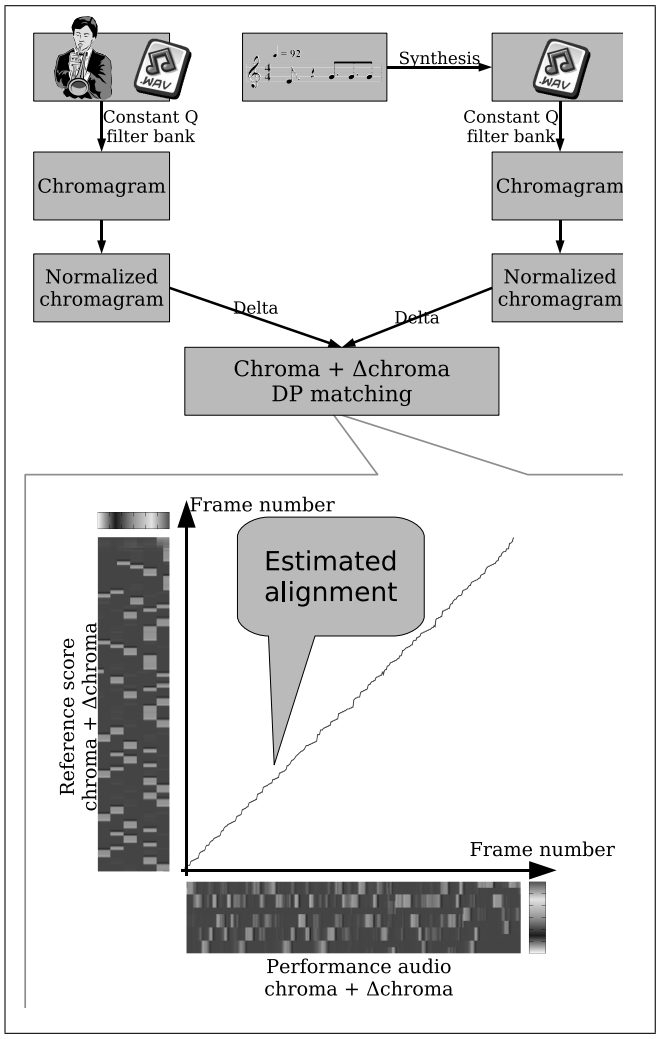

Figure 1. System overview and an example of alignment result with the reference database of MIREX 2006 (available at http://cosmal.ucsd.edu/arshia//mirex06-scofo/).

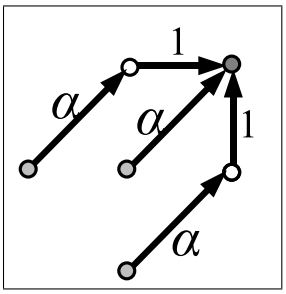

Figure 2. Constrained DTW path

by summing the normalized chroma and $\Delta\mathrm{chroma}$ with equal weighting.

3. MATCHING BY DTW

In this section we describe matching between the feature vectors based on DTW.

3.1 Local path of DTW

We used the following constrained DTW path:

$$D(i,j) = \min \left\{ \begin{array}{ccc} d(t,j) + \alpha d(t-1,j) & + & D(t-2,j-1) \\ \alpha d(t,j) & + & D(t-1,j-1) \\ d(t,j) + \alpha d(t,j-1) & + & D(t-1,j-2) \end{array} \right\}, \quad (1)$$

where t is the index of the current performance frame to be searched, j is the index of reference frame, $d(t,j)$ is the Euclidyan distance between the performance chroma of the t-th frame of and the reference chroma of the j-th frame, $D(t,j)$ is the accumulated distance when the performance signal is at the t-th frame and the reference signal is at the j-th frame, and α is the weighting of the diagonal steps. Normally α is set to be $1 \leq \alpha \leq 2$ and we used $\alpha = \sqrt{2}$. This constrained path inhibits successive occurrences of vertical or horizontal steps, and smooth path is estimated. The constrained path is shown in Figure 2.

3.2 Real-time DTW

Standard DTW assumes off-line search and the estimated path is obtained by backtracing of whole the signal. To extend DTW for the on-line search without backtracing, we simply select the reference frame which has the smallest accumulated distance with the current performance frame t. The configuration of the on-line DTW is shown in Fig. Figure 3.

4. RESULTS

We show the results of MIREX 2010 in Table 1.

	AW1	DP1	RVCC3	RVCC4	**SUROS**
Total precision	50.84%	49.11%	32.17%	32.44%	**73.97%**
Piecewise precision	50.33%	67.14%	62.79%	64.50%	**73.93%**

Table 1. Overall performance

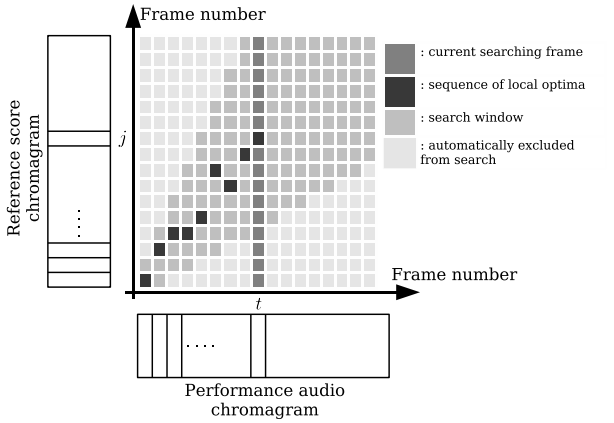

Figure 3. Configuration of the real-time version of DTW. For each performance frame t, the optimal reference frame j with the smallest accumulated distance is selected.

5. REFERENCES

[1] T. Fujishima, "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music," *Proc. the International Computer Music Conference*, pp. 464–467, 1999.