CLASSIFYING LEITMOTIFS IN RECORDINGS OF OPERAS BY RICHARD WAGNER

Michael Krause, Frank Zalkow, Julia Zalkow, Christof Weiß, Meinard Müller
International Audio Laboratories Erlangen, Germany
{michael.krause,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

From the 19th century on, several composers of Western opera made use of *leitmotifs* (short musical ideas referring to semantic entities such as characters, places, items, or feelings) for guiding the audience through the plot and illustrating the events on stage. A prime example of this compositional technique is Richard Wagner's four-opera cycle *Der Ring des Nibelungen*. Across its different occurrences in the score, a leitmotif may undergo considerable musical variations. Additionally, the concrete leitmotif instances in an audio recording are subject to acoustic variability. Our paper approaches the task of classifying such leitmotif instances in audio recordings. As our main contribution, we conduct a case study on a dataset covering 16 recorded performances of the *Ring* with annotations of ten central leitmotifs, leading to 2403 occurrences and 38448 instances in total. We build a neural network classification model and evaluate its ability to generalize across different performances and leitmotif occurrences. Our findings demonstrate the possibilities and limitations of leitmotif classification in audio recordings and pave the way towards the fully automated detection of leitmotifs in music recordings.

1. INTRODUCTION

Music has long been used to accompany storytelling, from Renaissance madrigals to contemporary movie soundtracks. A central compositional method is the association of a certain character, place, item, or feeling with its own musical idea. This technique culminated in 19th century opera where these ideas are denoted as *leitmotifs* [1, 2]. A major example for the use of leitmotifs is Richard Wagner's tetralogy *Der Ring des Nibelungen*, a cycle of four operas [1] with exceptional duration (a performance lasts up to 15 hours) and a continuous plot spanning all four operas. As many characters or concepts recur throughout the cycle, so do their corresponding leitmotifs. This allows the audience to identify these concepts not only through text or visuals, but also in a musical way. While all these different *occurrences* of a leitmotif in the score share a characteristic musical idea, they can appear in different musical contexts and may vary substantially in compositional aspects such as melody, harmony, key, tempo, rhythm, or instrumentation. When considering recorded *performances* of the *Ring*, another level of variability is introduced due to acoustic conditions and aspects of interpretation such as tempo, timbre, or intonation. In the following, we denote the concrete realization of a leitmotif in an audio recording as an *instance* of the motif. This paper approaches the problem of classifying such leitmotif instances in audio recordings, as illustrated in Figure 1. In particular, we study generalization across occurrences and performances.

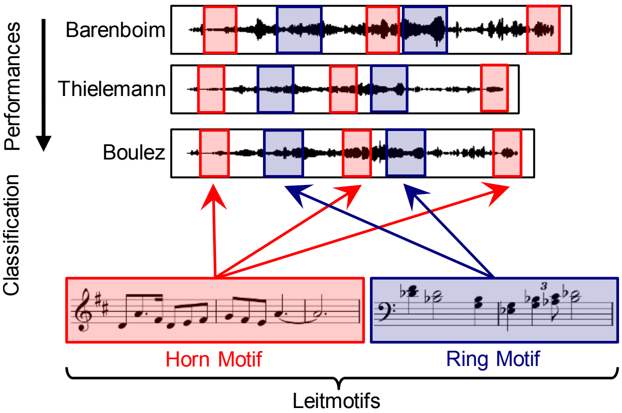


Figure 1. Illustration of example leitmotifs (red for the Horn motif, blue for the Ring motif) occurring several times in the Ring cycle and across different performances.

Cross-version studies on multiple performances have been conducted regarding the harmonic analysis of Beethoven sonatas [3] or Schubert songs [4], but also for the *Ring* [5, 6]. Beyond harmonic aspects, the *Ring* scenario was considered for capturing audience experience using body sensors and a live annotation procedure [7] or for studying the reliability of measure annotations [8, 9]. Regarding leitmotifs, several works have focused on the human ability to identify motifs [10–12]. In particular, [13] found that distance of chroma features correlates with difficulty for listeners in identifying leitmotifs. In [6], Zalkow et al. presented a framework for exploring relationships between leitmotif usage and tonal characteristics of the *Ring*.

[1] While Wagner referred to his works as *music dramas* instead of operas, we choose the more commonly used latter term.

Attribution: Michael Krause, Frank Zalkow, Julia Zalkow, Christof Weiß, Meinard Müller, "Classifying Leitmotifs in Recordings of Operas by Richard Wagner", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

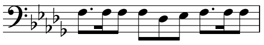
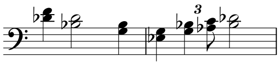
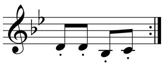
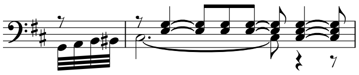
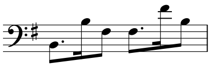
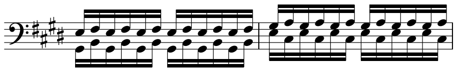
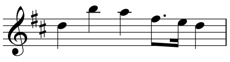
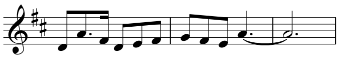
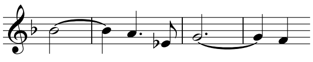
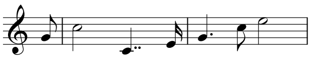
Name (English translation)	ID	Score	# Occurrences	Length Measures	 Seconds
Nibelungen (Nibelungs)	L-Ni		536	0.96 ± 0.23	1.72 ± 0.50
Ring (Ring)	L-Ri		286	1.49 ± 0.65	3.64 ± 2.30
Mime (Mime)	L-Mi		242	0.83 ± 0.25	0.87 ± 0.24
Nibelungenhass (Nibelungs' hate)	L-NH		237	0.96 ± 0.17	3.10 ± 1.11
Ritt (Ride)	L-RT		228	0.66 ± 0.17	1.24 ± 0.37
Waldweben (Forest murmurs)	L-Wa		223	1.10 ± 0.30	2.70 ± 0.76
Waberlohe (Swirling blaze)	L-WL		190	1.21 ± 0.39	4.39 ± 1.60
Horn (Horn)	L-Ho		172	1.38 ± 1.05	2.44 ± 1.57
Geschwisterliebe (Siblings' love)	L-Ge		155	1.31 ± 0.83	3.03 ± 2.55
Schwert (Sword)	L-Sc		134	1.89 ± 0.55	3.68 ± 1.88

Table 1. Overview of the leitmotifs used in this study. Lengths are given as mean and standard deviations over all annotated occurrences (in measures) or instances (in seconds) from all performances given in Table 2.

From a technical perspective, our scenario entails the task of automatically detecting leitmotifs within an audio recording. This paper represents a first step towards this goal by considering a simplified classification scenario with pre-segmented instances (see Figure 1).

Due to the multiple sources of variability described above, we opt for a data-driven approach. Neural networks have emerged as the dominant classification models. In particular, recurrent neural networks (RNNs) are able to handle input sequences of varying length. Our study shows that despite the difficulties of the scenario, an RNN classifier is surprisingly effective in dealing with the variability across occurrences and performances.

The main contributions of our work are as follows: We conduct a case study on classifying leitmotif instances in audio recordings of the *Ring*. For this, we describe the task of leitmotif classification and provide a dataset of more than 38000 annotated instances within 16 performances of the *Ring* (Section 2). We further build an RNN model for classifying leitmotifs in audio recordings (Section 3). We carefully evaluate our model with respect to variabilities across performances and leitmotif occurrences over the course of the *Ring*. Moreover, we investigate the effect of adding temporal context and critically discuss the potential limitations and generalization capabilities of our classifier (Section 4). Finally, we suggest new research directions that may continue our work (Section 5).

2. SCENARIO

We now discuss the dataset and leitmotif classification scenario underlying our experiments.

2.1 Leitmotifs in Wagner's Ring

While Wagner mentioned the importance of motifs for his compositional process [14], he did not explicitly specify the concrete leitmotifs appearing in the *Ring*. Whether a recurring musical idea constitutes a leitmotif—and how to name it—is a topic of debate even among musicologists, see, e. g., [15] where differences in leitmotif reception are discussed. In line with [6], we follow Julius Burghold's specification of more than 130 leitmotifs in the *Ring* [16].

For our experiments, we selected ten central motifs frequently occurring throughout the *Ring* (see Table 1 for an overview including the number of occurrences per motif). These motifs constitute the classes of our classification task. The selection comprises motifs associated with an item such as the sword (`L-Sc`), with characters such as the dwarf Mime (`L-Mi`), or with emotions such as love (`L-Ge`). All occurrences of these motifs were annotated by a musicologist using a vocal score of the *Ring* as a reference, resulting in 2403 occurrences.

As discussed in Section 1, a leitmotif may occur in different shapes over the course of a drama. These musical variations may be necessary to fit the musical context in which the occurrences appear and, thus, be adjusted to the

ID	Conductor	Year	hh:mm:ss
P-Ba	Barenboim	1991–92	14:54:55
P-Ha	Haitink	1988–91	14:27:10
P-Ka	Karajan	1967–70	14:58:08
P-Sa	Sawallisch	1989	14:06:50
P-So	Solti	1958–65	14:36:58
P-We	Weigle	2010–12	14:48:46
P-Bo	Boulez	1980–81	13:44:38
P-Bö	Böhm	1967–71	13:39:28
P-Fu	Furtwängler	1953	15:04:22
P-Ja	Janowski	1980–83	14:08:34
P-Ke	Keilberth/Furtwängler	1952–54	14:19:56
P-Kr	Krauss	1953	14:12:27
P-Le	Levine	1987–89	15:21:52
P-Ne	Neuhold	1993–95	14:04:35
P-Sw	Swarowsky	1968	14:56:34
P-Th	Thielemann	2011	14:31:13

Table 2. Recorded performances of the *Ring* used in this study (see also [6]). Measure positions have been annotated manually for the topmost three performances (P-Ba, P-Ha and P-Ka), 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.

current key, meter, or tempo. Moreover, occurrences of leitmotifs may appear in different registers, musical voices, or instruments. In addition to this, motifs can also occur in abridged or extended shape, with parts of the motif being repeated, altered, or left out. Despite these diverse musical variations across occurrences, listeners can often identify motifs easily when listening to a performance. This is in line with Wagner's intention of using the motifs as a guideline, thus forming the musical surface of the *Ring* [17].

2.2 Recorded Performances

As mentioned in the introduction, we do not attempt to classify leitmotifs within a score representation but on the basis of a performance given as an audio recording. To be more concrete, our work relies on 16 recorded performances of the *Ring* that have been used before in [6]. For three of these performances, the positions of measures from the score were manually annotated in the audio [8]. For the remaining 13 performances, the measure positions were transferred from the manually annotated performances using automatic audio-to-audio synchronization [9]. Table 2 specifies the performances. We automatically located the 2403 leitmotif occurrence regions from the score in each of the 16 recorded performances using linear interpolation between measure positions. This way, we obtained the 38448 instances used for our experiments. The occurrence and instance positions are made publicly available as a dataset for further research. [2]

2.3 Leitmotif Classification Task

In this paper, we consider the task of leitmotif classification. We define this as the problem of assigning a given audio excerpt to a class according to the occurring leitmotif. Here, we consider ten classes corresponding to the motifs in Table 1. We further make the simplifying assumption that only a single leitmotif is played at a time. Thus, we omit excerpts where multiple motifs occur simultaneously. Our classification task therefore becomes a multi-class, single-label problem.

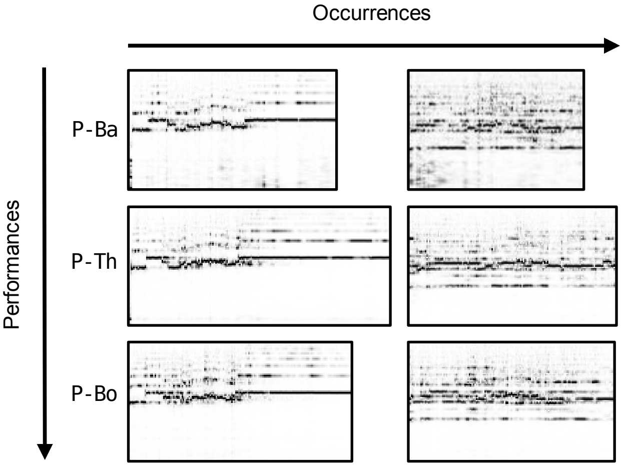


Figure 2. Variability of L-Ho across occurrences and performances. Six instances (two occurrences for three performances) are shown in a CQT representation, which is also used as input to our classification model.

Our dataset allows us to approach the leitmotif classification task from two perspectives, each of which incorporates its own types of variabilities. First, the *performance* perspective concerns variabilities across different performances, resulting from different instrumental timbres, tempi, or other decisions made by the artists. Furthermore, this perspective encompasses technical properties such as acoustic, recording, and mastering conditions, which can lead to the so-called "album effect" [18]. Second, the compositional or *occurrence* perspective concerns diverse musical variabilities of leitmotif occurrences in the score (as discussed in Section 2.1). Figure 2 shows the Horn motif L-Ho for different performances and occurrences. The variability is evident in different durations of the instances as well as different energy distributions due to other musical events sounding simultaneously. These variabilties make our classification task a challenging problem. In our experiments, we investigate the generalization across these two perspectives, similar to the study in [4].

3. RECURRENT NEURAL NETWORK FOR LEITMOTIF CLASSIFICATION

Neural networks have previously proven to be useful for classification tasks in the music domain, see, e. g., [19–21]. As we are dealing with variable length inputs (leitmotif instances may last from less than one to over ten seconds in a performance), recurrent neural networks (RNNs) are a natural choice for our scenario.

As input to our system, we take audio excerpts containing leitmotif instances from our 16 performances of the *Ring*, sampled at 22050 Hz. These excerpts are processed by a constant-Q-transform (CQT) [22, 23] with semitone resolution over six octaves and a hop length of 512 sam-

[2] https://www.audiolabs-erlangen.de/resources/MIR/2020-ISMIR-LeitmotifClassification

Layer	**Output Shape**	**Parameters**
Input	(V, 84)	
LSTM	(V, 84)	109056
LSTM	(V, 128)	131584
LSTM	(V, 128)	131584
Take last	(128)	
Batch normalization	(128)	512
Dense	(10)	1290
Output: Softmax	(10)	

Table 3. Architecture of our RNN for leitmotif classification. V indicates variable length.

ples, where we adjust for tuning deviations (estimated automatically per performance and opera act). These steps are implemented using librosa [24]. Finally, all CQT frames are normalized using the max-norm and the resulting representations serve as inputs to our network.

Table 3 gives an overview of the network structure. We use an RNN-variant, the long short-term memory (LSTM) proposed in [25]. We stack multiple LSTM layers and, after the final LSTM output, append batch normalization [26] as well as a single fully connected classification layer to obtain leitmotif predictions. We set the number of LSTM layers and the size of their internal representation to 3 and 128, respectively. We train this network for 50 epochs by minimizing the cross-entropy loss between predictions and correct classes using the Adam optimizer [27] with a learning rate of 0.001 on mini-batches of 32 excerpts. Since the excerpts in a batch may have different lengths, we need to zero-pad them to the maximum number of frames among excerpts in that batch. During computation, we then use masking to ignore zeros added to shorter inputs. We further avoid overfitting by selecting the weights of the epoch that yields the highest mean F-measure on the validation set (as described in Section 4.2). The network is implemented in Python using Tensorflow.

4. EXPERIMENTS

4.1 Setup and Splits

We follow the common machine learning approach of partitioning our dataset into training, validation, and test subsets to train, tune hyperparameters, and estimate the results on unseen samples, respectively. In contrast to standard procedures, we partition the data according to musical aspects as motivated in Section 2.3. We will consider two splits: the performance and occurrence splits.

For the performance split, we select the three recordings with manually annotated measure positions (`P-Ba`, `P-Ha` and `P-Ka`, see Table 2) for the test set and three performances with automatically transferred measure positions for the validation set (`P-Sa`, `P-So` and `P-We`). The remaining ten performances are used for training. In this split, all subsets comprise all occurrences of all motifs. Results on the performance split are given in Section 4.3.

In contrast, for the occurrence split, we randomly choose 80% of the occurrences for training and 10% each for the validation and test set. [3] We further ensure that the proportions of occurrences for each motif is the same in all subsets. In this split, each subset contains all instances of the occurrences in that subset. Results on the occurrence split are given in Section 4.4.

Context	***Strict***			***Variable***			***Fixed* (10 sec.)**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
`L-Ni`	0.94	0.95	0.94	0.90	0.95	0.92	0.93	0.93	0.93
`L-Ri`	0.93	0.92	0.93	0.84	0.93	0.88	0.86	0.89	0.87
`L-Mi`	0.96	0.95	0.96	0.95	0.93	0.94	0.92	0.98	0.95
`L-NH`	0.94	0.92	0.93	0.96	0.88	0.92	0.97	0.87	0.92
`L-RT`	0.95	0.94	0.95	0.94	0.90	0.92	0.96	0.95	0.96
`L-Wa`	0.94	0.98	0.96	0.98	0.96	0.97	0.96	0.99	0.98
`L-WL`	0.98	0.93	0.96	0.93	0.93	0.93	0.95	0.94	0.94
`L-Ho`	0.90	0.89	0.89	0.93	0.85	0.89	0.92	0.91	0.91
`L-Ge`	0.94	0.94	0.94	0.93	0.91	0.92	0.97	0.94	0.96
`L-Sc`	0.91	0.96	0.93	0.94	0.89	0.92	0.84	0.86	0.85
Mean	0.94	0.94	0.94	0.93	0.91	0.92	0.93	0.92	0.93

Table 4. Main results of our method on the test set of the performance split for different strategies of using temporal context.

4.2 Evaluation Measures

We adopt standard measures from information retrieval for evaluating our models. For a given class (i. e., motif), we treat the classification problem as a retrieval problem, yielding class-dependent precision (P), recall (R), and F-measure (F) as usual, see, e. g., [28].

We also report the mean precision, recall, and F-measure over all classes. This gives a general impression of the classification quality. Note that these averages are not affected by class imbalance. Therefore, low results on an infrequent class will influence the mean results as much as low results on a frequent class.

4.3 Results on the Performance Split

Basic Experiment. The left block in Table 4 (*Strict*) summarizes results for our model on the test subset of the performance split. We obtain high classification results with a mean F-measure of 0.94. Results are similar across motifs. Highest precision ($\mathrm{P} = 0.98$) is obtained for `L-WL`, while highest recall ($\mathrm{R} = 0.98$) is reached for `L-Wa`. Recall and precision per motif are often similar. We conclude that it is indeed possible to classify leitmotif instances in previously unseen performances, provided that all occurrences were seen before in other performances. In the following, we expand on this result by considering other classification and split scenarios.

Temporal Context. In our basic experiment, we considered isolated leitmotif instances as input to our classification model, i. e., the audio excerpts to be classified start and end strictly at instance boundaries. We therefore call this the *Strict* scenario. Identifying leitmotifs when instance boundaries are not known in advance could pose an additional challenge. However, the temporal context before and

[3] The same occurrences are chosen in all experiments for comparability.

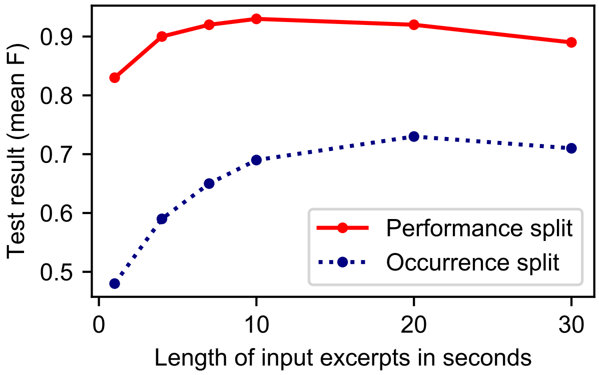


Figure 3. Mean F-measures for our model when using different input lengths in the *Fixed* scenario.

after the instance boundaries might also be helpful in identifying the class of an excerpt. Next, we analyze the effect of temporal context on the leitmotif classification results.

To this end, we compare the *Strict* scenario with an alternative, called *Variable*, where we add a randomly chosen amount of temporal context to the input excerpts. Context may be added before and after the motif instance. More specifically, the excerpt length is at most doubled and the instance in question is not constrained to be in the excerpt center. Such use of context also prevents our model from relying on length and boundary properties of the leitmotif instances. The center block in Table 4 gives results for this scenario. Compared to the *Strict* case, the mean F-measure decreases slightly to 0.92.

We also perform experiments on fixed input lengths, which we call the *Fixed* scenario. Here, we randomly take subsections of an instance if it is longer than the fixed input length or add context before and after in case it is shorter. Mean F-measure values for different fixed input lengths are shown in Figure 3 (solid red line). The plot indicates that results decrease for lengths that are shorter than most instances,[4] e. g., one second. When a fixed length of ten seconds is chosen, which encompasses almost all instances in the dataset, results are comparable to the *Strict* case (see also the right block in Table 4). Longer inputs again yield lower results, which may be attributed to the difficulty posed by additional context. However, one should note that for such large durations, input excerpts are no longer guaranteed to contain instances of a single motif only and thus, our initial assumption on a single label per input may be violated.

In Section 5, we discuss how the results for different amounts of temporal context may be interpreted in the context of a leitmotif detection scenario.

Potential for Overfitting. Deep learning models often rely on features of the input that would be deemed task-irrelevant by human experts, see, e. g., [29, 30]. In our case, the correct class for each input may be inferred not only from musically relevant aspects of leitmotifs such as melody or rhythm (as given in Table 1), but also from confounding features of the excerpts such as instrument activity or volume. This is especially true for the performance split, where a classification model may predict correct outputs on the test set by merely memorizing all occurrences during training instead of distinguishing musically relevant features of the leitmotifs (we will revisit this possibility in Section 4.6). In contrast, for the occurrence split, the model needs to generalize to previously unseen realizations of the leitmotif classes and therefore needs to rely on their common musical characteristics.

Context	*Strict*			*Variable*			*Fixed* (10 sec.)		
	P	R	F	P	R	F	P	R	F
L-Ni	0.67	0.80	0.73	0.67	0.86	0.75	0.80	0.91	0.85
L-Ri	0.36	0.41	0.38	0.44	0.43	0.43	0.49	0.67	0.56
L-Mi	0.79	0.87	0.83	0.82	0.80	0.81	0.97	0.96	0.97
L-NH	0.72	0.20	0.31	0.62	0.25	0.36	0.92	0.32	0.47
L-RT	0.57	0.65	0.61	0.60	0.77	0.68	0.71	0.91	0.80
L-Wa	0.87	0.80	0.84	0.81	0.88	0.84	0.95	0.95	0.95
L-WL	0.25	0.21	0.23	0.23	0.17	0.20	0.52	0.20	0.28
L-Ho	0.46	0.57	0.51	0.52	0.57	0.54	0.61	0.91	0.73
L-Ge	0.28	0.30	0.29	0.38	0.43	0.40	0.58	0.68	0.63
L-Sc	0.52	0.50	0.51	0.64	0.53	0.58	0.76	0.58	0.66
Mean	0.55	0.53	0.52	0.57	0.57	0.56	0.73	0.71	0.69

Table 5. Main results of our method on the test set of the occurrence split for different strategies of using temporal context.

4.4 Results on the Occurrence Split

Table 5 presents results for the occurrence split with different strategies for adding temporal context. Overall results are lower than for the performance split. In the *Strict* scenario, the obtained mean F-measure of 0.52 is substantially lower than for the performance split, but still well above chance (which corresponds to 0.1 mean F-measure). Results vary considerably among motifs, with F-measures ranging from 0.23 for L-WL to 0.84 for L-Wa. In addition, the differences between precision and recall per motif can be large as in the case of L-NH (P = 0.72 and R = 0.20). We conclude that classifying leitmotif instances for unknown occurrences is challenging but possible.

We further observe that—in contrast to the performance split—context is beneficial in the occurrence split. Mean F-measures of the *Variable* and *Fixed* scenarios increase to 0.56 and 0.69, respectively. Figure 3 shows F-measures for different amounts of context in the occurrence split (dotted blue line). Results increase for excerpt lengths up to ten seconds and then stabilize. We see two potential reasons for this. Firstly, by training with temporal context, the classifier may learn to identify features that indicate instance starts and ends, which could be helpful for identifying instances in the test set. Secondly, however, longer temporal context also means that instances from the training set may occur in the context added to validation and test instances. Indeed, we observed that for a context length of 10 seconds, 67% of test excerpts overlap with a training instance of the same class, while 8% overlap with a training instance of another class. Predicting the class of known training occurrences would therefore yield good results on

[4] Statistics on instance lengths are given in Table 1 (rightmost column).

Split	Performance			Occurrence		
	P	R	F	P	R	F
Noise	0.90	0.87	0.89	0.32	0.36	0.34
L-Ni	0.90	0.95	0.93	0.63	0.74	0.68
L-Ri	0.89	0.89	0.89	0.28	0.32	0.30
L-Mi	0.94	0.93	0.94	0.78	0.75	0.76
L-NH	0.95	0.88	0.91	0.52	0.28	0.37
L-RT	0.93	0.93	0.93	0.54	0.73	0.63
L-Wa	0.93	0.96	0.94	0.79	0.79	0.79
L-WL	0.94	0.93	0.94	0.17	0.12	0.14
L-Ho	0.89	0.87	0.88	0.40	0.45	0.42
L-Ge	0.91	0.91	0.91	0.20	0.18	0.19
L-Sc	0.90	0.95	0.93	0.68	0.38	0.49
Mean	0.92	0.92	0.92	0.48	0.46	0.46

Table 6. Results of our method when incorporating a noise class in the performance and the occurrence split. No temporal context is added (*Strict* scenario).

the test set. The results for adding temporal context may thus partly be explained by overfitting to the training set.

4.5 Noise Class

So far, we only considered excerpts that contain one of ten leitmotifs. However, the *Ring* also contains regions with other or with no leitmotifs at all. Because of this, we also perform experiments with an additional Noise class, denoting excerpts where none of the leitmotifs in our selection are being played. We evaluate whether our model is able to correctly classify our selection of leitmotifs in the presence of this noise class, both for the performance and the occurrence split. To do so, we randomly select 400 Noise occurrences from the *Ring*, leading to 6400 Noise instances. The model described in Section 3 remains unchanged except for the final classification layer, which now has eleven outputs.

Results are given in Table 6. For the performance split, the additional noise class does not change results by much. Leitmotif classes obtain somewhat lower results (e. g., P = 0.90 for L-Ni compared to P = 0.94 in Table 4) while the noise class yields an F-measure lower than most leitmotif classes (F = 0.89). For the occurrence split, results for the leitmotif classes again decrease slightly (e. g. P = 0.63 for L-Ni compared to P = 0.67 in Table 5), while the noise class itself is especially hard to distinguish (F = 0.34). In both splits, the noise class does not lead to a complete deterioration of results. Section 5 discusses the implications of this for the task of leitmotif detection.

4.6 Random Labels

In all experiments, our model has consistently obtained higher results on the performance than on the occurrence split. As discussed at the end of Section 4.3, the latter split requires generalizing to new musical realizations of a motif. In contrast, the performance split could be tackled by memorizing all leitmotif occurrences, which is not possible on the occurrence split.

To further investigate the gap in results between performance and occurrence split, we now evaluate our model's capability to memorize input features on the performance split. To do so, we create a variant of the performance split where we assign a random class label from one to ten to each occurrence. Thus, while occurrences are labeled consistently across performances, their classes no longer correspond to leitmotifs. In this variant of the performance split, the class of a test excerpt can only be obtained by memorizing classes for occurrences during training and not by learning common properties of all occurrences for a motif. This random-labeling experiment is inspired by [31].

When training our model on this variant, we obtain a mean F-measure of 0.54 on the test set after 50 epochs, which is much lower than the 0.94 obtained for the original labels (see Table 4). We observed that training for this experiment had not converged after 50 epochs and trained for an additional 75 epochs, leading to an F-measure of 0.57. The faster convergence and higher results on the original labels suggest that our model does learn some relevant characteristics of leitmotifs. Our experiment shows, however, that memorizing excerpts may also contribute to the results.

5. SUMMARY AND FUTURE WORK

In this work, we evaluated the capability of a neural network classification model for identifying leitmotifs in audio excerpts. Despite the complex musical variabilities in this scenario, our RNN-based classification model is able to differentiate between a fixed set of motifs and to distinguish them from non-motif excerpts. Generalization is strong across performances and—to a lesser extent—across occurrences. Using temporal context is helpful in the latter case, although the improvement may partly be the result of overfitting.

Our results encourage the development of a system for automated detection of motif instances in full performances. Unlike the classification task, no pre-segmented instance boundaries would be available for detection. We therefore expect this to be a more challenging scenario.

In our experiments, we have already explored the use of fixed input lengths. Using these, our model may be applied to all positions in an audio recording in a sliding window fashion [32]. This way, we can obtain leitmotif predictions for an entire performance of the *Ring* and not just individual excerpts. Additionally, a model used for automated leitmotif detection from audio would also need to deal with input excerpts that do not contain any leitmotifs at all. Our experiments with a noise class suggest that this may lead to somewhat lower but still useful results.

Furthermore, a detection system would need to handle a much larger number of motifs (around 130 for the complete *Ring*) as well as excerpts containing multiple motifs played simultaneously. Multi-label extensions of our model on fixed input lengths may be suitable for this.

As an even more advanced scenario, one may imagine an informed detection setting in which instances of a previously unseen motif must be identified given only a few exemplary instances of that motif.

Acknowledgments: We thank Vlora Arifi-Müller for her assistance in preparing the data. This work was supported by the German Research Foundation (DFG MU 2686/7-2, 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.

6. REFERENCES

[1] H. M. Brown, E. Rosand, R. Strohm, R. Parker, A. Whittall, R. Savage, and B. Millington, "Opera," in *The New Grove Dictionary of Music and Musicians*, 2nd ed., S. Sadie, Ed. London: Macmillian Publishers, 2001, pp. 416–471.

[2] M. Bribitzer-Stull, *Understanding the Leitmotif.* Cambridge University Press, 2015.

[3] V. Konz, M. Müller, and R. Kleinertz, "A cross-version chord labelling approach for exploring harmonic structures—a case study on Beethoven's Appassionata," *Journal of New Music Research*, vol. 42, no. 1, pp. 61–77, 2013.

[4] H. Schreiber, C. Weiß, and M. Müller, "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)*, Barcelona, Spain, 2020.

[5] C. Weiß, F. Zalkow, M. Müller, S. Klauk, and R. Kleinertz, "Computergestützte Visualisierung harmonischer Verläufe: Eine Fallstudie zu Wagners Ring," in *Proceedings of the Jahrestagung der Gesellschaft für Informatik (GI)*, Chemnitz, Germany, 2017, pp. 205–217.

[6] F. Zalkow, C. Weiß, and M. Müller, "Exploring tonal-dramatic relationships in Richard Wagner's Ring cycle," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Suzhou, China, 2017, pp. 642–648.

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

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

[9] F. Zalkow, C. Weiß, T. Prätzlich, V. Arifi-Müller, and M. Müller, "A multi-version approach for transferring measure annotations between music recordings," in *Proceedings of the AES International Conference on Semantic Audio*, Erlangen, Germany, 2017, pp. 148–155.

[10] D. J. Baker and D. Müllensiefen, "Perception of leitmotives in Richard Wagner's Der Ring des Nibelungen," *Frontiers in Psychology*, vol. 8, p. 662, 2017.

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

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

[13] D. Müllensiefen, D. Baker, C. Rhodes, T. Crawford, and L. Dreyfus, "Recognition of leitmotives in Richard Wagner's music: An item response theory approach," in *Analysis of Large and Complex Data.* Cham, Switzerland: Springer, 2016, pp. 473–483.

[14] R. Wagner, *Opera and Drama.* University of Nebraska Press, 1995, translation of the original edition from 1851.

[15] L. Dreyfus and C. Rindfleisch, "Using digital libraries in the research of the reception and interpretation of Richard Wagner's leitmotifs," in *Proceedings of the International Workshop on Digital Libraries for Musicology*, London, UK, 2014, p. 1–3.

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

[17] ——, "On the application of music to the drama," in *Prose Works.* Broude Brothers, New York, 1966, pp. 175–191, translation of the original edition from 1879.

[18] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, London, UK, 2005, pp. 628–633.

[19] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 637–644.

[20] F. Korzeniowski and G. Widmer, "Genre-agnostic key classification with convolutional neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 264–270.

[21] I. Jeong and K. Lee, "Learning temporal features using a deep neural network and its application to music genre classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, 2016, pp. 434–440.

[22] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[23] C. Schörkhuber and A. P. Klapuri, "Constant-Q transform toolbox for music processing," in *Proceedings of the Sound and Music Computing Conference (SMC)*, Barcelona, Spain, 2010, pp. 3–64.

[24] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *Proceedings the Python Science Conference*, Austin, Texas, USA, 2015, pp. 18–25.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

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

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

[29] B. L. Sturm, "The "horse" inside: Seeking causes behind the behaviors of music content analysis systems," *Computers in Entertainment*, vol. 14, no. 2, pp. 3:1–3:32, 2016.

[30] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 125–136.

[31] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[32] P. Grosche and M. Müller, "Toward characteristic audio shingles for efficient cross-version music retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 473–476.