EVALUATING SALIENCE REPRESENTATIONS FOR CROSS-MODAL RETRIEVAL OF WESTERN CLASSICAL MUSIC RECORDINGS

Frank Zalkow, Stefan Balke, Meinard Müller

International Audio Laboratories Erlangen, Germany

{frank.zalkow,stefan.balke,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

In this paper, we consider a cross-modal retrieval scenario of Western classical music. Given a short monophonic musical theme in symbolic notation as query, the objective is to find relevant audio recordings in a database. A major challenge of this retrieval task is the possible difference in the degree of polyphony between the monophonic query and the music recordings. Previous studies for popular music addressed this issue by performing the cross-modal comparison based on predominant melodies extracted from the recordings. For Western classical music, however, this approach is problematic since the underlying assumption of a single predominant melody is often violated. Instead of extracting the melody explicitly, another strategy is to perform the cross-modal comparison directly on the basis of melody-enhanced salience representations. As the main contribution of this paper, we evaluate several conceptually different salience representations for our cross-modal retrieval scenario. Our extensive experimental results, which have been made available on a website, comprise more than 2000 musical themes and 100 hours of audio recordings.

Index Terms— Music Information Retrieval, Evaluation, Feature Representations

1. INTRODUCTION

Ongoing digitization efforts create large amounts of music data in different modalities, such as audio and video recordings, symbolic representations, or graphical sheet music. Accessing this data in a convenient way requires flexible retrieval strategies that are able to cope with the different modalities. In the last decades, many systems for audio retrieval based on the query-by-example paradigm have been suggested. Given a fragment of a symbolic or acoustic music representation as query, the task is to automatically retrieve documents from a music database containing parts or aspects that are similar to the query [1–4]. One such retrieval scenario is known as *query-by-humming* [5,6], where the user specifies a query by singing or humming a part of a melody. The objective is to identify all audio recordings (or other music representations) that contain a melody similar to the specified query. In related retrieval scenarios, a short symbolic query is given, e.g., taken from a musical score, and the task is to identify another symbolic music representation [7, 8] or an audio recording [9–13].

Many pieces from Western classical music contain short melodies or musical gestures that are especially prominent and memorable (e.g., the famous "Fate Motif" at the beginning of Beethoven's Symphony No. 5). Finding such *musical themes* in audio recordings with computational methods constitutes a challenging retrieval scenario. In particular, given a symbolic representation of a musical theme as query, the retrieval task is to find all recordings within a database of classical music recordings that contain this theme. Major challenges are due to the differences in modality (symbolic vs. audio), tuning, transposition, tempo, and degree of polyphony between the query and the database documents [12].

In this paper, we built upon the results presented in [12], where the database and the queries are compared on the basis of chroma features. We take this work as a baseline for our follow-up study. In particular, we address a major issue of the previous work, which is the compensation of the difference in degree of polyphony between the queries and the database documents. Instead of deriving chroma features from the full spectral content, we consider in this paper several kinds of enhanced time–frequency (TF) representations, so called *salience representations*, which put emphasis on certain tonal frequency components [14, Chapter 8] and enhance melodic structures in the spectrogram [15–19]. Previous studies [6] first extracted the predominant melody and bass line of the audio on the basis of salience representations. Then these representations are mapped to chroma features for performing the retrieval. However, in Western classical music, the underlying assumption of a single predominant melody is often violated, which has a negative impact on the robustness of melody extraction algorithms [20]. We propose not to extract the melodies, but to directly map the salience representations to chroma features for the retrieval.

The main contribution of this paper is to evaluate several state-of-the-art salience representations—originally designed for melody extraction—for the given retrieval scenario, and to conduct an extensive quantitative study exploring the potential of these representations. In Section 2, we describe the data set, used for the experiments. The feature representations, used throughout this study, are introduced in Section 3. In Section 4 we describe the retrieval procedure, report on results, and discuss effects of the representations on the retrieval results by means of a new evaluation metric, called *separation indicator*. The results of our experiments have been made publicly available on an interactive website.[1] Finally, Section 5 presents a short conclusion.

2. BARLOW-MORGENSTERN DATA SET

The data set considered in this paper is inspired by the book "A Dictionary of Musical Themes" by Harold Barlow and Sam Morgenstern (BM) [21], which contains roughly 10,000 musical themes of

We thank Lena Krauß and Lukas Lamprecht for their assistance in generating annotations. This work has been supported by the German Research Foundation (MU 2686/11-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

[1] https://www.audiolabs-erlangen.de/resources/MIR/2019-ICASSP-BarlowMorgenstern

Queries			Database			Composers
#	Mean Dur.	Total Dur.	#	Mean Dur.	Total Dur.	#
2045	00:00:09	05:00:03	1114	00:06:25	119:15:19	52

Table 1. Overview of the data set. Duration format: hh:mm:ss.

instrumental Western classical music. Published in the year 1949, this dictionary is an early example of indexing music by its prominent themes. It was designed as a reference book for trained musicians and professional performers to identify musical pieces by a short query fragment. Most of the 10,000 themes listed in the book have also been available as machine-readable versions (MIDI) on the internet.[2]

In our experiments, we use the data set *BM-Medium*, which was also used in previous work [12]. As shown in Table 1, the data set consists of 2045 themes from the BM book. For this paper, we substantially extended the data set by annotating the occurrences of these themes in an audio collection, including the durations and possible transpositions. We designed the audio database in such a way that for each query, there is exactly one relevant music recording in the database. Note that there can be more queries for a given audio recording: e.g., for the first movement of Beethoven's Symphony No. 5 there are six themes. The newly annotated audio material enables us to perform large-scale retrieval experiments in a controlled and systematic fashion, focusing on the monophonic–polyphonic matching problem. All queries of the data set can be accessed through the accompanying website.[1]

3. FEATURE REPRESENTATIONS

In this paper, we consider various TF representations that emphasize specific tonal frequency components. The considered approaches, which are well-known in the literature, are listed in Table 2 with links pointing to implementations. Details and properties of the TF representations are discussed in the next paragraphs. We convert these representations to time–chroma representations by suitably mapping the frequency bins to the twelve chromatic pitch classes, see [14, 22, 23]. In our study, we use a consistent frame rate of 10 Hz (applying median aggregation for representations with a higher frame rate). Furthermore, all frames of the chroma features are ℓ^2-normalized. Figure 1 visualizes the TF representations (left column) and their derived chroma features (right column) for a music example.

In the case of a MIDI query, the feature extraction is straightforward, see Figure 1a. While using a single feature representation for the MIDI query, we compare several chroma variants for the audio recordings. As a baseline, we use a TF representation $\mathcal{S}_{\mathrm{IIR}}$ similar to a spectrogram with a logarithmically-spaced frequency axis by using a bank of elliptic IIR filters [24, Chapter 3]. This representation was also used for the experiments in [12]. Obviously, this approach is influenced by the complete spectral content that is present in the audio, such as harmonics or noise-like signal components. For instance, the beginning of $\mathcal{C}_{\mathrm{IIR}}$ in Figure 1b has strong energy in the G-band, which corresponds to the fundamental frequency of the first note, but also in the D-band and in the B-band, which correspond to the third and fifth harmonics, respectively. A well-known approach that puts more emphasis on the predominant melody's fundamental frequency is harmonic summation, which is, e.g., used in MELODIA [16]. As shown in Figure 1c (first column),

[2]Unfortunately, the page has been put 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`

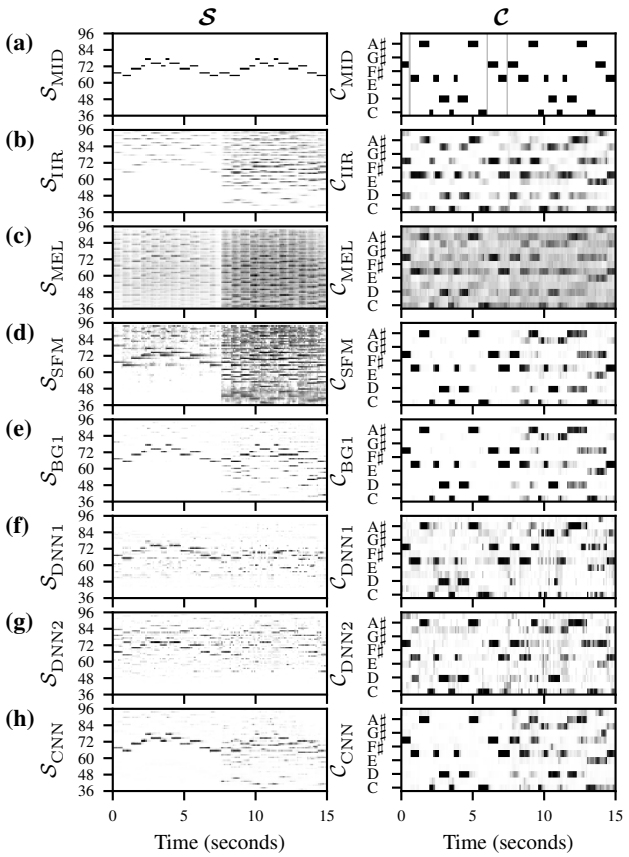

Fig. 1. Overview of different TF representations and their derived chroma features for the beginning of the first *Promenade* of Modest Mussorgsky's *Pictures at an Exhibition*. The first half of this example presents monophonic melody and the second half repeats it along with a homophonic brass section.

harmonic summation enhances the predominant melody in $\mathcal{S}_{\mathrm{MEL}}$, but also introduces additional noise. Note that this representation (as well as many of the other salience representations) was designed to serve as input to a subsequent melody extraction step. Further traditional signal processing approaches include a source-filter signal model $\mathcal{S}_{\mathrm{SFM}}$, introduced by Durrieu et al. [17], and a combination of MELODIA with Durrieu's source-filter model $\mathcal{S}_{\mathrm{BG1}}$, proposed by Bosch and Gómez [18]. For extracting $\mathcal{S}_{\mathrm{BG1}}$, we use the threshold parameters `BG1` that turned out to be specifically suited for orchestral music [25].

Recently, deep learning became ubiquitous for computing salience representations. For instance, Balke et al. [15] used a fully-connected neural network that was trained on jazz music for computing a salience representation specifically tailored to this kind of music, denoted as $\mathcal{S}_{\mathrm{DNN1}}$. Since deep learning is highly data-adaptive and the network was not trained on Western classical music at all, we re-trained the model, following the training procedure as described in [15], using the publicly available Orchset data set [28]. We applied pitch shifting as well as time scale modification to generate several versions of the data set in different tempi and keys. As a result of this augmentation process, the data used for training was increased from 23.5 minutes to 820.5 minutes. The resulting salience representation is denoted as $\mathcal{S}_{\mathrm{DNN2}}$. A more advanced deep learning approach for computing a salience representation $\mathcal{S}_{\mathrm{CNN}}$ was introduced by Bittner et al. [19]. They proposed a convolutional

	Ref.	Source for Implementation
$\mathcal{S}_{\mathrm{IIR}}$	[24, 26]	`http://github.com/librosa/librosa` [27]
$\mathcal{S}_{\mathrm{MEL}}$	[16]	`http://www.upf.edu/web/mtg/melodia`
$\mathcal{S}_{\mathrm{SFM}}$	[17]	`http://github.com/juanjobosch/`
$\mathcal{S}_{\mathrm{BG1}}$	[18]	`SourceFilterContoursMelody`
$\mathcal{S}_{\mathrm{DNN1}}$	[15]	`http://www.audiolabs-erlangen.de/resources/MIR/`
$\mathcal{S}_{\mathrm{DNN2}}$		`2017-ICASSP-SoloVoiceEnhancement`
$\mathcal{S}_{\mathrm{CNN}}$	[19]	`http://github.com/rabitt/ismir2017-deepsalience`

Table 2. Table of implementations, used for computing the salience representations.

neural network architecture that uses a custom feature representation as input, called *harmonic* CQT. The network was trained on classical, popular, as well as jazz music and outperformed state-of-the-art approaches in melody extraction.

Most salience representations were designed for a subsequent melody extraction step. We also perform this step on the three representations $\mathcal{S}_{\mathrm{MEL}}$, $\mathcal{S}_{\mathrm{BG1}}$ and $\mathcal{S}_{\mathrm{CNN}}$, with the respective methods proposed by the original literature [16, 18, 19]. We consider all frames as voiced, since this is the case for most queries.[3] The extracted melodies are mapped to chroma features, just like for the TF representations, resulting in $\mathcal{C}^{*}_{\mathrm{MEL}}$, $\mathcal{C}^{*}_{\mathrm{BG1}}$ and $\mathcal{C}^{*}_{\mathrm{CNN}}$. As we will see in our experiments in Section 4.2, the melody extraction step is not beneficial within our retrieval scenario.

4. EXPERIMENTS

In this section, we first summarize our retrieval procedure and describe our experiments. We then study how the different feature representations from Section 3 can cope with the difference in the degree of polyphony between monophonic symbolic theme (query) and polyphonic music recordings (database documents). Finally, we evaluate the approaches in depth.

4.1. Retrieval Procedure

We formalize our retrieval task following [12]. Similar procedures for synchronizing sheet music and audio recordings were described in the literature [9, 11, 14]. Let $\mathcal{Q}$ be a collection of musical themes, where each element $Q \in \mathcal{Q}$ is regarded as a *query*. Furthermore, let $\mathcal{D}$ be a set of audio recordings, which we regard as a database collection consisting of *documents* $D \in \mathcal{D}$. Given a query, the retrieval task is to identify the semantically corresponding documents. Note that in our experimental setting, there is only a single relevant document for each query. To compare a symbolic query $Q \in \mathcal{Q}$ to a database document $D \in \mathcal{D}$, we convert the query and the document into chroma sequences. Then, we use a standard technique known as *Subsequence Dynamic Time Warping* (SDTW) to compare the query with subsequences of the document, see [14, Chapter 7]. In particular, we use the cosine distance (for comparing ℓ^2-normalized chroma feature vectors), the step size condition $\Sigma := \{(2,1),(1,2),(1,1)\}$, as well as the weights $w_{\mathrm{vertical}} = 2$ and $w_{\mathrm{horizontal}} = w_{\mathrm{diagonal}} = 1$ in the SDTW.

As the result of SDTW, one obtains a matching function Δ^{Q}_{D} for a query Q and document D. Local minima of this function point to locations with a good match between the query Q and a subsequence of the document D. For a given query Q, the retrieval task can be solved by computing matching curves for all documents D and by taking the minimum $\delta^{Q}_{D} \in \mathbb{R}_{\geq 0}$ for each of the matching functions Δ^{Q}_{D}. The values of these minima yield a ranking of the database documents, which can then be presented in form of an ordered list. The position of a document $D \in \mathcal{D}$ in this list is called the *rank* of D. The rank of the relevant document is denoted as $r \in \mathbb{N}$.

Having a single relevant document for each query, the top-K evaluation metric gives the proportion of queries for which $r \leq K$ for a given $K \in \mathbb{N}$. Furthermore, we report the *mean reciprocal rank* (MRR), which is the average of $1/r$ over all queries.

[3] Additional experiments (not reported here) showed that automatic voicing estimation leads to a drastic drop in retrieval quality in all three cases.

	Top-01	Top-05	Top-10	Top-20	Top-50	MRR
$\mathcal{C}_{\mathrm{IIR}}$	0.470	0.593	0.648	0.699	0.792	0.531
$\mathcal{C}_{\mathrm{MEL}}$	0.231	0.363	0.430	0.500	0.599	0.299
$\mathcal{C}_{\mathrm{SFM}}$	0.742	0.818	0.839	0.863	0.894	0.779
$\mathcal{C}_{\mathrm{BG1}}$	0.754	0.835	0.861	0.885	0.913	0.792
$\mathcal{C}_{\mathrm{DNN1}}$	0.417	0.534	0.576	0.633	0.708	0.474
$\mathcal{C}_{\mathrm{DNN2}}$	0.552	0.661	0.701	0.748	0.800	0.605
$\mathcal{C}_{\mathrm{CNN}}$	0.693	0.788	0.823	0.853	0.896	0.739
$\mathcal{C}^{*}_{\mathrm{MEL}}$	0.421	0.522	0.574	0.630	0.714	0.474
$\mathcal{C}^{*}_{\mathrm{BG1}}$	0.734	0.816	0.843	0.867	0.899	0.774
$\mathcal{C}^{*}_{\mathrm{CNN}}$	0.680	0.773	0.802	0.837	0.881	0.724

Table 3. Retrieval results for *BM-Medium*.

4.2. Retrieval Results

In this study, we want to focus on the aspect of monophonic–polyphonic matching. In [12], it was shown that factors such as tuning, transposition, and query length have a major impact on the retrieval results. To reduce the effect of these factors, we modify each MIDI query such that its duration and key matches the corresponding audio excerpt in the database. We could reproduce the results reported in [12] for their smaller data set based on 177 queries (best top-1 rate 0.684) using the conventional chroma representation $\mathcal{C}_{\mathrm{IIR}}$. However, we could achieve a significant improvement reaching a top-1 rate of 0.876 by using the enhanced feature representation $\mathcal{C}_{\mathrm{BG1}}$. This means that about 19 % more queries achieve a top-1 rank compared to previous work for this subset due to the improved feature representation.

We now systematically analyze the impact of the various TF representations on the retrieval results by considering the much larger data set *BM-Medium*. The results are summarized in Table 3. The first row of the table shows the evaluation metrics for the baseline $\mathcal{C}_{\mathrm{IIR}}$. The top-1 rate (0.470) means that even for this simple approach nearly half of the themes achieve a rank of 1. About 70 % of the queries yield a rank of at least 20 (top-20: 0.699). Harmonic summation performs worse, yielding a top-1 rate of 0.231 for $\mathcal{C}_{\mathrm{MEL}}$. We want to note that it is not fair to directly compare this representation with the others, since harmonic summation amplifies many TF components that do not belong to the predominant melody. The source-filter model $\mathcal{C}_{\mathrm{SFM}}$ brings a major boost in performance with a top-1 rate of 0.742 and a top-20 rate of 0.863. $\mathcal{C}_{\mathrm{BG1}}$, which is a combination of harmonic summation and the source-filter model, further increases the top-1 rate to 0.754. About three quarter of the 2045 themes yield a rank of 1 and about 89 % achieve a rank of at least 20 (top-20: 0.885), which is a major improvement compared to the baseline approach. The fully-connected network $\mathcal{C}_{\mathrm{DNN1}}$, trained on jazz music, falls below the baseline for this task with a top-1 rate of 0.417. However, the version $\mathcal{C}_{\mathrm{DNN2}}$, which was re-trained on classical music, shows an increase of about 14 % in the top-1 rate (0.552) compared to the original version of the network. This reconfirms the obvious fact that such neural networks are highly data

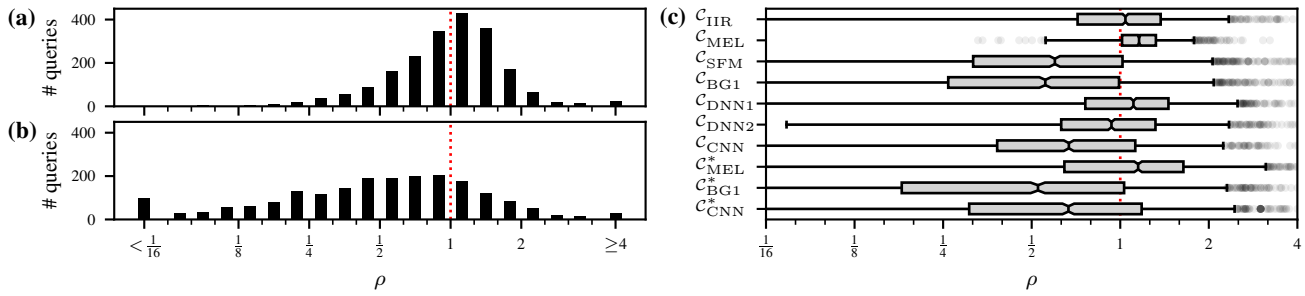

Fig. 2. **(a)** Histogram of ρ values for *BM-Medium* for $\mathcal{C}_{\mathrm{IIR}}$ and **(b)** $\mathcal{C}_{\mathrm{BG1}}$. **(c)** Boxplots of ρ values for all representations. Queries to the left of the red line ($\rho = 1$) yield a top-1 match.

dependent. Note that both networks have a simple architecture and do not exploit temporal context. The more advanced neural network approach $\mathcal{C}_{\mathrm{CNN}}$, which involves convolutional layers covering more audio context, performs better than the simple networks. $\mathcal{C}_{\mathrm{CNN}}$ achieves a top-1 rate of 0.693 and is close, but yet worse than the best model-based approaches $\mathcal{C}_{\mathrm{SFM}}$ (0.742) and $\mathcal{C}_{\mathrm{BG1}}$ (0.754).

In further experiments we also performed melody extraction as an intermediate step, which turned out to typically worsen the overall retrieval result for our Western classical music scenario. Only $\mathcal{C}_{\mathrm{MEL}}$ improved (increase of top-1 rate from 0.231 to 0.421 with $\mathcal{C}^*_{\mathrm{MEL}}$), but still this result is below the baseline. In case of the other representations the results slightly worsen, i.e. we observe a drop of top-1 rate from 0.754 ($\mathcal{C}_{\mathrm{BG1}}$) to 0.734 ($\mathcal{C}^*_{\mathrm{BG1}}$) or from 0.693 ($\mathcal{C}_{\mathrm{CNN}}$) to 0.680 ($\mathcal{C}^*_{\mathrm{CNN}}$).

As another contribution of this paper, we set up an extensive website presenting the ranks and sonifications for all 2045 queries, as well as visualizations of the corresponding feature representations.[1]

4.3. Discussion of Matching Quality

The rank-based evaluation metrics are rather coarse indicators of the matching quality. In the following, we want to examine it on a more fine-grained level. To this end, we introduce a novel evaluation metric called *separation indicator*. For a given query $Q \in \mathcal{Q}$, we obtain an ordered list of documents $(D_1, D_2, D_3, \dots)$. The position of the relevant document in this list is denoted as rank $r \in \mathbb{N}$. The matching quality is good when the matching cost of the relevant document $\delta^Q_{D_r}$ is significantly lower than the matching cost for the non-relevant document with the highest rank. The separation indicator $\rho \in \mathbb{R}_{\geq 0}$ quantifies the quality of the matching as follows:

$$\rho = \begin{cases} \delta^Q_{D_1} / \delta^Q_{D_2} & \text{if } r = 1, \\ \delta^Q_{D_r} / \delta^Q_{D_1} & \text{otherwise.} \end{cases} \tag{1}$$

Intuitively speaking, a low ρ below 1 implies a good matching quality since it indicates a distinct separation between the matching costs of the relevant document (with a rank of 1) and the first non-relevant document (with a rank of 2). The smaller ρ, the better the relevant document is separated from the non-relevant documents. A ρ close to 1 indicates that the top-1 decision is unstable for this query, and $\rho > 1$ implies that the query Q does not achieve rank 1. The distribution of ρ-values across all queries is a good indicator for the stability of top-1 decisions.

Figure 2a and b show histograms of ρ values for the baseline $\mathcal{C}_{\mathrm{IIR}}$ and the best performing representation $\mathcal{C}_{\mathrm{BG1}}$. For $\mathcal{C}_{\mathrm{IIR}}$, the ρ-values are centered around 1, indicating instability for the top-1 decisions. For $\mathcal{C}_{\mathrm{BG1}}$, the distribution is skewed to the left, which indicates that the increase in top-1 rate of $\mathcal{C}_{\mathrm{BG1}}$ is not due to small random-like changes of the matching costs, but to a substantial improvement in matching quality. Figure 2c shows boxplots of the distributions of ρ values for all considered feature representations. In this figure, the first and fourth row correspond to the histograms shown in Figure 2a and b, respectively. For $\mathcal{C}_{\mathrm{MEL}}$ (second row), the distribution is strongly centered just above 1.0, meaning that most queries do not achieve a rank of 1. The corresponding melody extraction $\mathcal{C}^*_{\mathrm{MEL}}$ (eighth row) has a strong effect on the distribution: Though the median is located to the right of the red line (i.e. less than half of the queries achieve rank 1), it shows that the queries cover a wider range of matching qualities. The best performing representations $\mathcal{C}_{\mathrm{SFM}}$, $\mathcal{C}_{\mathrm{BG1}}$, and $\mathcal{C}_{\mathrm{CNN}}$ (third, fourth, and seventh row) show similar tendencies: Most queries are located on the left side of the red line ($\rho < 1$) and they are covering a wide range of ρ values in that region including queries with high matching quality of $\rho < 1/2$. The comparison of $\mathcal{C}_{\mathrm{BG1}}$ and $\mathcal{C}^*_{\mathrm{BG1}}$ reveals that melody extraction improves the separation indicator for many queries. However, the top-K rates are worse for $\mathcal{C}^*_{\mathrm{BG1}}$. This may be explained as follows: For cases, where the music has a low degree of polyphony, the intermediate melody extraction step lowers the separation indicator significantly. But, in more complex cases the step does more harm than good.[4] As an effect, the overall retrieval result is better without melody extraction.

5. CONCLUSIONS

In this study, we considered a cross-modal retrieval scenario with the goal to find the relevant audio recording in a database, given a monophonic symbolic theme of Western classical music as query. Extending previous work [12], we showed that salience representations, originally designed for melody extraction, are suited for the given task. Furthermore, unlike related work [6], we showed that in our retrieval scenario it is beneficial to avoid an explicit melody extraction step and to perform the chroma reduction directly on the salience representations. In an extensive quantitative study, we compared various state-of-the-art salience representations and showed their benefits and limitations for the given task. Especially, the salience representation by Bosch and Gómez [18] turned out to be well-suited for monophonic–polyphonic matching. Additionally, we introduced a newly annotated data set consisting of more than 2000 queries and 100 hours of audio. The results of our experiments have been made available on an accompanying website.[1] Possible future research directions will deal with the design of specialized representations for further improving the retrieval results, including the adaptation of deep learning approaches.

[4] See `https://www.audiolabs-erlangen.de/resources/MIR/2019-ICASSP-BarlowMorgenstern/example` for two illustrative examples.

6. REFERENCES

[1] Michael A. Casey, Remco Veltkap, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[2] Peter Grosche, Meinard Müller, and Joan Serrà, "Audio content-based music retrieval," in *Multimodal Music Processing*, Meinard Müller, Masataka Goto, and Markus Schedl, Eds., vol. 3 of *Dagstuhl Follow-Ups*, pp. 157–174. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

[3] Colin Raffel and Daniel P. W. Ellis, "Large-scale content-based matching of MIDI and audio files," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 234–240.

[4] Rainer Typke, Frans Wiering, and Remco C. Veltkamp, "A survey of music information retrieval systems," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, London, UK, 2005, pp. 153–160.

[5] Matti Ryynänen and Anssi Klapuri, "Query by humming of MIDI and audio using locality sensitive hashing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008, pp. 2249–2252.

[6] Justin Salamon, Joan Serrà, and Emilia Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[7] Anna Lubiw and Luke Tanur, "Pattern matching in polyphonic music as a weighted geometric translation problem," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004, pp. 289–296.

[8] Kjell Lemström and Jorma Tarhio, "Searching monophonic patterns within polyphonic sources," in *Content-Based Multimedia Information Access – Volume 2*, Paris, France, 2000, RIAO '00, pp. 1261–1279.

[9] Christian Fremerey, Michael Clausen, Sebastian Ewert, and Meinard Müller, "Sheet music-audio identification," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 645–650.

[10] Jeremy Pickens, Juan Pablo Bello, Giuliano Monti, Tim Crawford, Matthew Dovey, Mark Sandler, and Don Byrd, "Polyphonic score retrieval using polyphonic audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[11] Iman S.H. Suyoto, Alexandra L. Uitdenbogerd, and Falk Scholer, "Searching musical audio using symbolic queries," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 372–381, 2008.

[12] Stefan Balke, Vlora Arifi-Müller, Lukas Lamprecht, and Meinard 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.

[13] Ning Hu, Roger B. Dannenberg, and George 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, Oct. 2003, pp. 185–188.

[14] Meinard Müller, *Fundamentals of Music Processing*, Springer Verlag, 2015.

[15] Stefan Balke, Christian Dittmar, Jakob Abeßer, and Meinard Müller, "Data-driven solo voice enhancement for jazz music retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 196–200.

[16] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[17] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[18] Juan J. Bosch and Emilia 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.

[19] Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan 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.

[20] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[21] Harold Barlow and Sam Morgenstern, *A Dictionary of Musical Themes*, Crown Publishers, Inc., revised edition edition, 1975.

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

[23] Emilia Gómez, *Tonal Description of Music Audio Signals*, PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[24] Meinard Müller, *Information Retrieval for Music and Motion*, Springer Verlag, 2007.

[25] Juan J. Bosch and Emilia Gómez, "Melody extraction for MIREX 2016," in *Music Information Retrieval Evaluation eXchange (MIREX) System Abstracts*. 2016.

[26] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Miami, Florida, USA, 2011, pp. 215–220.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, Austin, Texas, USA, 2015, pp. 18–25.

[28] Juan J. Bosch, Ricard Marxer, and Emilia Gómez, "Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music," *Journal of New Music Research*, vol. 45, no. 2, pp. 101–117, 2016.