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

AUDIO LABS

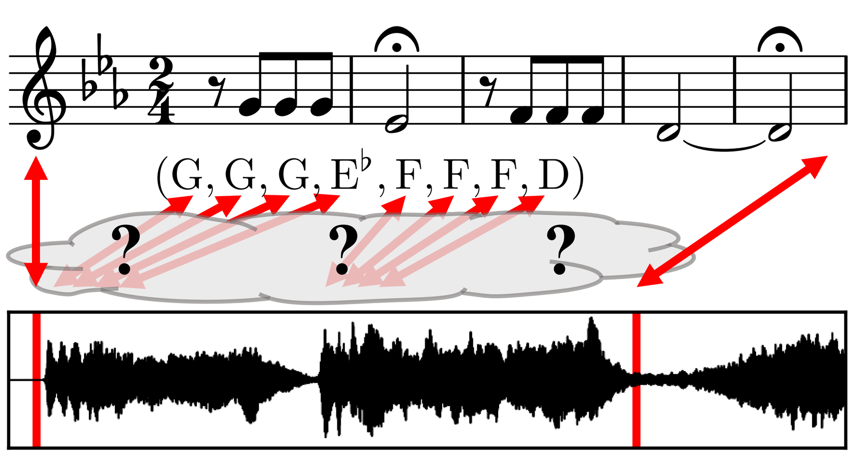
ISMIR MTL2020

Using Weakly Aligned Score–Audio Pairs to Train Deep Chroma Models for Cross-Modal Music Retrieval

Frank Zalkow and Meinard Müller, International Audio Laboratories Erlangen

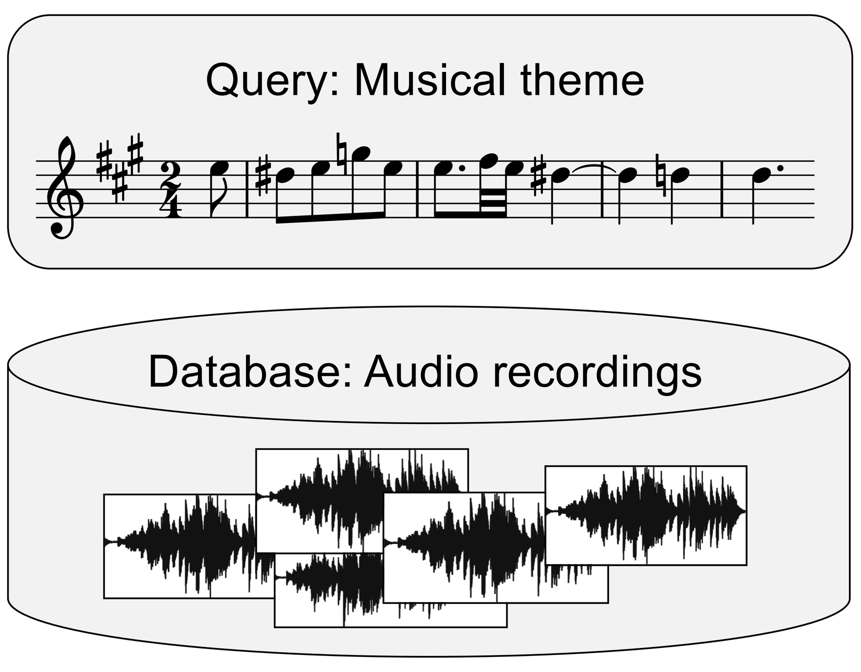
Summary

- **Cross-modal retrieval application** with monophonic symbolic musical themes as queries and audio database of polyphonic music
- Learn **enhanced chroma variant** for musical themes
- Only use **weakly aligned** training pairs
- Adapt a **deep salience model** and train with the **Connectionist Temporal Classification (CTC)** loss for computing chroma features
- Improved **state-of-the-art results** for cross-modal retrieval application



1. Cross-Modal Retrieval

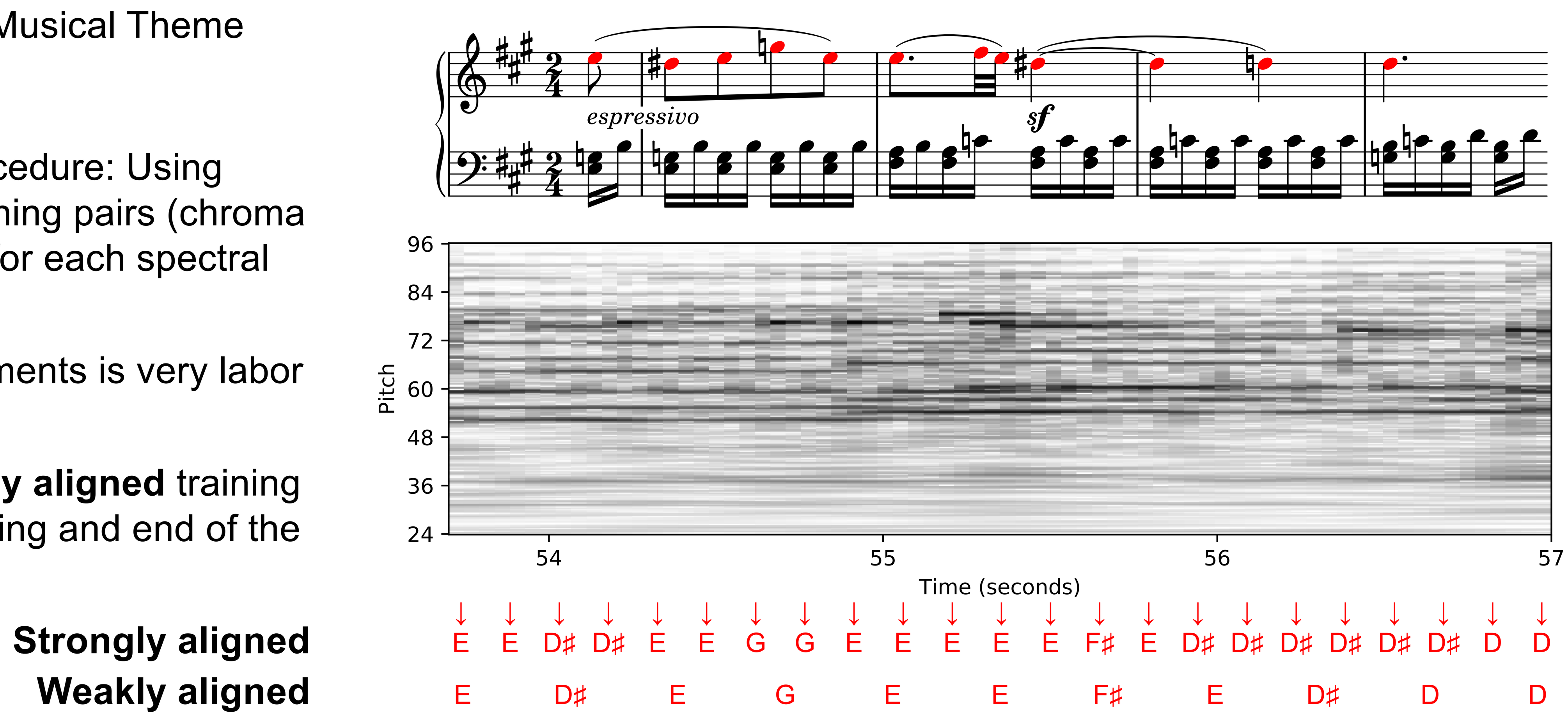
- **Query:** Monophonic theme in symbolic encoding
- **Database:** Polyphonic audio recordings of Western classical music
- **Aim:** Find relevant recordings where theme is played
- **Approach:** Retrieval based on subsequence dynamic time warping and chroma features [1]
- **Problem:** Monophonic–polyphonic difference between query and database
- **Solution:** Learn enhanced audio chroma features for musical themes



2. Data: Weakly vs. Strongly Aligned

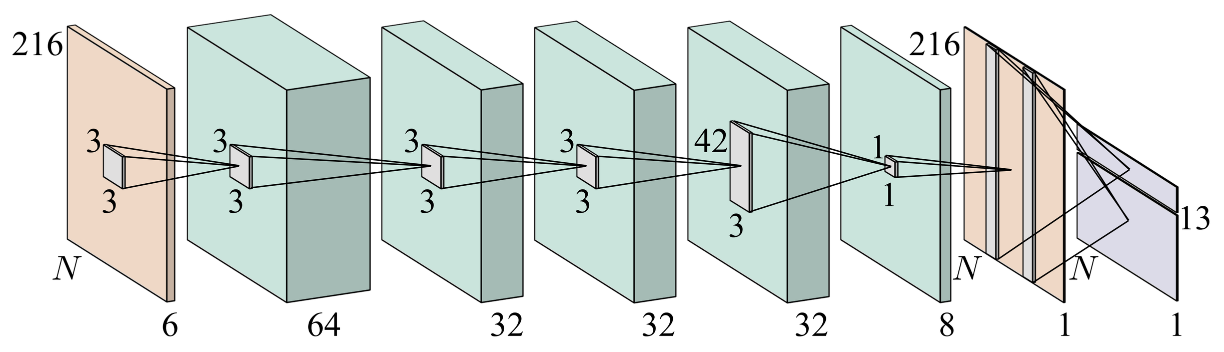
- Dataset with more than 2000 musical themes
- Based on dictionary by Barlow and Morgenstern from 1948 [2]
- Publicly available as Musical Theme Dataset (MTD) [8]
- Standard training procedure: Using **strongly aligned** training pairs (chroma labels are annotated for each spectral frame)
- Creating strong alignments is very labor intensive
- Our aim: Using **weakly aligned** training pairs (only the beginning and end of the theme is annotated)

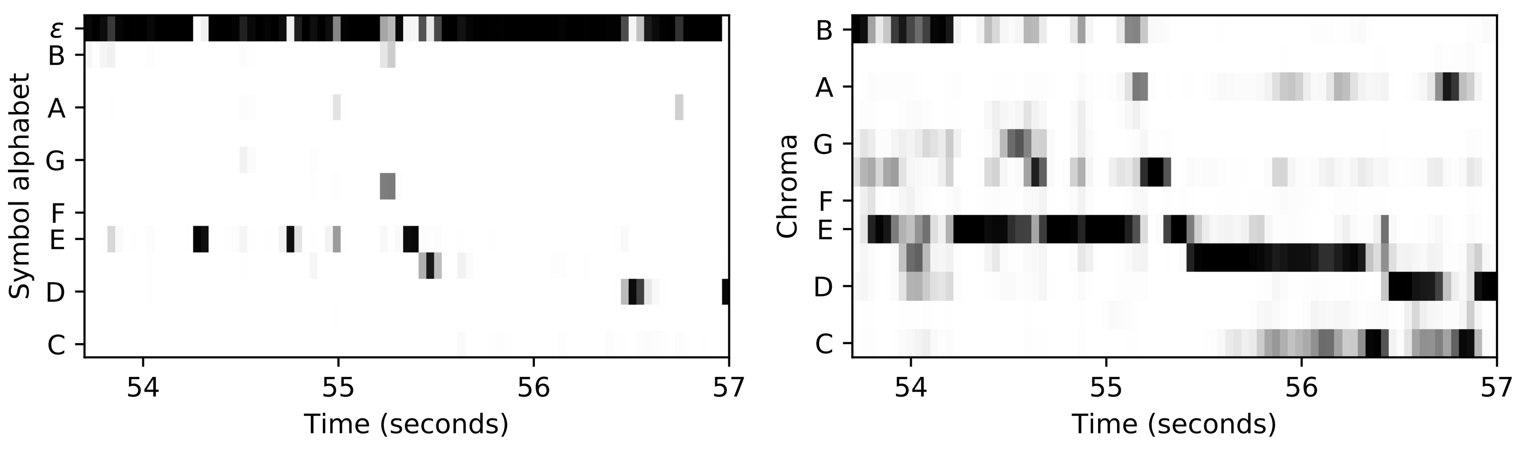
	#	Mean Dur.	Total Dur.
Queries	2048	00:00:09	04:54:58
Database	1114	00:06:26	119:28:27



3. Approach: Connectionist Temporal Classification

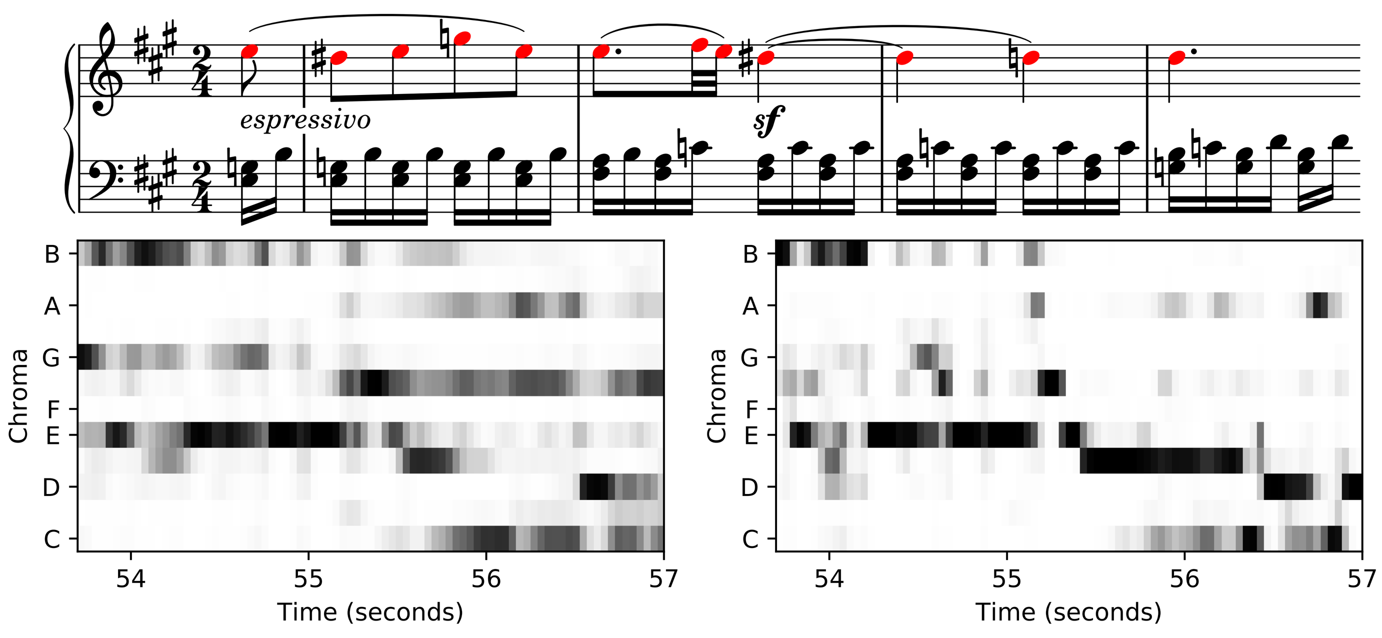
- Adapt **deep salience** model [3] to have fewer parameters and to output a chroma representation
- Training with **Connectionist Temporal Classification** (CTC) loss [5]
- Input: HCQT tensor of theme recording
- Output: probability distribution over symbol alphabet of chroma labels and blank symbol ε (*left matrix*)
- Given for training: Weakly aligned chroma label sequence
- Training aim: Maximize output probability for all possible alignments between input features and chroma label sequence
- After training: Remove blank symbol probabilities, normalize frames (*right matrix*)





4. Results

Qualitative comparison of standard chroma features (*left matrix*) and CTC-based chroma features (*right matrix*)



Quantitative comparison against state-of-the-art baselines [7] using retrieval-based evaluation

Measure	CTC	Bittner et al. [3]	Bosch/Gómez [4]
Top-01	0.865	0.763	0.820
Top-05	0.925	0.844	0.892
Top-10	0.941	0.867	0.910
MRR	0.893	0.802	0.854

References, Acknowledgments, Demos, and Code

[1] S. Balke, V. Arifi-Müller, L. Lamprecht, and M. Müller, *Retrieving audio recordings using musical themes*, ICASSP 2016.

[2] H. Barlow and S. Morgenstern, *A Dictionary of Musical Themes*, 3rd edition, Crown Publishers, Inc., 1975.

[3] R. Bittner, B. McFee, J. Salamon, P. Li, and J. Bello, *Deep salience representations for F0 tracking in polyphonic music*, ISMIR 2017.

[4] J. Bosch and E. Gómez, *Melody extraction based on a source-filter model using pitch contour selection*, SMC 2016.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*, ICML 2006.

[6] F. Korzeniowski and G. Widmer, *Feature learning for chord recognition: The deep chroma extractor*, ISMIR 2016.

[7] F. Zalkow, S. Balke, and M. Müller, *Evaluating salience representations for cross-modal retrieval of Western classical music recordings*, ICASSP 2019.

[8] F. Zalkow, S. Balke, V. Arifi-Müller, M. Müller, *MTD: A multimodal dataset of musical themes for MIR research*, accepted at TISMIR, to appear.

https://www.audiolabs-erlangen.de/resources/MIR/2020-ISMIR-ctc-chroma

Frank Zalkow and Meinard Müller are supported by the German Research Foundation (DFG-MU 2686/11-1, MU 2686/12-1). We thank Daniel Stoller for fruitful discussions on the CTC loss, and Michael Krause for proof-reading the manuscript. We also thank Stefan Balke and Vlora Arifi-Müller as well as all students involved in the annotation work, especially Lena Krauß and Quirin Seilbeck. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS. The authors gratefully acknowledge the compute resources and support provided by the Erlangen Regional Computing Center (RRZE).

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS