MIPFrontiers

Queen Mary University of London

centre for digital music

Structure-Aware Audio-to-Score Alignment using Progressively Dilated Convolutional Neural Networks

Ruchit Agrawal, Queen Mary University of London
Daniel Wolff, IRCAM Paris
Simon Dixon, Queen Mary University of London

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

2021 IEEE International Conference on Acoustics, Speech and Signal Processing

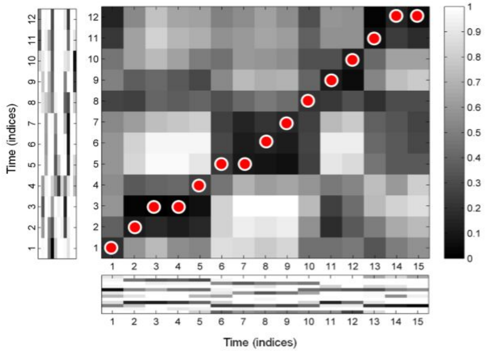
6-11 June 2021 • Toronto, Ontario, Canada

IEEE Signal Processing Society

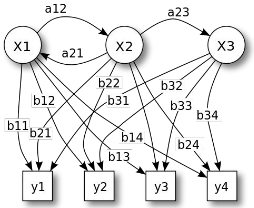
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

Introduction

- Audio-score alignment → important task with applications in performance analysis, score following, page turning, audio editing and so on.
- Aim → Map corresponding/matching positions in the two input sequences (could be different modalities)
- Traditionally done using Dynamic Time Warping (DTW) or Hidden Markov Models (HMMs)

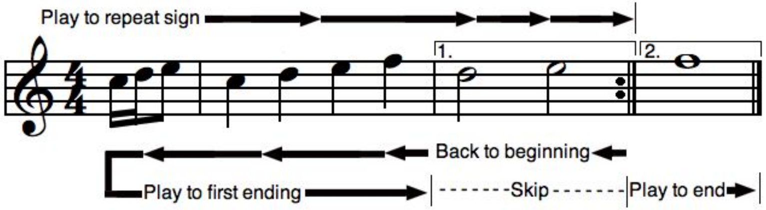


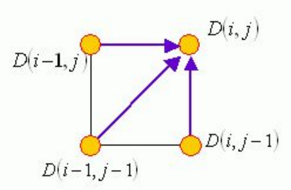
$$D(i,j) = d(i,j) + min \begin{cases} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{cases}$$



Motivation

- Repeats and jumps are an integral part of (classical) music performance
- Capturing structural differences is essential for effective alignment
- DTW/HMM based models do not typically account for structural deviations



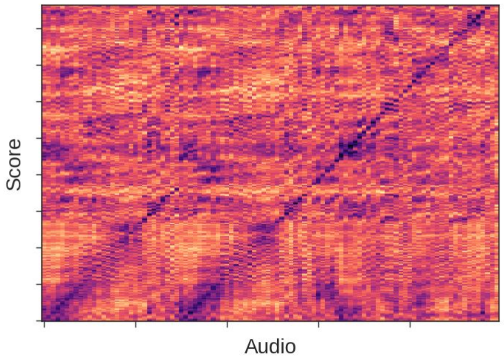


P(future | present, past) = P(future | present, ~~past~~)

Markov property

Existing approaches

- **Classical dynamic time warping (DTW)**
  - Does not handle structural differences
- **JumpDTW** [1]
  - Audio-to-score alignment
  - Identifies the "block sequence" taken by a performer along the score (based on a priori jump locations)
- **Needleman-Wunsch time warping (NWTW)** [2]
  - Audio-to-audio alignment
  - DP method with added "waiting mechanism"



Limitations of existing approaches

- **JumpDTW** requires manually specified block boundaries which are accurate at the frame level - not readily available in practical scenarios.
- **JumpDTW** works only with blocks
  - Unable to capture intra-block changes
  - Cannot deal with deviations not foreseeable from the score
  - Relies on OMR, which doesn't always detect the jump/repeat directives

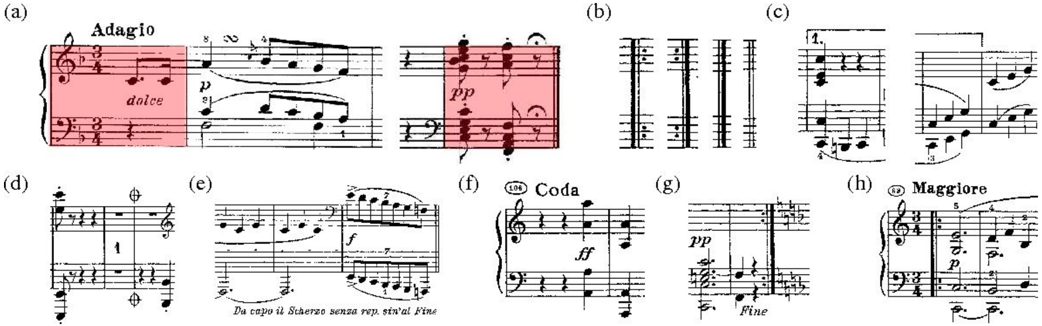
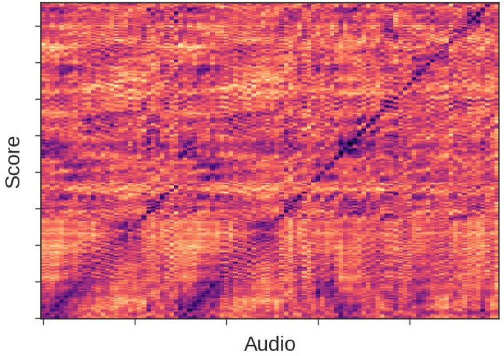


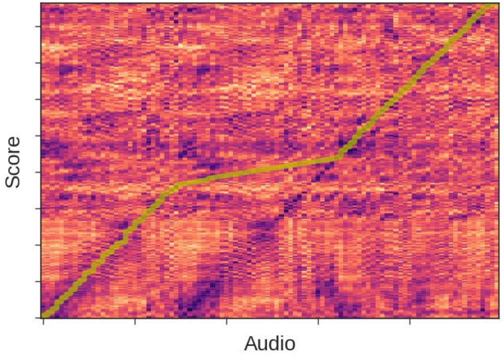
Figure from Fremery et al, 2010

Limitations of existing approaches

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

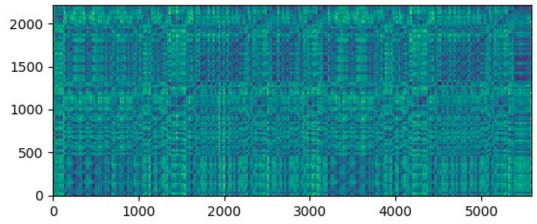
- **NWTW** does not align repeated segments
  - This is due to its “waiting mechanism”
  - Skips unmatchable parts of either sequence
- **NWTW** does not incorporate any score information
- Multiple deviations and interruptions possible in the practice scenario

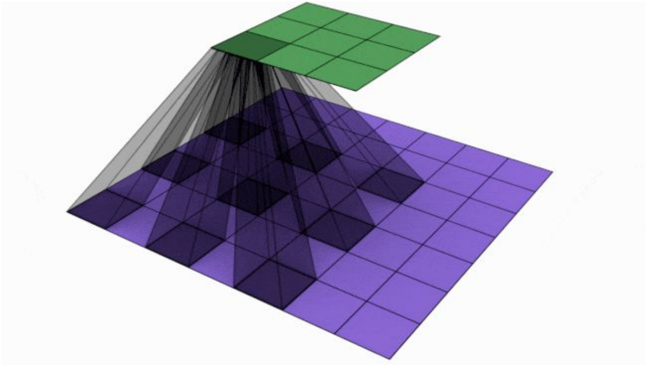




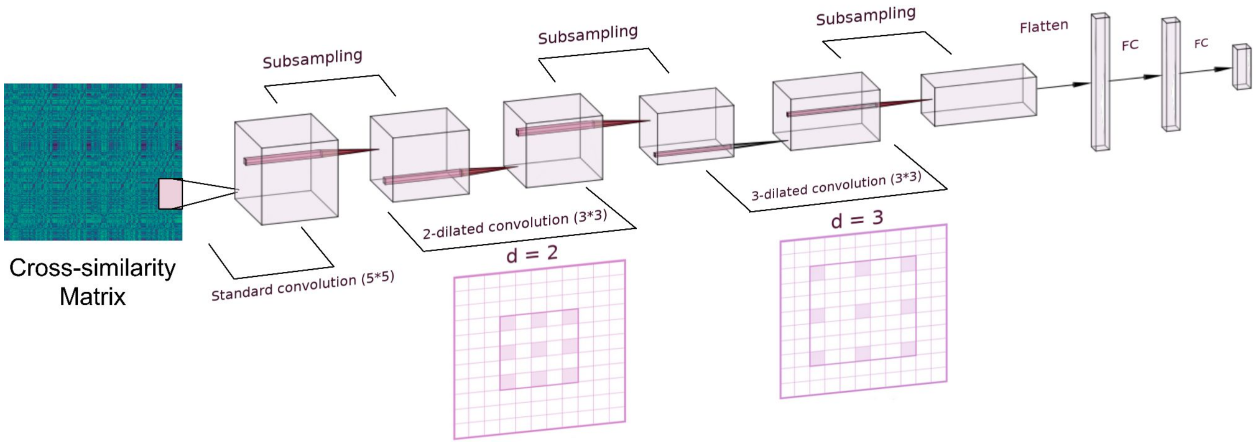
Proposed Method

- Custom CNN-based architecture + flexible DTW
- Standard + Dilated convolutions
- Detect synchronous subpaths between the score and performance by means of ‘inflection points’
- Incorporate varying dilation rates at different layers
- Dilation allows us to capture short and long-term context
- Inflection points passed on to flexible DTW framework to generate fine alignments





Model architecture



$$(F *_d f)(\mathbf{p}) = \sum_{\mathbf{t}} F(\mathbf{p} - d\mathbf{t}) f(\mathbf{t})$$

$$m' = m + (d-1) * (m-1)$$

Generating fine alignments

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

- We train our models to detect synchronous subpaths between the score and performance
- Incorporate varying dilation rates at different layers, predict inflection points
- The inflection points are passed to an extended-DTW framework

$$D(m,n) = e(m,n) + min \begin{cases} D(m, n-1) \\ D(m-1, n) \\ D(m-1, n-1) \\ D(a_{i-1}, b_{i-1}) \ \forall (m,n) = (a_i, b_i), \\ \qquad\qquad i \in \{2, 4, ..., N\} \end{cases}$$

Here, e(m, n) → Euclidean distance between points x_m and y_n,
D(m, n) → Total cost to be minimized for the path until the cell (m, n)
(a_i, b_i) → (x, y) co-ordinates of the i^{th} inflection point

Experimental Setup

- Model inputs: Performance-score cross-similarity matrices (computed using Euclidean distance between the chromagrams)
- Training
    - Generated synthetic data containing jumps and repeats
    - 495 performance-score pairs from the **MSMD** dataset, each utilized 5 times for varying number of repeats/jumps, in total 2475 audio pairs
    - Hand annotated data (150 audio pairs) from Tido UK Ltd.
    - Trained using the L2 regression loss
- Testing
    - Models tested on the **Mazurka** dataset and the **Tido** dataset
    - Results are compared with *MATCH [4], JumpDTW [1] and NWTW [2];* and a baseline CNN model without dilation (CNN_{1+1}).

Results on the Mazurka dataset

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

- Model nomenclature: **DCNN**$_{m+n}$, where m and n correspond to the dilation rates at the second and third layer respectively

Model	*On Mazurka dataset* <25ms	<50ms	<100ms	<200ms
MATCH [3]	64.8	72.1	77.6	83.7
JumpDTW [5]	65.8	75.2	79.8	85.7
NWTW [6]	67.6	75.5	80.1	86.2
CNN_{1+1}	68.2	75.7	80.5	87.1
$DCNN_{2+2}$	**69.9**	76.4	81.6	88.9
$DCNN_{2+3}$	69.7	**77.2**	**82.4**	**89.8**
$DCNN_{3+3}$	69.2	76.1	81.2	88.7
$DCNNsyn_{2+3}$	68.1	75.9	80.7	87.5

Alignment accuracy in %

Results on the Tido dataset

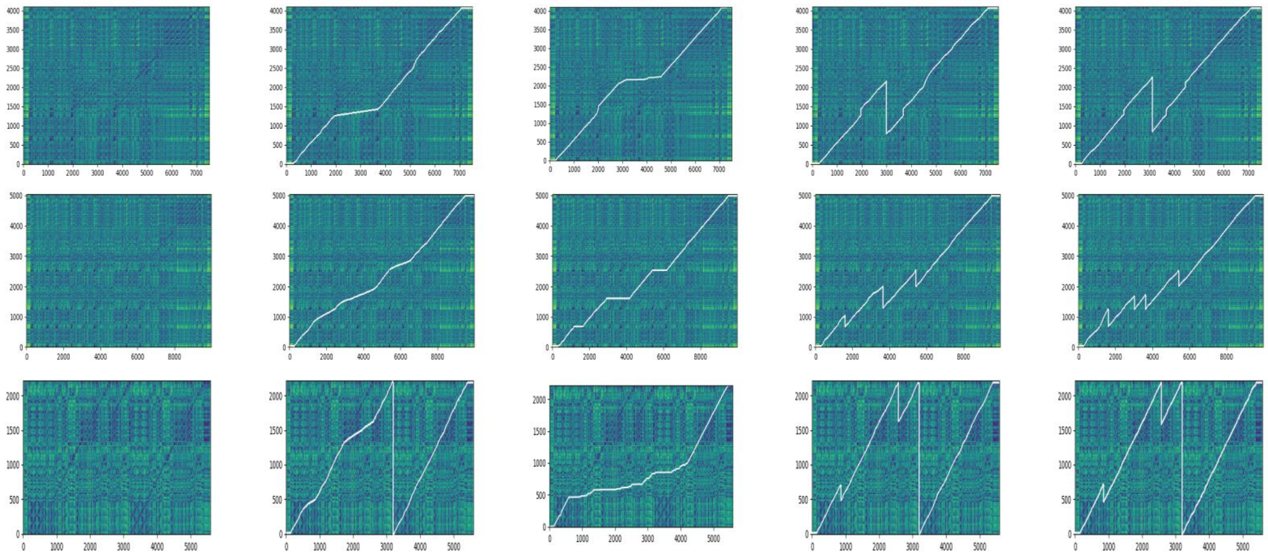
ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

- Separate testing for "structural" and "non-structural" alignment

Model	*With structural differences (Tido)*				*Without structural differences (Tido)*			
	<25ms	**<50ms**	**<100ms**	**<200ms**	**<25ms**	**<50ms**	**<100ms**	**<200ms**
MATCH [3]	61.5	70.4	74.6	80.7	70.2	78.4	84.7	90.3
JumpDTW [5]	69.1	77.2	82.0	88.4	68.7	77.5	82.1	88.9
NWTW [6]	68.6	75.8	80.7	87.5	68.4	77.1	82.8	89.4
CNN_{1+1}	70.4	78.3	83.4	90.1	69.3	78.0	84.1	89.3
$DCNN_{2+2}$	72.7	80.1	84.5	91.4	**71.4**	79.5	85.3	90.5
$DCNN_{2+3}$	**73.9**	**81.3**	**85.6**	**92.8**	71.0	**80.3**	**85.8**	**91.8**
$DCNN_{3+3}$	72.3	79.5	84.2	90.4	70.6	78.8	84.9	91.2
$DCNNsyn_{2+3}$	70.5	78.6	83.8	90.5	69.2	78.3	84.6	89.8

Alignment accuracy in %

Qualitative results



(a) Input (b) *JumpDTW* (c) *NWTW* (d) $DCNN_{2+3}$ (e) Ground Truth

- *DCNN* can handle forward jumps as well as unforeseeable deviations
- Struggles with multiple deviations within a short time span

Discussion

- **DCNN** models show:
  - 2-5% increase in alignment accuracy over JumpDTW and NWTW for test set containing structural differences
  - 1-3% increase over JumpDTW and NWTW on the test set not containing structural differences
  - 4-6% overall increase over MATCH (9-10% on the subset with structural differences) and 1-4% overall increase over JumpDTW and NWTW

- Our method is applicable in real-world scenarios
  - Can work with largely synthetic data
  - Limited hand-annotated data improves performance further
  - Doesn't require jump locations *a priori*
  - Compatible with other feature representations, such as learnt frame similarities [4] and multimodal embeddings [5], and also with non-DTW based methods.

Conclusion and Future Work

ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

- Progressively dilated convolutional neural networks are effective at structure aware audio-to-score alignment
- Noticeable improvement in capturing structural differences over previous approaches, and doesn't impair "non-structural" alignment
- Our method can also be used with raw or scanned images of sheet music using learnt features
- Inflection points could be used by non-DTW based methods as well
- Future work
    - Parallel dilation and merging
    - Handling of trills and cadenzas

Thank you for your attention! Questions?

References

[1] Fremerey, Christian, Meinard Müller, and Michael Clausen. "Handling Repeats and Jumps in Score-performance Synchronization." *ISMIR*. 2010.

[2] Grachten, Maarten, Martin Gasser, Andreas Arzt, and Gerhard Widmer. "Automatic alignment of music performances with structural differences." (2013).

[3] Dixon, Simon, and Gerhard Widmer. "MATCH: A Music Alignment Tool Chest." *ISMIR*. 2005.

[4] Agrawal, Ruchit, and Simon Dixon. "Learning frame similarity using Siamese Networks for Audio-to-Score Alignment." *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[5] Dorfer, Matthias, Jan Hajič jr., Andreas Arzt, Harald Frostel, and Gerhard Widmer. "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification." *Transactions of the International Society for Music Information Retrieval* 1.1 (2018).

[6] Waloschek, Simon, Aristotelis Hadjakos, and Alexander Pacha. "Identification and Cross Document Alignment of Measures in Music Score Images." *20th International Society for Music Information Retrieval Conference*. 2019.