CTC-Based Learning of Chroma Features for Score–Audio Music Retrieval

Frank Zalkow and Meinard Müller, *Fellow, IEEE*

***Abstract*—This paper deals with a score–audio music retrieval task where the aim is to find relevant audio recordings of Western classical music, given a short monophonic musical theme in symbolic notation as a query. Strategies for comparing score and audio data are often based on a common mid-level representation, such as chroma features, which capture melodic and harmonic properties. Recent studies demonstrated the effectiveness of neural networks that learn task-specific mid-level representations. Usually, such supervised learning approaches require score–audio pairs where the score's individual note events are aligned to the corresponding time positions of the audio excerpt. However, in practice, it is tedious to generate such strongly aligned training pairs. As one contribution, we show how to apply the Connectionist Temporal Classification (CTC) loss in the training procedure, which only uses weakly aligned training pairs. In such a pair, only the time positions of the beginning and end of a theme occurrence are annotated in an audio recording, rather than requiring local alignment annotations. We evaluate the resulting features in our theme retrieval scenario and show that they improve the state of the art for this task. As a main result, we demonstrate that with the CTC-based training procedure using weakly annotated data, we can achieve results almost as good as with strongly annotated data. Furthermore, we assess our chroma features in depth by inspecting their temporal smoothness or granularity as an important property and by analyzing the impact of different degrees of musical complexity on the features.**

***Index Terms*—Alignment, audio, chroma features, CTC loss, deep learning, music retrieval, musical themes.**

I. Introduction

MUSIC data is available in many different modalities, for example, in the form of audio or video recordings, symbolic representations, or as graphically encoded sheet music [1]. In particular, audio recordings and symbolic scores are important in many music information retrieval (MIR) tasks. One of the typical score–audio retrieval applications is a scenario, where a symbolic score is given as a query, and the task is to identify relevant audio recordings [2]–[6]. In this paper, we use monophonic musical themes in symbolic encodings as queries. For a given query, the aim is to find a relevant recording (i.e., a recording where this theme is played) in an audio database of Western classical music [5], [6]. A famous theme is, for example, the beginning of Beethoven's Fifth Symphony, which is shown in Fig. 1 in a score representation and as a waveform of a performance. One important aspect is that the queries are monophonic, but the themes usually appear with additional musical voices in the audio recordings.

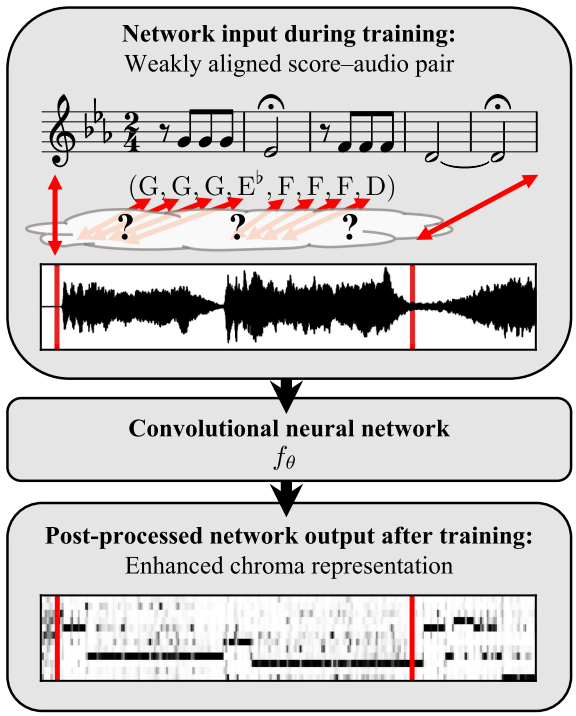


Fig. 1. Illustration of a weakly aligned score–audio pair as the input to a convolutional neural network during training, and an enhanced chroma representation as the post-processed output of the network after training. Music example: First movement of Beethoven's Fifth Symphony, first theme.

Typical cross-modal retrieval strategies (e.g., for score–audio retrieval) employ a common mid-level representation to compare the different modalities. In traditional music processing, chroma features are widely used as mid-level representation [1], [7], [8]. These features, which capture the energy in the twelve bands corresponding to the pitch classes of the chromatic scale, are robust to a certain degree against changes in instrumentation and timbre. In general, computing chroma features with traditional signal processing techniques involves many design choices. As our main contribution, we learn a task-specific chroma representation from training data.

Manuscript received April 6, 2021; revised June 14, 2021 and August 18, 2021; accepted August 27, 2021. Date of publication September 8, 2021; date of current version September 22, 2021. The work of Frank Zalkow and Meinard Müller are supported by German Research Foundation under Grant DFG-MU 2686/11-1, MU 2686/12-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee.

The authors are with the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany (e-mail: frank.zalkow@audiolabs-erlangen.de; meinard.mueller@audiolabs-erlangen.de).

Digital Object Identifier 10.1109/TASLP.2021.3110137

Several studies have shown the benefits of deep-learning models to compute enhanced mid-level representations [9]–[12]. These learned features have proven their effectiveness in many scenarios, such as audio–audio retrieval [13]–[15], chord recognition [11], [12], [16], or melodic pitch tracking [9], [17], [18]. Usually, training deep neural networks (DNNs) for these tasks requires aligned training pairs of audio recordings and corresponding annotations where, for each time position (or frame) of the audio recording, one has an annotation (or class label) to be learned by the model. We denote such pairs as "strongly aligned." For example, recordings with strongly aligned chord annotations have been used to train DNNs for computing chroma-like mid-level features for chord recognition [11], [12]. In general, strongly annotated datasets are crucial for learning meaningful music representations [19], [20]. However, creating strongly aligned training pairs is labor-intensive, and, for many music scenarios, data of this kind is hardly available. Rather than providing local alignments, it is much easier to annotate global correspondences, i.e., the beginning and end time positions of an annotated segment. We denote globally corresponding training pairs without local alignment as "weakly aligned." Fig. 1 illustrates such a weakly aligned score–audio pair for our Beethoven example, where the beginning and end time positions of a theme occurrence are annotated in an audio recording without local alignment in between.

To utilize such weakly aligned training data, we use the *Connectionist Temporal Classification* (CTC) loss to train a neural network. Graves *et al.* [21] originally introduced this loss for labeling unsegmented feature sequences with recurrent DNNs in the context of speech recognition. In the CTC training procedure, a kind of local alignment is computed as part of the loss function rather than having alignment annotations in the training data. We train a neural network with the CTC loss to compute enhanced audio chroma features (see Fig. 1), which are as close as possible to the chroma representations of the symbolic themes. The features are then used in our theme-based retrieval application. In systematic retrieval experiments, we examine if the chroma variant computed by the CTC-based approach, only employing weakly aligned training data, is able to outperform baseline chroma variants from the literature and chroma variants derived from training approaches that employ strongly aligned training data. Our network architecture is inspired by a convolutional neural network that was originally used to compute a pitch salience representation, which is a time–frequency representation and is used to measure the saliency of frequencies over time for tasks such as melody or multi-pitch tracking [9]. In our adaption of this so-called deep salience model, we aim to compute chroma features measuring the saliency of theme-like melodic structures.

This article is a substantially extended version of a previously published conference paper [22]. As a main result, in line with [22], we demonstrate that the CTC-based features improve the state of the art for our theme-based retrieval application. As a major contribution beyond [22], we compare our CTC-based results (using weakly annotated data) with results obtained by a standard training strategy (using strongly annotated data). We show that the retrieval quality achieved with our CTC-based model is almost as high as with a model that we trained with strongly aligned data. We conclude that one can save a lot of tedious annotation work and access much more training data easily by using the CTC loss. To get a deeper understanding of these results, we present several quantitative and qualitative analyses of our CTC-based chroma representation. We show that different features have distinct properties in terms of temporal granularity (or smoothness) and that these properties have an impact on the retrieval results. We also investigate how the musical texture (such as monophonic, homophonic, and polyphonic) affects the retrieval results.

In modern MIR research, reproducibility is an important aspect [23]. To make our results transparent and accessible, we provide a website with various tools and interfaces.[1] First, we make all details of our retrieval results available on an interactive web interface. Second, we provide pre-trained models and code to apply them. Third, our training data is publicly accessible [24].

The remainder of the paper is organized as follows. In Section II, we review related work on score–audio retrieval, deep salience and deep chroma models, as well as on musical applications of the CTC loss. Then, in Section III, we discuss methodological aspects of our study, including our dataset, our retrieval procedure, the adapted neural network architecture, the CTC loss used to train our network, and our approach to derive chroma features. Next, in Section IV, we present an evaluation of our CTC-based features in the context of our retrieval application. As a further main contribution, we analyze in Section V the effect of our CTC strategy by comparing it with standard approaches to train neural networks. Furthermore, in Section VI, we analyze the features' temporal granularity. Then, in Section VII, we come back to our retrieval application and analyze the impact of musical complexity on the retrieval results. In Section VIII, we conduct a small experiment to verify the generalizability of our approach to other datasets. Finally, we conclude with Section IX.

II. Related Work

In the following, we discuss related work on theme-based retrieval, which is our motivating scenario for learning enhanced chroma features. Then, we review deep salience and deep chroma models as used in previous work, as well as musical applications of the CTC loss.

A. Score–Audio Retrieval

In cross-modal music retrieval scenarios, the aim is to find correspondences between different types of music representations [25], such as audio or video recordings, symbolic representations, or graphical sheet music. For example, an audio-visual retrieval task is to find audio excerpts that match a given graphical sheet music representation (or vice versa) [26]. In our score–audio retrieval application, the aim is to find relevant audio recordings for a given symbolically encoded musical theme as a query. Several studies already addressed theme-based music retrieval [5], [6], [22], [27]. A previous study [5] pointed out the challenges of the task, which are due to the differences in modality (symbolic vs. audio), tuning, transposition, tempo, and musical texture between the query and the recordings. The

[1] https://www.audiolabs-erlangen.de/resources/MIR/2021_TASLP-ctc-chroma

difference in musical texture is a major challenge because the themes are monophonic, but they usually appear in a polyphonic context in the recordings. Previous work [6] has shown that pitch salience representations are suitable mid-level features to compare the audio recordings with the symbolic themes. In this paper, building upon these findings, we introduce an approach for learning a task-specific feature representation for our theme-based retrieval task.

B. Deep Salience and Deep Chroma Models

In MIR, many studies have demonstrated the effectiveness of using deep-learning models to compute task-specific feature representations [28]. One example is the use of deep salience models to compute enhanced time–frequency representations (measuring the saliency of frequencies over time) for tasks such as melody [9], [17], [18] or multi-pitch tracking [9]. Another example is the use of deep chroma models for computing enhanced chroma features (encoding the energy in the twelve chromatic pitch class bands) for chord recognition [11], [12], [16], [29]. This paper is inspired by the deep salience approach by Bittner *et al.* [9], who introduced a feature representation named harmonic CQT (HCQT) as input for a convolutional DNN. The HCQT is a three-dimensional tensor, where the three dimensions are time, frequency (logarithmic scaling), and harmonics. The third dimension ensures that harmonically related frequency bins are neighbors across the depth of the tensor. This way, the convolutional kernels of the network can easily exploit harmonic frequency relationships. Assuming a reference frequency $f_{\mathrm{ref}} = 32.7$ Hz (corresponding to the pitch of C_1), the harmonic dimension consists of six CQT representations, where the respective lowest frequency bin corresponds to a frequency of $h \cdot f_{\mathrm{ref}}$, using $h \in \{0.5, 1, 2, 3, 4, 5\}$. Many studies use the deep salience representation as a baseline [18], [30] or build upon this model for diverse tasks such as polyphonic fundamental frequency estimation [31], dominant melody estimation [10], instrument recognition [32], tempo estimation [33], or chord recognition [34].

The study of Wu *et al.* [34] is related to ours in two respects. First, they also use the HCQT representation, and, second, they use weakly aligned training data. However, they aim for chord recognition instead of learning a mid-level representation for score–audio retrieval. In contrast to our contribution, they take a three-step approach: First, they use a pre-trained deep chroma extractor to compute features. Second, they automatically align their chord labels to the chroma features using a hidden Markov model (HMM). Third, they use a frame-wise DNN classifier for chord recognition. In our paper, we present a single-step approach to realize the alignment within the DNN training procedure.

C. Musical Applications of the CTC Loss

The Connectionist Temporal Classification (CTC) loss refers to a loss function used to train neural networks where a sequence of target labels is temporally aligned to the network's output during the loss computation. We describe the computation procedure in Section III-D.

TABLE I
DATASET OVERVIEW. DURATION FORMAT: HH:MM:SS. SD: STANDARD DEVIATION

	#	Mean Dur.	SD Dur.	Total Dur.
Themes	2067	00:00:09	00:00:06	04:57:48
Audio Recordings	1126	00:06:24	00:04:12	120:03:03

Originally proposed for the task of speech recognition [21], CTC has been adopted to several MIR applications, including optical music recognition [35]–[38], monophonic audio-to-score transcription [39], lyrics alignment [40]–[42], and audio tagging [43]. An alternative to CTC for sequence learning without aligned training data is the usage of an attention mechanism, which was used for, e.g., monophonic singing voice transcription [44].

III. METHODS

In this section, we present various methodological aspects relevant to the investigations of this paper. First, we describe the dataset used in the experiments (Section III-A). Second, we outline the basic retrieval pipeline, which is used later in the experiments to evaluate our learned features (Section III-B). Third, we explain our adaptation of a deep salience model to compute an enhanced chroma representation for our score–audio music retrieval application (Section III-C). Fourth, we introduce the CTC loss used to train the adapted model (Section III-D). Fifth, we describe how to derive chroma features from the output of the CTC-based network (Section III-E).

A. Dataset

For our experiments, we use a dataset based on "A Dictionary of Musical Themes" by Barlow and Morgenstern [45]. This book from 1948 contains roughly 10000 musical themes. The dataset covers a subset of 2067 of these themes and is publicly available as Musical Theme Dataset (MTD) [24]. For each theme, the MTD provides a symbolic encoding and an occurrence in an audio recording. Furthermore, it comprises annotations about differences in transposition between the symbolic and audio versions. For our retrieval experiments, we also use the entire audio recordings, where the occurrences have been annotated. In total, the audio database consists of 1126 audio recordings with a duration of about 120 hours. A theme corresponds to precisely one recording, which, in turn, can contain the rendition of several themes. Table I shows some statistics for the dataset.

B. Basic Retrieval Procedure

Closely following [5], [6], we describe our retrieval pipeline and our evaluation measures. First, we have a set $\mathcal{Q}$ of symbolic encodings of musical themes, which serve as *queries*. Furthermore, we have a collection of audio recordings, which we denote as database *documents*.

Throughout this paper, we focus on the challenges due to the difference in musical texture (monophonic queries and audio recordings of polyphonic music). Accordingly, we use a controlled retrieval scenario where, for each query, there is

precisely one audio document that contains a globally corresponding rendition of the query theme. I.e., we ensure matching transpositions and scale the query using a tempo that results in approximately the same duration as the theme occurrence in the recording. For a fixed symbolic query $Q \in \mathcal{Q}$, the aim is to retrieve the corresponding audio document. To compare the query with a document, we convert both into chroma sequences. For the symbolic query, we simply compute a binary chroma representation. For converting the audio recording, we employ an enhanced chroma representation (from our CTC or a baseline approach, as described later). Then, we use Subsequence Dynamic Time Warping (SDTW) to compare the query with subsequences of the document [1]. Inspired by [5], [6], we use the cosine distance, the step size condition $\Sigma := \{(2,1),(1,2),(1,1)\}$, as well as the weights $w_{\text{vertical}} = 2$ and $w_{\text{horizontal}} = w_{\text{diagonal}} = 1$. As a result of SDTW, one obtains a matching function, where local minima point to locations with a good match between the query and a document subsequence. We consider the minimal value of the matching function as the distance between query and document.

To obtain the retrieval result, we compute distances between all documents and the query. Note that in our scenario, there is only one relevant document for a given query $Q \in \mathcal{Q}$. We order the documents according to ascending distance values. In the resulting ordered list, the relevant document's position (or *rank*) is denoted by $r_Q \in \mathbb{N}$. We evaluate the retrieval procedure by assessing only the top $K \in \mathbb{N}$ documents of the ranked list. For a given K, the retrieval for a query is considered successful if its relevant recording is among the top K matches (i.e., $r_Q \leq K$). Employing all queries $Q \in \mathcal{Q}$, we then compute the top-K recall rate $\rho_K \in [0,1]$ (in short: recall@K or top-K rate) as the proportion of queries with a successfully retrieved document:

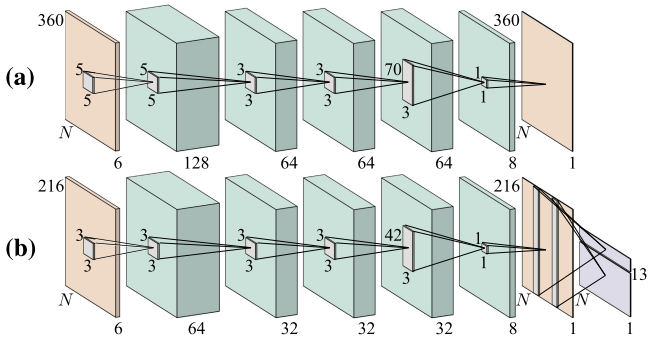
$$\rho_K = |\{Q \in \mathcal{Q} : r_Q \leq K\}| \,/\, |\mathcal{Q}|. \tag{1}$$

Furthermore, we report the *mean reciprocal rank* (MRR), which is the average of $1/r_Q$ across all queries.

C. Deep Salience Model Adaptation

In this section, we explain our adaption of the deep salience model by Bittner *et al.* [9], who approached the task of melody and multi-pitch tracking using a strongly aligned dataset of 10 hours. In our case, we aim to learn an enhanced chroma representation for score–audio retrieval, employing our weakly aligned 5-hour dataset of 2067 themes. To avoid overfitting to the smaller dataset, we simplified the original model in several ways by reducing the number of parameters and memory requirements. Additionally, we adapted the network such that it can be trained with the CTC loss and used as a deep chroma extractor.

Figs. 2(a) and (b) illustrate the original network architecture and our adapted version, and the table in Fig. 2(c) gives further details for our version. Compared to the model by Bittner *et al.* [9], we introduce the following modifications: First, we use a feature rate of 25 Hz (i.e., 25 feature vectors per second) instead of 86 Hz. Second, we use a frequency resolution of a third semitone instead of a fifth semitone. The high time and frequency resolutions of the original model may be beneficial in the application of melody estimation, but are not needed for



(c)

Layer	Output Shape	Activation	Parameters
Input	(N, 216, 6)		
Conv2D 64 × (3, 3, 6)	(N, 216, 64)	LReLU	3520
Conv2D 32 × (3, 3, 64)	(N, 216, 32)	LReLU	18 464
Conv2D 32 × (3, 3, 32)	(N, 216, 32)	LReLU	9248
Conv2D 32 × (3, 3, 32)	(N, 216, 32)	LReLU	9248
Conv2D 8 × (3, 42, 32)	(N, 216, 8)	LReLU	32 264
Conv2D 1 × (1, 1, 8)	(N, 216, 1)	Sigmoid	9
Pooling	(N, 13)	Softmax	217

Fig. 2. Network architectures. **(a)** Illustrations of the original architecture proposed by Bittner *et al.* [9] and **(b)** our adapted architecture used in this paper. Illustrations inspired by [9]. **(c)** Details for our adapted architecture (72970 parameters in total).

our task (learning chroma features for retrieval). The decreased frequency resolution results in 216 instead of 360 frequency bins. As third modification, we reduced the number of filter kernels as well as the size of some of the filter kernels. The latter reduction accounts for the decreased frequency resolution. Fourth, we use leaky ReLU activations instead of ReLU activations to avoid zero gradients [46], [47]. Fifth, we do not use batch normalization, which was used at the input to each layer in the original model. Instead, we ℓ^2-normalize all columns of the input to the network for being invariant to dynamics. Sixth, we add a pooling layer at the end, which we explain in the next paragraph.

After the last convolutional layer (with sigmoid activation), we obtain a representation that we could interpret as a kind of pitch salience representation of size $N \times 216$, where $N \in \mathbb{N}$ is the number of time steps. In our retrieval scenario, where we want to learn chroma-based mid-level features, we aim for an output size of $N \times 13$. Here, each of the N columns encodes a probability vector over the set of the twelve chroma labels and an additional blank symbol, which means that no chroma label is active (more details in Section III-D). Let us consider a single column of size 216 as input, which we want to transform to a probability vector of size 13. To compute the first twelve entries, we add up all pitch bins corresponding to the respective chroma bins. This fixed pooling has no learnable parameters. To compute the last entry for the blank symbol, we apply a standard dense layer (linear activation) to the input column. This layer has 217 learnable parameters (216 weights and a bias). Finally, we apply the softmax function to the resulting 13-dimensional vector. We transform all columns of the input using this pooling procedure.

In summary, our adapted model differs from the original model [9] in two important aspects: First, we reduced the number of parameters from 406921 to 72970. Second, the model's output is a sequence of probability vectors over 13 dimensions

(encoding chroma vectors at a semitone-resolution) rather than 360 dimensions (encoding a pitch salience at a fifth semitone resolution).

D. CTC Loss

In the following, we present the main idea of the CTC loss function introduced by Graves *et al.* [21]. Beyond the original article [21], we also recommend the review on the CTC loss in the thesis by Hannun [48]. We describe the CTC loss computation for a single pair consisting of an audio feature sequence and a label sequence. Let

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \tag{2}$$

denote the feature sequence of length $N \in \mathbb{N}$, which consists of feature vectors $\boldsymbol{x}_n \in \mathbb{R}^D$ for $n \in [1:N] := \{1, 2, \ldots, N\}$ and dimensionality $D \in \mathbb{N}$. The second sequence of the pair is a label sequence

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M) \tag{3}$$

of length $M \in \mathbb{N}$, where typically $M \ll N$. This sequence consists of elements $\boldsymbol{y}_m \in \mathbb{A}$ for $m \in [1:M]$. The alphabet $\mathbb{A}$ is the set of symbols that can occur in the label sequence. For example, in the case of lyrics alignment, the alphabet is the set of all possible characters [40]. In our case, it is the set of the twelve different chroma labels:

$$\mathbb{A} := \{\mathrm{C}, \mathrm{C}^{\#}, \mathrm{D}, \ldots, \mathrm{B}\}. \tag{4}$$

A DNN f_θ with parameters θ transforms the feature sequence $\boldsymbol{X}$ to a sequence of probability distributions

$$f_\theta(\boldsymbol{X}) = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N) \tag{5}$$

having the same length N as the feature sequence. Each element of the sequence $\boldsymbol{p}_n : \mathbb{A}' \to [0,1]$ maps a symbol from the modified alphabet $\mathbb{A}'$ to a probability value. The modified alphabet

$$\mathbb{A}' := \mathbb{A} \cup \{\epsilon\} \tag{6}$$

contains an additional blank symbol ϵ, which encodes that no symbol is active. We further explain the role of this symbol later in this section.

When we align the feature sequence $\boldsymbol{X}$ and the label sequence $\boldsymbol{Y}$, we assign a suitable symbol to each time frame. Intuitively, we can consider this alignment as an expansion of the label sequence $\boldsymbol{Y}$ to the length of the feature sequence. More formally, an alignment is represented by a sequence

$$\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N) \tag{7}$$

of elements $\boldsymbol{a}_n \in \mathbb{A}'$ and length N that satisfies the following condition. When removing all consecutive duplicates and then all blank symbols ϵ, the alignment sequence $\boldsymbol{A}$ is reduced to the label sequence $\boldsymbol{Y}$. Given an alignment $\boldsymbol{A}$ and the sequence of probability distributions, we can compute the probability

$$P(\boldsymbol{A}|\boldsymbol{X}) = \prod_{n=1}^{N} \boldsymbol{p}_n(\boldsymbol{a}_n) \tag{8}$$

of the alignment. When computing the CTC loss, we do not explicitly know the correct alignment, but only the label sequence. Because a label sequence can correspond to multiple alignments, all possible alignments between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are considered. Let us denote the sum of probabilities for all these alignments as $P(\boldsymbol{Y}|\boldsymbol{X})$. The final CTC loss for a single training pair is

$$L_\theta(\boldsymbol{X}, \boldsymbol{Y}) = -\log P(\boldsymbol{Y}|\boldsymbol{X}). \tag{9}$$

This loss function is used in mini-batch gradient descent to update the parameters θ by averaging the loss value over all training pairs in a mini batch. By this procedure, the network's parameters improve to produce probability sequences that make the ground-truth label sequences of the training set more probable. Graves *et al.* [21] described how to compute $P(\boldsymbol{Y}|\boldsymbol{X})$ in a differentiable and efficient way using dynamic programming, similar to the forward algorithm for HMMs [48], [49].

Finally, we want to clarify the role of the blank symbol ϵ. We already mentioned one role for ϵ, namely the indication of no active symbol, i.e., a rest in our music application. But as an additional role, the symbol also indicates repetitions. Let us illustrate this role using the first theme of Beethoven's famous Fifth Symphony as an example (Fig. 3a). We only consider the beginning of the label sequence for brevity, i.e., $\boldsymbol{Y} = (\mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{E}^\flat)$. An alignment of $\boldsymbol{A} = (\mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{E}^\flat)$ would not correspond to this label sequence because we remove the consecutive duplicates when converting an alignment to a label sequence. Rather, this alignment corresponds to the label sequence $\boldsymbol{Y} = (\mathrm{G}, \mathrm{E}^\flat)$. To represent repeated symbols in the label sequence, we need to

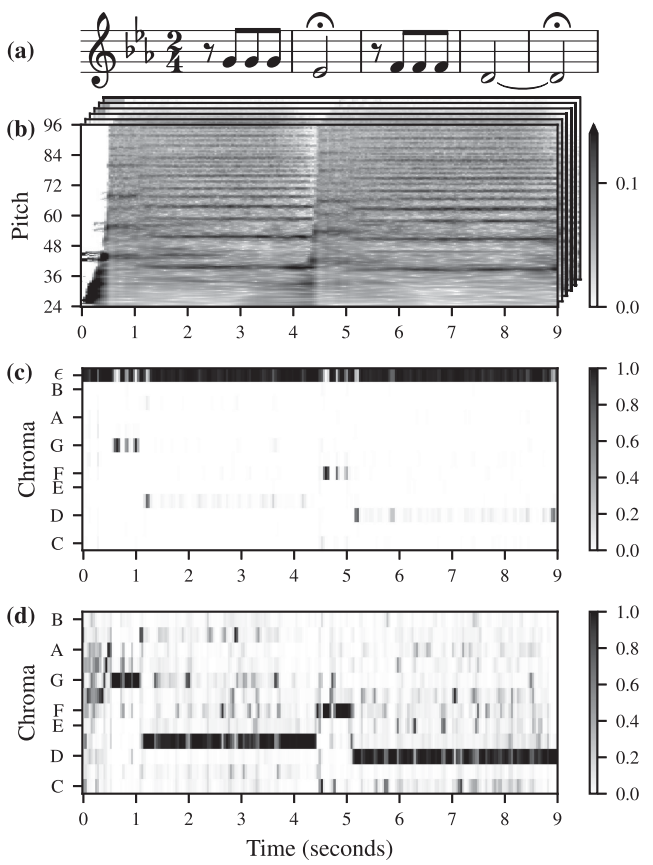


Fig. 3. Representations for the first theme of Beethoven's Fifth Symphony. **(a)** Score of monophonic theme. **(b)** HCQT input representation $\boldsymbol{X}$ (front slice corresponding to the first harmonic). **(c)** Network output. **(d)** Chroma features used for matching.

separate them by a blank. A valid alignment for the Beethoven excerpt is, e.g., $\boldsymbol{A} = (\mathrm{G}, \epsilon, \mathrm{G}, \epsilon, \mathrm{G}, \mathrm{E}^{\flat})$.

E. CTC-Based Chroma Features

We train the model described in Section III-C with the CTC loss. The network's input is an HCQT tensor computed for an excerpt from an audio recording, where a musical theme is played. As an example, Fig. 3a shows the score of the Beethoven theme. Fig. 3b shows a slice of the HCQT features for a recording of this theme. The corresponding label sequence is the sequence of chroma labels of the theme with neither rhythmic information nor temporal alignment to the input (see also Fig. 1). For this Beethoven example, the sequence is $\boldsymbol{Y} = (\mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{E}^{\flat}, \mathrm{F}, \mathrm{F}, \mathrm{F}, \mathrm{D})$. The network's output is a sequence of probability distributions, visualized in Fig. 3c for the Beethoven example. We see that the ϵ symbol has the largest probability most of the time, and the chroma labels only have large probabilities at the beginning of the corresponding note events. To use the network output as a feature representation, we remove the row corresponding to the ϵ symbol and interpret the resulting matrix as a sequence of chroma features. Finally, we ℓ^2-normalize the 12-dimensional chroma vectors to compensate for the removed ϵ symbol. The ℓ^2 norm was chosen because the features are used later for SDTW, where the cosine distance is applied. Fig. 3d shows the normalized chroma features for the Beethoven example, which correspond well to the label sequence.

IV. Retrieval Experiments

We now evaluate the CTC-based chroma features in the context of our retrieval application. This task allows us to evaluate the features with quantitative evaluation measures. Using such measures, we also compare our learned representation with other features, computed by a traditional method and a deep-learning approach not adapted to the theme retrieval task.

A. Baseline Experiments

In the experiments throughout this paper, we consider various chroma variants referred to by the symbol $\mathcal{C}$ (with an additional subscript to specify a particular variant or without subscript when referring to standard chroma features using the full spectral content). To compare our approach with prior work on score–audio retrieval, we use the best-performing chroma representations from a previous evaluation study [6]. The authors proposed converting pitch salience representations to chroma features by a simple pooling strategy, where the energies of the frequency bins that correspond to the same chroma are summed. The first baseline chroma variant ($\mathcal{C}_{\mathrm{BG1}}$) is based on a traditional pitch salience representation by Bosch and Gómez [50], which is computed by combining a source-filter model with harmonic summation, using threshold parameters that are particularly suited for orchestral music [51].[2] The second baseline chroma variant ($\mathcal{C}_{\mathrm{Bit}}$) is based on the original deep salience representation for melody estimation by Bittner *et al.* [9]. To compute this representation, we use the original network[3] that was not adapted to our task and trained with a standard strongly aligned approach (using the public MedleyDB [20] in combination with additional non-public training data).

TABLE II
Retrieval Results of the Baseline Methods (a) Using a Feature Rate of 10 Hz as Reported in Previous Work [6], (b) Using a Feature Rate of 25 Hz

(a)

	Top-1	Top-5	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{BG1}}$	0.754	0.835	0.861	0.885	0.913	0.792
$\mathcal{C}_{\mathrm{Bit}}$	0.693	0.788	0.823	0.853	0.896	0.739

(b)

	Top-1	Top-5	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{BG1}}$	0.824	0.894	0.911	0.925	0.952	0.857
$\mathcal{C}_{\mathrm{Bit}}$	0.767	0.846	0.868	0.895	0.930	0.805

Table II(a) cites the results from the previous study [6], where a 10 Hz feature rate was used. According to this study, three quarters of the themes (75.4 %) yielded the relevant document on the first rank using $\mathcal{C}_{\mathrm{BG1}}$. For $\mathcal{C}_{\mathrm{Bit}}$, this is the case for 69.3 % of the themes. We checked whether the feature rate is appropriate for the given retrieval task and found that an increased time resolution is beneficial for the retrieval quality. A rate of 25 Hz turned out to be a good trade-off between retrieval accuracy and efficiency. We repeated the experiments for $\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}_{\mathrm{Bit}}$ with the increased feature rate and show the results in Table II(b). Just by changing the temporal resolution, we see a substantial improvement in the results. For example, for $\mathcal{C}_{\mathrm{Bit}}$, the top-1 rate increases from 0.693 to 0.767, which means that the number of query themes with a correct top match increased by about 7 %.[4] The reason for this may be the following: A fast tempo of *Presto* corresponds to up to 200 BPM. Having a quarter-note beat in such a tempo, a sixteenth note has a duration of 75 ms, which is shorter than the length of a frame given the feature rate of 10 Hz. In such cases, the increased feature rate is necessary to represent the musical content in a more meaningful way. In all subsequent experiments, we use the feature rate of 25 Hz.

For both feature rates, the representation $\mathcal{C}_{\mathrm{BG1}}$ performs better than $\mathcal{C}_{\mathrm{Bit}}$. For example, the respective top-1 rates are 0.824 and 0.767 for the 25 Hz rate. The results for $\mathcal{C}_{\mathrm{Bit}}$ may be lower because the training data of the underlying DNN consisted mainly of popular music (for overall 240 training tracks, only 22 are tagged as "classical" in version 1 of MedleyDB [20]). Another possible reason is that the saliency characteristics in the training data (coming from the "Melody 2" definition of MedleyDB) are different from the characteristics of musical themes.

B. Training Details

We split the 2067 score–audio pairs of our dataset (see Section III-A) into five folds, where we use three folds for training, one for validation, and another one for testing. We ensure that all themes by a composer are part of precisely one fold. As a

[2]The specific parameter setting is named "BG1" in [51].

[3]Original weights ("Melody 2"). In [6], $\mathcal{C}_{\mathrm{Bit}}$ was denoted by $\mathcal{C}_{\mathrm{CNN}}$ and $\mathcal{C}$ was denoted by $\mathcal{C}_{\mathrm{IIR}}$.

[4]Compared to the previous experiment [6], the dataset was revised and slightly extended, accounting for up to 2 % improvements in the accuracy. Still, the main improvements are due to the increased time resolution.

TABLE III
RETRIEVAL RESULTS FOR $\mathcal{C}_{\mathrm{CTC}}$

Fold	Queries	**Top-1**	**Top-5**	**Top-10**	**Top-20**	**Top-50**	**MRR**
1	559	0.875	0.941	0.957	0.964	0.970	0.903
2	377	0.828	0.891	0.905	0.926	0.947	0.859
3	377	0.859	0.918	0.934	0.952	0.966	0.884
4	377	0.899	0.947	0.958	0.968	0.981	0.922
5	377	0.873	0.931	0.950	0.958	0.981	0.900
∅		0.867	0.927	0.942	0.955	0.969	0.894

consequence, we do not use themes from the same composer for training and evaluation in order to avoid overfitting to the characteristics of particular composers. The first fold contains more themes (559) than the others because it contains all MTD themes by Ludwig van Beethoven, which is the most prominent composer of the dataset, see [24]. The other folds have fewer themes (377) and are more diverse in terms of composers, having 12 to 14 different composers each.

During training, we apply circular shifts along the chroma axis as data augmentation to simulate transpositions (up to a minor third upwards and downwards). We perform mini-batch gradient descent with a mini-batch size of eight using the Adam optimizer [52] and a learning rate annealing procedure. In the first phase of this procedure, the initial learning rate is 0.001, and we train the model until the loss for the validation fold does not improve for five epochs. In the next phase, we halve the learning rate and continue the training with the model that has the lowest validation loss among the models of all previous epochs. We repeat ten such phases. After the training is finished, we use the model with the lowest validation loss as a chroma feature extractor, and evaluate its effectiveness in the retrieval scenario. We only use the query themes from the respective test fold and all 1126 documents of our database for retrieval. The reported average evaluation measures (∅) are weighted with the number of queries from the respective test fold.

C. CTC-Based Results

We now discuss the results we achieved with our CTC-based approach $\mathcal{C}_{\mathrm{CTC}}$. Table III shows the evaluation results for the five cross-validation iterations. The second column gives the number of query themes in the respective test fold. The retrieval results vary, ranging from a top-1 rate of 0.828 for the second test fold up to 0.899 for the fourth test fold. The last row of the table shows an average of the results, weighted by the number of queries used. Overall, we see a substantial improvement compared to the baseline approaches (Table II b). For example, the average top-1 rate is 0.867 for $\mathcal{C}_{\mathrm{CTC}}$, compared to 0.767 for $\mathcal{C}_{\mathrm{Bit}}$ and 0.824 for $\mathcal{C}_{\mathrm{BG1}}$. There are also improvements for larger ranks, such as in the top-50 rate (0.969 compared to 0.930 and 0.952, respectively). When repeating the training and evaluation procedures with different random initializations of the network weights, we only observed minor variations in the average evaluation measures (below 1 %). The results show that our approach is able to outperform the baselines, which are the state of the art for the given retrieval task [6].

In our experiments, we used the CTC-based chroma features in a retrieval pipeline based on SDTW. As an alternative, one might also directly apply the CTC loss for retrieval. We address this idea in the appendix of this article.

TABLE IV
RETRIEVAL RESULTS (∅) FOR AN ORACLE OF THE BASELINE BY BOSCH AND GÓMEZ [50] AND OUR CTC APPROACH

	Top-1	**Top-5**	**Top-10**	**Top-20**	**Top-50**	**MRR**
$\mathcal{C}_{\mathrm{BG1}}$	0.824	0.894	0.911	0.925	0.952	0.857
$\mathcal{C}_{\mathrm{CTC}}$	0.867	0.927	0.942	0.955	0.969	0.894
Oracle	0.907	0.947	0.960	0.970	0.983	0.925

D. Oracle Experiment

The traditional approach $\mathcal{C}_{\mathrm{BG1}}$ also shows excellent performance for this task. To investigate the relationship between $\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}_{\mathrm{CTC}}$, we evaluated both strategies with an oracle fusion procedure. Given a query, let us denote the rank from by the baseline by $r_{\mathrm{BG1}} \in \mathbb{N}$ and from the CTC-based approach by $r_{\mathrm{CTC}} \in \mathbb{N}$. In our oracle procedure, we consider the better rank $\min(r_{\mathrm{BG1}}, r_{\mathrm{CTC}})$ when computing the evaluation measures. The oracle's average evaluation measures over all queries indicate the results of an optimal fusion of both methods.

Table IV again shows the results for $\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}_{\mathrm{CTC}}$ for convenience and shows the oracle results in the third row. The oracle further improves the results for $\mathcal{C}_{\mathrm{CTC}}$. For example, the top-1 rate is 4 % larger (0.907 instead of 0.867). For top-K rates with larger K, there are still some small improvements. The oracle indicates that $\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}_{\mathrm{CTC}}$ capture different aspects for certain queries. $\mathcal{C}_{\mathrm{BG1}}$ is a slightly better feature representation for some queries than the CTC-based approach. We conclude that there is still some room for improvement in the given retrieval scenario for future work, e.g., by combining different feature representations.

E. Representative Example

To illustrate the properties of our CTC-based features, we close this section by comparing various mid-level representations for a representative example.

Fig. 4a shows the full score and the chroma sequence for the second theme in the first movement of Beethoven's Piano Sonata Op. 2, No. 2. In this case, the theme is played by the right hand (upper staff), and the left hand (lower staff) plays an accompaniment. The sixteenth notes of the accompaniment present a minor triad $(\mathrm{E}, \mathrm{G}, \mathrm{B})$ in the pickup and first measure, and a diminished triad $(\mathrm{F}^\sharp, \mathrm{A}, \mathrm{C})$ in the second and third measure. Ideally, for our retrieval scenario, we aim for a chroma representation that only captures energy from the theme and not from the accompaniment. Figs. 4b–e show chroma features for the full spectral content $\mathcal{C}$, the baseline approaches $\mathcal{C}_{\mathrm{BG1}}$, $\mathcal{C}_{\mathrm{Bit}}$, and our CTC strategy $\mathcal{C}_{\mathrm{CTC}}$, respectively. The accompaniment is strongly represented in the representation using the full spectral content (Fig. 4b). For example, in the beginning, most energy is in the E, G, and B bands, which correspond to the accompaniment's E minor triad. In the representation $\mathcal{C}_{\mathrm{BG1}}$ (Fig. 4c), the main notes of the theme are well represented. However, some shorter notes of the theme (e.g., fourth note G or seventh note $\mathrm{F}^\sharp$) are not salient in this representation. $\mathcal{C}_{\mathrm{Bit}}$ (Fig. 4d) does not capture the theme well. This is especially the case in the

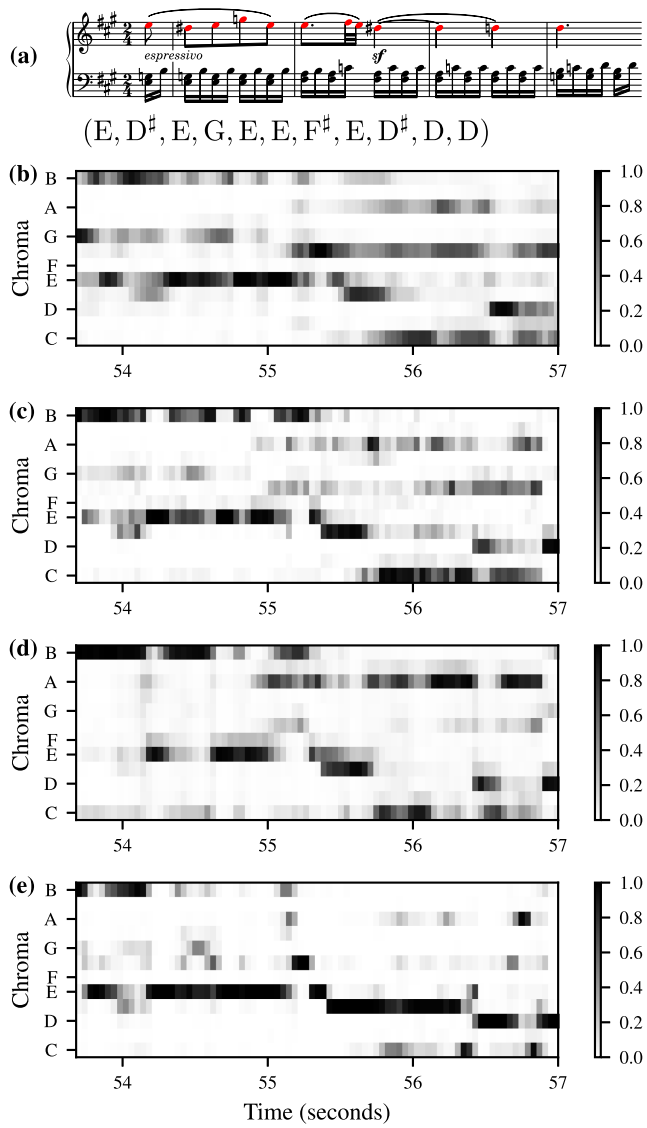


Fig. 4. Second theme of Beethoven's Piano Sonata Op. 2, No. 2, first movement. **(a)** Full score with the theme's notes colored in red, along with the chroma sequence of the theme. **(b)** $\mathcal{C}$. **(c)** $\mathcal{C}_{\mathrm{BG1}}$. **(d)** $\mathcal{C}_{\mathrm{Bit}}$. **(e)** $\mathcal{C}_{\mathrm{CTC}}$.

second half, where the chroma bin A has the highest energy, which is part of the accompaniment. Among all representations, the theme is most evident in $\mathcal{C}_{\mathrm{CTC}}$ (Fig. 4e). In general, $\mathcal{C}_{\mathrm{CTC}}$ attenuates the energy in the chroma bands corresponding to the accompaniment. The ability to represent the chroma energy of a musical theme is the main reason why our CTC-based features are a powerful tool for score–audio music retrieval.

V. Effect of CTC Loss

We showed that our CTC-based chroma representation outperforms the baseline approaches in our retrieval application. To further analyze the effect of the CTC strategy, we performed additional experiments, where we learned features without the CTC loss function. Instead, we trained our adapted DNN model using a standard loss function (categorical cross-entropy). This training procedure requires strongly annotated training data and implies classifying each spectral frame with respect to the twelve chroma labels.

TABLE V
Retrieval Results (∅) Using Cross-Entropy

	Top-1	Top-5	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{CTC}}$	0.867	0.927	0.942	0.955	0.969	0.894
$\mathcal{C}_{\mathrm{linear}}$	0.829	0.897	0.914	0.929	0.953	0.863
$\mathcal{C}_{\mathrm{strong}}$	0.882	0.927	0.939	0.945	0.961	0.904

A. Comparison With Linear Scaling

To treat our task as a classification problem, we have to assume a particular temporal alignment between the symbolic themes and the corresponding excerpts in the audio recordings. As a first alignment approach ($\mathcal{C}_{\mathrm{linear}}$), we assume a constant tempo throughout the theme occurrence. We used binary chroma representations to encode the output labels by temporally scaling the symbolic themes to the same length as the corresponding audio excerpts in a linear way (using nearest-neighbor interpolation). As a result of the scaling procedure, we technically obtain a one-to-one correspondence between the input representations (spectral frames) and the target labels (binary chroma vectors), facilitating a frame-wise training of our neural network (without the need for an alignment at the loss stage). Similar to the CTC strategy, this approach uses weakly aligned data, but we also used the rhythm information and note durations from the MIDI files (which we did not use for $\mathcal{C}_{\mathrm{CTC}}$). The blank symbol now only indicates rests in a theme. Since the assumption of a constant tempo is not realistic, we expect this to yield a sort of lower limit for the performance of the CTC-based approach.

We trained the model with the linearly scaled training data, used it as a chroma extractor, and then evaluated this strategy in the theme retrieval context. The first row of Table V repeats the average evaluation measures from Table III for convenience, and the second row presents the average results for the classification strategy with linear scaling ($\mathcal{C}_{\mathrm{linear}}$). For $\mathcal{C}_{\mathrm{linear}}$, the evaluation measures are lower than the CTC-based results, e.g., having a top-1 rate of 0.829 compared to 0.867. This difference is due to the non-linear temporal correspondence between the audio recordings and the symbolic themes.

The evaluation measures for $\mathcal{C}_{\mathrm{linear}}$ are in a similar range as the results for $\mathcal{C}_{\mathrm{BG1}}$ (given in Table II b), which highlights that the learning procedure yields a decent representation despite the linear scaling.

B. Comparison With Strongly Aligned Data

As a second alignment approach ($\mathcal{C}_{\mathrm{strong}}$), we use manually aligned correspondences between the symbolic themes and the audio occurrences. In contrast to the previous strategies, this is a standard strongly aligned approach to train a neural network, similar to the training procedure used for the original deep salience model [9]. Creating the strong alignment annotations was highly labor-intensive because it implied annotating the onset time position for every note in each theme. We expect that training with strongly aligned data yields an upper limit for the performance of the CTC-based approach.

The third row of Table V shows the results for the classification approach using the manual alignments ($\mathcal{C}_{\text{strong}}$). The strongly aligned approach slightly improves the results compared to the CTC strategy. For example, the top-1 rate is 0.882 compared to 0.867. For higher ranks, both strategies are on par with each other, e.g., yielding top-10 rates of 0.939 and 0.942, respectively. The fact that our CTC-based results are closer to the upper limit ($\mathcal{C}_{\text{strong}}$) than the lower limit ($\mathcal{C}_{\text{linear}}$) demonstrates that the CTC loss implicitly handles the alignment problem well in the training procedure. Without CTC, one has to take care of the alignment at the input level, using annotations, which are often not available or hard to generate.

C. Qualitative Comparison

Fig. 5 shows the score and five feature variants for a theme by A. Dvořák. The first representation (Fig. 5b) is based on a symbolic encoding, manually aligned to the audio occurrence. Fig. 5c shows standard chroma features $\mathcal{C}$, using the full spectral content. As expected, the theme and the accompanying voices influence the features. While the theme's quarter notes of the first bar are clearly visible, the theme's sixteenth notes of the second and third bar are not captured as well. Fig. 5d shows our CTC-based representation. Despite some noise in the first half, we see that $\mathcal{C}_{\text{CTC}}$ mainly captures the theme's notes. Fig. 5e shows the representation $\mathcal{C}_{\text{linear}}$ from the weakly aligned classification approach, which is smoother compared to $\mathcal{C}_{\text{CTC}}$. In fact, the representation $\mathcal{C}_{\text{linear}}$ is strongly over-smoothed. The features have this property because there is no accurate temporal correspondence between the input and output representations in the corresponding training strategy. As a consequence, the model temporally smears the active chroma bins. In the next section, we will further analyze the feature's temporal granularity and smoothness. The features from the strongly aligned classification approach $\mathcal{C}_{\text{strong}}$ (Fig. 5f) are cleaner and sharper compared to the other audio representations.

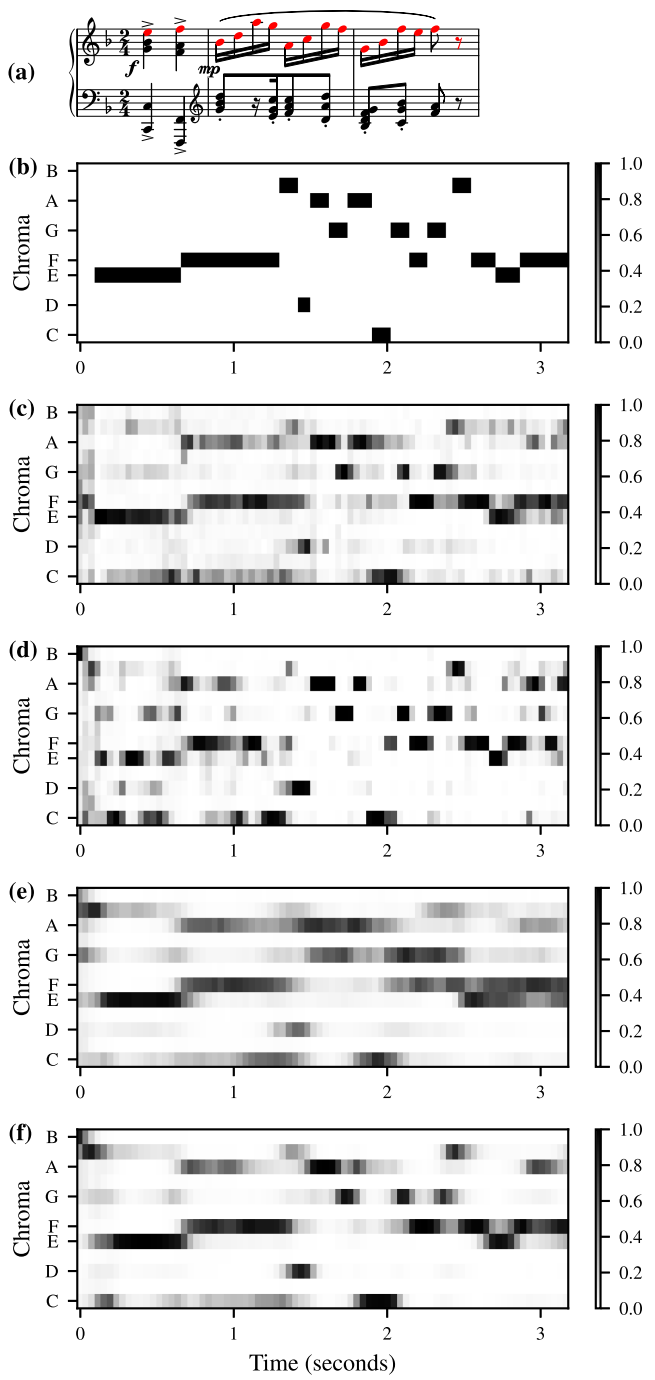


Fig. 5. Representations for the first theme of A. Dvořák's Slavonic Dance in F major Op. 72, No. 3. **(a)** Score of piano reduction with the theme's notes colored in red. **(b)** $\mathcal{C}_{\text{MID}}$. **(c)** $\mathcal{C}$. **(d)** $\mathcal{C}_{\text{CTC}}$. **(e)** $\mathcal{C}_{\text{linear}}$. **(f)** $\mathcal{C}_{\text{strong}}$.

VI. Temporal Granularity

In the previous section, we showed that differently learned features have distinct properties. In particular, we observed different degrees of temporal granularity and smoothness in the feature representations. To better understand how these differences impact our retrieval results, we now compare the chroma representations $\mathcal{C}_{\text{CTC}}$ and $\mathcal{C}_{\text{BG1}}$ in terms of temporal granularity.

A. Granularity Measure

We now introduce a measure to quantify the temporal granularity of a feature representation on a scale from low granularity (i.e., smooth) to high granularity (i.e., fine-grained). To this end, we compare a sequence of ℓ^2-normalized chroma feature vectors

$$\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N) \tag{10}$$

of length $N \in \mathbb{N}$, having elements $\boldsymbol{c}_n \in \mathbb{R}^{12}$ for $n \in [1:N]$, with a smoothed variant of $\boldsymbol{C}$. To compute this variant, we apply a temporal average filter of 6 frames (corresponding to 240 ms) and then again ℓ^2-normalize each vector. We apply the smoothing in a centric way, using suitable zero-padding conventions. As a result, the smoothed sequence

$$\boldsymbol{C}^{\text{smooth}} = (\boldsymbol{c}_1^{\text{smooth}}, \boldsymbol{c}_2^{\text{smooth}}, \ldots, \boldsymbol{c}_N^{\text{smooth}}) \tag{11}$$

has the same length N as the original sequence. Then, inspired by the distance measure used for SDTW, we compute as a measure of temporal granularity $\Gamma : \mathbb{R}^{12\times N} \to [0,1]$ the average cosine

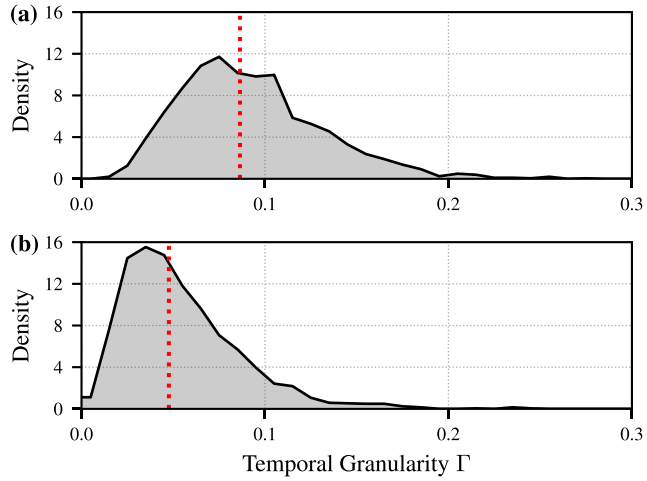


Fig. 6. Distribution of granularity values for **(a)** $\mathcal{C}_{\mathrm{CTC}}$ and **(b)** $\mathcal{C}_{\mathrm{BG1}}$. The red dotted lines indicate the median value of the respective distribution.

distance

$$\Gamma(\boldsymbol{C}) = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \frac{\langle \boldsymbol{c}_n, \boldsymbol{c}_n^{\mathrm{smooth}}\rangle}{\|\boldsymbol{c}_n\| \cdot \|\boldsymbol{c}_n^{\mathrm{smooth}}\|}\right). \tag{12}$$

High Γ-values indicate fine-grained features since the features are dissimilar to their smoothed variants. Low Γ-values indicate smooth features since they are similar to their smoothed variants. Fig. 6 shows the distribution of Γ-values for the 2067 themes of our dataset. For the CTC approach (Fig. 6a), the median Γ is around 0.09, and for $\mathcal{C}_{\mathrm{BG1}}$ (Fig. 6b), it is approximately 0.05. This difference shows that the CTC-based features are more fine-grained.

B. Effect on Retrieval

How does the difference in temporal granularity of our features impact the retrieval results? A possible disadvantage of smooth features is that they hardly capture short note events. Since such events only span a short period, they may not be represented well in smooth features. Therefore, the CTC model's fine-grained features may perform better for themes with short note events. To examine this hypothesis, we relate the theme's note durations to the retrieval results. We compute the median note duration (using the manual alignments) for each theme and show the resulting distribution in Fig. 7a. Most of the themes have a median note duration between 0.1 and 0.6 seconds.

Both feature representations $\mathcal{C}_{\mathrm{CTC}}$ and $\mathcal{C}_{\mathrm{BG1}}$ perform well in the retrieval application for many themes. For 1623 of the 2067 themes, the relevant document is ranked at the top (rank 1) in both approaches. We now consider only the remaining 444 themes yielding a non-relevant top match in at least one of the approaches. These themes constitute the more difficult part of our dataset. Only using this part, we again show the distributions of median note durations in Figs. 7b and c. The distribution shape corresponding to the 444 themes is similar to the distribution shape corresponding to the total dataset (Fig. 7a). The histogram's colors and hatches indicate the rank that a theme yielded in the two strategies $\mathcal{C}_{\mathrm{CTC}}$ (Fig. 7b) and $\mathcal{C}_{\mathrm{BG1}}$ (Fig. 7c). We see that most themes with a short median note duration of

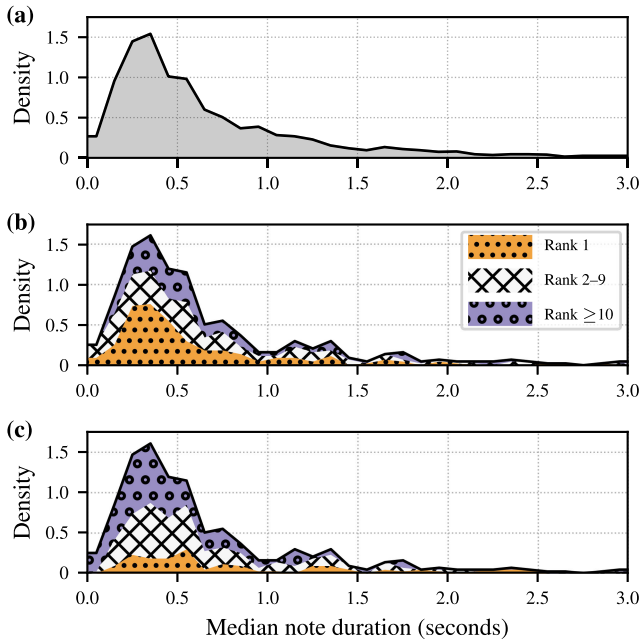


Fig. 7. Distribution of median note durations for **(a)** all 2067 themes, **(b)** the set of difficult themes (color and hatches based on $\mathcal{C}_{\mathrm{CTC}}$), and **(b)** the set of difficult themes (color and hatches based on $\mathcal{C}_{\mathrm{BG1}}$).

below 0.5 have better ranks in the CTC-based approach than for $\mathcal{C}_{\mathrm{BG1}}$. But, we have no substantial differences between the procedures for themes with a median note duration of above 0.5. Apparently, the fine-grained CTC-based features better represent themes with short note durations.

VII. Musical Evaluation

After having analyzed the effect of the CTC loss and the features' properties, we now come back to our music retrieval application. A main challenge of the retrieval scenario is the fact that the queries are monophonic, and the audio recordings are polyphonic. In this section, we review categories of musical texture that help us to specify this monophony–polyphony discrepancy. Then, we analyze our retrieval results in terms of musical texture.

A. Musical Texture

We categorize the themes according to their musical texture, using the standard texture categories of monophony, homophony, and polyphony. We expect that more complex musical textures go along with decreased retrieval evaluation measures.

Closely following the MTD article [24] and the textbook by Benward and Saker [53], we review our used musical terminology, ordered by increasing complexity. A monophonic texture consists of a single melodic line (possibly doubled by octaves). We already showed a monophonic example in Fig. 3, which has no monophony–polyphony discrepancy between the query and the corresponding audio occurrence for this theme. As a consequence, the retrieval was successful (i.e., correct top match) for all chroma representations considered.

A homophonic texture is made up of a melody and an accompaniment. Themes with a similar rhythm in all voices are

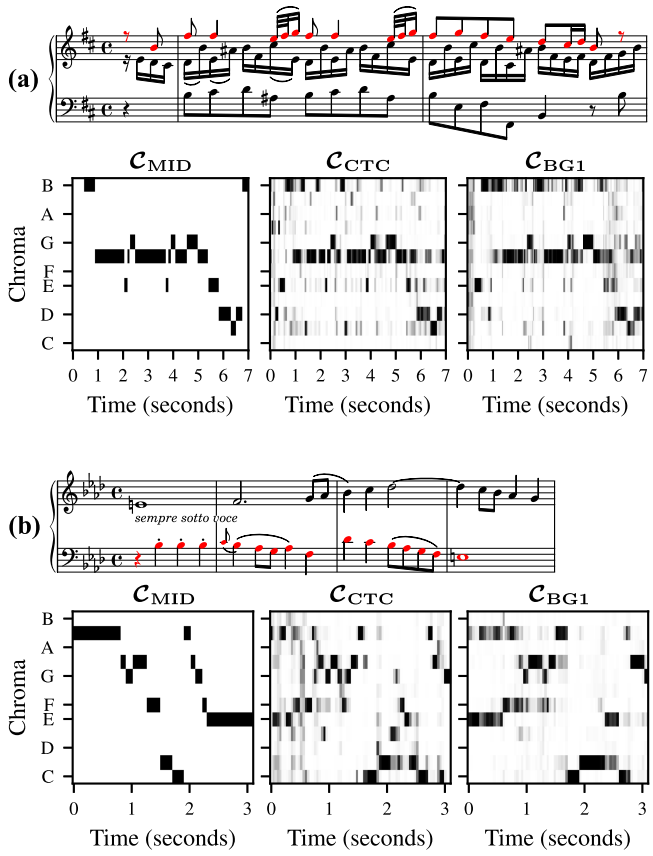


Fig. 8. Sheet music of piano reduction and chroma feature representations for two polyphonic music examples. The theme's notes are colored in red. In case that the audio occurrence has a different transposition than the sheet music, we cyclically shifted the feature representations to match the sheet music. **(a)** J. S. Bach: Flute Sonata in B minor, BWV 1030, beginning. **(b)** J. Haydn: Quartet in F minor, Hob. III:35, Finale, second fugue subject.

also included in this category. We already showed homophonic examples in Figs. 4 and 5 and discussed that our CTC strategy is able to attenuate the energy in the chroma bands corresponding to the accompaniment. As for the theme in Fig. 4, $\mathcal{C}_{\mathrm{CTC}}$ performed better in the retrieval-based evaluation (rank 1) than $\mathcal{C}_{\mathrm{BG1}}$ (rank 37). The theme of Fig. 5 yielded a top match in both approaches $\mathcal{C}_{\mathrm{CTC}}$ and $\mathcal{C}_{\mathrm{BG1}}$.

A polyphonic texture comprises two or more musically independent melodic lines. We show two complex polyphonic examples in Fig. 8, where the sheet music displays the theme's notes colored in red and all other voices in black. Along with the score, we show a symbolic representation ($\mathcal{C}_{\mathrm{MID}}$), which is manually aligned to the respective audio occurrence and, therefore, temporally corresponds to the audio features. Furthermore, we show the audio feature representations $\mathcal{C}_{\mathrm{CTC}}$ and $\mathcal{C}_{\mathrm{BG1}}$. The first polyphonic example (Fig. 8 a) comes from a flute sonata by Bach. Though some energy from the other voices is present in the features (especially in the B band), the theme is well preserved in the feature representations. The theme yielded a relevant top match in both approaches $\mathcal{C}_{\mathrm{CTC}}$ and $\mathcal{C}_{\mathrm{BG1}}$.

The second polyphonic example (Fig. 8 b) is a theme from a string quartet fugue by Haydn. This is a complex case because two fugue subjects are overlapping. In this example, we consider the second subject played by the cello while the second violin presents the first subject (the second violin starts two measures before the beginning of this example). The feature representations contain both subjects to a certain degree, but the first subject is more strongly represented. This theme yielded the ranks of 42 and 6 in the approaches $\mathcal{C}_{\mathrm{CTC}}$ and $\mathcal{C}_{\mathrm{BG1}}$, respectively.

For more musical examples and a discussion on the texture categories, we refer to the corresponding dataset article [24]. Furthermore, on our accompanying website, we provide visualizations of every theme in all feature representations, which gives a comprehensive overview of all queries used in the retrieval-based evaluation.

B. Analysis of Retrieval Results

Table VI shows the evaluation results according to the categories monophony (M), homophony (H), and polyphony (P). For convenience, we also add results for the total dataset (T), which was already shown in Table III. If different textures appear in a particular theme (e.g., starting monophonic and getting polyphonic later), we assign a single category to this query, where we always use the most complex texture that occurs for the theme.

We start our discussion with the first row that reports on results for the first fold. This fold contains 40 monophonic, 418 homophonic, and 101 polyphonic themes. In total (T), 559 themes are in this fold. For 37 out of 40 monophonic themes, the relevant document was on the first rank (top-1: 0.93). 367 of the 418 homophonic queries had a relevant top match (top-1: 0.88). For the 101 polyphonic themes, this is the case for 85 themes (top-1: 0.84). For all 559 themes of this fold, the top-1 measure is 0.87. The evaluation measures negatively correlate with the complexity of the musical texture of the themes. This trend also holds for the average results ($\varnothing$), where the monophonic themes have a higher top-1 measure (0.93) than the homophonic themes (0.88) and the polyphonic themes (0.84). We can observe the same correlation for evaluation measures that take more than only the top rank into account. For example, the MRR is 0.94, 0.90, and 0.87 for the monophonic, homophonic, and polyphonic themes, respectively.

Similar trends can also be observed for $\mathcal{C}_{\mathrm{BG1}}$ (last row). For example, the top-1 rates are 0.95, 0.83, and 0.79 for the monophonic, homophonic, and polyphonic themes, respectively. The CTC approach does not improves over $\mathcal{C}_{\mathrm{BG1}}$ for monophonic themes (top-1: 0.93 and 0.95), but for the homophonic (top-1: 0.88 and 0.83) and polyphonic (top-1: 0.84 and 0.79) themes.

At first sight, it may be surprising that the difference between homophonic and polyphonic texture does not lead to greater differences in retrieval results. Though the difference in both categories is essential from a musical point of view, it may not be as important from a signal-processing perspective. In both texture categories, the theme appears with additional voices that make an audio occurrence dissimilar to its corresponding monophonic query. It may only have a limited impact on our retrieval results if these voices constitute an accompaniment (homophonic theme) or musically independent melodies (polyphonic theme). It is more critical if the additional voices (independent or not) contribute a lot of energy to the theme occurrence in the audio recording.

TABLE VI
EVALUATION ON DATASET PARTS WITH DIFFERENT MUSICAL CHARACTERISTICS: MONOPHONY (M), HOMOPHONY (H), POLYPHONY (P), AND TOTAL (T)

		Queries				Top-1				Top-10				MRR			
	Fold	M	H	P	T	M	H	P	T	M	H	P	T	M	H	P	T
$\mathcal{C}_{\mathrm{CTC}}$	1	40	418	101	559	0.93	0.88	0.84	0.87	0.97	0.97	0.90	0.96	0.94	0.91	0.87	0.90
	2	11	308	58	377	0.82	0.84	0.74	0.83	0.91	0.91	0.86	0.90	0.87	0.87	0.79	0.86
	3	42	162	173	377	0.98	0.87	0.82	0.86	1.00	0.93	0.92	0.93	0.98	0.89	0.85	0.88
	4	16	236	125	377	0.88	0.91	0.88	0.90	0.94	0.95	0.97	0.96	0.89	0.92	0.92	0.92
	5	6	269	102	377	1.00	0.88	0.85	0.87	1.00	0.94	0.96	0.95	1.00	0.90	0.89	0.90
$\mathcal{C}_{\mathrm{CTC}}$	⌀	115	1393	559	2067	0.93	0.88	0.84	0.87	0.97	0.94	0.93	0.94	0.94	0.90	0.87	0.89
$\mathcal{C}_{\mathrm{BG1}}$	⌀	115	1393	559	2067	0.95	0.83	0.79	0.82	0.97	0.91	0.89	0.91	0.96	0.86	0.83	0.86

TABLE VII
RETRIEVAL RESULTS (TOP-1) FOR THE ORCHSET

	$\lambda = 3\,\mathrm{s}$	$\lambda = 6\,\mathrm{s}$	$\lambda = 9\,\mathrm{s}$
$\mathcal{C}_{\mathrm{Bit}}$	0.785	0.892	0.929
$\mathcal{C}_{\mathrm{BG1}}$	0.819	0.924	0.950
$\mathcal{C}_{\mathrm{CTC}}$	0.845	0.935	0.951

VIII. DATASET GENERALIZATION

To verify if our learned features generalize beyond the MTD dataset, which was used for training our CTC-based model (using cross-validation with a "composer-split," as described in Section IV-B), we report in this section on an additional retrieval experiment with a separate dataset. We use the Orchset, which is a publicly available dataset of orchestral music recordings with main melody annotations [54]. The dataset contains 64 audio excerpts having durations between 9.5 to 32.6 seconds and a total duration of about 23.5 minutes. As pointed out in the MTD article [24], the musical concepts of themes and main melodies are not identical but closely related. Since both concepts refer to monophonic salient elements in an excerpt of music, we can use the main melody annotations from the Orchset as queries as we did with the themes in our experiments with the MTD.[5]

To generate queries from the Orchset, we sample segments of length $\lambda \in \mathbb{R}$ with a hop size of 1 s from the main melody annotations. Then, we use the melody segment as a query, utilize a certain chroma variant for the audio recordings, apply our SDTW-based retrieval procedure to rank the 64 audio recordings of the dataset, and evaluate the results as described in Section III-B. Note that the smaller dataset (64 recordings of less than half an hour instead of 1126 recordings of more than 120 hours) constitutes a less challenging retrieval task. In the following, we only consider the top-1 recall rate, which is the "strictest" of our evaluation measures.

The resulting top-1 rates are presented in Table VII for different chroma variants and query durations λ. The second column shows results for a query duration of $\lambda = 3$ seconds (leading to 1242 queries). Here, the baseline representations $\mathcal{C}_{\mathrm{Bit}}$ and $\mathcal{C}_{\mathrm{BG1}}$ yield top-1 rates of 0.785 and 0.819, respectively. To obtain an evaluation measure for our CTC-based approach, we perform five iterations of the retrieval procedure using the five different versions of our CTC-based model (trained with the five different folds of the MTD, as described in Section IV-B). The resulting top-1 rates cover a range between 0.820 and 0.858. The average for the five rates, presented in Table VII, is 0.845.

With increasing query duration λ, the retrieval results improve for all approaches. For example, using a query duration of $\lambda = 6$ seconds (leading to 1050 queries), the corresponding top-1 rates are 0.892 for $\mathcal{C}_{\mathrm{Bit}}$, 0.924 for $\mathcal{C}_{\mathrm{BG1}}$, and 0.935 for $\mathcal{C}_{\mathrm{CTC}}$. For $\lambda = 9$ seconds (leading to 858 queries), the evaluation measures still increase while the differences in the results between the approaches become smaller. It is not surprising that increased query durations λ go along with improved retrieval results because longer queries are more characteristic, thus making the retrieval of the relevant recordings easier.

Overall, the retrieval experiments with the Orchset reveal similar tendencies as our experiments with the MTD. Among the baseline approaches, $\mathcal{C}_{\mathrm{BG1}}$ yields better results than $\mathcal{C}_{\mathrm{Bit}}$. The evaluation measures for $\mathcal{C}_{\mathrm{CTC}}$ are still higher than those for the baselines, which confirms that our approach is applicable to new datasets and new musical scenarios (e.g., main-melody-based instead of theme-based retrieval).

IX. CONCLUSION

In our paper, we showed how to apply the CTC loss to train deep chroma models with weakly aligned training data. In our theme retrieval scenario, we improved the state of the art by using features learned by such a model. These improvements were obtained with a standard neural network architecture, which may be further optimized in future work.

Previous work on speech recognition [21] and lyrics alignment [40] already realized the goal of using weak rather than strong annotations for training by using CTC. As a main contribution of our study, we conducted explicit experiments for comparing the CTC strategy with standard classification approaches (where we use weakly and strongly aligned training data) in our theme-based music application. Our results verified that the CTC strategy is superior to standard DNN training procedures for weakly aligned training data. Furthermore, our experiments showed that the CTC results are only slightly worse than the results obtained by training approaches using strongly aligned data, which can be considered the ideal case in our scenario. This finding is of major importance because it is highly labor-intensive to create strong alignments. A CTC-based strategy allows for saving a lot of annotation work and only leads to a slight drop in retrieval quality. This potential of CTC is also relevant for other MIR tasks, where annotations are hardly available, such as melody estimation [55] or chord recognition [56].

A primary challenge of our theme-based retrieval task is the difference in musical texture between the monophonic

[5] Although the recordings are entirely different, around two-thirds of the Orchset's audio excerpts correspond to musical works that are also represented in the MTD. However, these excerpts usually correspond to different musical passages within the pieces compared to the MTD's theme occurrences.

queries and audio recordings of polyphonic music. Through various examples and quantitative measures, we showed that our task-specific chroma features implicitly reduce the degree of polyphony of the audio content, which makes them well-suited for our theme-based retrieval scenario and other related applications (e.g., main-melody-based retrieval). Furthermore, our analyses revealed that the CTC-based approach avoids a temporal over-smoothing of the chroma representations and captures short note events that may be characteristic elements for certain musical themes.

We make our contributions reproducible and accessible in three different ways.[6] First, we provide an interactive web interface that shows detailed retrieval results for each theme and includes visualizations of the various feature representations. A tabular view allows ordering the themes according to the corresponding retrieval ranks, making it easy to find well-behaved and problematic queries. Second, we make pre-trained models and code to apply them available, which allows computing the CTC-based chroma features for arbitrary audio files. Third, our training data is publicly accessible, including an overview website with score visualizations and sonifications [24].

We think that the findings of this paper are relevant beyond our retrieval scenario. The trend towards using ever-increasing annotated datasets is considered critical among researchers. Data efficiency is still an important topic in the age of deep learning [57]. Some MIR researchers critically consider that certain large annotated datasets are not equally accessible by industry and academia [58]. A step towards solving this problem could be to use weakly annotated datasets, which are much easier to obtain. Our work shows that procedures for weakly aligned annotations can achieve results nearly as good as approaches using strongly aligned annotations, encouraging further adaptations and developments of such procedures.

APPENDIX

In our retrieval pipeline, as described in Section III-B, we used SDTW to compare a query (symbolically encoded monophonic musical theme) with subsequences of each document (audio recordings of polyphonic music). Similar to SDTW, the CTC loss can also be used to align a query with a document, thus being an alternative in our retrieval pipeline. In this section, we describe a CTC-based retrieval approach and compare it with the SDTW-based strategy.

A. Matching Functions

SDTW allows for comparing a short query sequence with subsequences of a longer document. As a result of SDTW, we obtain a matching function (the uppermost row of the accumulated cost matrix), denoted by $\Delta_{\mathrm{SDTW}} : [1:N] \to \mathbb{R}$, where $N \in \mathbb{N}$ is the number of time steps (see [1] for more details). For normalization, we divide each cost value $\Delta_{\mathrm{SDTW}}(n)$ for $n \in [1:N]$ by the query length. Each value $\Delta_{\mathrm{SDTW}}(n)$ encodes the cost of a cost-minimizing alignment between the query and a variable-length subsequence of the document that ends with index n. In the matching function, local minima indicate a good match between the query and a subsequence of the document.

As an alternative to SDTW, we compute a CTC-based matching function Δ_{CTC} having the same interpretation as Δ_{SDTW}. CTC allows for globally comparing two sequences. To simulate subsequence matching, we use a sliding window approach, where we segment the document into overlapping subsequences of a fixed length (e.g., given by the duration of the annotated ground truth match). Then, we compute the CTC loss (see Equation 9 in Section III-D) between the query and all considered subsequences separately. The CTC loss $\Delta_{\mathrm{CTC}}(n)$ integrates the cost measures for all possible alignments between the query and the considered document subsequence.

Note that SDTW is much more efficient because the subsequence matching is done implicitly in the dynamic programming algorithm, without a need for explicitly segmenting the document into subsequences (as we do for CTC).

B. Feature Representations

In our comparison, both approaches use a feature representation computed by our neural network trained with the CTC loss. For computing the SDTW-based matching functions, we post-process the network's output (removal of ϵ-values and ℓ^2-normalization), which results in 12-dimensional chroma features (see Section III-E). For the CTC-based matching functions, we directly use the network's output (13-dimensional probability vectors).

C. Experiments

In Fig. 9, we show SDTW- and CTC-based matching functions for three queries and their respective relevant audio documents. The red dotted lines indicate the annotated ground truth match. Note that our dataset only contains annotations for a single occurrence of each theme, even if the theme occurs several times. In our first example, we use the first theme of Beethoven's Fifth Symphony as a query and a recording of the entire first movement of this musical piece as a database document. Fig. 9a shows Δ_{SDTW}, and Fig. 9b shows Δ_{CTC}. The theme occurs various times in this movement, which is reflected by several low-cost values in the matching functions (e.g., around seconds 20, 100, and 460). For Δ_{SDTW} and Δ_{CTC}, the matching quality is associated with the values of the local minima relative to the matching function's overall level. In general, both functions show similar tendencies, especially for local minima. Similar tendencies can also be seen in the other examples (Figs. 9(c)–(f)).

We performed a CTC-based retrieval experiment with the MTD, using the same cross-validation strategy as before. In Table A1, we show the resulting evaluation measures for this experiment. Furthermore, we repeat the average SDTW-based results (from Table III) for convenience. Both strategies perform similarly, e.g., yielding top-1 rates of 0.867 (SDTW) and 0.871 (CTC), respectively.

TABLE A1
RETRIEVAL RESULTS (∅) FOR THE SDTW-BASED AND CTC-BASED RETRIEVAL APPROACH

Approach	Top-01	Top-05	Top-10	Top-20	Top-50	MRR
SDTW	0.867	0.927	0.942	0.955	0.969	0.894
CTC	0.871	0.933	0.950	0.963	0.977	0.900

[6]https://www.audiolabs-erlangen.de/resources/MIR/2021_TASLP-ctc-chroma

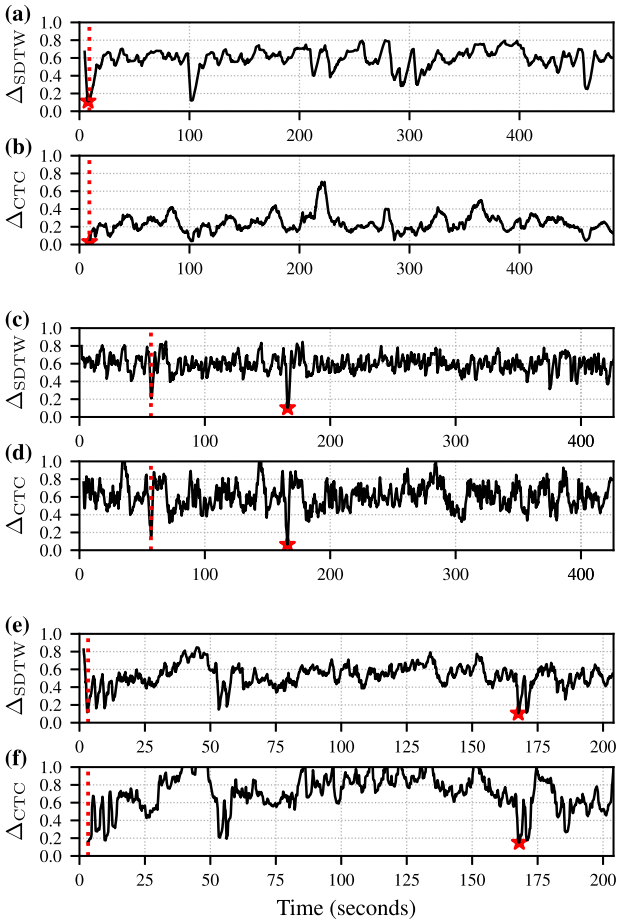


Fig. 9. SDTW-based and CTC-based matching functions for various examples. The dotted red vertical line denotes the ground truth match. The red star denotes the minimum of the function (estimated match). **(a)**–**(b)** First theme of Beethoven's Fifth Symphony Op. 67, first movement (see Fig. 3). **(c)**–**(d)** Second theme of Beethoven's Piano Sonata Op. 2, No. 2, first movement (see Fig. 4). **(e)**–**(f)** First theme of A. Dvořák's Slavonic Dance in F major Op. 72, No. 3 (see Fig. 5).

D. Summary and Discussion

As an alternative to SDTW, we used the CTC loss for our retrieval pipeline. The results of both approaches turned out to be similar. Because of the major benefits offered by SDTW (described below), we used this approach throughout our paper.

First, the CTC-based retrieval strategy can only be applied for the CTC-based representation and not for the baselines ($\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}_{\mathrm{Bit}}$). In contrast to that, we can compute SDTW-based matching functions for all representations considered. Second, we need to explicitly estimate the duration of the ground truth match for the CTC-based approach, providing the fixed length of the sliding window. SDTW is more flexible because it implicitly computes the best match between the query and a variable-length subsequence from the document. Third, SDTW is much more efficient than our "naive" sliding window approach employing the CTC loss (differing approximately by a factor in the order of the query length). In our implementations, the runtime for computing CTC-based matching curves increases by a factor of about 10 (for an average query length). To compensate for this shortcoming, one may devise a subsequence variant of CTC in the future.

ACKNOWLEDGMENT

We thank Christof Weiß for proof-reading the manuscript. Many people have been involved in the preparation of the dataset used in our experiments. In particular, we thank Stefan Balke, Vlora Arifi-Müller, 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).

REFERENCES

[1] M. Müller, *Fundamentals of Music Processing*. Berlin, Germany: Springer Verlag, 2015.
[2] J. Pickens *et al.*, "Polyphonic score retrieval using polyphonic audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2002. [Online]. Available: https://doi.org/10.5281/zenodo.1418091
[3] I. S. Suyoto, A. L. Uitdenbogerd, and F. Scholer, "Searching musical audio using symbolic queries," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 372–381, Feb. 2008.
[4] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2003, pp. 185–188.
[5] S. Balke, V. Arifi-Müller, L. Lamprecht, and M. Müller, "Retrieving audio recordings using musical themes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 281–285.
[6] F. Zalkow, S. Balke, and M. Müller, "Evaluating salience representations for cross-modal retrieval of western classical music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 331–335.
[7] E. Gómez, "Tonal description of music audio signals," PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.
[8] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.
[9] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 tracking in polyphonic music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 63–70.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.
[11] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 37–43.
[12] F. Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," in *Proc. 26th IEEE Int. Workshop Mach. Learn. Signal Process.*, 2016.
[13] F. Zalkow and M. Müller, "Learning low-dimensional embeddings of audio shingles for cross-version retrieval of classical music," *Appl. Sci.*, vol. 10, no. 1, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/7738895
[14] G. Doras and G. Peeters, "Cover detection using dominant melody embeddings," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 107–114.
[15] F. Yesiler, J. Serrà, and E. Gómez, "Accurate and scalable version identification using musically-motivated embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 21–25.
[16] Y. Wu and W. Li, "Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 2, pp. 355–366, Feb. 2019.
[17] S. Balke, C. Dittmar, J. Abeßer, and M. Müller, "Data-driven solo voice enhancement for jazz music retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 196–200.
[18] D. Basaran, S. Essid, and G. Peeters, "Main melody estimation with source-filter NMF and CRNN," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 82–89.
[19] E. J. Humphrey and J. P. Bello, "Four timely insights on automatic chord estimation," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2015, pp. 673–679.
[20] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 155–160.

[21] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[22] F. Zalkow and M. Müller, "Using weakly aligned score-audio pairs to train deep chroma models for cross-modal music retrieval," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 184–191.
[23] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 128–137, Jan. 2019.
[24] F. Zalkow, S. Balke, V. Arifi-Müller, and M. Müller, "MTD: A multimodal dataset of musical themes for MIR research," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 180–192, 2020.
[25] M. Müller, A. Arzt, S. Balke, M. Dorfer, and G. Widmer, "Cross-modal music retrieval and applications: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 52–62, Jan. 2019.
[26] M. Dorfer, J. Hajič Jr., A. Arzt, H. Frostel, and G. Widmer, "Learning audio-sheet music correspondences for cross-modal retrieval and piece identification," *Trans. Int. Soc. Music Inf.*, vol. 1, no. 1, pp. 22–31, 2018.
[27] L. Prechelt and R. Typke, "An interface for melody input," *ACM Trans. Comput.- Hum. Interaction*, vol. 8, no. 2, pp. 133–149, 2001.
[28] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2012, pp. 403–408.
[29] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 188–194.
[30] L. Su, "Vocal melody extraction using patch-based CNN," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 371–375.
[31] H. Cuesta, B. McFee, and E. Gómez, "Multiple F0 estimation in vocal ensembles using convolutional neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 302–309.
[32] Y.-N. Hung and Y.-H. Yang, "Frame-level instrument recognition by timbre and pitch," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 135–142.
[33] H. F. Aarabi and G. Peeters, "Deep-rhythm for global tempo estimation in music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 636–643.
[34] Y. Wu, T. Carsault, and K. Yoshii, "Automatic chord estimation based on a frame-wise convolutional recurrent neural network with non-aligned annotations," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.
[35] J. Calvo-Zaragoza, J. J. Valero-Mas, and A. Pertusa, "End-to-end optical music recognition using neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2017, pp. 472–477.
[36] J. Calvo-Zaragoza and D. Rizo, "End-to-end neural optical music recognition of monophonic scores," *Appl. Sci.*, vol. 8, no. 4, 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/4/606
[37] D. Schneider, N. Korfhage, M. Mühling, P. Lüttig, and B. Freisleben, "Automatic transcription of organ tablature music notation with deep neural networks," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 4, no. 1, pp. 14–28, 2021.
[38] C. Wick and F. Puppe, "Experiments and detailed error-analysis of automatic square notation transcription of medieval music manuscripts using CNN/LSTM-networks and a neume dictionary," *J. New Music Res.*, vol. 50, no. 1, pp. 18–36, 2021.
[39] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "An end-to-end framework for audio-to-score music transcription on monophonic excerpts," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 34–41.
[40] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 181–185.
[41] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 496–500.
[42] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché-Buc, "Multilingual lyrics-to-audio alignment," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2020, pp. 512–519.
[43] Y. Hou, Q. Kong, and S. Li, "Audio tagging with connectionist temporal classification model using sequentially labelled data," in *Proc. Int. Conf. Commun., Signal Process., Syst.*, 2019, pp. 955–964.
[44] R. Nishikimi, E. Nakamura, S. Fukayama, M. Goto, and K. Yoshii, "Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 161–165.
[45] H. Barlow and S. Morgenstern, *A Dictionary of Musical Themes*, *revised edition third printing ed*. New York, NY, USA: Crown Publishers, Inc., 1975.
[46] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop Deep Learn. Audio, Speech, Lang. Process. (WDLASL).*, 2013.
[47] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, "Dying ReLU and initialization: Theory and numerical examples," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1671–1706, 2020.
[48] A. Hannun, "Transcribing real-valued sequences with deep neural network," Ph.D. dissertation, Stanford University, 2018.
[49] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[50] J. J. Bosch and E. Gómez, "Melody extraction based on a source-filter model using pitch contour selection," in *Proc. 13th Sound Music Comput. Conf.*, 2016, pp. 67–74.
[51] J. J. Bosch and E. Gómez, "Melody extraction for MIREX 2016," in *Music Inf. Retrieval Eval. eXchange (MIREX) System Abstr.*, 2016. [Online]. Available: https://www.music-ir.org/mirex/abstracts/2016/BG1.pdf
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[53] B. Benward and M. Saker, *Music in Theory and Practice*, *8th ed.* New York, NY, USA: McGraw Hill, 2009.
[54] J. J. Bosch, R. Marxer, and E. Gómez, "Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music," *J. New Music Res.*, vol. 45, no. 2, pp. 101–117, 2016.
[55] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.
[56] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 54–63.
[57] H. D. Hlynsson, A. N. Escalante-B., and L. Wiskott, "Measuring the data efficiency of deep learning methods," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2019, pp. 691–698.
[58] W. Chen *et al.*, "Data usage in MIR: History & future recommendations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2019, pp. 25–32.

Frank Zalkow studied Musicology (B.A.) and Music Informatics (M.A.) at the University of Music Karlsruhe, Germany. He received the Ph.D. degree from the Friedrich-Alexander-Universität Erlangen-Nürnberg, in 2021, for his thesis on *Learning Audio Representations for Cross-Version Retrieval of Western Classical Music*. He was with the Institute for Music and Acoustics in the ZKM Center for Art and Media Karlsruhe and also Max-Reger-Institute Karlsruhe. In 2015, he became a Research Fellow with the Institute of Musicology, Saarland University, Germany. In 2016, he joined the International Audio Laboratories Erlangen, to work on his doctorate under the supervision of Prof. Meinard Müller. His main research interests include content-based music retrieval, musical applications of machine learning and signal processing, and cross-connections between musicology and computer science.

Meinard Müller (Fellow, IEEE) received the Diploma degree in mathematics and the Ph.D. degree in computer science from the University of Bonn, Germany, in 1997 and 2001, respectively. In 2007, he finished his Habilitation in the field of multimedia retrieval. From 2007 to 2012, he was a Member of Saarland University, Saarbrücken, Germany, and the Max-Planck Institut für Informatik. Since 2012, he holds a professorship for Semantic Audio Processing with International Audio Laboratories Erlangen. He has coauthored more than 150 peer-reviewed scientific papers, wrote a monograph titled Information Retrieval for Music and Motion (Springer-Verlag, 2007) and also a text-book titled *Fundamentals of Music Processing* (Springer-Verlag, 2015, www.music-processing.de). His recent research interests include music processing, music information retrieval, audio signal processing, multimedia retrieval, and motion processing. He is currently a Member of the Senior Editorial Board of the IEEE Signal Processing Magazine and the President of the International Society for Music Information Retrieval. In 2020, he was elevated to IEEE Fellow for contributions to music signal processing. Recently, he released a comprehensive collection of educational Python notebooks designed for teaching and learning audio signal processing using music as an instructive application domain.