INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen Nürnberg

ISMIR Tutorial
Daejeon, Korea, September 21, 2025

Differentiable Alignment Techniques for Music Processing: Techniques and Applications

Part 1: Introduction to Alignment Techniques

Meinard Müller, Johannes Zeitler
International Audio Laboratories Erlangen
{meinard.mueller, johannes.zeitler}@audiolabs-erlangen.de

Overview

Part 0: Overview

Part 1: Introduction to Alignment Techniques

Coffee Break

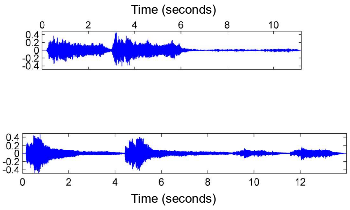
Part 2: Theoretical Foundations & Implementation

Motivation: Audio-Audio Alignment

Beethoven's Fifth

Karajan (Orchester)

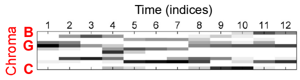
Gould (Piano)



Motivation: Audio-Audio Alignment

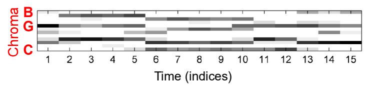
Beethoven's Fifth

Karajan (Orchester)



Time–chroma representations

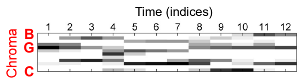
Gould (Piano)



Motivation: Audio-Audio Alignment

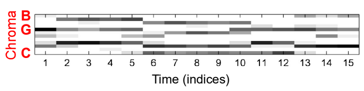
Beethoven's Fifth

Karajan (Orchester)



Time–chroma representations

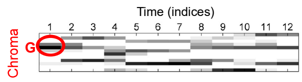
Gould (Piano)



Motivation: Audio-Audio Alignment

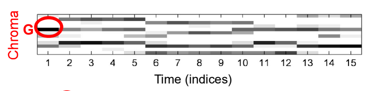
Beethoven's Fifth

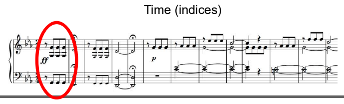
Karajan (Orchester)



Time–chroma representations

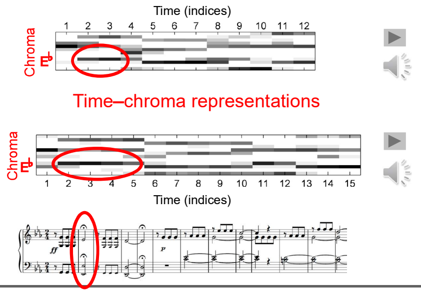
Gould (Piano)





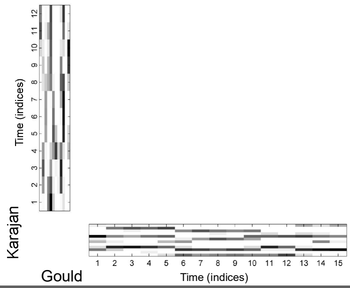
Motivation: Audio-Audio Alignment
Beethoven's Fifth

Karajan (Orchester)

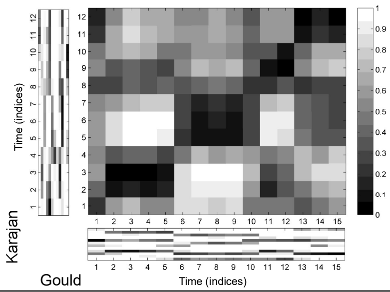


Gould (Piano)

Motivation: Audio-Audio Alignment
Beethoven's Fifth

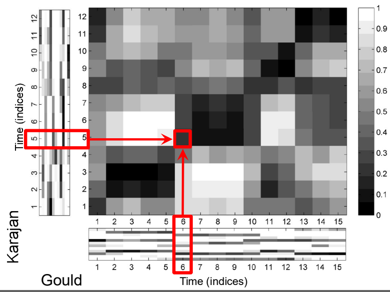


Motivation: Audio-Audio Alignment
Beethoven's Fifth



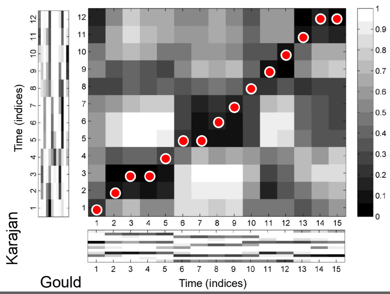
Cost matrix

Motivation: Audio-Audio Alignment
Beethoven's Fifth



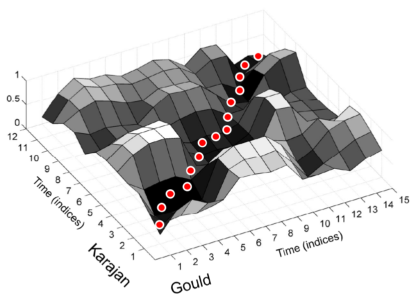
Cost matrix

Motivation: Audio-Audio Alignment
Beethoven's Fifth



Cost-minimizing warping path

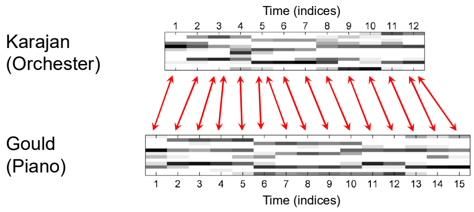
Motivation: Audio-Audio Alignment
Beethoven's Fifth



Cost-minimizing warping path

Motivation: Audio-Audio Alignment

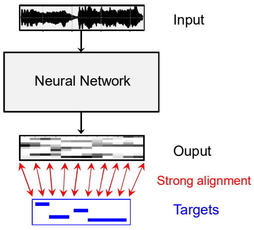
Beethoven's Fifth



Cost-minimizing warping path

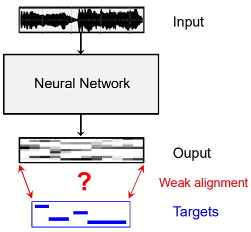
→ Strong alignment

Feature Learning



- Task: Learn audio features using a neural network
- Loss: Binary cross-entropy
  - framewise loss
  - requires strongly aligned targets
  - hard to obtain

Feature Learning



- Task: Learn audio features using a neural network
- Loss: Binary cross-entropy
  - framewise loss
  - requires strongly aligned targets
  - hard to obtain
- Alignment as part of loss function
  - requires only weakly aligned targets
  - needs to be differentiable
- Problem: DTW is not differentiable
  → Soft DTW

Dynamic Time Warping (DTW)

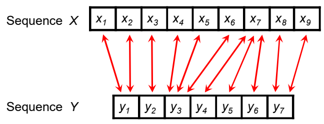
$X := (x_1, x_2, \ldots, x_N)$

$Y := (y_1, y_2, \ldots, y_M)$

$x_n, y_m \in \mathcal{F},\ n \in [1:N],\ m \in [1:M]$

$\mathcal{F}$ = Feature space

Alignment

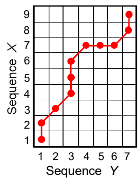


Alignment matrix

$A \in \{0,1\}^{N\times M}$

Set of all possible alignment matrices

$\mathcal{A}_{N,M} \subset \{0,1\}^{N\times M}$



Dynamic Time Warping (DTW)

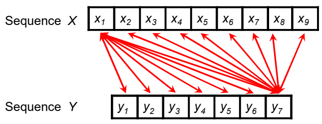
$X := (x_1, x_2, \ldots, x_N)$

$Y := (y_1, y_2, \ldots, y_M)$

$x_n, y_m \in \mathcal{F},\ n \in [1:N],\ m \in [1:M]$

$\mathcal{F}$ = Feature space

Alignment

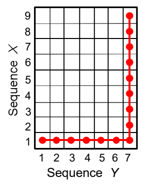


Alignment matrix

$A \in \{0,1\}^{N\times M}$

Set of all possible alignment matrices

$\mathcal{A}_{N,M} \subset \{0,1\}^{N\times M}$



Dynamic Time Warping (DTW)

$X := (x_1, x_2, \ldots, x_N)$

$Y := (y_1, y_2, \ldots, y_M)$

$x_n, y_m \in \mathcal{F},\ n \in [1:N],\ m \in [1:M]$

$\mathcal{F}$ = Feature space

Alignment matrix

$A \in \{0,1\}^{N\times M}$

Set of all possible alignment matrices

$\mathcal{A}_{N,M} \subset \{0,1\}^{N\times M}$

Cost measure: $c : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$

Cost matrix: $C \in \mathbb{R}^{N\times M}$ with $C(n,m) := c(x_n, y_m)$

Cost of alignment: $\langle A, C \rangle$

DTW cost: $\mathrm{DTW}(C) = \min\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

Optimal alignment: $A^* = \mathrm{argmin}\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

Dynamic Time Warping (DTW)

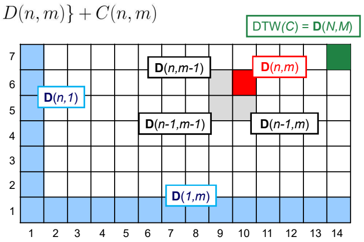
DTW cost: $\mathrm{DTW}(C) = \min\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

- Efficient computation via Bellman's recursion in O(*NM*)

  $D(n,m) = \min\{D(n-1,m), D(n,m-1), D(n,m)\} + C(n,m)$

  for *n>1* and *m>1* and suitable initialization.

  $\mathrm{DTW}(C) = D(N,M)$



Dynamic Time Warping (DTW)

DTW cost: $\mathrm{DTW}(C) = \min\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

- Efficient computation via Bellman's recursion in O(*NM*)

  $D(n,m) = \min\{D(n-1,m), D(n,m-1), D(n,m)\} + C(n,m)$

  for *n>1* and *m>1* and suitable initialization.

  $\mathrm{DTW}(C) = D(N,M)$

- Problem: DTW(*C*) is not differentiable with regard to *C*
- Idea: Replace min-function by a smooth version

$$\min^{\gamma}(\mathcal{S}) = -\gamma \log \sum_{s\in\mathcal{S}} \exp(-s/\gamma)$$

for set $\mathcal{S} \subset \mathbb{R}$ and temperature parameter $\gamma \in \mathbb{R}$

Soft Dynamic Time Warping (SDTW)

SDTW cost: $\mathrm{SDTW}^{\gamma}(C) = \min^{\gamma}\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

- Efficient computation via Bellman's recursion in O(*NM*) still works:

  $D^{\gamma}(n,m) = \min^{\gamma}\{D^{\gamma}(n-1,m), D^{\gamma}(n,m-1), D^{\gamma}(n,m)\} + C(n,m)$

  for *n>1* and *m>1* and suitable initialization.

  $\mathrm{SDTW}^{\gamma}(C) = D^{\gamma}(N,M)$

- Limit case: $\mathrm{SDTW}^{\gamma}(C) \xrightarrow{\gamma\to 0} \mathrm{DTW}(C)$
- SDTW(*C*) is differentiable with regard to *C*
- Questions:
  - How does the gradient look like?
  - Can it be computed efficiently?
  - How does SDTW generalize the alignment concept?

Soft Dynamic Time Warping (SDTW)

Soft-DTW
Cuturi, Blondel: Soft-DTW: A Differentiable Loss Function for Time-Series. ICML, 2017

SDTW cost: $\mathrm{SDTW}^{\gamma}(C) = \min^{\gamma}\left(\{\langle A, C\rangle \mid A \in \mathcal{A}_{N,M}\}\right)$

- Define $p^{\gamma}(C)$ as the following "probability" distribution over $\mathcal{A}_{N,M}$:

$$p^{\gamma}(C)_A = \frac{\exp(-\langle A, C\rangle/\gamma)}{\sum_{A'\in\mathcal{A}_{N,M}} \exp(-\langle A', C\rangle/\gamma)} \quad \text{for } A \in \mathcal{A}_{N,M}$$

- The expected alignment with respect to $p^{\gamma}(C)$ is given by:

$$E^{\gamma}(C) = \sum_{A\in\mathcal{A}_{N,M}} p^{\gamma}(C)_A A \quad \in \mathbb{R}^{N\times M}$$

- The gradient is given by:

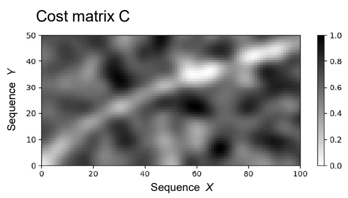
$$\nabla_C \mathrm{SDTW}^{\gamma}(C) = E^{\gamma}(C)$$

- The gradient can be computed efficiently in O(*NM*) via a recursive algorithm.

Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^{\gamma}(C) = \sum_{A\in\mathcal{A}_{N,M}} p^{\gamma}(C)_A A \quad \in \mathbb{R}^{N\times M}$

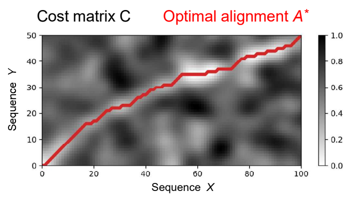
- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ



Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^{\gamma}(C) = \sum_{A\in\mathcal{A}_{N,M}} p^{\gamma}(C)_A A \quad \in \mathbb{R}^{N\times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

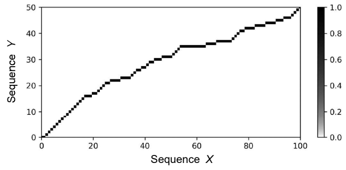


Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^\gamma(C) = \sum_{A \in \mathcal{A}_{N,M}} p^\gamma(C)_A A \quad \in \mathbb{R}^{N \times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

$E^\gamma(C)$ with $\gamma = 0$ $(= A^*)$



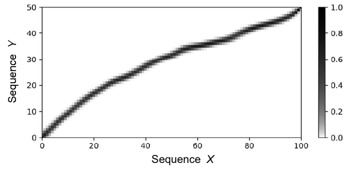
© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^\gamma(C) = \sum_{A \in \mathcal{A}_{N,M}} p^\gamma(C)_A A \quad \in \mathbb{R}^{N \times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

$E^\gamma(C)$ with $\gamma = 0.1$



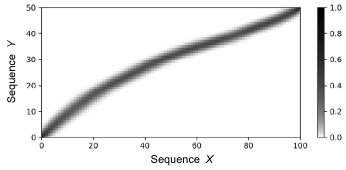
© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^\gamma(C) = \sum_{A \in \mathcal{A}_{N,M}} p^\gamma(C)_A A \quad \in \mathbb{R}^{N \times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

$E^\gamma(C)$ with $\gamma = 1$



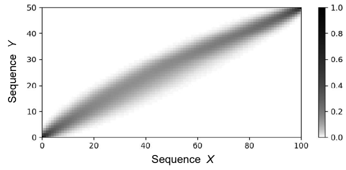
© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^\gamma(C) = \sum_{A \in \mathcal{A}_{N,M}} p^\gamma(C)_A A \quad \in \mathbb{R}^{N \times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

$E^\gamma(C)$ with $\gamma = 10$



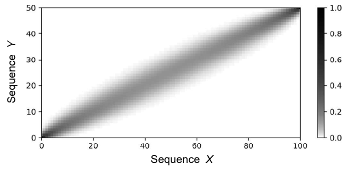
© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)

Expected alignment : $E^\gamma(C) = \sum_{A \in \mathcal{A}_{N,M}} p^\gamma(C)_A A \quad \in \mathbb{R}^{N \times M}$

- Can be interpreted as a smoothed version of an alignment
- Degree of smoothing depends on temperature parameter γ

$E^\gamma(C)$ with $\gamma = 100$



© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)
Conclusions

- Direct generalization of DTW (replacing min by smooth variant)
- Gradient is given by expected alignment
- Fast forward algorithm: O(NM)
- Fast gradient computation: O(NM)
- SDTW yields a (typically) poor lower bound for DTW
- Can be used as loss function to learn from weakly aligned sequences

© AudioLabs, 2025

Soft Dynamic Time Warping (SDTW)
References

- Marco Cuturi, Mathieu Blondel: Soft-DTW: A Differentiable Loss Function for Time-Series. ICML, pages 894–903, 2017.
- Arthur Mensch, Mathieu Blondel: Differentiable Dynamic Programming for Structured Prediction and Attention, ICML, 2018.
- Michael Krause, Christof Weiß, Meinard Müller: Soft Dynamic Time Warping for Multi-Pitch Estimation and Beyond. IEEE ICASSP, 2023.

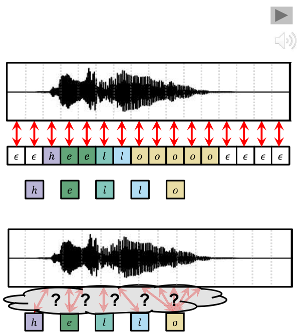
Thanks:
Michale Krause (Ph.D. 2023)
Johannes Zeitler (Ph.D.)

CTC Loss: Introduction

- Connectionist Temporal Classification (CTC)
- Graves, Fernández, Gomez, and Schmidhuber:
  Connectionist Temporal Classification:
  Labelling Unsegmented Sequence Data with Recurrent Neural Networks.
  ICML, 2006.
- **Temporal Classification**: Labelling unsegmented data sequences
- **Connectionist**: Refers to the use of deep learning

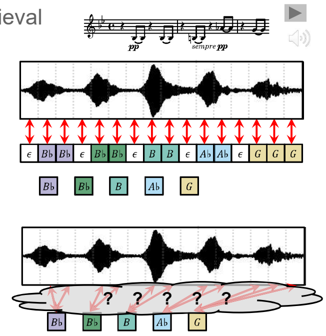
CTC Loss: Introduction
Training data in speech recognition

- Strongly aligned training data
  - Character annotations (labels) for each time step
  - Can be used for training in a standard classification setup
  - Tedious to annotate
- Weakly aligned training data
  - Globally corresponding character sequence without local alignment
  - Cannot be used for training in a standard classification setup
  - Easier to annotate
- Aim of CTC: Employ only weakly aligned data for training

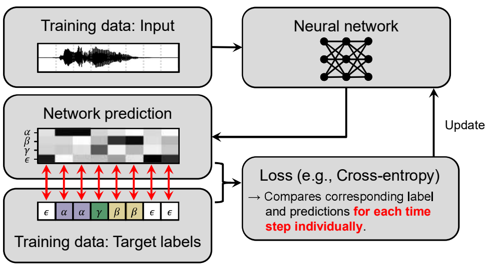


CTC Loss: Introduction
Training data in theme-based music retrieval

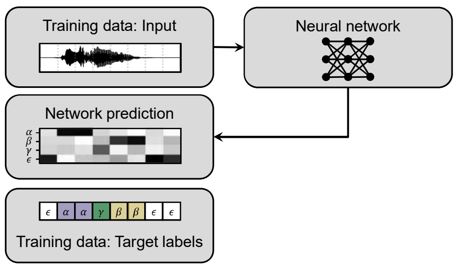
- Strongly aligned training data
  - Character annotations (labels) for each time step
  - Can be used for training in a standard classification setup
  - Tedious to annotate
- Weakly aligned training data
  - Globally corresponding character sequence without local alignment
  - Cannot be used for training in a standard classification setup
  - Easier to annotate
- Aim of CTC: Employ only weakly aligned data for training



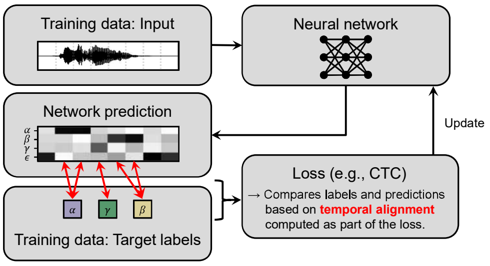
CTC Loss: Introduction
Standard deep learning setup: Strongly aligned training data



CTC Loss: Introduction
Standard deep learning setup: Strongly aligned training data

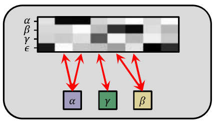


CTC Loss: Introduction
Non-standard deep learning setup: Weakly aligned training data

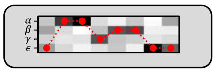


CTC Loss: Introduction
Alignment Representations

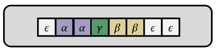
"Arrow" representation



"Point" representation

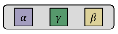


"Unfolded" representation

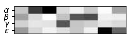


CTC Loss: Introduction

- Alphabet $\mathbb{A} = \{\alpha, \beta, \gamma\}$
- Label sequence $\boldsymbol{Y} = (\alpha, \gamma, \beta)$



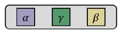
- Network output $f_\theta(\boldsymbol{X}) =$



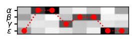
- Alignment $\boldsymbol{A}$ is "expansion" of $\boldsymbol{Y}$ to length of $f_\theta(\boldsymbol{X})$ (possibly consecutive duplicates and blank symbols ϵ)

CTC Loss: Introduction

- Alphabet $\mathbb{A} = \{\alpha, \beta, \gamma\}$
- Label sequence $\boldsymbol{Y} = (\alpha, \gamma, \beta)$

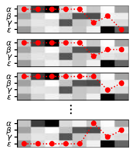


- Naive idea: "Hard" alignment (Related: Viterbi decoding)



$P(\epsilon, \alpha, \alpha, \gamma, \beta, \beta, \epsilon, \epsilon) \approx 0.015$

- Not suitable for gradient-descent-based training (not differentiable)
- Therefore: "Soft" alignment (Related: Forward algorithm)



$P(\alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \gamma, \beta) \approx 7.14 \cdot 10^{-7}$

$P(\alpha, \alpha, \alpha, \alpha, \alpha, \gamma, \beta, \beta) \approx 3.98 \cdot 10^{-7}$

$P(\alpha, \alpha, \alpha, \alpha, \alpha, \gamma, \beta, \epsilon) \approx 3.23 \cdot 10^{-6}$

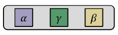
$\vdots$

$P(\epsilon, \epsilon, \epsilon, \epsilon, \epsilon, \alpha, \gamma, \beta) \approx 5.82 \cdot 10^{-6}$

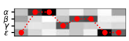
$\sum_{A} P(A) \approx 0.069$

CTC Loss: Introduction

- Alphabet $\mathbb{A} = \{\alpha, \beta, \gamma\}$
- Label sequence $\boldsymbol{Y} = (\alpha, \gamma, \beta)$

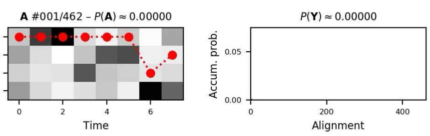


- Naive idea: "Hard" alignment (Related: Viterbi decoding)

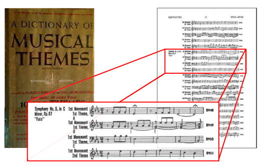


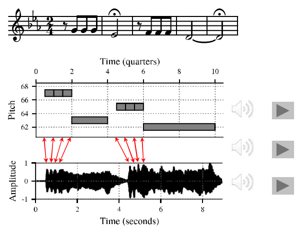
$P(\epsilon, \alpha, \alpha, \gamma, \beta, \beta, \epsilon, \epsilon) \approx 0.015$

- Not suitable for gradient-descent-based training (not differentiable)
- Therefore: "Soft" alignment (Related: Forward algorithm)



Theme-Based Audio Retrieval
Barlow & Morgenstern (1949): A Dictionary of Musical Themes





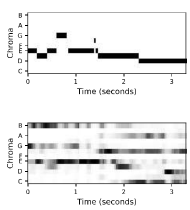
- 2067 themes by 54 different composers
- Recordings (1126 recordings, ~ 120 hours)
- Theme occurences (~ 5 hours)

Theme-Based Audio Retrieval
Monophony–Polyphony Challenge

Monophonic symbolic musical theme

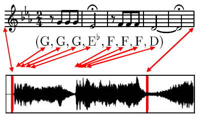
Audio recording of polyphonic music

Chromagram



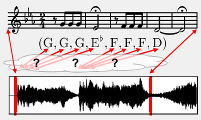
Goal: Compute "enhanced" chromagram from polyphonic audio recording that better matches the symbolic monophonic theme

Theme-Based Audio Retrieval
Strongly Aligned Training Data



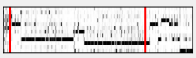
Theme-Based Audio Retrieval
Weakly Aligned Training Data

Network input for training:
Weakly aligned score–audio pairs

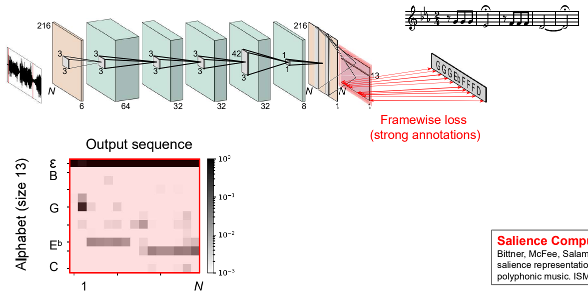


Convolutional neural network:
Trained with the CTC loss

Post-processed output of trained network:
Enhanced chroma representation

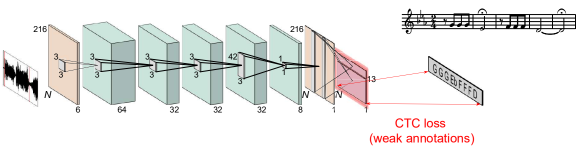


Theme-Based Audio Retrieval



Salience Computation
Bittner, McFee, Salamon, Li, Bello: Deep salience representations for F0 tracking in polyphonic music. ISMIR, 2017.

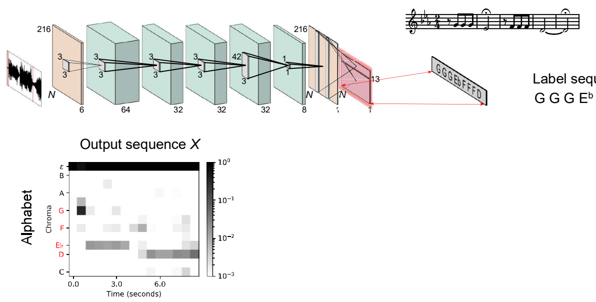
Theme-Based Audio Retrieval



- Idea of CTC loss similar to SDTW
- Theme is given as label sequence over finite alphabet (size 13 including blank symbol)
- Expand label sequence to match audio feature sequence → valid alignment
- CTC loss considers probability over **all** valid alignments → differentiable

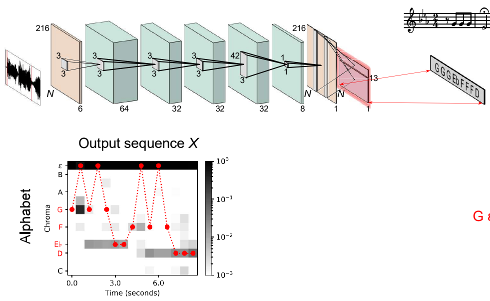
CTC Loss
Graves, Fernández, Gomez, Schmidhuber: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML, 2006.

Theme-Based Audio Retrieval
CTC-Based Training



Label sequence Y
G G G E♭ F F F D

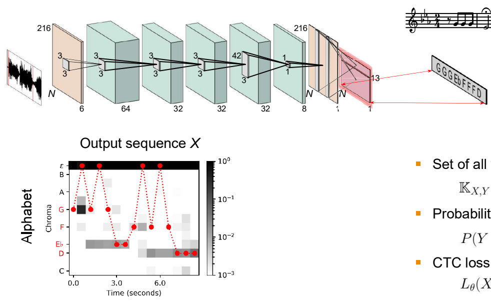
Theme-Based Audio Retrieval
CTC-Based Training



Label sequence Y
G G G E♭ F F F D

Valid alignment
G ε G ε G E♭ E♭ F ε F ε F D D D
→ matches sequence X

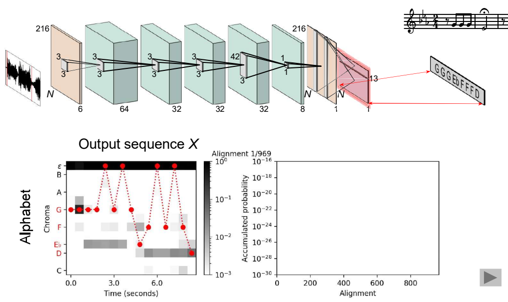
Theme-Based Audio Retrieval
CTC-Based Training



Label sequence Y
G G G E♭ F F F D

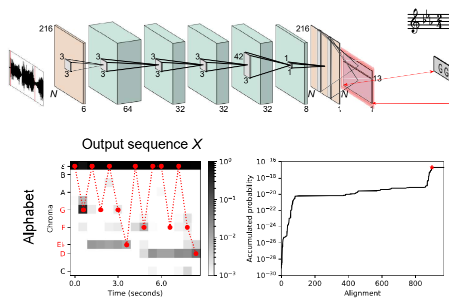
- Set of all valid alignments
  $\mathbb{K}_{X,Y} = \{A \in (\mathbb{A}')^N : \kappa(A) = Y\}$
- Probability of label sequence
  $P(Y \mid X) = \sum_{A \in \mathbb{K}_{X,Y}} P(A \mid X)$
- CTC loss
  $L_\theta(X, Y) = -\log P(Y \mid X)$

Theme-Based Audio Retrieval
CTC-Based Training



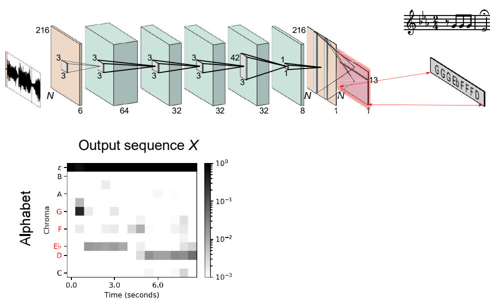
Label sequence Y
G G G E♭ F F F D

Theme-Based Audio Retrieval
CTC-Based Training



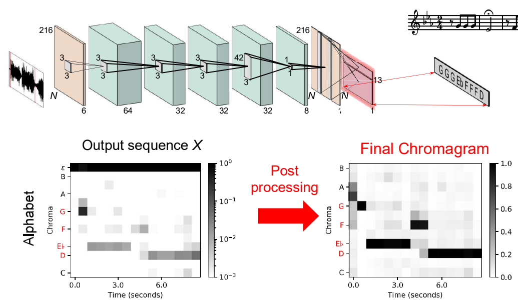
Label sequence Y
G G G E♭ F F F D

Theme-Based Audio Retrieval
CTC-Based Training



Label sequence Y
G G G E♭ F F F D

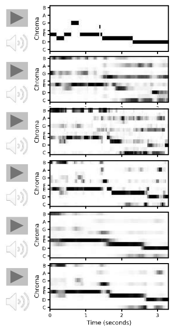
Theme-Based Audio Retrieval
CTC-Based Training



Label sequence Y
G G G E♭ F F F D

Theme-Based Audio Retrieval
Evaluation Results

Chroma Variant	Top-1	Top-10
Standard chromagram	0.561	0.723
Enhanced chromagram (baseline)	0.824	0.861
DNN-based chromagram (CTC)	0.867	0.942
DNN-based chromagram (linear scaling)	0.829	0.914
DNN-based chromagram (strong alignment)	0.882	0.939



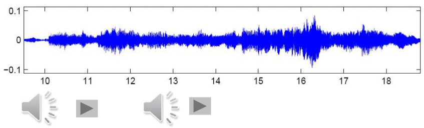
Theme-Based Audio Retrieval
References

- R. Bittner, B. McFee, J. Salamon, P. Li, and J. Bello: Deep salience representations for F0 tracking in polyphonic music. Proc. ISMIR, pages 63–70, 2017.
- A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. ICML, 2006.
- F. Zalkow, S. Balke, V. Arifi-Müller, and M. Müller. MTD: A multimodal dataset of musical themes for MIR research. TISMIR, 3(1), 2020.
- F. Zalkow, S. Balke, and M. Müller. Evaluating salience representations for cross-modal retrieval of Western classical music recordings. Proc. ICASSP, 2019.
- F. Zalkow and M. Müller. CTC-based learning of deep chroma features for score-audio music retrieval. 2021. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 29, pages 2957–2971, 2021.

Thanks:
Frank Zalkow (Ph.D. 2021)
Stefan Balke (Ph.D. 2018)

Lyrics–Audio Alignment

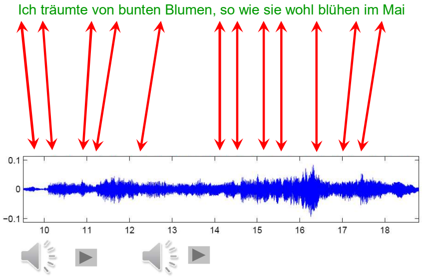
Ich träumte von bunten Blumen, so wie sie wohl blühen im Mai



CTC Loss for Lyrics Alignment
Stoller, Durand, Ewert: End-to-end Lyrics Alignment for Polyphonic Music Using an Audio-To-Character Recognition Model. ICASSP 2019.

Lyrics–Audio Alignment

Ich träumte von bunten Blumen, so wie sie wohl blühen im Mai

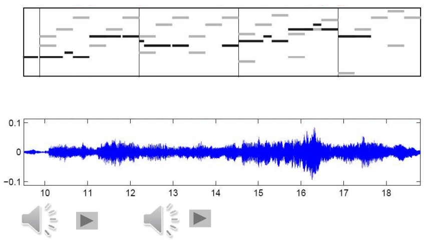


Lyrics-Audio

CTC Loss for Lyrics Alignment
Stoller, Durand, Ewert: End-to-end Lyrics Alignment for Polyphonic Music Using an Audio-To-Character Recognition Model. ICASSP 2019.

Lyrics–Audio Alignment

Ich träumte von bunten Blumen, so wie sie wohl blühen im Mai

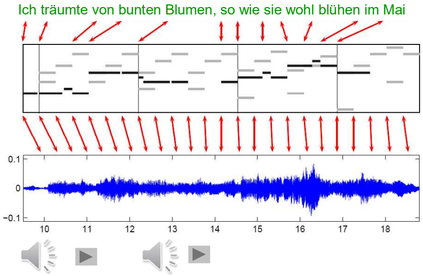


Lyrics-Audio

Multimodal Lyrics Alignment
Müller, Kurth, Damm, Fremerey, Clausen: Lyrics-based Audio Retrieval and Multimodal Navigation in Music Collections. ECDL 2007.

Lyrics–Audio Alignment

Ich träumte von bunten Blumen, so wie sie wohl blühen im Mai



Lyrics-MIDI

Lyrics-Audio

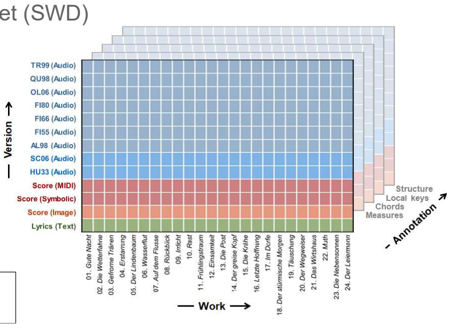
MIDI-Audio

Multimodal Lyrics Alignment
Müller, Kurth, Damm, Fremerey, Clausen: Lyrics-based Audio Retrieval and Multimodal Navigation in Music Collections. ECDL 2007.

Datasets
Schubert Winterreise Dataset (SWD)

- Song cycle by Franz Schubert
- 24 songs
- 9 performances (versions)
- Annotations
  - Lyrics
  - Chords
  - Local keys
  - Structure

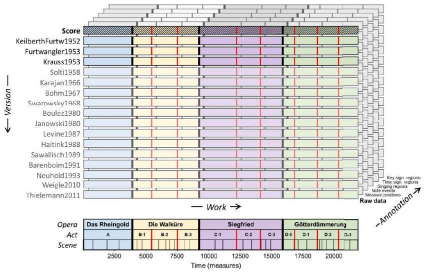
Weiß et al.: Schubert Winterreise Dataset: A Multimodal Scenario for Music Analysis ACM J. Computing & Cultural Heritage, 2021.



Datasets
Wagner Ring Dataset (WRD)

- Opera cycle by Richard Wagner
- 4 operas (ca. 22,000 meaures)
- 16 performances (versions)
- Annotations
  - Lyrics
  - Measure positions
  - Aligned reduced score
  - ...

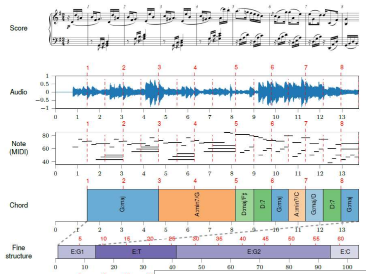
Weiß et al.: Wagner Ring Dataset: A Complex Opera Scenario for Music Processing and Computational Musicology, TISMIR 2023.



Datasets
Beethoven Piano Sonata Dataset

- Piano Sontats by Beethoven
- 32 first movments
- 11 performances (versions)
- Annotations
  - Notes
  - Measures and beats
  - Chords, local & global keys
  - Musical structures

Zeitler et al.: BPSD: A Coherent Multi-Version Dataset for Analyzing the First Movements of Beethoven's Piano Sonatas. TISMIR 2024.

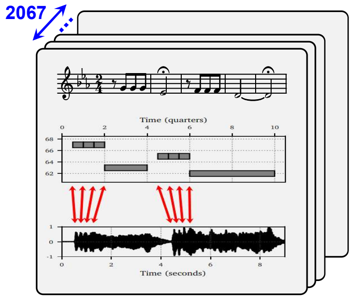


Musical Motifs:
Hsiao, Hung, Chen, Su: BPS-Motif: A Dataset for Repeated Pattern Discovery of Polyphonic Symbolic Music. ISMIR 2023: 281-288

Datasets
Musical Theme Dataset (MTD)

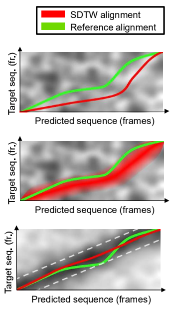
- Western classical music
- Inspired by Barlow & Morgenstern (1948)
- 2067 themes
  - Symbolic encodings
  - Audio excerpts
  - Strong alignments
  - ...

Zalkow et al.: MTD: A Multimodal Dataset of Musical Themes for MIR Research. TISMIR 2020.



Soft Dynamic Time Warping (SDTW)
Stabilizing Training

- Standard SDTW often unstable
  - Unstable training in early stages
  - Degenerate output alignment
- Hyperparameter adjustment
  - High temperature to smooth alignments
  - Temperature annealing
- Diagonal prior
- Modified step size condition



Soft Dynamic Time Warping (SDTW)
Representation Learning

- Symmetric application
  - Learn representation of both sequences
  - Needs a contrastive loss term
- Assymmetric application
  - Use fixed (e.g., binary) encoding of target
  - Learn representation of only one sequences
  - No contrastive loss term need
- Simulation of CTC-loss using SDTW possible
- Many DTW variants also possible for SDTW