TISMIR

Krause, M., et al. (2021). Towards Leitmotif Activity Detection in Opera Recordings. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 127–140. DOI: https://doi.org/10.5334/tismir.116

RESEARCH ARTICLE

Towards Leitmotif Activity Detection in Opera Recordings

Michael Krause, Meinard Müller and Christof Weiß

This paper approaches the automatic detection of musical patterns in audio recordings with a particular focus on leitmotifs, which are specific types of patterns associated with certain characters, places, items, or feelings occurring in an opera or movie soundtrack. The detection of such leitmotifs is particularly challenging since their appearance can change substantially over the course of a musical work. In our case study, we consider a self-contained yet comprehensive scenario comprising 16 recorded performances of Richard Wagner's four-opera cycle *Der Ring des Nibelungen*, which is a prime example for the use of leitmotifs. Within this scenario, we introduce and formalize the novel task of leitmotif activity detection. Based on a dataset of 200 hours of audio with over 50 000 annotated leitmotif instances, we explore the benefits and limitations of deep-learning techniques for detecting leitmotifs. To this end, we adapt two common deep-learning strategies based on recurrent and convolutional neural networks, respectively. To investigate the robustness of the trained systems, we test their sensitivity to different modifications of the input. We find that our deep-learning systems work well in general but capture confounding factors, such as pitch distributions in leitmotif regions, instead of characteristic musical properties, such as rhythm and melody. Thus, our in-depth analysis demonstrates some challenges that may arise from applying deep-learning approaches for detecting complex musical patterns in audio recordings.

Keywords: leitmotifs; opera; musical patterns; deep neural networks; sound event detection

1. Introduction

Within music information retrieval (MIR), detecting musical patterns in audio recordings is a fundamental task. These patterns can be characterized by any musical property, including rhythmic phrases, melodic shapes, or harmonic progressions. Across different occurrences, a pattern may vary considerably both in musical aspects and acoustic realization and may appear within different accompanying parts and other musical voices, thus being embedded in varying sound mixtures. In Western music tradition, such patterns play a crucial role for the narration, interpretation, and enrichment of dramatic plots in many genres—from Renaissance madrigals to movie soundtracks. In this context, composers have found creative ways of associating certain characters, places, items, or feelings with specific musical ideas, thus guiding their audience through the story. The use of such compositional techniques culminated in 19th century opera where these ideas became known as *leitmotifs* (Bribitzer-Stull, 2015), later adopted by movie soundtracks. A central role is attributed to Richard Wagner's operas with their extensive usage of leitmotifs. In his theoretical writings, Wagner intended these motifs to be particularly memorable and to guide the listeners through the work (Wagner, 1995). Knowing, rediscovering, and understanding the usage of leitmotifs may therefore enrich the experience of an audience (Baker and Müllensiefen, 2017) and help musicologists analyze the compositional structure of the works (Zalkow et al., 2017a). In this context, automated methods for detecting leitmotifs over the course of an opera (as illustrated by **Figure 1**) are of high interest for various applications such as the augmentation of recorded, virtual, and live performances and may serve commercial, didactic, and musicological research purposes. For instance, an automated leitmotif detection procedure may be used to display leitmotif names alongside a recorded performance of the work, thus enhancing the audience's experience of the composition.

In this paper, we study leitmotif detection in the context of Richard Wagner's four-opera cycle *Der Ring des Nibelungen*, for which a typical performance lasts about 15 hours. To the best of our knowledge, this is the first work dealing with automated leitmotif detection. We explore this task using a novel dataset of the *Ring* involving over 50000 annotated leitmotif instances. We design two typical

International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany

Corresponding author: Michael Krause (michael.krause@audiolabs-erlangen.de)

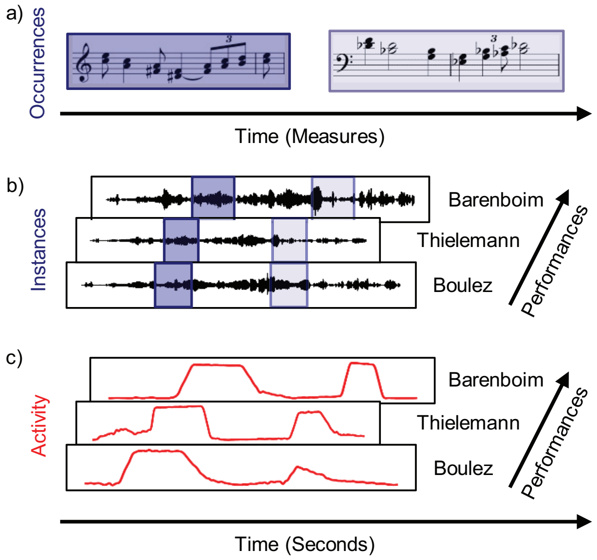


Figure 1: Illustration of a leitmotif (here the *Ring* motif `L-Ri`) and its manifestations as (a) leitmotif occurrences in the score, (b) leitmotif instances in several recorded performances (audio), (c) continuous leitmotif activity output by a detection system.

deep-learning systems for detecting the activity of several leitmotifs in recordings of the *Ring* and investigate their robustness under different modifications of the input, thus simulating different types of musical variability. We find evidence that despite good numerical results on a held-out test set, our models capture confounding factors rather than relying on characteristic musical properties. By analyzing our systems in this complex leitmotif scenario, we aim for a deeper understanding of their properties and explore some of the challenges that may arise from applying standard deep-learning systems for detecting musical patterns in audio recordings.

A leitmotif may be subject to several musical variations across its different *occurrences* in the musical score (see **Figure 1a**), such as transposition, tempo changes, abridgment, prolongation, as well as melodic, harmonic, or rhythmic changes. Due to this variety, systems generally need to be informed about the specific leitmotifs to detect. Possible application scenarios may have different degrees of such side information. In the main scenario considered in this paper, we have annotations of all *instances* of the relevant leitmotifs (see **Figure 1b**) for a specific recording. Based on this input, a system needs to detect the leitmotifs in other performances.

The remainder of the paper is organized as follows. In Section 2, we introduce the musical scenario of the *Ring*, outline our cross-performance dataset, and formalize the leitmotif activity detection task. In Section 3, we summarize related work, outline our deep-learning approaches and evaluation procedure, and present first results. In Section 4, we analyze our models with regard to different input modifications. Section 5 presents an outlook to less-informed scenarios. Section 6 summarizes our findings.

2 Musical Scenario and Task Specification

This section outlines our musical scenario consisting of Wagner's *Ring* cycle and its specific use of leitmotifs. We present an overview of our cross-performance dataset and provide a formalization of the leitmotif activity detection task.

2.1 Leitmotifs in Wagner's Ring

The scenario of our case study is centered around Richard Wagner's tetralogy *Der Ring des Nibelungen*, a musical work of extraordinary dimensions. As indicated by **Figure 2**, the *Ring* consists of the four operas *Das Rheingold*, *Die Walküre*, *Siegfried*, and *Götterdämmerung*, spanning a continuous plot. Comprising 21 941 measures, this large work has been considered for several tasks within MIR such as audio-based harmony analysis (Zalkow et al., 2017a), symbolic pattern search (Kornstädt, 2001) or meta-analyses of audience experience (Page et al., 2015). For organizing this comprehensive material, we consider eleven parts of the *Ring* (first row in **Figure 2**), which usually correspond to acts of individual operas (thus hereafter denoted as *acts*) with continuous measure count in the score.

The *Ring* cycle is well-known for its frequent use of leitmotifs—characteristic musical ideas associated with characters, places, items, or feelings. Most motifs are characterized by their melodic and rhythmic shape but are interwoven into the compositional structure. Therefore, a leitmotif may appear in different musical contexts, thereby varying in compositional aspects (such as melody, harmony, or rhythm) in order to fit the current key, meter, or tempo. Zalkow et al. (2017a) explored relationships between leitmotif usage and tonal characteristics of the *Ring*. Beyond that, leitmotifs may occur in different registers, voices, or instruments, and in abridged or extended versions with parts of the motif being repeated,

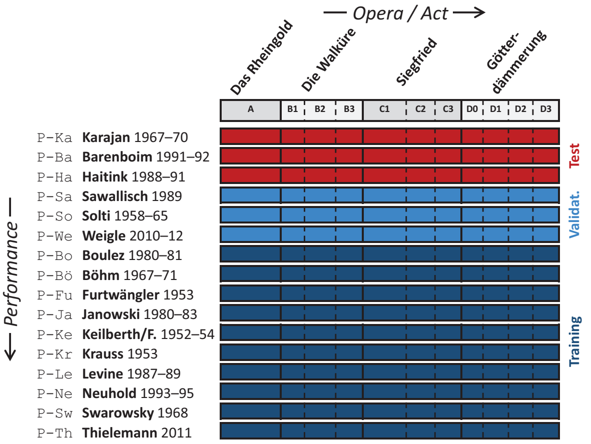


Figure 2: Structure of Richard Wagner's *Ring* cycle and overview of 16 recorded performances, see also Zalkow et al. (2017a). Measure positions have been annotated manually for the topmost three performances (`P-Ka`, `P-Ba`, and `P-Ha`), which also constitute the test set in our performance split. The three middle performances (`P-Sa`, `P-So`, and `P-We`) constitute the validation set. All other performances are used for training.

altered, or left out. Despite these musical variations, listeners can often identify motifs when listening to a performance. This is in line with Wagner's intention of using the motifs as a guideline and, thus, employing them in a clearly perceivable way (Wagner, 1995). This human ability to identify motifs has been analyzed from a psychological perspective (Baker and Müllensiefen, 2017; Morimoto et al., 2009; Albrecht and Frieler, 2014).

While Wagner mentioned the importance of such motifs for his compositional process (Wagner, 1995), there is no explicit specification of concrete leitmotifs by the composer. Whether a recurring musical idea constitutes a leitmotif or not is topic of debate among musicologists (Dreyfus and Rindfleisch, 2014). In line with our prior works (Zalkow et al., 2017a; Krause et al., 2020), we follow the specification of 130 leitmotifs in the *Ring* by Julius Burghold (Wagner, 2013). A musicologist annotated the score-based segments (in measures/beats) for all occurrences of these motifs in the *Ring*. Contiguous repetitions of motifs are considered as individual segments, and abridged, extended, or varied occurrences are also included (with our annotator deciding on the amount of variation that can be considered as the same motif). Since many leitmotifs occur rarely or are musically ambiguous, we pursue a pragmatic approach, restricting ourselves to 20 characteristic and frequent motifs, which are specified in **Table 1**. The motif L-Ho, for example, is associated with the hero Siegfried and is often used as a narrative device. It appears in its full heroic form when Siegfried is first introduced, changes to a diminished chord as the hero is fighting a great beast and is played again as other characters remember him following his demise. In total, our annotations comprise 3569 occurrences of these 20 motifs.

2.2 Cross-performance dataset

As a peculiarity of Western classical music, several recorded performances of a work are usually available, varying in interpretation aspects (tempo, dynamics, intonation), timbral aspects of instruments and singers, and production aspects (mastering, acoustic conditions). For certain music analysis tasks that are independent of such aspects, the availability of multiple performances allows for systematically studying the robustness of MIR systems in *cross-performance* (also called *cross-version*) experiments, as done by Schreiber et al. (2020) and Zalkow et al. (2017a).

In this paper, we make use of a cross-performance dataset of the *Ring*, comprising the 16 audio recordings (both live and studio) listed in **Figure 2**. Their duration varies between 13.5 and 15.5 hours. For the performances P-Ka, P-Ba, and P-Ha, the measure positions were manually annotated in the audio recordings (Weiß et al., 2016). For the remaining 13 performances, we made use of an automated transfer of measure positions from the manually annotated performances relying on highly accurate audio–audio synchronization methods (Zalkow et al., 2017b). As an indicator for this high accuracy, we analyzed measure positions obtained for one performance (P-Ba) using this transfer procedure and found that they deviate only marginally from the manually annotated measure positions (by 0.137 seconds on average).

Relying on these measure positions, we transferred the 3569 leitmotif occurrence regions from the score to the 16 recorded performances. For leitmotif boundaries not lying on measure boundaries, we used linear interpolation between measure positions. The resulting 57 104 *leitmotif instance regions* in the different recordings (see **Figure 1**) represent the reference annotations for our detection task. We provide our annotations of occurrence and instance positions as a publicly available dataset.[1]

In a previous study (Krause et al., 2020), we used these instances (for ten selected motifs) for evaluating a leitmotif *classification* task, where presegmentation of relevant audio excerpts (containing a leitmotif) is assumed to be given. In this paper, we aim for detecting the *activity* of the leitmotifs in a continuous fashion (**Figure 1c**) without assuming any presegmentation. In consequence, our detection problem is substantially harder than the classification problem studied in Krause et al. (2020). Moreover, we extend the task to 20 leitmotifs in total.

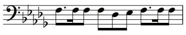
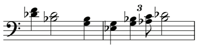
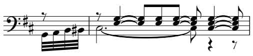
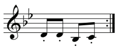
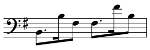
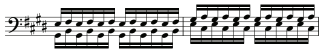
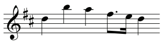
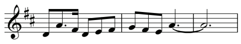
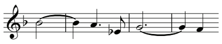
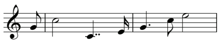
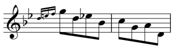
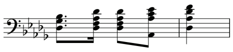
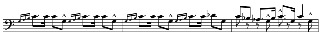
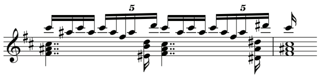
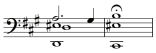
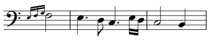
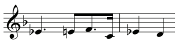
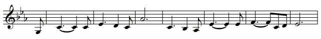
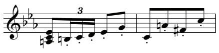
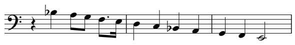
To systematically test the generalization capabilities of MIR systems, musical datasets can be split across different dimensions. For example, Schreiber et al. (2020) observed differences between systems for detecting local key when generalizing to unknown performances versus unknown songs. For most experiments in this paper, we make use of a *performance split* (see **Figure 2**), using the three recordings with manually annotated measure positions (P-Ka, P-Ba, P-Ha) for testing. The synchronization-based measure transfer may introduce small deviations for the other performances, which may be unproblematic for training but quite relevant for testing purposes. The validation set comprises the performances P-Sa, P-So, and P-We. The remaining ten performances are used for training. In Section 5, we report preliminary results for detecting leitmotifs in unknown musical material using an *opera split*.

2.3 Leitmotif activity detection

We now want to formalize the leitmotif activity detection task motivated in the introduction. To this end, we consider a set of leitmotifs $\mathcal{L}$ that is indexed by $\ell \in [1:L]: = \{1,2,..., L\}$ with $\mathrm{L} = |\mathcal{L}|$. In our dataset described in Section 2.2, we have $\mathcal{L} = \{$L-Ni, L-Ho,...$\}$ with $L = 20$, see **Table 1**. We further consider an audio recording with a discretized time axis given by the index set $[1:N]$. Due to variations in tempo, the time axis $[1:N]$ is performance-specific and the value of N varies between performances of the same act. Then, a leitmotif activity function φ_ℓ outputs probabilities for motif ℓ being active at each frame $n \in [1:N]$ of a specific performance, thus $\varphi_\ell: [1:N] \rightarrow [0,1]$.

In our dataset, we consider audio recordings from 16 performances of the eleven acts in the *Ring* (see **Figure 2**). As described in Section 2.2, the reference leitmotif annotations are given on a musical time axis specified in measures. For an act with S measures, we represent our reference annotations as a binary matrix $\mathcal{A}^{\mathrm{Ref}} \in \mathbb{B}^{L \times M}$, for $\mathbb{B} = \{0,1\}$ and $M = S \cdot B$ (see **Figure 3** for an illustration of an excerpt of such a matrix). Here, B is

Table 1: Overview of the 20 leitmotifs used in this study (the first ten of these motifs were previously used in Krause et al. (2020)). Score examples shown are adapted from Wagner (2013). Lengths are given as means and standard deviations over all annotated occurrences (in measures) or instances (in seconds) from all performances given in Figure 2. Counts and lengths differ from Krause et al. (2020), because we allow for concurrent motif activity in this study.

Name (English translation)	**ID**	**Score**	**# Occurrences**	**Length**	
				Measures	**Seconds**
Nibelungen (Nibelungs)	`L-Ni`		562	0.95 ± 0.24	1.72 ± 0.50
Ring (Ring)	`L-Ri`		297	1.50 ± 0.66	3.77 ± 2.46
Nibelungenhass (Nibelungs' hate)	`L-NH`		252	0.96 ± 0.17	3.22 ± 1.20
Mime (Mime)	`L-Mi`		243	0.83 ± 0.25	0.84 ± 0.20
Ritt (Ride)	`L-RT`		228	0.66 ± 0.17	1.26 ± 0.38
Waldweben (Forest murmurs)	`L-Wa`		228	1.10 ± 0.30	2.65 ± 0.73
Waberlohe (Swirling blaze)	`L-WL`		194	1.21 ± 0.39	4.59 ± 1.70
Horn (Horn)	`L-Ho`		195	1.30 ± 1.02	2.34 ± 1.51
Geschwisterliebe (Siblings' love)	`L-Ge`		158	1.32 ± 0.84	3.13 ± 2.65
Schwert (Sword)	`L-Sc`		148	1.88 ± 0.63	3.73 ± 1.99
Jugendkraft (Youthful vigor)	`L-Ju`		146	1.23 ± 0.57	0.96 ± 0.38
Walhall-b (Valhalla-b)	`L-WH`		143	1.10 ± 0.47	3.53 ± 2.14
Riesen (Giants)	`L-RS`		136	0.95 ± 0.39	2.83 ± 1.96
Feuerzauber (Magic fire)	`L-Fe`		112	1.18 ± 0.40	3.57 ± 1.09
Schicksal (Fate)	`L-SK`		94	2.02 ± 0.47	8.11 ± 2.64
Unmuth (Upset)	`L-Un`		92	1.87 ± 0.70	5.85 ± 3.21
Liebe (Love)	`L-Li`		89	1.78 ± 0.51	5.54 ± 2.47
Siegfried (Siegfried)	`L-Si`		86	2.88 ± 1.60	8.03 ± 5.46
Mannen (Men)	`L-Ma`		83	1.15 ± 0.50	1.37 ± 0.70
Vertrag (Contract)	`L-Ve`		83	2.29 ± 0.65	5.72 ± 2.12

a discretization factor. Setting $B = 1$, we evaluate on the level of whole measures. Setting $B = 16$, we subdivide each measure into 16 equidistant sub-segments and evaluate on sixteenth of a measure (e.g., in a 4/4 time signature, each sub-segment would correspond to a 16th note). M is then the total number of such measure sub-segments in the act and $m \in [1{:}M]$ are indices on our musical time axis. We set $B = 16$ for all our experiments. $\mathcal{A}^{\text{Ref}}$ can now be constructed from the annotations by assigning $\mathcal{A}^{\text{Ref}}_{\ell m} = 1$ if and only if an occurrence of motif ℓ covers measure sub-segment m.

In contrast to our reference annotations $\mathcal{A}^{\text{Ref}}$, which are defined on the musical time axis $[1 : M]$ of an act, we define our leitmotif activity functions $\varphi\ell$ on the physical time axis $[1 : N]$ of an audio recording. Therefore, to evaluate a leitmotif activity function, we first transfer its outputs onto a musical time axis by taking the maximum over all outputs for a measure sub-segment. Here, the

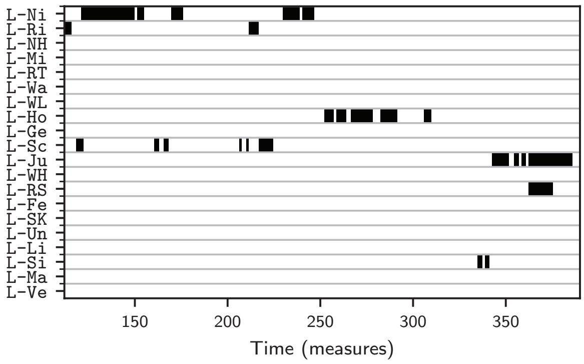


Figure 3: Illustration of our ground truth occurrence annotations. Measures 112 to 390 from the first act of *Siegfried* are shown. For instance, `L-Ni` is active around measure 150, whereas `L-SK` is never active throughout this excerpt.

correspondence between physical and musical time axes is given by our measure annotations refined with linear interpolation. Since φ_ℓ has a continuous output, we then use a thresholding procedure (described in Section 3.2) to also obtain a binary matrix $\mathcal{A}^{\mathrm{Est}} \in \mathbb{B}^{L \times M}$. This matrix can be evaluated against $\mathcal{A}^{\mathrm{Ref}}$ using standard measures such as precision, recall, and F-measure (see Section 3.3).

Evaluating detection results on a musical time axis has two advantages: first, it allows us to quantitatively compare results obtained on different performances (for which the physical time axes might differ, but the musical time axis does not). Second, by defining our evaluation metrics in terms of measure sub-segments, we are able to relate evaluation scores to musical material rather than physical duration (thus e.g. equally considering faster and slower sections) and to introduce a musically informed tolerance parameter in our evaluation (see Section 3.4).

Conceptually, our leitmotif activity detection task can be considered as a special case of polyphonic sound event detection as illustrated by Virtanen et al. (2018, Fig 8.1d). For example, the task of environmental sound detection consists of detecting the activity of multiple parallel sound sources within an environmental sound scene. Similarly, multiple different leitmotifs may be active at the same time. However, the activity functions of different environmental sounds are typically independent from each other, i.e., uncorrelated, and from any other sound in the mixture. As opposed to this, we can expect correlations between motif activities.[2] Furthermore, our leitmotifs are not independent of other musical parts (such as accompaniment or other motifs), since all musical parts have to fit into the larger harmonic context. These characteristics distinguish our task from other, more general sound event detection scenarios.

Concerning a coarsely related problem, the Music Information Retrieval Evaluation eXchange (MIREX)[3] has run a task on *Discovery of Repeated Themes and Sections*, but this was limited to synthesized audio and prominent themes with little variation. In contrast, we deal with real-world orchestral recordings and our leitmotifs may vary considerably or appear in the accompaniment.

3 Deep-learning-based Leitmotif Activity Detection

In this section, we present two approaches to leitmotif activity detection based on neural networks, introduce the evaluation measures used and report first results using our models. We start with a short discussion of related work on sound event detection.

3.1 Related work

Some years ago, traditional techniques such as non-negative matrix factorization dominated the field of sound event detection (Stowell et al., 2015). In recent years, deep neural networks have become the dominant approaches for such tasks. Network architectures that have been considered include feed-forward, recurrent or convolutional neural networks, as well as combinations of these (Çakir et al., 2017). More recent approaches make use of techniques such as dilated convolutions (Li et al., 2020). Novel systems are proposed frequently and evaluated for standard (non-musical) sound event detection scenarios at the yearly DCASE challenges.[4] We refer to a recent survey for a comprehensive overview of neural networks for sound event detection (Xia et al., 2019).

As for *musical* sound event detection, an example task considered in the literature is singing voice activity detection, where regions of singer activity constitute the sound events to be detected. This task has been approached through frame-wise classification using convolutional neural networks by Schlüter and Lehner (2018). Another task is beat tracking, where musical beats are considered as sound events. For this task, Böck et al. (2016) proposed a recurrent neural network that jointly detects beat and downbeat positions. Finally, one may also consider music transcription tasks as a variant of sound event detection. For instance, in drum transcription, individual drum hits are considered as sound events to be detected. For a comprehensive overview of recent drum transcription approaches, including ones using convolutional neural networks, we refer to Wu et al. (2018). For most of the mentioned tasks, sound events are usually short and only depend on very local context. In contrast, the leitmotif instances considered in our paper can last several seconds and a detection system must be able to process the appropriate amount of temporal context to identify them. Therefore, to implement a leitmotif activity detection function, an RNN-based approach can be considered appropriate. Such an architecture can, at least in theory, detect entities of arbitrary lengths (such as our leitmotif instances). As discussed above, however, convolutional architectures have been used more frequently in recent years. As a second approach, we therefore consider a CNN-based system, paying special attention to the appropriate receptive field in time.

3.2 Methods

We begin by extracting audio excerpts of ten seconds' length (containing leitmotif instances, but also excerpts where none of our motifs occur) from the ten training performances of the *Ring* described in Section 2.2. Here, ten second excerpts are long enough to completely cover the full leitmotif instance for nearly all instances in our dataset. For the 3569 leitmotif occurrence regions, we randomly add context before and after the instance in case the motif is shorter than ten seconds or randomly remove parts of the beginning and end of the instance in case it is longer. We further include 4000 examples where no motif occurs.

The audio excerpts are sampled at 22 050 Hz and converted to mono. Subsequently, we process the excerpts by a constant-Q transform (CQT) with twelve semitones per octave from C1 to B7 and a hop length of 512 samples, adjusted for tuning deviations (estimated automatically per performance and opera act). These steps are implemented using librosa.[5] We only take the magnitude of the CQT. The resulting CQT frames with a frame rate of 43.1 Hz are then max-normalized individually (in order to obtain normalized network input and achieve some degree of loudness invariance) and used as input to our networks. Both networks process CQT frames and output frame-wise predictions per leitmotif.

3.2.1 RNN-based approach

For our experiments, adapting the approach from Krause et al. (2020), we use the network architecture as specified in **Table 2**. The input consists of 431 CQT frames (obtained from a ten-second audio excerpt), every frame being a vector of *84* CQT bins (one for each semitone in seven octaves), resulting in the input shape (431,84). The input is processed by three stacked long short-term memory (LSTM) layers, which are variants of RNN layers designed to be easily trainable (Goodfellow et al., 2016). Each LSTM uses 128 units for its internal operations. The third LSTM layer is followed by batch normalization and a dense layer (applied at each frame individually), which outputs one prediction per motif as well as an additional output indicating no motif activity (leading to 21 outputs in total). These predictions (logits) are converted to probabilities through a standard sigmoid activation. Based on these frame-wise outputs, our network models leitmotif activity functions φ_ℓ for each motif $\ell \in \mathcal{L}$. Since this corresponds to a frame-wise multi-label classification problem, multiple outputs may be activated for the same frame (corresponding to simultaneous motif activity). Moreover, the procedure is causal, meaning that the output at any frame depends only on this frame and the preceding frames. We did not observe improvements for increasing the number of stacked LSTM layers, increasing their number of units, replacing them with gated recurrent unit (GRU) layers, or applying regularization such as weight decay or dropout.

Table 2: Network architecture used for our RNN-based leitmotif activity detection system (adapted from Krause et al. (2020)).

Layer	Output Shape	Parameters
Input	(431, 84)	
LSTM	(431, 128)	109 056
LSTM	(431, 128)	131 584
LSTM	(431, 128)	131 584
Batch normalization	(431, 128)	512
Dense (per frame)	(431, 21)	2 709
Output: Sigmoid	(431, 21)	

3.2.2 CNN-based approach

As our second network, we consider a convolutional architecture as illustrated in **Table 3**. The input of shape (431,84) is identical to the RNN input. The subsequent architecture follows the paradigm of stacking convolution

Table 3: Network architecture used for our CNN-based leitmotif activity detection system (inspired by Schlüter and Lehner (2018)). Note that all operations have stride one in time and pitch, except for MaxPool2D, which has stride three in the pitch direction. Dilation rates in time increase after each max-pooling operation.

Layer (Kernel size), (Strides), (Dilations)	Output Shape	Parameters
Input	(431, 84)	
Expand	(431, 84, 1)	
Conv2D (3, 3), (1, 1), (1, 1)	(431, 84, 128)	1 152
Batch normalization	(431, 84, 128)	512
Conv2D (3, 3), (1, 1), (1, 1)	(431, 84, 64)	73 728
Batch normalization	(431, 84, 64)	256
MaxPool2D (3, 3), (1, 3), (1, 1)	(431, 29, 64)	
Conv2D (3, 3), (1, 1), (3, 1)	(431, 29, 128)	73 728
Batch normalization	(431, 29, 128)	512
Conv2D (3, 3), (1, 1), (3, 1)	(431, 29, 64)	73 728
Batch normalization	(431, 29, 64)	256
MaxPool2D (3, 3), (1, 3), (3, 1)	(431, 10, 64)	
Conv2D (3, 3), (1, 1), (9, 1)	(431, 10, 128)	73 728
Batch normalization	(431, 10, 128)	512
Conv2D (3, 3), (1, 1), (9, 1)	(431, 10, 64)	73 728
Batch normalization	(431, 10, 64)	256
MaxPool2D (3, 3), (1, 3), (9, 1)	(431, 4, 64)	
Conv2D (1, 4), (1, 1), (1, 1)	(431, 1, 64)	16 384
Batch normalization	(431, 1, 64)	256
Squeeze	(431, 64)	
Conv1D (3), (1), (27)	(431, 128)	24 576
Batch normalization	(431, 128)	512
Conv1D (3), (1), (27)	(431, 64)	24 576
Batch normalization	(431, 64)	256
MaxPool1D (3), (1), (27)	(431, 64)	
Dense (per frame)	(431, 21)	1 365
Output: Sigmoid	(431, 21)	

and max-pooling operations (Goodfellow et al., 2016) and is inspired by the network used by Schlüter and Lehner (2018) for singing voice detection. In order to obtain a frame-wise output and a receptive field of appropriate size, we made two adjustments: first, all max-pooling operations have a stride of one in time such that the final output consists of 431 frames (same as the input). Consequently, all layers following the max pooling operations have appropriate dilation factors in time. Second, after the pitch axis has been pooled out, we add one-dimensional convolutions to increase the receptive field in time. Ultimately, the network has a receptive field covering the full pitch axis (all 84 CQT bins) and around 5.5 seconds on the time axis (encompassing most motif instances in our dataset, see **Table 1**). All convolutional layers use a leaky ReLU activation function with $\alpha = 0.2$. After the final convolution and max-pooling stage, we apply a dense layer at each frame and obtain leitmotif activity functions φ_ℓ in the same fashion as the RNN system. Unlike the RNN, however, this system is not causal but operates in a centric fashion, so the output at any frame depends on the frame itself and an equal number of preceding and subsequent frames.

3.2.3 Training and post-processing

We consider both networks as representatives for their respective architectural paradigms (recurrent vs. convolutional). Thus, we abstain from proposing complicated improvement strategies to either model. For the same reason, we take care to keep the number of parameters in the same order of magnitude (375 445 for the RNN and 440 021 for the CNN). This allows us to attribute any differences in network behavior to the architectural paradigms rather than the network size.

We train the networks by minimizing the average binary cross-entropy loss between predicted probabilities and correct labels at all frames using the Adam optimizer with a learning rate of 0.002 on mini-batches of 32 excerpts. We use the validation loss as a monitor for early stopping. After 30 epochs without decreasing loss, we reset the weights to the optimal epoch. These operations are implemented in Python using Tensorflow 2.[6]

After training, we obtain leitmotif activity predictions by pre-processing the test recordings and passing the resulting CQT frames through the model (from start to finish, i.e., including parts not containing leitmotifs). Essentially, the network layers are operating on entire test recordings, without restrictions due to their input shape (431,84). For the RNN-based model, this is achieved by passing on the internal LSTM states from frame to frame. Regarding the CNN-based model, we apply it on overlapping chunks of the test recordings with the overlap equal to its receptive field in time. This way, we can obtain predictions that are not affected by zero padding at the input edges. This yields the frame-wise activity functions φ_ℓ for each ℓ. Then, we post-process φ_ℓ using a median filter of length 0.5 seconds (applied in a centric fashion). Median filtering removes outliers (such as gaps and spikes) from φ_ℓ that are much shorter than the typical length of a leitmotif instance (see **Table 1**). Such a post-processing step is common for other detection procedures, e.g., for detecting singing voice (Schlüter and Lehner, 2018). Finally, we apply binarization with an individual binarization threshold per motif (tuned to maximize motif F-measure on the validation set using grid search). We proceed with the post-processed network outputs as described in Section 2.3 (transferring predictions from a physical to a musical time axis) to obtain $\mathcal{A}^{\mathrm{RNN}}$ and $\mathcal{A}^{\mathrm{CNN}}$.

3.3 Evaluation measures

After this conversion to a musical time axis, it is straightforward to use the resulting matrix $\mathcal{A}^{\mathrm{Est}}$ (i.e. $\mathcal{A}^{\mathrm{RNN}}$,$\mathcal{A}^{\mathrm{CNN}}$, or any other model output) and the reference $\mathcal{A}^{\mathrm{Ref}}$ for computing the number of true positive, false positive, and false negative predictions for a motif $\ell \in [1:L]$:

$$\mathrm{TP}_\ell = \sum_{m=1}^{M} \mathcal{A}_{\ell m}^{\mathrm{Ref}} \mathcal{A}_{\ell m}^{\mathrm{Est}} \tag{1}$$

$$\mathrm{FP}_\ell = \sum_{m=1}^{M} (1-\mathcal{A}_{\ell m}^{\mathrm{Ref}}) \mathcal{A}_{\ell m}^{\mathrm{Est}} \tag{2}$$

$$\mathrm{FN}_\ell = \sum_{m=1}^{M} \mathcal{A}_{\ell m}^{\mathrm{Ref}} (1-\mathcal{A}_{\ell m}^{\mathrm{Est}}) \tag{3}$$

Based on these numbers, we derive standard metrics such as precision (P), recall (R), and F-measure (F) for motif ℓ. Finally, we take the mean over these values for all motifs in order to obtain what we call the *class mean* evaluation measures. Thus, for these mean values, all classes (i.e. motifs) are counted equally, regardless of the amount of leitmotif activity per class.

Furthermore, we also compute

$$\mathrm{TP} = \sum_{\ell \in [1:L]} \mathrm{TP}_\ell \tag{4}$$

(likewise for FP, FN) and then obtain precision, recall, and F-measure based on TP, FP, and FN, instead. Since we aggregate values from the whole matrices here (regardless of class), we call these the *matrix mean* evaluation measures. These values are subject to class imbalance on the level of measure sub-segments: motifs with more (and longer) activity regions affect the result more than rare (and short) motifs. For these values, all leitmotif activity is counted equally, regardless of class.

The metrics described here correspond to segment-based precision, recall, and F-measure in their class-based (macro-averaged) and instance-based (micro-averaged) variant (Mesaros et al., 2016), respectively.

3.4 Evaluation with tolerance

Many applications of leitmotif activity detection may not require a very fine temporal granularity. For example, indicating a leitmotif one measure in advance may be sufficient for an application that draws a listener's attention to a forthcoming leitmotif. Furthermore, our automated annotation transfer with linear interpolation described in Section 2.2 may have introduced small errors, which should be accounted for in the evaluation. Motivated by such requirements, we introduce an additional tolerance parameter K in our evaluation. When comparing $\mathcal{A}^{\text{Ref}}$ and $\mathcal{A}^{\text{Est}}$, we filter both matrices prior to thresholding using a moving maximum filter of length K for each motif. In the subsequent experiments, we set $K = B$ so that the filter length corresponds to one measure. Thus, short interruption of a motif's activity (less than a measure long) are considered as the motif still being active. As another consequence, each false positive sub-segment leads to a *minimal penalty* in the evaluation, since the maximum filter enlarges false positive predictions to a duration of at least one measure (even if they are shorter). The same applies to false negative sub-segments, since any leitmotif activity in $\mathcal{A}^{\text{Ref}}$ is also enlarged to a duration of at least one measure. In a similar fashion, each true positive prediction is enlarged to a duration of at least one measure, which can be thought of as a *minimal reward* for true positives. In this context, it is important to note that the median filter applied to the model outputs already eliminates very short positive predictions (of less than roughly 0.25 seconds).

3.5 Experimental results

We evaluate the trained models on the three test performances (see **Figure 2**), post-process the output, and apply the evaluation procedure and metrics as described above. For the RNN-based system, we obtain the results given in the left block of **Table 4**. Precision, recall, and F-measure are given for each motif, e.g., for `L-RT`, P = 0.85, R = 0.86, and F = 0.85. In this experiment based on the RNN model, precision values are usually higher than recall values, especially for `L-Ju`, where P = 0.82 and R = 0.68. The effect is also evident in the class mean, where P = 0.83 and R = 0.79, implying that our model has more difficulties with false negatives than false positives.

We obtain the highest F-measure for `L-Wa` with F = 0.92, while the lowest is F = 0.73 for `L-Sc`. The class mean F-measure (F = 0.81) and the matrix mean F-measure (F = 0.80) are close to each other, which indicates that results for frequent and infrequent motifs (in terms of active measure sub-segments per motif) are similar. Overall, evaluation metrics for our RNN-based system for all motifs are above 0.7, with the mean results at around 0.8 for all evaluation metrics.

In **Figure 4**, we visualize results for our RNN-based model on an excerpt of the first act of *Siegfried*. Here, black regions correspond to true positive predictions of our model (after thresholding), while light and dark red regions indicate false negative and false positives, respectively. White color indicates true negative predictions. In the excerpt in **Figure 4**, most regions of leitmotif activity (and inactivity) are predicted correctly (black and white regions). Sometimes, only parts of a leitmotif instance are predicted as active (see, e.g., for `L-Ri` around measure 220). There are also some clear outliers such as the false positive predictions for `L-Ni` at measure 300 and `L-Ju` around measure 310. Overall, the correctly predicted regions dominate the visualization.

Table 4: Results for our deep learning-based leitmotif activity detection systems on the test set.

	RNN			CNN		
	P	R	F	P	R	F
`L-Ni`	0.87	0.76	0.81	0.85	0.79	0.82
`L-Ri`	0.80	0.73	0.76	0.82	0.76	0.79
`L-NH`	0.89	0.78	0.83	0.91	0.82	0.86
`L-Mi`	0.86	0.86	0.86	0.87	0.79	0.83
`L-RT`	0.85	0.86	0.85	0.80	0.83	0.82
`L-Wa`	0.94	0.90	0.92	0.93	0.95	0.94
`L-WL`	0.86	0.85	0.85	0.83	0.85	0.84
`L-Ho`	0.80	0.76	0.78	0.82	0.80	0.81
`L-Ge`	0.89	0.81	0.85	0.85	0.81	0.83
`L-Sc`	0.74	0.72	0.73	0.83	0.72	0.77
`L-Ju`	0.82	0.68	0.74	0.87	0.78	0.82
`L-WH`	0.79	0.77	0.78	0.78	0.76	0.77
`L-RS`	0.87	0.84	0.86	0.86	0.81	0.84
`L-Fe`	0.87	0.88	0.88	0.93	0.86	0.89
`L-SK`	0.75	0.72	0.74	0.81	0.75	0.78
`L-Un`	0.79	0.75	0.77	0.84	0.81	0.83
`L-Li`	0.89	0.81	0.85	0.82	0.84	0.83
`L-Si`	0.78	0.75	0.76	0.83	0.80	0.81
`L-Ma`	0.79	0.81	0.80	0.87	0.79	0.83
`L-Ve`	0.84	0.73	0.78	0.83	0.83	0.83
Class mean	0.83	0.79	0.81	0.85	0.81	0.83
Matrix mean	0.83	0.78	0.80	0.85	0.80	0.82

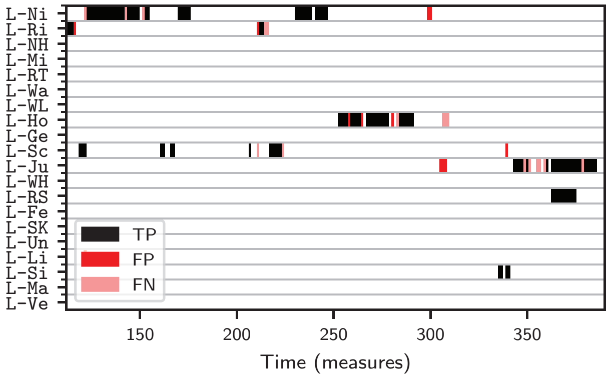


Figure 4: Illustration of results for our RNN-based leitmotif activity detection system (shown for measures 112 to 390 from the first act of *Siegfried* in `P-Ba`).

The right block of **Table 4** shows our results obtained with the CNN-based system. Overall, results are slightly better than for the RNN (see e.g. the class mean F-measure F = 0.83 compared to F = 0.81 for the RNN). Aside from this, we observe similar behavior as for the RNN. For example, `L-Wa` again yields the highest F-measure among motifs with F = 0.94. We conclude that it is unlikely that either architecture is strongly superior to the other in terms of evaluation scores on the test set.

4 Robustness to Input Modifications

We now want to gain a deeper understanding of the properties learned by our neural network-based models. To do so, we systematically modify the input to our models in different ways and investigate the impact this has on the model outputs.

Figure 5, upper row, gives a qualitative overview of the modifications we consider in this section. Besides the unmodified model input (a), these modifications encompass (b) tempo changes, (c) pitch shifts, (d) replacement of leitmotif frames by noise, and (e) shuffling of leitmotif frames. The lower row of **Figure 5** illustrates the activity functions resulting from the RNN for an example (solid red line), together with the reference annotation (dashed blue line). From a musical point of view, we would expect our activity detection approach to be robust against tempo changes and pitch shifts, while it should be sensitive to shuffling and noise replacement of frames. Strikingly, however, we see that tempo change and shuffling do not seem to change the results much, while pitch shifting and noise affect them strongly. We can also observe that our model anticipates the motif instance before it actually begins (**Figure 5a**). Very similar behavior can be observed for the CNN (not shown here in the interest of space), although the CNN does not anticipate the motif instance in this example.

In the following we examine these qualitative findings in a quantitative fashion. To do so, we apply the modifications to all acts of all performances in the test set, detect leitmotif activity in these modified inputs using our networks, and then evaluate with our usual procedure.

4.1 Tempo changes

First, we simulate global tempo changes in our test recordings by stretching or compressing our CQT representation along the time axis using bilinear filtering (see also **Figure 5b**).[7] **Figure 6a** shows the matrix mean F-measure obtained by the RNN on the test set for different tempo changes. For example, at 50% tempo, the input is stretched to twice its original length (i.e. slower), whereas for 200% tempo, the input is compressed to half its original length (i.e. faster). The solid red curve in **Figure 6** demonstrates the effect of this transformation on our model. The resulting F-measure steadily decreases for slower inputs (from F = 0.80 at 100% to F = 0.69 at 50%). For faster inputs, the F-measure remains higher compared to slower inputs (e.g. F = 0.76 at 200%). Nevertheless, most results are above F = 0.70, meaning that our model can deal even with considerable tempo changes. It should be noted that all test performances are longer (i.e. slower) than an average performance in

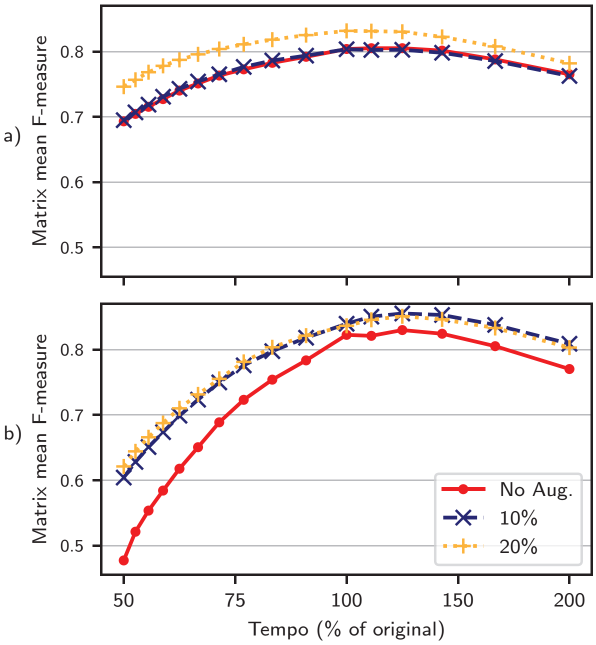


Figure 6: Results for our (a) RNN-based and (b) CNN-based leitmotif activity detection systems on the test set under tempo changes. The CQT input is stretched in time (using bilinear resampling) by the given percentage.

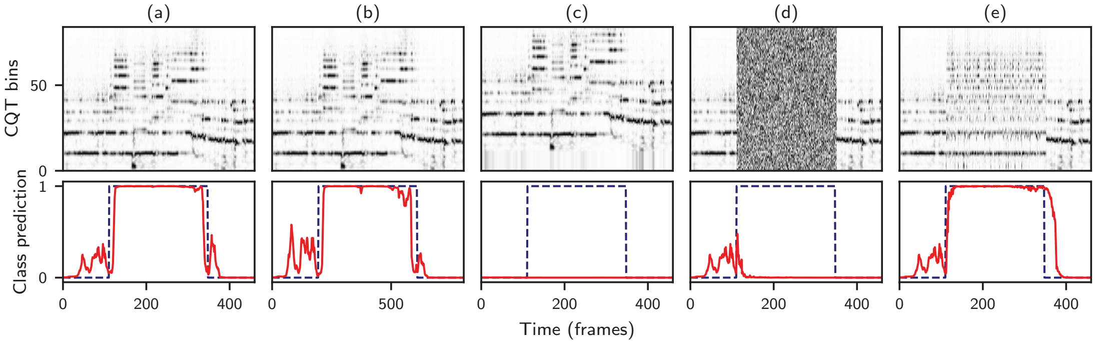


Figure 5: Results for our RNN-based leitmotif activity detection system on measures 117 to 123.5 of the first act of *Siegfried* in P-Ba (see also Figure 3 and Figure 4; outputs of the CNN-based model are similar). A prominent instance of L-Sc is being played in the higher registers, accompanied by low-frequency tremolo. The model input is shown in the upper row. The respective output activations for the L-Sc class are plotted underneath in red (solid line). The dashed blue line corresponds to the ground truth annotations for L-Sc. The input is given to the network (a) unchanged, (b) slowed down to 175% of the original length, (c) with a pitch shift of eleven semitones, (d) with motif frames replaced by noise, and (e) with motif frames shuffled along the time axis.

the training set. This may be the reason why our activity detection procedure is more robust to speeding up test performances while being more sensitive towards slowing them down.

A similar trend can be observed for the CNN-based model in **Figure 6b**. Here, we observe a stronger drop in results for slower inputs (from F = 0.82 at 100% to F = 0.48 at 50%). We hypothesize that this is due to the fixed size of the CNN's receptive field, which means that its predictions are based on less musical content for inputs at slower tempos and on more musical content for inputs at faster tempos.

We now conduct the same experiment with an additional data augmentation strategy, as is common practice in deep learning (Goodfellow et al., 2016), by also simulating global tempo changes during training. The dashed blue curve in **Figure 6a** shows the RNN's results in this experiment. This way, training examples are randomly stretched or compressed to be at most 10% slower or faster. The solid red and dashed blue curves are almost identical, meaning that this augmentation does not affect results much. We repeat this experiment with training augmentations of up to 20% change in tempo, indicated by the dotted orange curve. Here, test F-measure increases for all amounts of tempo changes (including F = 0.83 at 100%). For the CNN, we observe a similar behavior in **Figure 6b**. Here, both augmentation experiments yield improved results, although there is still a drop for very slow inputs (F = 0.62 at 50% for augmentations up to 20%). From these experiments, we conclude that training on ten different performances of the *Ring* already introduces some robustness to minor tempo changes in our model, which may further be enhanced through augmentations.

4.2 Pitch shifts

Second, we simulate transpositions in our test recordings by shifting our CQT representations along the pitch axis (using nearest-neighbor padding at the boundaries, i.e., the value for the lowest/highest CQT bin is replicated), see also **Figure 5c**. **Figure 7a** (solid red curve) shows matrix mean F-measures obtained with the RNN after modifying the test recordings in this way. This curve demonstrates that pitch shifts have a dramatic effect. For example, the test results drop to F = 0.11 for a shift of one semitone upwards. Shifting by more semitones, the F-measure drops further. We conclude that our model crucially relies on absolute pitch information. Even though leitmotif instances of the same motif appear in different registers and keys, the model has not learned their properties in a transposition-invariant way. As such, the model can only detect transposed motifs seen during training and would fail to generalize to new, unseen transpositions.

Convolutional architectures such as our CNN-based model are usually ascribed a certain degree of translation-invariance due to the weight-sharing and pooling operations (Goodfellow et al., 2016). Performing the pitch shift experiment for our CNN (**Figure 7b**), we can indeed observe better results than for the RNN when applying pitch shifting to the model input. For example, a shift of one semitone upwards now yields F = 0.26 and F-measures never drop below 0.1 for any considered shift. However, all shifts yield F-measures below 0.3, meaning that absolute pitch information is still highly important for our CNN-based model.

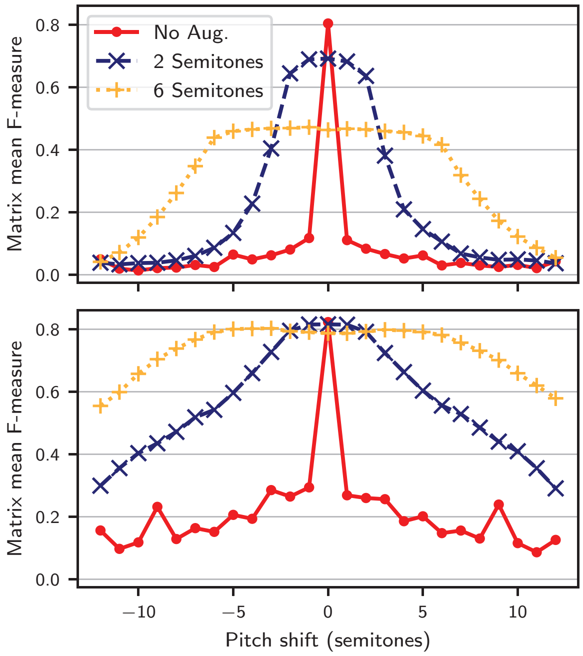


Figure 7: Results for our (a) RNN-based and (b) CNN-based leitmotif activity detection systems on the test set under pitch shifts. The CQT input has been shifted (using nearest-neighbor padding) on the pitch axis by the given number of semitones (corresponding to CQT bins).

We repeat this experiment with an augmentation strategy, using pitch shifting also for the training set. Here, training examples are randomly shifted at most two semitones in either direction along the pitch axis. The dashed blue curve in **Figure 7a** shows the corresponding results for the RNN. We observe that applying this augmentation decreases results for the unmodified test inputs (i.e. F = 0.69 for a shift of 0), but increases results for transformations considered during training (shifts of -2 to +2 semitones). Larger shifts still cause the model to fail. The same effect is seen in the dotted orange curve, where shifts of up to ±6 semitones were applied as augmentation during training. Here, the result for unmodified model input drops to F = 0.46, but the model can now cope with pitch shifts within the same range as used for augmentation (e.g. F = 0.42 for a shift of +6 semitones). In addition, the slopes of the F-measure curve are less steep, implying better generalization (e.g. F = 0.26 for a shift of minus eight semitones, even though only shifts up to ±6 semitones were included during training).

Figure 7b shows the corresponding curves for the CNN. Here, results for unmodified model input (shift of 0 semitones) drop only slightly when adding augmentations (e.g. F = 0.79 for up to ±6 semitones pitch shift augmentation compared to F = 0.82 without augmentation). Additionally, the slopes of the F-measure curves are even less steep (e.g. F = 0.74 for a shift of minus eight semitones and up to ±6 semitones as augmentation).

4.3 Noise

Third, we study the effect of completely removing all information in leitmotif regions from our test set. To do so, we replace all frames within a leitmotif instance by uniform noise (see **Figure 5d**). The impact of this modification on the RNN's results is shown in **Figure 8a (1)**. When replacing all leitmotif frames (denoted as "All"), we obtain a much lower F-measure (F = 0.13) compared to the original model input ("Unchanged," F = 0.80). In order to see whether our model responds to certain parts of leitmotif instances, we further modify only the first ("Start"), the middle ("Middle"), or the last third of frames ("End") for each leitmotif instance. The drop in F-measure is most pronounced for the beginning of motif instances (leading to F = 0.42 when replacing the first third but preserving the rest, compared to F = 0.63 for the last third). Yet, the overall F-measure does not drop entirely even when replacing all frames by noise. This implies that context around the leitmotif instances can help in identifying motifs even when the actual motif frames are absent. Again, we observe similar results for the CNN in **Figure 8b (1)**. Here, frames in the middle of each leitmotif affect results more strongly and results drop even further when replacing all leitmotif frames (F = 0.07). Overall, we can conclude that our CNN-based model exploits context around leitmotif regions in a similar fashion as the RNN does.

4.4 Shuffling

Fourth, we study the effect of removing the temporal order from the leitmotif activity regions. To do so, we shuffle the frames within a leitmotif instance along the time axis, see also **Figure 5e**. The impact of this modification on the RNN is shown in **Figure 8a (2)**, again for different parts of a leitmotif instance. We can observe that shuffling has only a minor impact on results (giving F = 0.79 when shuffling only the first third or F = 0.67 for all frames). Since shuffling along the time axis destroys any rhythmic information as well as the temporal aspects of melody (the order of notes), we conclude that such rhythmic or melodic cues are largely ignored by our model. We hypothesize that our model instead captures the pitch distributions in leitmotif instances, which are related to harmony. These distributions are mostly preserved when shuffling leitmotif frames, explaining the high results even for shuffling all frames of a leitmotif instance. Our experiments on pitch shifting (see Section 4.2) further suggest that the model depends on absolute pitch distributions rather than relative harmonic relationships (since pitch shifting preserves relative pitch relationships but changes absolute pitch distributions, leading to worse results).

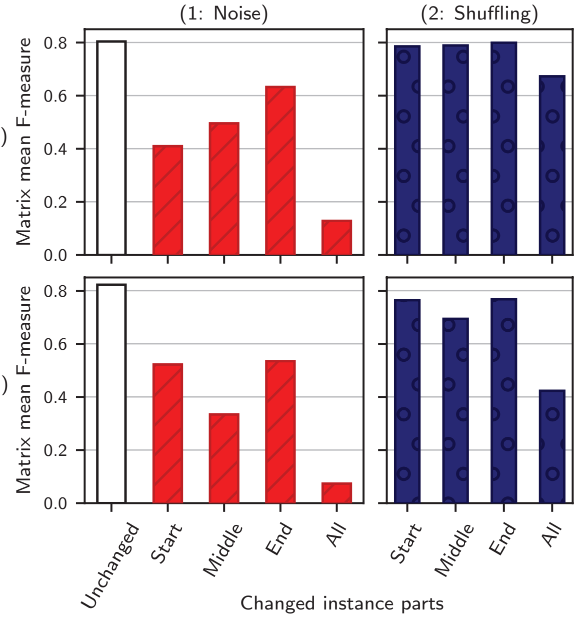


Figure 8: Results for our (a) RNN-based and (b) CNN-based leitmotif activity detection systems on the test set when (1) replacing leitmotif frames by noise or (2) shuffling them along the time axis. The modifications have been applied to either the first, middle, or last third of each leitmotif instance (Start, Middle, End), for none (Unchanged), or for all leitmotif frames (All).

The CNN reacts more strongly to this input modification, see **Figure 8b (2)**. When shuffling all frames, for example, the F-measure drops to 0.42. F-measures remain high when only individual parts of the instances are shuffled (e.g. F = 0.77 when shuffling only the end). Therefore, we hypothesize that our CNN only weakly reacts to temporal relationships.

Summarizing the insights obtained from the input modifications, we find that our models are to some degree robust to global tempo changes, which is a desirable property. However, we also found that they rely on pitch distributions within leitmotif instances (which is undesirable since these distributions can be affected by other musical parts) instead of capturing many musical cues that human listeners would associate with specific leitmotifs (such as temporal aspects of melody and rhythm). We further found that our recurrent and convolutional architectures behave similarly under input modifications, with some slight differences. While the CNN is affected more strongly by slowed down input, it is more robust to pitch shifts, especially when using additional augmentation. In addition, the CNN is affected slightly more strongly by shuffling of leitmotif frames than the RNN.

5 Towards Less Informed Scenarios

This paper considers the task of detecting leitmotif activity in a continuous (frame-wise) fashion over the course of entire opera recordings. As a more informed scenario, our previous study considered classification of pre-segmented audio excerpts according to the leitmotif played (Krause et al., 2020). Additionally, we ruled out excerpts where multiple leitmotifs were played simultaneously. Compared to this constrained scenario, the leitmotif activity detection task is more challenging since no pre-segmented instances are given and inputs may contain no motif or simultaneously active motifs. In Krause et al. (2020), we report F-measures of about 0.9 for a leitmotif classification setting with the first ten motifs of **Table 1**. While our results cannot be compared directly (especially since we evaluate on a frame level instead of an excerpt level as in Krause et al. (2020)), we can see that the detection F-measures obtained with our deep-learning systems (**Table 4**) are lower, at roughly 0.8 on average.

To approach scenarios with an even lower degree of side information, our systems must be able to deal with previously unseen leitmotif occurrences. The classification experiments reported in Krause et al. (2020) demonstrate that generalizing to unseen leitmotif occurrences is more challenging than generalizing to unseen performances of known occurrences. To this end, different splits of the dataset were considered in Krause et al. (2020). In a similar way, we performed a preliminary experiment where we split the dataset across operas instead of performances. Here, we trained on all operas except for *Das Rheingold* in all 16 performances (**Figure 2**). We then evaluated on a test set containing only *Das Rheingold*, again in all 16 performances. From this experiment, we obtained low evaluation measures with P = 0.17, R = 0.07 and F = 0.10 (matrix mean) for the RNN-based system, as well as P = 0.18, R = 0.13 and F = 0.15 for the CNN-based system.

The discrepancy between the performance and the opera split's results may be explained with the models relying on confounding factors such as pitch distributions in leitmotif instances, while ignoring musically relevant aspects of leitmotifs such as rhythmic or melodic progressions. In other words, our models can be said to be overfitted towards the specific motif instances in the training set. In order to approach less informed scenarios such as the opera split (i.e. generalizing to unseen pattern occurrences) or the discovery of unknown leitmotifs (i.e. discovering unknown patterns in an unsupervised fashion), it becomes important to limit the impact of confounding factors. For this purpose, using more diverse data is recommended in the machine learning literature (Goodfellow et al., 2016). This could be realized, e.g., by adding more performances, considering data augmentation strategies, or utilizing artificial training data to expose the models to a larger variety of tempo, key, or timbre. As a different approach, one might annotate additional musical works and utilize transfer-learning techniques (Choi et al., 2017). Another improvement strategy could be the use of more elaborate neural network architectures by increasing the number of network parameters or by using convolutional-recurrent architectures (Çakir et al., 2017) and other recent models proposed for sound event detection tasks (Li et al., 2020). Additionally, dedicated architectures introducing invariance to tempo (Di Giorgi et al., 2020), key (Elowsson and Friberg, 2019) or other properties (Lattner et al., 2019) may be useful.

6. Conclusion

In this paper, we approached the task of detecting leitmotif activity in opera recordings as a case study for the detection of complex musical patterns in audio. For our experiments, we considered a scenario comprising 3569 annotated occurrences of 20 characteristic leitmotifs in Wagner's *Ring* cycle, realized in 16 different performances and, thus, summing up to 57 104 activity regions within more than 200 hours of audio material. As our main contributions, we tested two deep-learning models for leitmotif activity detection and analyzed their behavior under different input modifications. Our deep-learning models obtained good numerical results on a held-out test set but captured confounding factors such as absolute pitch distributions, rather than relying on characteristic musical properties of leitmotifs such as rhythmic or melodic patterns. Thus, our study demonstrates the challenges faced by neural networks for detecting musical patterns. Future work may employ elaborate model architectures and dedicated training strategies in order to handle this task in a more robust way and to proceed towards approaching other, less-informed scenarios.

Notes

[1] https://www.audiolabs-erlangen.de/resources/MIR/2021-TISMIR-TowardsLeitmotifDetection.

[2] An example of motifs whose occurrences are possibly correlated are the motif for the horn of the hero Siegfried (`L-Ho`) and the motif for the character himself (`L-Si`).

[3] https://www.music-ir.org/mirex/wiki/2017:Discovery_of_Repeated_Themes_%26_Sections.

[4] http://dcase.community/challenge2020/.

[5] https://librosa.org/.

[6] https://www.tensorflow.org/.

[7] The experiments in this and the following section yield similar trends and conclusions when performed using a phase vocoding technique for time-scale modification.

Acknowledgements

We thank Julia Zalkow and Vlora Arifi-Müller for their assistance in preparing the data. This work was supported by the German Research Foundation (DFG MU 2686/7-2). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen (IIS).

Competing Interests

The authors have no competing interests to declare.

References

Albrecht, H., and **Frieler, K.** (2014). The perception and recognition of Wagnerian leitmotifs in multimodal conditions. In *Proceedings of the International Conference of Students of Systematic Musicology (SysMus)*, London, UK.

Baker, D. J., and **Müllensiefen, D.** (2017). Perception of leitmotives in Richard Wagner's Der Ring des Nibelungen. *Frontiers in Psychology*, 8: 662. DOI: https://doi.org/10.3389/fpsyg.2017.00662

Böck, S., Krebs, F., and **Widmer, G.** (2016). Joint beat and downbeat tracking with recurrent neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–261, New York City, USA.

Bribitzer-Stull, M. (2015). *Understanding the Leitmotif*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781316161678

Çakir, E., Parascandolo, G., Heittola, T., Huttunen, H., and **Virtanen, T.** (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6): 1291–1303. DOI: https://doi.org/10.1109/TASLP.2017.2690575

Choi, K., Fazekas, G., Sandler, M. B., and **Cho, K.** (2017). Transfer learning for music classification and regression tasks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 141–149, Suzhou, China.

Di Giorgi, B., Mauch, M., and **Levy, M.** (2020). Downbeat tracking with tempo-invariant convolutional neural networks. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, pages 216–222, Montréal, Canada.

Dreyfus, L., and **Rindfleisch, C.** (2014). Using digital libraries in the research of the reception and interpretation of Richard Wagner's leitmotifs. In *Proceedings of the InternationalWorkshop on Digital Libraries for Musicology*, pages 1–3, London, UK. DOI: https://doi.org/10.1145/2660168.2660181

Elowsson, A., and **Friberg, A.** (2019). Modeling music modality with a key-class invariant pitch chroma CNN. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 541–548, Delft, The Netherlands.

Goodfellow, I., Bengio, Y., and **Courville, A.** (2016). *Deep Learning*. MIT Press, Cambridge and London. http://www.deeplearningbook.org.

Kornstädt, A. (2001). The JRing system for computerassisted musicological analysis. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 93–98, Bloomington, Indiana, USA.

Krause, M., Zalkow, F., Zalkow, J., Weiß, C., and **Müller, M.** (2020). Classifying leitmotifs in recordings of operas by Richard Wagner. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 473–480, Montréal, Canada.

Lattner, S., Dörfler, M., and **Arzt, A.** (2019). Learning complex basis functions for invariant representations of audio. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 700–707, Delft, The Netherlands.

Li, Y., Liu, M., Drossos, K., and **Virtanen, T.** (2020). Sound event detection via dilated convolutional recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 286–290. DOI: https://doi.org/10.1109/ICASSP40776.2020.9054433

Mesaros, A., Heittola, T., and **Virtanen, T.** (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6): 162. DOI: https://doi.org/10.3390/app6060162

Morimoto, Y., Kamekawa, T., and **Marui, A.** (2009). Verbal effect on memorisation and recognition of Wagner's leitmotifs. In *Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*.

Page, K. R., Nurmikko-Fuller, T., Rindfleisch, C., Weigl, D. M., Lewis, R., Dreyfus, L., and **De Roure, D.** (2015). A toolkit for live annotation of opera performance: Experiences capturing Wagner's Ring cycle. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 211–217, Málaga, Spain.

Schlüter, J., and **Lehner, B.** (2018). Zero-mean convolutions for level-invariant singing voice detection. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 321–326, Paris, France.

Schreiber, H., Weiß, C., and **Müller, M.** (2020). Local key estimation in classical music recordings: A cross-version study on Schubert's Winterreise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 501–505, Barcelona, Spain. DOI: https://doi.org/10.1109/ICASSP40776.2020.9054642

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and **Plumbley, M. D.** (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10): 1733–1746. DOI: https://doi.org/10.1109/TMM.2015.2428998

Virtanen, T., Plumbley, M. D., and **Ellis, D.** (2018). *Computational Analysis of Sound Scenes and Events*. Springer. DOI: https://doi.org/10.1007/978-3-319-63450-0

Wagner, R. (1995). *Opera and Drama*. University of Nebraska Press. Translation of the original edition from 1851.

Wagner, R. (2013). *Der Ring des Nibelungen. Vollständiger Text mit Notentafeln der Leitmotive*. Schott Music, Mainz. Reprint of the original edition from 1913 (Ed. Julius Burghold).

Weiß, C., Arifi-Müller, V., Prätzlich, T., Kleinertz, R., and **Müller, M.** (2016). Analyzing measure annotations for Western classical music recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 517–523, New York, USA.

Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Müller, M., and **Lerch, A.** (2018). A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9): 1457–1483. DOI: https://doi.org/10.1109/TASLP.2018.2830113

Xia, X., Togneri, R., Sohel, F., Zhao, Y., and **Huang, D.** (2019). A survey: Neural network-based deep learning for acoustic event detection. *Circuits, Systems, and Signal Processing*, 38(8): 3433–3453. DOI: https://doi.org/10.1007/s00034-019-01094-1

Zalkow, F.,Weiß, C., and **Müller, M.** (2017a). Exploring tonal-dramatic relationships in Richard Wagner's Ring cycle. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 642–648, Suzhou, China.

Zalkow, F., Weiß, C., Prätzlich, T., Arifi-Müller, V., and **Müller, M.** (2017b). A multi-version approach for transferring measure annotations between music recordings. In *Proceedings of the AES International Conference on Semantic Audio*, pages 148–155, Erlangen, Germany.

How to cite this article: Krause, M., Müller, M., and Weiß, C. (2021). Towards Leitmotif Activity Detection in Opera Recordings. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 127–140. DOI: https://doi.org/10.5334/tismir.116

Submitted: 26 May 2021 **Accepted:** 13 September 2021 **Published:** 02 November 2021

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS