Real-time Audio to Score Alignment

Kosuke Suzuki

Graduate School of Information Science and Technology,
The University of Tokyo

suzuki@hil.t.u-tokyo.ac.jp

Outline

- Background
- Previous Works in This Area
- MIREX 2010 Implementation and Evaluation Results
- Summary and Conclusion
- Future Works

- Background
- Previous Works in This Area
- MIREX 2010 Implementation and Evaluation Results
- Summary and Conclusion
- Future Works

Background

- Audio to score alignment
  - Aligning performance audio signal to its score
- Off-line audio to score alignment
    - Query by hamming
    - Database
- On-line (Real-time) audio to score alignment
  - Cannot use the 'future' performance information
  - Scores can be treated off-line
    - Automatic page turner
    - Automatic music accompaniment system
      - One of the greatest success

Automatic Accompaniment System

- Components
  - Feature extraction
    - Extracting features from performance audio music signal
  - Score following
    - Matching audio signal to its score
    - Telling the current position
  - Accompaniment control
    - Inference of the notes in score based on the results of score following
  - Music synthesis
    - Generating accompaniment sounds

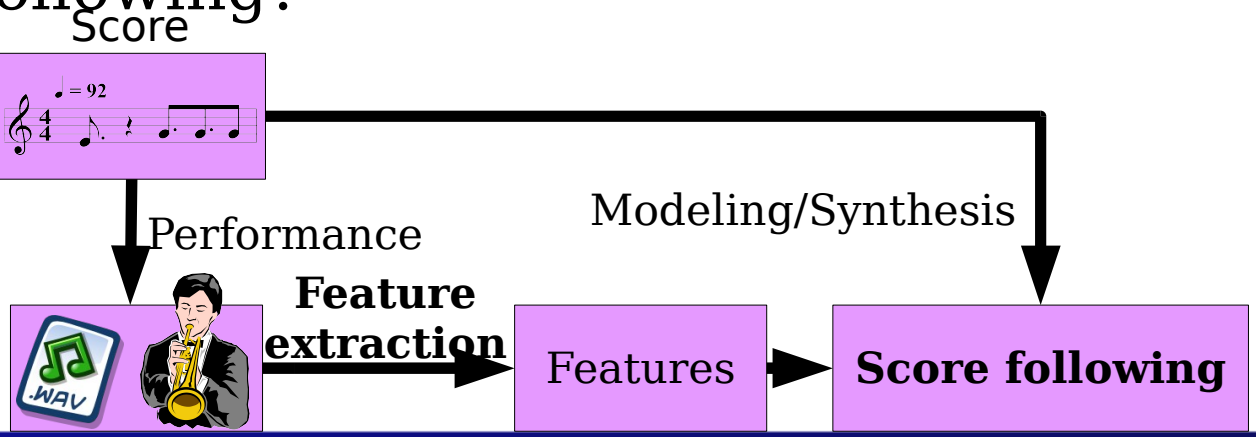
Automatic Accompaniment System

- Components
  - Feature extraction
    - Extracting features from performance audio music signal
  - Score following
    - Matching audio signal to its score
    - Telling the current position
  - Accompaniment control
    - Inference of the notes in score based on the results of score following
  - Music synthesis
    - Generating accompaniment sounds

- Background
- Previous Works in This Area
- MIREX 2010 Implementation and Evaluation Results
- Summary and Conclusion
- Future Works

Problems

- Feature extraction
  - Audio music signal from a performer
    - Problem:
      - Which feature is efficient?
- Score following
  - Matching audio signal to its score
    - Problems:
      - How to treat a score? Modeling or synthesis?
      - What kind of algorithm is suitable for the score following?



Feature Extraction

- Extracting features from audio performance data
- Feature candidates:
  - STFT spectrum
  - Pitch, multi-pitch
  - Power, Delta power
  - Logarithmic spectrum
  - Chroma
  - Delta chroma
    - Combination as a feature vector

Score Following

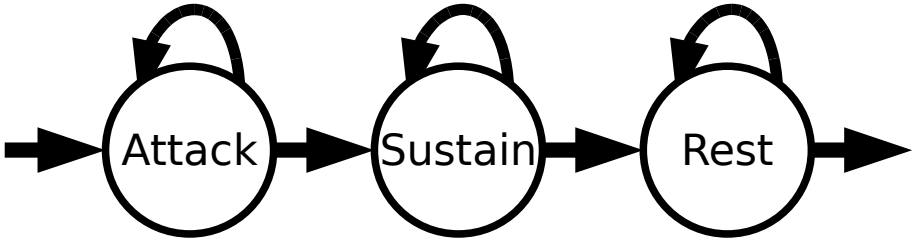
- HMM and training
  - Cont, *et al. MIREX,* 2006.
    - High precision rate

Total Precision:	**82.90%**
Piecewise Precision:	**90.06%**

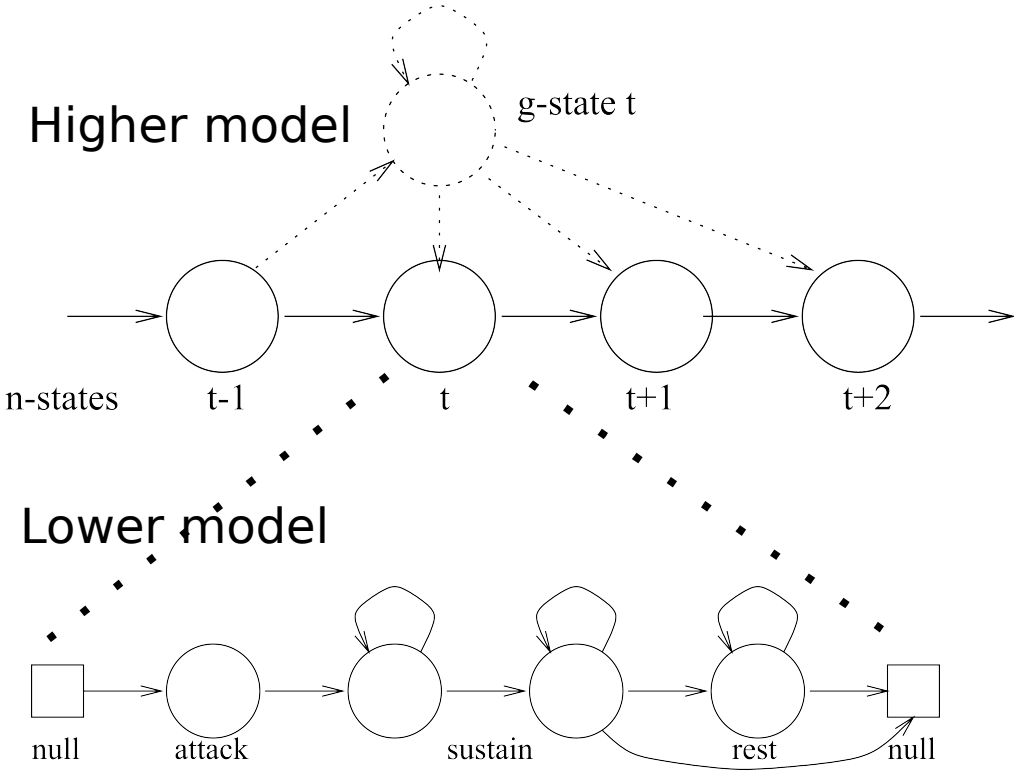
  - Adaptable to jumping and repeating
    - Probabilistic model of arbitrary performances
  - Training is necessary
- Matching by DTW
  - Dixon. *DAFx,* 2005.
  - Score following can be done without training
  - Not so high precision rate as HMM-based one

Using HMMs

- Hidden Markov model (HMM)
  - Probabilistic model
- Probabilistic way
  - Note model
    - Hidden states: states of an instrument
    - Observations: features
  - Tune model with two-level HMM
    - Higher: models for arbitrary performances
    - Lower: models for one note



One of the simplest model of a note

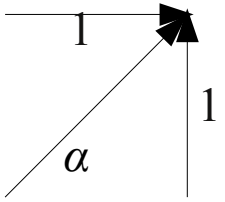


N. Orio, et al. *proc. ICMC,* 2001.

Matching by DTW

- Dynamic time warping (DTW)
  - Based on Dynamic programming (DP)
  - Often used in off-line alignment
    - Backtracing
- Audio – Audio matching
  - After score → audio synthesis
  - Similarity measure
    - Euclidean distance
    - Cosine similarity
  - Local path

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



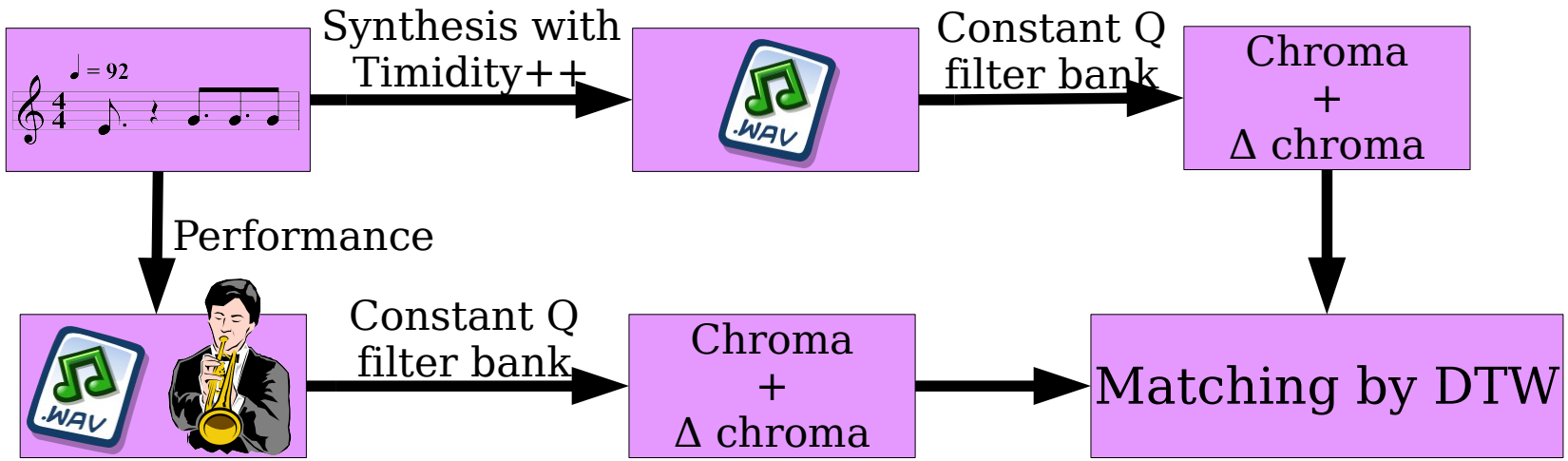
$$1 \leq \alpha \leq 2$$

$d(i,j)$: difference between performance frame i and reference j
$D(i,j)$: accumulated distance from current (i,j)
α:diagonal weighting

- Background
- Previous Works in This Area
- MIREX 2010 Implementation and Evaluation Results
- Summary and Conclusion
- Future Works

Implementation For MIREX 2010

- Feature:
  - Normalized chroma + Δ chroma
- Score following:
  - MIDI → audio synthesis
  - Matching by DTW
    - Local path: constrained path
    - Similarity: euclidean distance



Chroma-based Features

- Chroma: the sum of power of octave
  - Frame-by-frame extraction with constant Q filter bank
- Δ chroma: differentiation of chroma

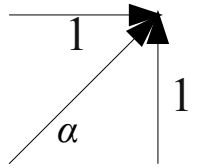
$$\Delta\mathrm{chroma}(t) = \mathrm{chroma}(t) - \mathrm{chroma}(t-1)$$

- Normalization of chroma
  - Adaption to the difference of the amplitude between performances and synthesized score
  - Drawback: enhancement of low-amplitude chroma
- Normalized chroma + delta chroma

Constrained Local Path

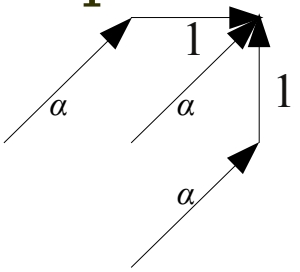
- Assumption: Performances are
  - not faster than twice the references
  - not slower than half the references
- Local path
  - Normal

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



  - Constrained
    - Exclude successive vertical or horizontal steps

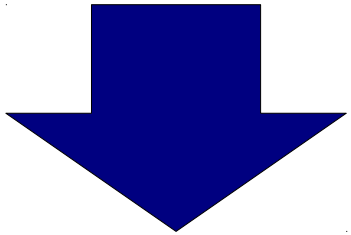
$$D(i,j) = \min \left\{ \begin{array}{rcl} d(i,j) + \alpha d(i-1,j) & + & D(i-2,j-1) \\ \alpha d(i,j) & + & D(i-1,j-1) \\ d(i,j) + \alpha d(i,j-1) & + & D(i-1,j-2) \end{array} \right\}$$



$$\alpha = \sqrt{2}$$

On-line Implementation

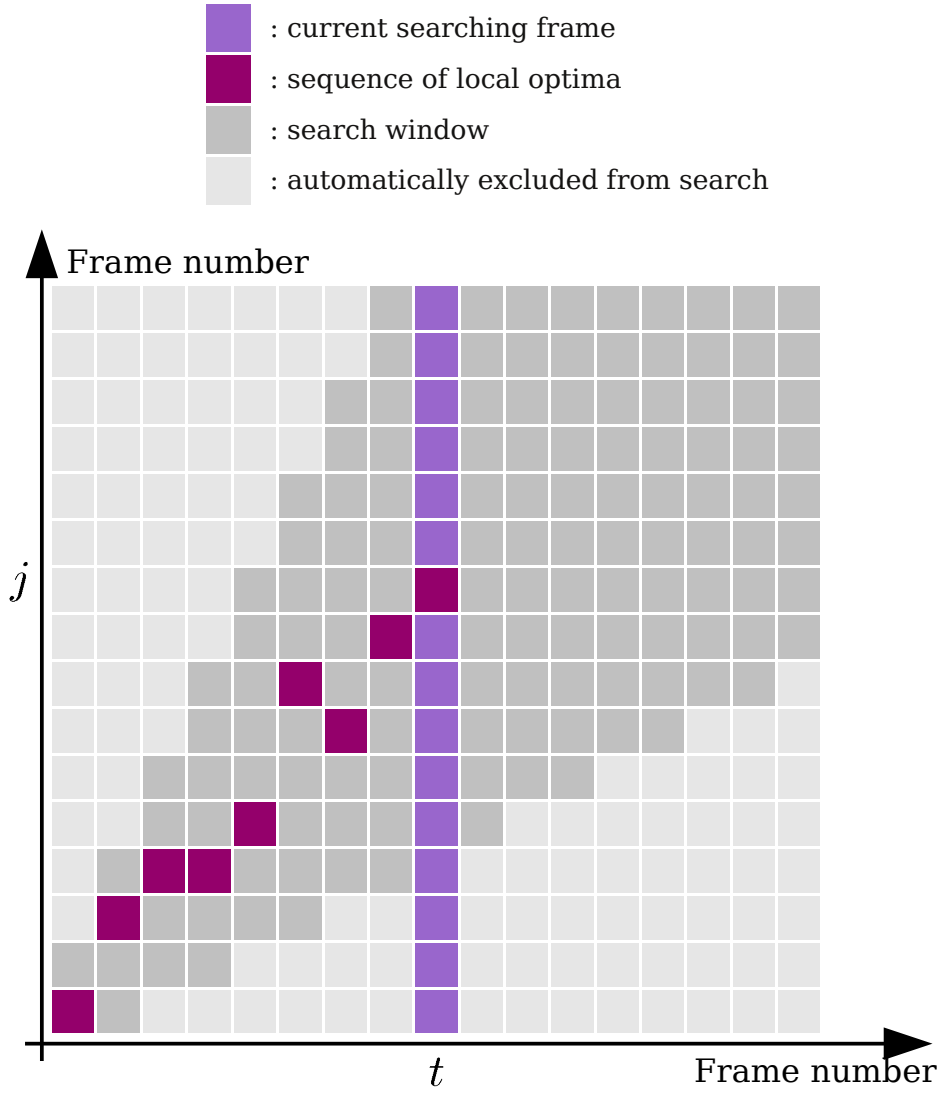
- Causal system cannot use 'future' information



- Calculate current optimal path for each t
  - No backtracing
  - Optimal j is found by

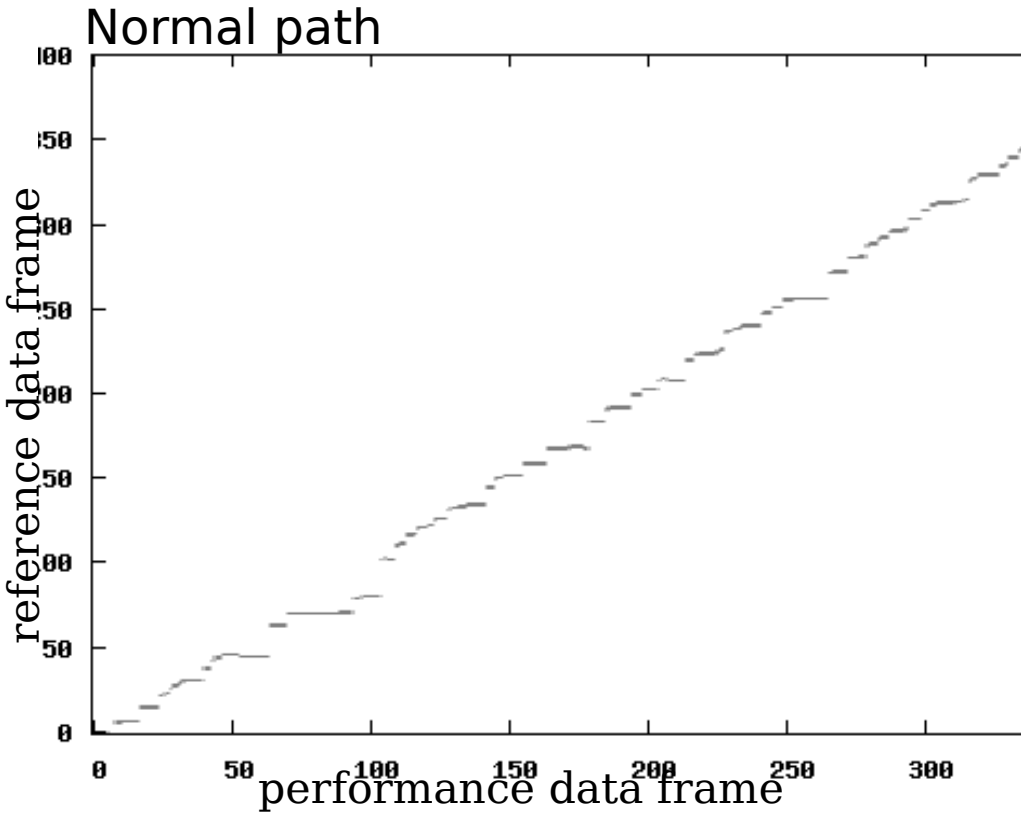
$$j = \operatorname{argmin} D(t, j)$$

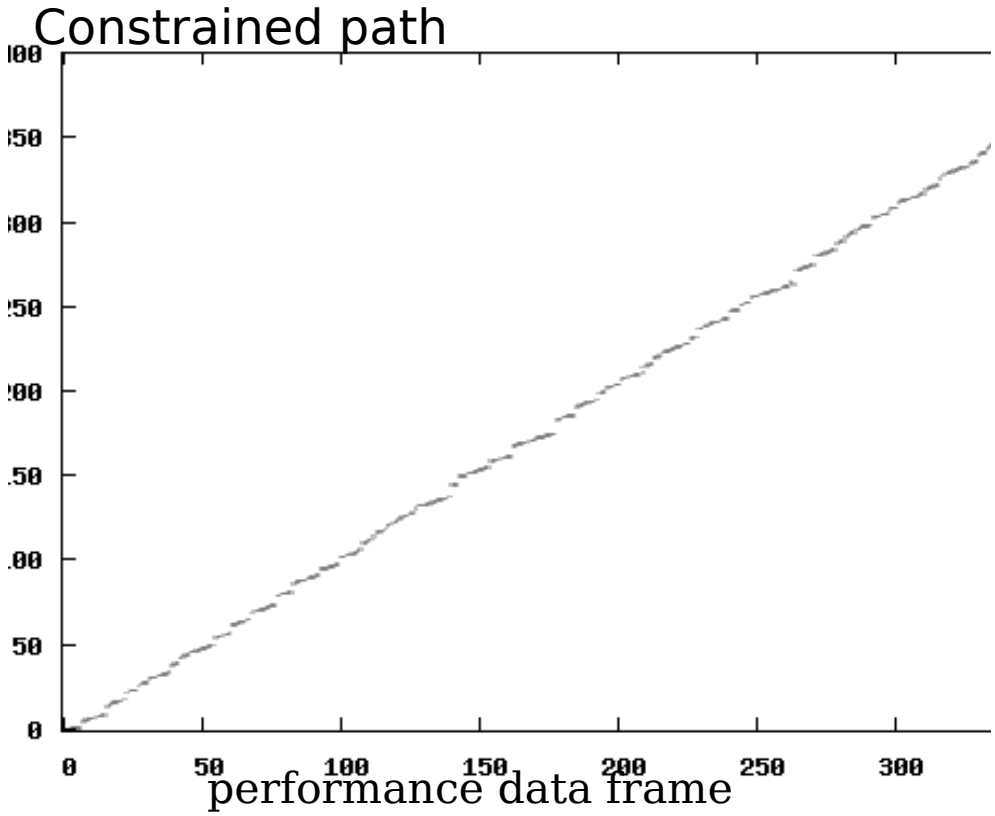
    - The problem about complexity is ignored



Comparison of Local Path

- Promenade (MIREX example test data)
  - Ground truth data is made by manual labeling





	Normal path	Constrained path
Mean offset [ms]	165	101
Standard offset [ms]	257	97

MIREX 2010 Results

- Real-time Audio to Score Alignment (a.k.a. Score Following)
  - Overall performance

	AW1	DP1	RVCC3	RVCC4	**SUROS1**	Cont (2006)
Total Precision	50.84%	49.11%	32.17%	32.44%	**73.97%**	82.90%
Piecewise Precision	50.33%	67.14%	62.79%	64.50%	**73.93%**	90.06%

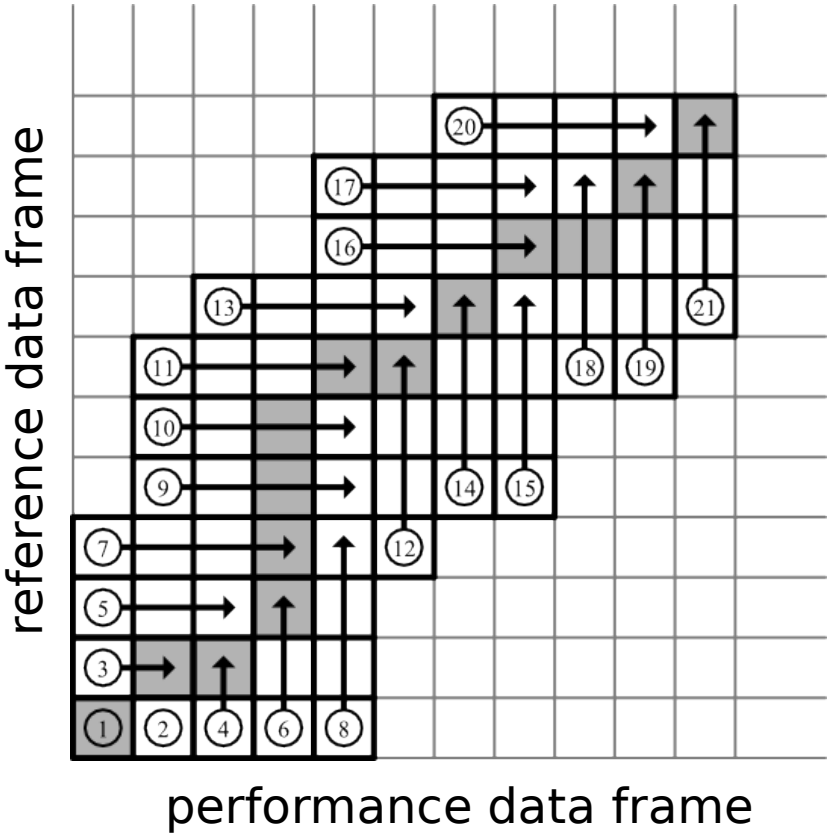
    - The best performance among 2010 participants
    - Cont (*MIREX*, 2006) is better than mine
    - Still some room for improvement

Other MIREX Participants

- Arzt, *et al.* (AW1)
  - Feature: based on STFT spectrum
  - Following: On-line Time Warping (Improved DTW)
    - Using normal local path
- Duan, *et al.* (DP1)
  - Feature: multi-pitches (with observation model)
  - Following: Inference by particle filtering
    - Process model with score position and tempo
- Rodriguez-Serrano, *et al.* (RVCC3, RVCC4)
  - Feature: multi-pitch estimation
    - Nonlinear least-squares method (NLS)
  - Following: DTW (sub-optimal approach)
    - Locally constrained path (but not as mine)

On-line Time Warping

- On-line time warping
  - Arzt, *et al.* (AW1) used
  - Move search window dynamically
  - Make complexity linear to the length of performance data frame

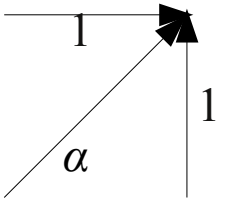


S. Dixon. *DAFx,* 2005.

Another Constrained Path

- Normal path (No constraining)

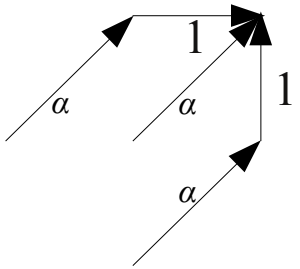
$$D(i,j) = \min \left\{ \begin{array}{ccc} d(i,j) & + & D(i-1,j) \\ \alpha d(i,j) & + & D(i-1,j-1) \\ d(i,j) & + & D(i,j-1) \end{array} \right\}$$



$1 \leq \alpha \leq 2$

- Constrained path (which we used)

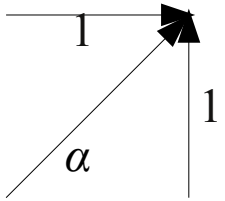
$$D(i,j) = \min \left\{ \begin{array}{ccc} d(i,j) + \alpha d(i-1,j) & + & D(i-2,j-1) \\ \alpha d(i,j) & + & D(i-1,j-1) \\ d(i,j) + \alpha d(i,j-1) & + & D(i-1,j-2) \end{array} \right\}$$



$1 \leq \alpha \leq 2$

- Another constrained path
  - Set a higher cost to vertical and horizontal steps
    - To make the system choose diagonal steps

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



$\boxed{\alpha < 1}$

F. J. Rodriguez-Serrano, et al. *MIREX*, 2010.

- Background
- Previous Works in This Area
- MIREX 2010 Implementation and Evaluation Results
- Summary and Conclusion
- Future Works

Summary and Conclusion

- Implementation of real-time audio to score alignment by DTW with chroma-based features
- Chroma-based feature is more robust than STFT spectrum and multi-pitch
  - Compared to other MIREX 2010 participants
- Alignment by DTW without any training can achieve good results with constrained local path

Future Works

- Comparison among the ways of other MIREX participants
- Quantitative evaluation
  - Experimental data containing ground truth
    - Offline matching and human correcting
- Detail comparison among feature candidates
- Accompaniment control
  - Inference of where a performer is performing