USING WEAKLY ALIGNED SCORE–AUDIO PAIRS TO TRAIN DEEP CHROMA MODELS FOR CROSS-MODAL MUSIC RETRIEVAL

Frank Zalkow, Meinard Müller
International Audio Laboratories Erlangen, Germany
{frank.zalkow,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Many music information retrieval tasks involve the comparison of a symbolic score representation with an audio recording. A typical strategy is to compare score–audio pairs based on a common mid-level representation, such as chroma features. Several recent studies demonstrated the effectiveness of deep learning models that learn task-specific mid-level representations from temporally aligned training pairs. However, in practice, there is often a lack of strongly aligned training data, in particular for real-world scenarios. In our study, we use weakly aligned score–audio pairs for training, where only the beginning and end of a score excerpt is annotated in an audio recording, without aligned correspondences in between. To exploit such weakly aligned data, we employ the Connectionist Temporal Classification (CTC) loss to train a deep learning model for computing an enhanced chroma representation. We then apply this model to a cross-modal retrieval task, where we aim at finding relevant audio recordings of Western classical music, given a short monophonic musical theme in symbolic notation as a query. We present systematic experiments that show the effectiveness of the CTC-based model for this theme-based retrieval task.

1. INTRODUCTION

Music appears in many different modalities, for example, as audio or video recordings, in the form of symbolic representations, or as graphical sheet music [1]. In particular, audio recordings and symbolic representations are of great importance in many music information retrieval (MIR) tasks. An example is cross-modal retrieval, where a symbolic score is given as a query, and the task is to identify relevant audio recordings [2–4]. A general strategy for matching such different modalities is to use a common mid-level representation. In music processing, chroma features are widely used as mid-level [1, 5, 6]. These features, which capture the energy in the twelve chromatic pitch class bands, are robust against changes in octave, instrumentation, and timbre.

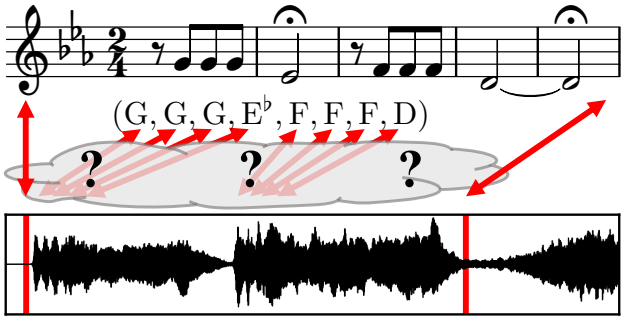

Figure 1. Illustration of a weakly aligned score–audio pair.

In recent years, many studies have shown the benefits of deep learning models to compute task-specific mid-level representations [7–10]. These learned features have proven their effectiveness in many scenarios, for example, audio–audio retrieval [11–13], chord recognition [9, 10, 14], or pitch tracking [7, 15, 16]. Training deep neural networks (DNNs) usually requires aligned training pairs, i.e., in MIR, music recordings with temporally aligned annotations. For example, the training pairs for a deep salience model by Bittner et al. [7] consist of time–frequency representations (more details in Section 2.2) with fundamental frequency annotations, where inputs and annotations correspond to each other for all time frames. For popular music, annotated data sets [17] have led to significant advances in research on pitch salience representations. However, creating such strongly aligned training pairs is labor-intensive, and, for many music scenarios, such data is hardly available. In contrast to the difficulty in annotating local alignments, it may be much easier to annotate global correspondences. In this paper, we use training pairs, where only global correspondences have been annotated. We denote these pairs as weakly aligned.

In our contribution, we use a deep learning model to compute enhanced chroma features, which we then use as a mid-level representation for a cross-modal retrieval task. Given a symbolic representation of a monophonic musical theme as a query and an audio database of Western classical music, the task is to find all audio recordings in which the theme is played [18, 19]. To obtain a task-specific chroma variant, we train a deep learning model with weakly aligned score–audio pairs, where only the beginning and end of a musical theme is annotated in an audio recording. Figure 1 illustrates such a pair for the famous first theme of Beethoven's Symphony No. 5. As our

Themes			Audio Recordings		
#	Mean Dur.	Total Dur.	#	Mean Dur.	Total Dur.
2048	00:00:09	04:54:58	1114	00:06:26	119:28:27

Table 1. Dat set overview. Duration format: hh:mm:ss.

main contribution, we combine a deep salience model [7] with a training procedure for weakly aligned data. [1] This procedure, called *Connectionist Temporal Classification* (CTC) [20], allows us to use training pairs of audio excerpts (in the form of spectral features) as input and musical themes (as sequences of chroma labels) as output. Using this CTC-based strategy, we train a model to compute enhanced chroma features for musical themes. We evaluate these features using more than 2000 themes and 1000 audio recordings and show that they improve the state of the art for our cross-modal retrieval scenario.

In Section 2, we review several prerequisites, such as cross-modal retrieval (Section 2.1), deep salience and deep chroma models (Section 2.2), and the CTC loss (Section 2.3). Then, in Section 3, we describe our adaption of the deep salience model, which computes chroma features and can be trained with the CTC loss. We present our experiments in Section 4 and conclude with Section 5.

2. PRIOR WORK AND PREREQUISITES

2.1 Cross-Modal Retrieval

For our retrieval scenario, we use a data set based on "A Dictionary of Musical Themes" by Barlow and Morgenstern (BM) [21], which contains roughly 10000 musical themes of instrumental Western classical music. Most of these themes have also been available as symbolic versions (MIDI) on the internet. [2] For a subset of the themes, we annotated their occurrences in audio recordings. In these annotations, a theme corresponds to exactly one recording, which, in turn, can correspond to several themes. The annotations comprise global correspondences, i.e., the beginning and end of the occurrences, as well as transpositions. Table 1 shows some statistics for our data set, which consists of 2048 themes from the BM book and 1114 corresponding recordings. The BM book already inspired several MIR studies [22, 23]. Some of them [18, 19] used the same subset for retrieval. We slightly corrected some annotations for this paper. A previous study [18] pointed out the challenges of the task, which are due to the differences in modality (symbolic vs. audio), tuning, transposition, tempo, and polyphony between the query and the recordings. The last point means that the themes are monophonic, but they usually appear in polyphonic context in the recordings (further discussion in Section 5). Previous work [19] has shown that pitch salience representations are capable of overcoming the differences in polyphony. In this paper, building upon these findings, we introduce an approach for learning a task-specific salience representation.

[1] Pre-trained models and code to apply them are available at `https://www.audiolabs-erlangen.de/resources/MIR/2020-ISMIR-ctc-chroma`.

[2] Unfortunately, the page is now offline. It is still reachable with the Wayback Machine without access to the MIDI files: `https://web.archive.org/web/20160209045946/http://www.multimedialibrary.com/barlow/index.asp`

2.2 Deep Salience and Deep Chroma Models

Many studies have demonstrated the effectiveness of using deep learning models to compute task-specific feature representations. 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 or multi-pitch tracking [7, 15, 16]. 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 [9, 10, 14].

This paper is inspired by the deep salience approach by Bittner et al. [7]. They 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 relationships. Many studies use this deep salience representation as a baseline [16,24] or build upon this model for diverse tasks such as dominant melody estimation [8], instrument recognition [25], tempo estimation [26], or chord recognition [27]. In Section 3, we describe how we adapt the deep salience model for computing enhanced chroma features.

The study of Wu et al. [27] 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 cross-modal retrieval. Unlike us, they take a three-step approach: First, they use a pre-trained deep chroma extractor to compute features. Second, they strongly align their annotations to the chroma features using a hidden Markov model. 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.

2.3 CTC Loss

Graves et al. [20] originally introduced *Connectionist Temporal Classification* (CTC) as the task of labeling unsegmented feature sequences with recurrent DNNs in the context of speech recognition. However, their training technique can be used with any DNN architecture. Furthermore, the task can be generalized to any scenario, where the aim is to map feature sequences to sequences of symbols. If the training data consisted of strongly aligned pairs of feature and symbol sequences (i.e., each vector of the feature sequences is labeled with a symbol), then a standard classification approach could be taken. The key aspect of CTC is that there is no need for strongly aligned training data, i.e., the feature and symbol sequences may be of

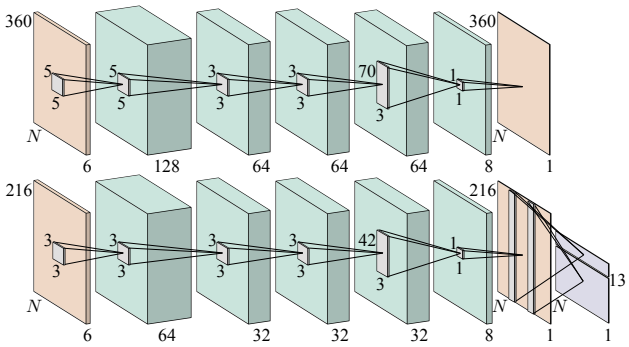

Figure 2. Network architectures. **Upper:** Original architecture proposed by Bittner et al. [7]. **Lower**: Adapted architecture used in this paper. Illustration inspired by [7].

different length, and the temporal correspondence between both sequences is unknown and may be non-linear.

Several studies from the MIR community used CTC, e.g., for optical music recognition [28], monophonic audio-to-score transcription [29], lyrics alignment [30], and audio tagging [31]. An alternative to CTC for sequence learning without aligned training data is the usage of an attention mechanism, which, e.g., was used for monophonic singing voice transcription [32].

In the following, we give the main idea of the CTC loss function introduced by Graves et al. [20] We describe the computation of the CTC loss for a single pair consisting of an audio feature sequence and a symbol sequence. Let

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \tag{1}$$

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

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M) \tag{2}$$

of length $M \in \mathbb{N}$, which consists of elements $\boldsymbol{y}_m \in \mathbb{A}$ for $m \in [1:M]$. The alphabet $\mathbb{A}$ of size $A \coloneqq |\mathbb{A}|$ is the set of symbols that can occur in the symbol sequence. Typically $M \ll N$. For example, in the case of lyrics alignment, the alphabet is the set of all considered characters [30]. In our case, it is the set of the twelve different chroma labels. The feature sequence $\boldsymbol{X}$ is transformed by a DNN f_θ with parameters θ to a sequence of probability vectors

$$f_\theta(\boldsymbol{X}) = \boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N) \tag{3}$$

having the same length N as the feature sequence and consisting of probability vectors $\boldsymbol{p}_n \in [0,1]^A$. We interpret the probability vector element $p_{n,a}$ for $a \in [1:A]$ as the probability that the n^{th} feature vector $\boldsymbol{x}_n$ corresponds to the a^{th} symbol in $\mathbb{A}$ (assuming an order of the set).

We can now compute the probability of the symbol sequence $\boldsymbol{Y}$ given the feature sequence $\boldsymbol{X}$. For a fixed alignment between $\boldsymbol{X}$ and $\boldsymbol{Y}$, one multiplies all values of the probability sequence $\boldsymbol{P}$ that correspond to that alignment. Since the alignment is unknown, instead of a specific one, all possible alignments between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are taken into account. Let us denote this overall probability as $\hat{p} \in \mathbb{R}$. Graves et al. [20] described how to compute $\hat{p}$ in a differentiable and efficient way using dynamic programming similar to the forward algorithm for hidden Markov models [33]. The final CTC loss for a single training pair is

$$L_\theta(\boldsymbol{X}, \boldsymbol{Y}) = -\log \hat{p}. \tag{4}$$

This loss function is used in batch gradient descent to update the parameters θ by averaging the loss value over multiple training pairs in a batch. By this procedure, the parameters of the network improve to produce probability sequences that make the ground-truth symbol sequences more probable.

In our explanation, we left out a crucial detail of the procedure. For a fixed alignment between $\boldsymbol{X}$ and $\boldsymbol{Y}$, the aligned symbol sequence can be represented by an "unfolded" sequence of length N that contains the active symbol for each time step. Let us consider the case of an unfolded sequence having multiple neighboring time steps with the same active symbol. So far, we cannot tell if this means one symbol occurrence with a long duration or multiple successive occurrences of the same symbol with shorter durations. To solve this ambiguity, an additional symbol named blank ϵ is part of the alphabet $\mathbb{A}$. This symbol serves two purposes: First, it means that no symbol is active. Second, it indicates a repeated occurrence of the same symbol if a succession of the same active symbol in the unfolded sequence is only interrupted by ϵ.

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	18464
Conv2D 32 × (3, 3, 32)	(N, 216, 32)	LReLU	9248
Conv2D 32 × (3, 3, 32)	(N, 216, 32)	LReLU	9248
Conv2D 8 × (42, 3, 32)	(N, 216, 8)	LReLU	32264
Conv2D 1 × (1, 1, 8)	(N, 216, 1)	Sigmoid	9
Pooling	(N, 13)	Softmax	217

Table 2. Details of the used DNN model (72970 parameters in total).

3. DEEP SALIENCE MODEL ADAPTATION

The DNN model used in this paper is inspired by the deep salience model proposed by Bittner et al. [7]. In this section, we explain our adaption of the model.

Bittner et al. [7] approached the task of melody and multi-pitch tracking, using a strongly aligned data set of 10 hours. In our case, we aim to learn an enhanced chroma representation for cross-modal retrieval, employing our weakly aligned 5-hour data set of 2048 themes. We simplified the original model in several ways to reduce the number of parameters and memory requirements. Additionally, we adapted the network so that it can be trained with the CTC loss and used as a deep chroma extractor. Figure 2 illustrates the original network architecture and our adapted version, and Table 2 gives further details for our version. Compared to the model by Bittner et al. [7],

we introduce the following modifications: First, we use a frame rate of 25 Hz instead of 86 Hz. Second, we use a frequency resolution of a third semitone instead of a fifth semitone. This resolution results in 216 instead of 360 frequency bins. Third, 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. Forth, we use leaky ReLU activations instead of ReLU activations to avoid zero gradients [34]. Fifth, we do not use batch normalization at all, 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 dynamic changes. 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 of size $N \times 216$. In our case, we aim for an output size of $N \times 13$, where the N columns are probability vectors over the set of the twelve chroma labels and an additional ϵ symbol:

$$\mathbb{A} := \{\mathrm{C}, \mathrm{C}^{\#}, \mathrm{D}, \ldots, \mathrm{B}\} \cup \{\epsilon\}. \tag{5}$$

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 ϵ 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 repeat this process for all columns of the input.

In summary, our adapted model differs from the original model [7] in two important aspects: First, we reduced the number of parameters from 407 thousand to 73 thousand. Second, the output of the model is a probability matrix over the set $\mathbb{A}$ instead of a pitch salience representation.

We train this adapted model with the CTC loss, as described in Section 2.3. The input to the network is an HCQT tensor computed for an excerpt from an audio recording, where a musical theme is played. Figure 3a shows a slice of the HCQT features for a recording of the first theme of Beethoven's Fifth Symphony. The corresponding symbol sequence is the sequence of chroma labels of the theme with neither any rhythmic information nor any temporal alignment to the input. For our Beethoven example (see also Figure 1), this sequence is $\boldsymbol{Y} = (\mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{E}^{\flat}, \mathrm{F}, \mathrm{F}, \mathrm{F}, \mathrm{D})$. Figure 3b visualizes the probability sequence for the Beethoven example after training. We see that the ϵ symbol has the largest probability for 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 chroma features. Finally, we ℓ^2-normalize the 12-dimensional chroma vectors to increase the energies in the time segments, where ϵ was dominating. Figure 3c shows the normalized chroma features, which correspond well with the symbol sequence.

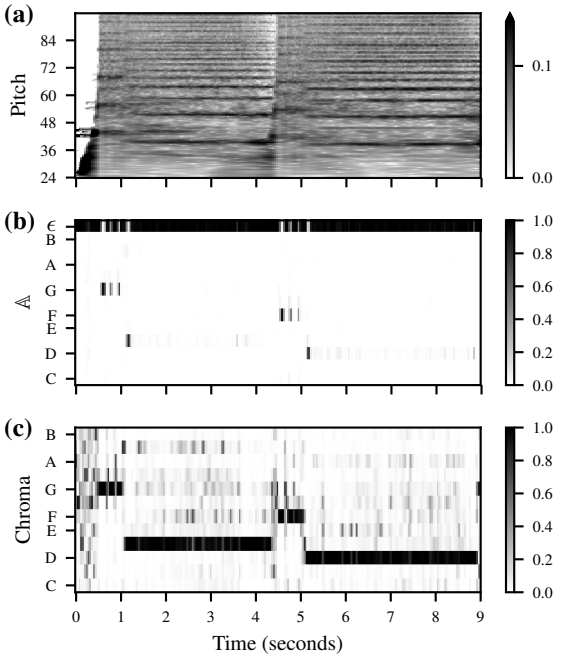

Figure 3. Representations for the first theme of Beethoven's Fifth Symphony. **(a)** HCQT input representation $\boldsymbol{X}$ (slice corresponding to the first harmonic). **(b)** Network output $\boldsymbol{P}$. **(c)** Features used for matching.

4. EXPERIMENTS

4.1 Training Details

We split our data set 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 consequence, we do not use themes from the same composer for training and evaluation, thus avoiding a "composer overfitting." For the training folds, we perform transpositions (up to a minor third upwards and downwards) as data augmentation. We perform batch gradient descent with a batch size of eight using the Adam optimizer [35] 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. When we finished training, we use the model with the lowest validation loss as a chroma feature extractor, and evaluate its effectiveness in the retrieval scenario, using the query themes from the test fold.

4.2 Retrieval-Based Evaluation

We shortly describe our retrieval pipeline and our evaluation measures following [18, 19]. First, we have a set $\mathbb{Q}$ of symbolic (MIDI) encodings of musical themes, which serve as *queries*. Furthermore, we have a collection of audio recordings, which we denote as database *documents*. These are actual recordings, not synthesized MIDI files. For each query, there is exactly one audio document that

(a)

	Top-01	Top-05	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-01	Top-05	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{BG1}}$	0.820	0.892	0.910	0.925	0.952	0.854
$\mathcal{C}_{\mathrm{Bit}}$	0.763	0.844	0.867	0.895	0.931	0.802

Table 3. Retrieval results of the baseline methods **(a)** using a feature rate of 10 Hz as reported in previous work [19], **(b)** using a feature rate of 25 Hz.

contains a globally corresponding rendition of the query theme (i.e., matching duration and transposition). For a fixed symbolic query, 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 a salience representation (from our CTC or a baseline approach). Then, we use Subsequence Dynamic Time Warping (SDTW) to compare the query with subsequences of the document [1]. In particular, we use the cosine distance, the step size condition $\Sigma := \{(2,1),(1,2),(1,1)\}$, as well as the weights $w_{\mathrm{vertical}} = 2$ and $w_{\mathrm{horizontal}} = w_{\mathrm{diagonal}} = 1$. 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 solve the retrieval task, we compute distances between all documents and the query. We then order the documents according to ascending distance values. The document's position in this ordered list is called the *rank* $r \in \mathbb{N}$ of the document. The top-K evaluation metric yields a value of one if the relevant document is among the top K matches, i.e., $r \leq K$. We then average this metric across all queries. Furthermore, we report the *mean reciprocal rank* (MRR), which is the average of $1/r$ across all queries.

In the cross-validation iterations of our evaluation, we only use the query themes from the respective test fold to search within the 1114 documents of our database. The reported average evaluation measures ($\varnothing$) are weighted with the number of queries from the respective test fold.

4.3 Baseline

As for our baselines, we consider the best-performing representations from a previous study [19], namely $\mathcal{C}_{\mathrm{Bit}}$, using the original deep salience model for melody estimation by Bittner et al. [7] [3], and $\mathcal{C}_{\mathrm{BG1}}$, using a model-based salience representation by Bosch and Gómez [36]. The latter one is a combination of a source-filter model with harmonic summation, using threshold parameters (named "BG1") that are particularly suited for orchestral music [37].

Table 3a cites the results from the previous study [19], where a 10 Hz feature rate was used. Since we use an increased feature rate of 25 Hz in this paper, we reproduced the experiments with this rate. The results are shown in Table 3b. Just by changing the feature rate, we see a substantial improvement of the results. For example, for $\mathcal{C}_{\mathrm{Bit}}$, the top-1 rate increases from 0.693 to 0.763, which means that 7 % more themes achieved a rank of 1. Since we corrected some errors in the data set, an improvement of up to 2 % may be due to the revision, but the main improvements are due to the increased time resolution. 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.

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 0.820 and 0.763 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 [17]). 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.

[3] Original weights ("Melody 2"). $\mathcal{C}_{\mathrm{Bit}}$ was denoted by $\mathcal{C}_{\mathrm{CNN}}$ in [19].

	$\lvert\mathbb{Q}\rvert$	Top-01	Top-05	Top-10	Top-20	Top-50	MRR
1	559	0.891	0.946	0.961	0.971	0.977	0.918
2	373	0.823	0.887	0.917	0.938	0.954	0.855
3	372	0.839	0.911	0.933	0.944	0.954	0.872
4	372	0.903	0.949	0.952	0.962	0.976	0.922
5	372	0.855	0.919	0.935	0.949	0.976	0.885
$\varnothing$		0.865	0.925	0.941	0.955	0.968	0.893

Table 4. Retrieval results for $\mathcal{C}_{\mathrm{CTC}}$.

4.4 CTC-Based Results

We now discuss the results we achieved with our CTC-based approach $\mathcal{C}_{\mathrm{CTC}}$. Table 4 shows the evaluation results for the five cross-validation iterations. The second column ($|\mathbb{Q}|$) gives the number of query themes in the respective test fold. This number is larger in the first fold (559) because this fold contains all BM themes by Ludwig van Beethoven, which is the most prominent composer of our data set. The other folds have fewer queries (372 or 373) and are more diverse in terms of composers, having 12 or 13 different composers each. The retrieval results have some diversity, ranging from a top-1 rate of 0.823 for test fold 2 up to 0.903 for test fold 4. 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 3b). For example, the average top-1 rate is 0.865 for $\mathcal{C}_{\mathrm{CTC}}$, compared to 0.763 for $\mathcal{C}_{\mathrm{Bit}}$ and 0.820 for $\mathcal{C}_{\mathrm{BG1}}$. Improvements for larger ranks can also be seen, such as in the top-50 rate (0.968 compared to 0.931 and 0.952, respectively). The results show that our approach is able to outperform the baselines, which have been the state of the art for the task [19].

	Top-01	Top-05	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{CTC}}$	0.865	0.925	0.941	0.955	0.968	0.893
$\mathcal{C}_{\mathrm{CCE}}$	0.814	0.890	0.907	0.929	0.951	0.849

Table 5. Retrieval results ($\varnothing$) using cross-entropy.

	Top-01	Top-05	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{BG1}}$	0.820	0.892	0.910	0.925	0.952	0.854
$\mathcal{C}_{\mathrm{CTC}}$	0.865	0.925	0.941	0.955	0.968	0.893
Oracle	0.904	0.947	0.958	0.967	0.983	0.924

Table 6. Retrieval results ($\varnothing$) for an oracle of the baseline by Bosch and Gómez [36] and our CTC approach.

4.5 Importance of CTC-Alignment

To verify the need for the CTC procedure in our scenario, we performed an additional experiment, where we assumed a linear temporal alignment between the symbolic themes and the corresponding excerpts in the audio recordings. Here, we changed the training procedure from our CTC strategy to a standard classification approach, using categorical cross-entropy (CCE). As output labels, we used binary chroma representations that we obtained by linearly scaling the symbolic themes to the same length as the corresponding audio excerpts. The ϵ symbol here only indicates rests in a theme. Note that, in this experiment, we used the rhythm information and note durations from the MIDI files, which we did not use in the CTC approach.

The trained model was used as chroma extractor and then evaluated in the theme retrieval context. The first row of Table 5 repeats the average evaluation measures from Table 4 for convenience and the second row presents the average results for the CCE approach. The evaluation measures are lower compared to the CTC-based results, e.g., having a top-1 rate of 0.814 compared to 0.865. This difference is due to the non-linear temporal correspondence between audio recordings and the symbolic themes.

4.6 Oracle Experiment

The model-based approach $\mathcal{C}_{\mathrm{BG1}}$ also shows excellent performance for this task. To investigate the relationship between $\mathcal{C}_{\mathrm{BG1}}$ and the $\mathcal{C}_{\mathrm{CTC}}$, we evaluated both strategies with an oracle procedure. For each query, we took the better rank: either achieved with $\mathcal{C}_{\mathrm{BG1}}$ or $\mathcal{C}_{\mathrm{CTC}}$. Table 6 repeats the results for the baseline and CTC approaches 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 (0.904 instead of 0.865). For top-K rates with larger K, there are still some small improvements. The oracle indicates that for some queries, $\mathcal{C}_{\mathrm{BG1}}$ is a slightly better feature representation than $\mathcal{C}_{\mathrm{CTC}}$.

5. CONCLUSION

In this paper, we showed the potential of CTC [20] for training a deep salience model with weakly aligned data. Adapting a model by Bittner et al. [7] to compute a task-specific mid-level representation, we improved state-of-the-art results for a cross-modal retrieval task for musical

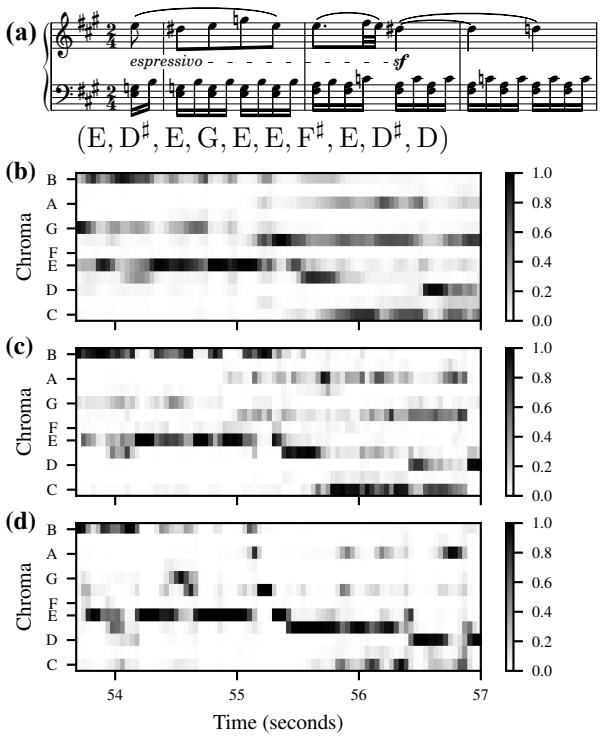

Figure 4. Second theme of Beethoven's Piano Sonata Op. 2, No. 2, first movement. **(a)** Full score with the chroma sequence of the theme, **(b)** standard chroma features using the full spectral content, **(c)** $\mathcal{C}_{\mathrm{BG1}}$, **(d)** $\mathcal{C}_{\mathrm{CTC}}$.

themes. To achieve these improvements, the feature computation procedure has to reduce the potential polyphony of the audio recording, which is a major challenge. We close our paper with a qualitative example to show the feature's properties for a representative polyphonic example.

Figure 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 first half and a diminished triad $(\mathrm{F}^\sharp, \mathrm{A}, \mathrm{C})$ in the second half. Ideally, for our retrieval scenario, we aim for a chroma representation that only captures energy from the theme and not from the accompaniment. Figures 4b, 4c, and 4d show chroma features for the full spectral content, the baseline salience approach $\mathcal{C}_{\mathrm{BG1}}$, and our CTC strategy $\mathcal{C}_{\mathrm{CTC}}$, respectively. In all representations, 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 most evident in $\mathcal{C}_{\mathrm{CTC}}$. 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 cross-modal retrieval.

In this study, we excluded the challenges due to differences in transposition. This could be taken into account by circularly shifting the chroma features [18], or by incorporating it into the learning procedure [38]. Furthermore, our oracle experiment suggests a possible next step of combining our strategy with traditional salience approaches [36].

Acknowledgments: 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).

6. REFERENCES

[1] M. Müller, *Fundamentals of Music Processing*. Springer Verlag, 2015.

[2] J. Pickens, J. P. Bello, G. Monti, T. Crawford, M. Dovey, M. B. Sandler, and D. Byrd, "Polyphonic score retrieval using polyphonic audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[3] I. S. Suyoto, A. L. Uitdenbogerd, and F. Scholer, "Searching musical audio using symbolic queries," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 372–381, 2008.

[4] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proceedings of the Workshop on Applications of Signal Processing (WASPAA)*, New Paltz, New York, USA, 2003, pp. 185–188.

[5] E. Gómez, "Tonal description of music audio signals," PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[6] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[7] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 tracking in polyphonic music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 63–70.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI*, Munich, Germany, 2015, pp. 234–241.

[9] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, 2016, pp. 37–43.

[10] ——, "A fully convolutional deep auditory model for musical chord recognition," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Salerno, Italy, 2016.

[11] F. Zalkow and M. Müller, "Learning low-dimensional embeddings of audio shingles for cross-version retrieval of classical music," *Applied Sciences*, vol. 10, no. 1, 2020.

[12] G. Doras and G. Peeters, "Cover detection using dominant melody embeddings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 107–114.

[13] F. Yesiler, J. Serrà, and E. Gómez, "Accurate and scalable version identification using musically-motivated embeddings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 21–25.

[14] Y. Wu and W. Li, "Automatic audio chord recognition with midi-trained deep feature and BLSTM-CRF sequence decoding model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 355–366, 2019.

[15] S. Balke, C. Dittmar, J. Abeßer, and M. Müller, "Data-driven solo voice enhancement for jazz music retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, 2017, pp. 196–200.

[16] D. Basaran, S. Essid, and G. Peeters, "Main melody estimation with source-filter NMF and CRNN," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 82–89.

[17] 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 *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014, pp. 155–160.

[18] S. Balke, V. Arifi-Müller, L. Lamprecht, and M. Müller, "Retrieving audio recordings using musical themes," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 281–285.

[19] F. Zalkow, S. Balke, and M. Müller, "Evaluating salience representations for cross-modal retrieval of western classical music recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 311–335.

[20] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA, 2006, pp. 369–376.

[21] H. Barlow and S. Morgenstern, *A Dictionary of Musical Themes*, revised edition third printing ed. Crown Publishers, Inc., 1975.

[22] L. Prechelt and R. Typke, "An interface for melody input," *ACM Transactions on Computer-Human Interaction*, vol. 8, no. 2, pp. 133–149, 2001.

[23] S. Balke, S. P. Achankunju, and M. Müller, "Matching musical themes based on noisy OCR and OMR input," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 703–707.

[24] L. Su, "Vocal melody extraction using patch-based cnn," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 371–375.

[25] Y.-N. Hung and Y.-H. Yang, "Frame-level instrument recognition by timbre and pitch," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 135–142.

[26] H. F. Aarabi and G. Peeters, "Deep-rhythm for global tempo estimation in music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 636–643.

[27] Y. Wu, T. Carsault, and K. Yoshii, "Automatic chord estimation based on a frame-wise convolutional recurrent neural network with non-aligned annotations," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[28] J. Calvo-Zaragoza, J. J. Valero-Mas, and A. Pertusa, "End-to-end optical music recognition using neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 472–477.

[29] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "An end-to-end framework for audio-to-score music transcription on monophonic excerpts," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 34–41.

[30] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 181–185.

[31] Y. Hou, Q. Kong, and S. Li, "Audio tagging with connectionist temporal classification model using sequentially labelled data," in *Proceedings of the International Conference in Communications, Signal Processing, and Systems (CSPS)*, Dalian, China, 2019, pp. 955–964.

[32] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 161–165.

[33] A. Hannun, "Transcribing real-valued sequences with deep neural network," Ph.D. dissertation, Stanford University, 2018.

[34] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning (ICML)*, Atlanta, Georgia, USA, 2013.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015.

[36] J. J. Bosch and E. Gómez, "Melody extraction based on a source-filter model using pitch contour selection," in *Proceedings of the 13th Sound and Music Computing Conference (SMC)*, Hamburg, Germany, 2016, pp. 67–74.

[37] ——, "Melody extraction for MIREX 2016," in *Music Information Retrieval Evaluation eXchange (MIREX) System Abstracts*, 2016.

[38] A. Arzt and S. Lattner, "Audio-to-score alignment using transposition-invariant features," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 592–599.