With recent advances in the field of music informatics, approaches to audio-based music structural analysis have matured considerably, allowing researchers to reassess the challenges posed by the task, and reimagine potential applications. We review the latest breakthroughs on this topic and discuss the challenges that may arise when applying these techniques in real-world applications. Specifically, we argue that it could be beneficial for these systems to be application-dependent in order to increase their usability. Moreover, in certain scenarios, a user may wish to decide which version of a structure to use, calling for systems with multiple outputs, or where the output adapts in a user-dependent fashion. In reviewing the state of the art and discussing the current challenges on this timely topic, we highlight the subjectivity, ambiguity, and hierarchical nature of musical structure as essential factors to address in future work.

Keywords: Music Informatics Research; Music Structure Analysis; Music Signal Processing

1. Introduction
The field of Music Informatics Research (MIR) has experienced significant advances in recent years, helped by more powerful machine learning techniques (Humphrey et al., 2012), greater computation (Dieleman et al., 2018), larger and richer datasets (Gemmeke et al., 2017), and increased interest in applications (Schedl et al., 2014; Murthy and Koolagudi, 2018). Thanks to this, several areas in MIR have advanced quickly, allowing researchers to reconsider what is possible from a more mature perspective. One such area is the widely discussed audio-based Music Structure Analysis (MSA) (Paulus et al., 2010).

The basic premise of MSA is that any song can be divided into non-overlapping segments, each with a label defining its segment type, and that this segmentation and labelling can characterize a human's perception or analysis of the song. The task originates from a common practice in Western music theory: analyzing the form of a piece of music by identifying important segments, whether at short time scales (e.g., motives, which are short musical ideas that tend to recur across a piece) or longer parts (e.g., the exposition of a sonata, or the verse, chorus and bridge sections of a pop song). Music experts and non-musicians alike perceive music as consisting of distinct segments, and while identifying these segments is a highly subjective task—several structural annotations might be valid for a given piece—there is often broad agreement between listeners about which segment boundaries are more important (Bruderer et al., 2009; Wang et al., 2017). Accordingly, those studying music perception and cognition have proposed theoretical models of how segments are perceived (Lerdahl and Jackendoff, 1983), some of which have been implemented as algorithms (Cambourousopoulos, 2001; Hamanaka et al., 2006; Groves, 2016).

Within MIR, many are interested in automating the task of MSA for the sake of testing and refining music theoretic and music perceptual models. But the interests of MIR go further, and a structural analysis of audio content representing a music track, as a high-level ‘map’ or ‘outline’ of the content of a song, has many applications. Given the unprecedented size of several commercial and independent music catalogs, the task of MSA has the potential to enhance the final user experience when understanding, navigating, and discovering large-scale collections. On the other hand, this promise has been advertised for over a decade (Goto, 2006b), but so far the only obvious commercial applications of MSA have been for thumbnail creation and music-related video games. This motivates the focus of this work on the most persistent challenges of MSA, since we believe that if properly addressed, further applications could be derived from this topic in several other areas such as music
creation and production, music recommendation, music generation, and musicology. Moreover, computational MSA may help to improve many MIR tasks, such as making chord and downbeat estimates more robust (Mauch et al., 2009b; Fuentes et al., 2019b).

In this article we review the state of the art of this timely topic and discuss current open challenges, with an emphasis on subjectivity, ambiguity, and structural hierarchies. While we acknowledge that MSA may also be performed on symbolic representations of music (Janssen et al., 2013), in this work we focus exclusively on the audio-based approaches, since they have dramatically advanced in the past two decades and their applicability to real-world scenarios is broader. In the subsequent sections, the term MSA is used specifically to refer to computational audio-based MSA. Moreover, we present a list of applications that mature MSA algorithms could help to realize. It is our hope that this article inspires new and seasoned researchers in this field to focus on the areas that may advance this task even further.

The outline of this paper is as follows: Section 2 reviews the state of the art for audio-based MSA, including methods, principles, evaluation, datasets, and state-of-the-art performance; Section 3 discusses current trends and challenges with a special attention on subjectivity, ambiguity, and hierarchy; and in Section 4 we review the potential applications that MSA could enhance and/or inspire. Lastly, we draw conclusions in Section 5.

2. The Music Structure Analysis Problem

Audio-based MSA aims to identify the contiguous, non-overlapping musical segments that compose a given audio signal, and to label them according to their musical similarity. These segments may be identified at different time scales: a short motive may only last a few seconds, while a large-scale section encompassing several long fragments may last longer than a minute. When an analysis consists of multiple segmentations, describing the structure at more than one time scale, we call it a hierarchical analysis. Deeper levels of such hierarchies tend to subdivide the segments of the levels above, but defining a completely different (and finer) set of segments in lower levels is also considered valid. While we can formally define this problem (see 2.1), MSA is often regarded as challenging due to the ambiguity of both the exact placement of the boundaries that define such segments (Bruderer, 2008; Wang et al., 2017), and the quantification of the degree of similarity amongst them (Nieto and Bello, 2016). Given that only the most recent approaches have focused on identifying hierarchical structures, in this section we exclusively focus on the flat (i.e., non-hierarchical) ones, leaving the discussion of hierarchy for Section 3.3.

Figure 1 depicts a visual example of the flat structure analysis of track number 10 from the SALAMI dataset (Smith et al., 2011). In this case, four different types of large-scale segments (i.e., A, B, C, C’) plus the additional “Silence” label at the end of the track have been identified by the expert who annotated it. Using letters to label the sections is a common practice in music theory, as is the practice of using a prime symbol (’) to denote a repetition that is varied in an important way. Thus, this expert deemed that C and C’ are related, but not similar enough to be the same segment. Segments A and B each repeat three times, while segments C and C’ appear only once each. To simplify the problem, researchers mandate that these segments be non-overlapping for a given hierarchical level, even though in some musicological approaches these could, theoretically, overlap (Lerdahl and Jackendoff, 1983).

2.1 Problem Definition

MSA has been an ill-defined problem since its very introduction: subjectivity, ambiguity, and lack of data contribute to make this task particularly hard to define. We will discuss such drawbacks in Section 3. Nevertheless, as its current form, this problem can be formally framed...
as follows: a flat structural analysis is defined as a set of temporally contiguous, non-overlapping, labeled time intervals, which span the duration of an audio signal. Given a musical recording of T audio samples, a flat segmentation is fully defined by a set of segment boundaries \(B \subseteq \{2, \ldots, T\}\), a set of \(k\) unique labels \(Y = \{y_1, \ldots, y_k\}\), and a mapping of segment starting points to labels \(S: \{1\} \cup B \rightarrow Y\). From this, we can derive the label assignment \(L: \{1, \ldots, T\} \rightarrow Y\) that assigns a label to each time point: \(L(t) = S(\max(\{1\} \cup \{t, \in B | t \leq t\}))\). Note that segment boundaries do not imply a label change; there can well be consecutive segments with the same label.

It is standard to set the sampling rate of the time points as points of resolution and computational efficiency (Raffel et al., 2014); this value is employed in the rest of this work.

Following the example of Figure 1, we have \(k = 5\) unique labels: \(Y = \{A, B, C, C', \text{Silence}\}\). Moreover, we can identify various time points that share the same assigned label, such as \(L(1) = L(1000) = L(2000) = A\) or \(L(700) = L(1500) = L(2500) = B^3\). In this example, segment boundaries are found whenever two consecutive time points are labeled differently, e.g., \(L(530) \neq L(531) \Rightarrow 531 \in B\). Naturally, boundaries attract special interest: each indicates a critical moment in the music signal where a perceived cue determines the actual segmentation. Such cues may be directly related to sonic events occurring at or in the vicinity of the boundary: in the top right spectrogram, it is clear that the sound evolves differently on either side of the boundary. However, many cues are not directly relatable to the spectrogram; for example, the boundary between two sections with the same label may be inferred by the listener through parallelism, rather than some abrupt change in the signal. Moreover, sonic cues may conflict with each other, some suggesting continuity, others a boundary; the problem of ambiguity is further discussed in Section 3.1.

### 2.2 Segmentation Principles

Three main principles were initially identified when segmenting music: homogeneity, novelty, and repetition (Paulus et al., 2010). Later on, Sargent et al. (2011) employed a fourth principle: regularity. In this subsection we discuss them and argue that homogeneity and novelty can be, in practice, exploited similarly.

We make use of a standard tool in MSA to represent a track such that its structure might become more visually apparent: a self-similarity matrix (SSM). Each position of the SSM represents the degree of similarity between two audio frames, thus resulting in a square matrix whose diagonal always contains the highest degree of similarity. In Figure 2 we depict an SSM mock up of the track discussed above that will help us illustrate the main segmentation principles.

#### 2.2.1 Homogeneity and Novelty

In the homogeneity approach, it is assumed that musical segments are relatively homogeneous with respect to some musical attribute (e.g., key or instrumentation), meaning that boundaries between dissimilar segments are detectable as points of novelty. Therefore, the novelty principle implicitly assumes that the music is locally homogeneous on either side of a boundary (Ullrich et al., 2014). In this work we treat these two principles indistinctly.

Boundaries of homogeneous segments may be straightforward to identify if sudden musical changes are present in the track to be analyzed. These differences in terms of continuation of a given musical attribute might appear as blocks in the SSM. In Figure 2, we see several blocks that tend to correlate with the annotated boundaries. Note that repeated segments (e.g., the first two of the example—A, A), will not be retrievable by this principle. Homogeneous blocks can be hard to subdivide unless one makes use of recurrence sequences, as described next.

#### 2.2.2 Repetition

The repetition principle assumes that segments with the same label are similar sequences—again, with respect to some musical attribute. The boundaries of repeated segments may only be discovered as a by-product of such repetitions; i.e., the boundaries would be defined by the start and end points of the repeated sequences. As opposed to the previous principle, the repeated segments are not necessarily consistent across their entire segment, but with their respective repeated segment(s) in a sequential fashion.

In Figure 2, the repeated sequences appear as perfect paths (i.e., diagonals). Approaches that aim to identify repeated segments focus on such paths, which often only become prominent if analyzing a specific musical attribute (unless, of course, there are acoustically exact repetitions in the track). As it can be seen in the B segments of the example, there are often cases where segments can not be extracted using the repetition principle.

#### 2.2.3 Regularity

Frequently, musical segments hold a certain degree of regularity that this principle aims at exploiting, as described by Sargent et al. (2011). For instance, the duration of two equally labelled segments tends to span an integer ratio number of beats. Moreover, segment lengths tend to be
log-normally distributed across the track, regardless of the musical genre or the level of annotation in a potentially hierarchical structure (Smith and Goto, 2016).

Back to our example in Figure 2, we see that the regularity principle would be helpful at identifying the sequence B, C. Without this principle, the annotator might define such a sequence as a single segment, which does not follow the regularities discussed above. Note that it is not apparent how to employ the homogeneous and/or repetitive approaches to fully identify these two segments.

### 2.2.4 Combining Principles

These three principles are useful, since in real examples segments may be homogeneous with respect to one attribute (e.g., instrumentation), also characterized by a unique sequence (e.g., a distinct chord progression for the chorus that is different from the verse), and hold certain regularities across their segment lengths (e.g., a particularly long bridge section that could potentially be subdivided into several segments). Thus, it is not uncommon to employ a combination of such principles to determine the structure of a piece of music.

In the example above (readers are encouraged to listen to it), the annotator placed a boundary between A and B at time 0:53.1, perhaps based on the novel appearance of vocals and the drastic transient of loud cymbals (see right side of Figure 1); in addition, the B segments have consistent harmony (a single cycle of chords), and although this cycle is the same as in A, the instrumentation throughout B is consistent and different from A. All of this fits with the homogeneity/novelty principle. As for the repetition principle, the repeated melody played by the synthesized strings in the C and C’ parts was likely influential in this annotation. Finally, we see that segments B and C could easily be grouped as a single segment, but due to the regularity principle, the annotator likely decided to split this potentially longer segment into two.

These principles are not necessarily exhaustive, and others may be identified in the future to better narrow the definition of this task. Moreover, these discussed principles have not been formally defined, which makes this task highly subjective and ambiguous. While these topics will be addressed in Section 3, what follows is a list of the main computational approaches to MSA, some of which clearly employ a combination of such principles.

### 2.3 Computational Methods

Here we review the standard approaches to MSA, largely focused on the advances of the last 10 years. For a survey of more classical approaches from the mid and early 2000s, we suggest the work by Dannenberg and Goto (2008). In this section we divide the methods into the identification of boundaries and the labeling of the segments they define. As we will see, the principles described above can be used to address either of these two subtasks. In practice, the starting point of all these methods involves feature extraction from raw audio signals. These methods are typically tuned and optimized employing manually annotated datasets (discussed in Section 2.5). While most early methods focused on harmonic (e.g., chromagram and/or timbral (e.g., mel-frequency cepstral coefficient) features, it has been more recently shown that compacted spectral representations (e.g., constant-Q transforms, log-mel spectrograms) tend to yield superior results when used (if possible) in any of the algorithms described below (Nieto and Bello, 2016). As previously mentioned, the following methods focus on flat segmentations exclusively.

#### 2.3.1 Music Segmentation Methods

The checkerboard kernel technique is, despite being one of the first proposed for this problem, relevant due to its simplicity and effectiveness. It is based on the homogeneity principle (Foote, 2000), where a kernel with a checkerboard-like structure (i.e., four quadrants: two positive and two negative, whose duration will determine the amount of context) is convolved over the main diagonal of an SSM. This yields a novelty curve highlighting sudden changes in the selected musical features from which to extract the boundaries by identifying its more prominent peaks. Such a checkerboard kernel may be binary or Gaussian depending on the desired novelty curve smoothness. The peaks in the novelty curve tend to correlate with annotated segment boundaries. As an example, Figure 3 shows an SSM computed from a mel-scaled spectrogram and its associated novelty curve, both marked with the annotated boundaries. As it can be seen, these boundaries tend to follow the structure of the SSM and also the peaks in its novelty curve. More sophisticated approaches based on homogeneity include the use of supervised learning (Turnbull et al., 2007) or lag matrices (Goto, 2003). Lag matrices represent the similarity of each time step to each of the K previous time steps, i.e., their rows correspond to K appropriately padded diagonals above the main diagonal of an SSM. This allows to detect repetitions (appearing as horizontal lines, instead of the diagonals in the SSM) within a limited context. A more recent technique that yielded state-of-the-art results in certain metrics combines the homogeneity and repetition principles by a simple rotation of the lag matrix, yielding the so-called structural features (Serrà et al., 2014). These features can be used to produce a novelty curve from which to extract the segment boundaries.

Due to the often required pre-processing of the features, the checkerboard kernel and structural feature techniques estimate boundaries that may be located within one or more seconds away from the reference ones. This problem can be addressed by employing features that are synchronized to estimated beats (i.e., the main rhythmic units of a given music piece), thus yielding one feature vector per beat. Having such beat-synchronous features makes the repeated sequences easier to extract from the SSM, since they become perfect diagonals. Figure 3 depicts such features, which yield prominent diagonal structures that can be seen in the C, C’ blocks.

Such beat-synchronous features can be helpful in several scenarios. For example, the supervised technique of ordinal linear discriminative analysis is directly applied to the structural features and yields more precise results...
Accordingly, it has been shown that the combination of the structural features with the lag matrix techniques improves results even further (Peeters and Bisot 2014). Moreover, beat-synchronous features can be more helpful when employing the regularity principle. For example, Maezawa (2019) employs this principle in combination with the rest to identify boundaries by explicitly incorporating statistical properties regarding the regularity of the detected segments. Furthermore, Pauwels et al. (2013) introduce a method that makes use of these beat-synchronous features to jointly estimate key, chords, and segment boundaries.

Finally, deep convolutional architectures have proven highly successful for this task, yielding superior scores in most metrics, mostly employing the homogeneity principle (Ullrich et al., 2014; Grill and Schlüter, 2015a). Furthermore, a recent work by McCallum (2019) reached state-of-the-art results by learning audio features in an unsupervised, representation learning fashion with convolutional neural networks optimized with a triplet loss. Once these deep audio features are learned, the traditional checkerboard kernel technique defined above is used to identify segment boundaries, thus showing the importance of having high-quality audio features when analyzing music structure, even when only the homogeneity segmentation principle is employed. It is still to be explored how to combine such a principle with all the rest when using deep architectures to identify segment boundaries. Moreover, all of these methods focus on the retrieval of a single set of boundaries per track, which significantly differs with the degree of variability when several annotators analyze a given track. Such ambiguities will be discussed in Section 3.

2.3.2 Music Segment Similarity Methods
Quantifying the similarity between segments defined by structural boundaries can be framed under the audio similarity problem. As such, several methods have been proposed, some of the most relevant using Gaussian mixture models on a pool of low-level audio features in a supervised approach (Wang et al., 2011), a variant of nearest neighbor search on a multidimensional Gaussian space applied to timbral features (Schnitzer et al., 2011), and non-negative matrix factorization (NMF) to cluster the homogeneous blocks of an SSM (Kaiser and Sikora, 2010). The current state-of-the-art when employing reference annotations for this task is the technique based on 2D Fourier Magnitude Coefficients (2DFMC) (Nieto and Bello, 2014). This method projects harmonic features into the 2DFMC representation, which allows for both key and temporal shift invariance, and can be further clustered efficiently via adaptive k-means to label the resulting segments. State-of-the-art segment similarity when estimating boundaries is achieved by combining the structural features described above with techniques employed for cover song identification (Serrà et al., 2009), as proposed by Serrà et al. (2014). As in the case of boundary retrieval, identifying the degree of similarity between segments is a highly subjective and ambiguous task, and thus none of these methods obtain close to perfect scores in any of the publicly available datasets as we will see in Section 2.6. We will further discuss these challenges in Section 3.

2.3.3 Segmentation and Labeling Methods
While many methods first estimate boundaries and then label segments, this can be susceptible to error propagation if mistakes are made in the boundary retrieval step. This observation has motivated methods to jointly identify boundaries and segment labels, and here we mention the most relevant ones.

Levy and Sandler (2008) propose to encode audio frames as states of a hidden Markov model (HMM) trained on audio features. The most likely sequence of states is later clustered into segments, thus obtaining both boundaries and labels simultaneously by employing the homogeneity and
regularity principles. Other prior work employing HMMs can be found by Logan and Chu (2000) and Peeters et al. (2002). Paulus and Klapiuri (2009) present a probabilistic fitness measure that is able to yield the likelihood of having a specific segmentation given a music track. They apply a greedy search algorithm to avoid the intractable problem of computing all possible segmentation combinations. Kaiser and Peeters (2013) present a method that fuses homogeneity-based representations with repetition-based ones and yields an SSM that can later be used to both extract boundaries and cluster labels with a mixture of techniques like the checkerboard kernel and NMF described above. Weiss and Bello (2011) demonstrate that a probabilistic version of convolutive NMF can be successfully applied to identify boundaries and label segments. Nieto and Jehan (2013) propose a convex variant of NMF that surpasses previous NMF-based approaches. The principle of homogeneity tends to be favored in these matrix factorization techniques, since it is common to apply certain degrees of aggregation (e.g., median, mean) of the audio features for a given segment in order to capture their similarity.

A method that exploits the repetition principle by converting the blocks of an SSM into paths has also been presented (Cheng et al., 2018). Furthermore, Panagakis and Kotropoulos (2012) discuss the fusion of several audio features by employing ridge regression and then obtaining the final segmentation with spectral clustering. McFee and Ellis (2014a) also make use of spectral clustering, in this case applied to a set of features optimized to enhance the repetition in the piece and the affinity of different musical attributes such as harmony and timbre. Such spectral clustering techniques tend to favor the repetition principle.

Additionally, Sargent et al. (2017) propose a technique that specifically employs the regularity principle by constraining the sizes of the final segmentation. Moreover, a method that employs the homogeneity, repetition, and regularity principles using a Bayesian approach yields state-of-the-art results when combining the two subtasks of segmentation and labeling (Shibata et al., 2019). Finally, certain techniques such as spectral clustering (McFee and Ellis, 2014a) are also capable of discovering smaller segments such as riffs and motives, and therefore producing hierarchical outputs. These will be further discussed in Section 3.3.

Interestingly, the authors are not aware of any end-to-end methods employing the latest advances in deep architectures to jointly identify boundaries and label the segments. While such architectures have become the trend in several MIR tasks (e.g., music tagging (Pons et al., 2018), onset detection (Kim and Bello, 2019), beat-tracking (Fuentes et al., 2019a), chord recognition (Chen and Su, 2019)), it remains to be seen how the latest advances in machine learning will be applied to MSA.

2.4 Evaluation
In this section we review standard techniques to evaluate boundary or labeling agreement between a flat reference segmentation and its respective estimation. All the metrics described here are implemented in the open source package mir_eval (Raffel et al., 2014). Note that these metrics make use of a single set of segments as reference, which collides with the fact that multiple annotators might yield different segmentations. Therefore, these evaluations are inherently limited given the ambiguity and subjective nature of this task, which we will discuss in Section 3.2.

2.4.1 Segment Boundaries
The most established metric to assess the quality of a set \( B^* \) of estimated segment boundaries against reference boundaries \( B^\mathcal{R} \) is the so-called Hit Rate measure (Turnbull et al., 2007; Levy and Sandler, 2008). The set of hits \( \mathcal{H} \) may be defined as follows:

\[
\mathcal{H}(B^*, B^\mathcal{R}) = \{(i, j) \in B^* \times B^\mathcal{R} \mid |i-j| \leq \varepsilon \}
\]

\[
\text{s.t. } (i, j) \in \mathcal{H} \Rightarrow \forall (k, j) \notin \mathcal{H}
\]

\[
\text{and } (i, j) \in \mathcal{H} \Rightarrow \forall (i, k) \notin \mathcal{H}
\]

where \( \varepsilon \) is a tolerance parameter typically set to 0.5 (Turnbull et al., 2007) or 3 (Levy and Sandler, 2008) seconds. The hit rate combines two different statistics: (i) the precision \( P_{\mathcal{H}} \), representing the proportion of estimated boundaries that constitute a hit, formally \( P_{\mathcal{H}} = |\mathcal{H}(B^*, B^\mathcal{R})|/|B^*| \), and (ii) the recall \( R_{\mathcal{H}} \), which is the proportion of reference boundaries that were hit, \( R_{\mathcal{H}} = |\mathcal{H}(B^*, B^\mathcal{R})|/|B^\mathcal{R}| \). These two values are further combined using the harmonic mean, also known as the F1 measure:

\[
F_1(P, R) = 2 \frac{P \cdot R}{P + R}
\]

which the hit rate parametrizes as \( F_1(P_{\mathcal{H}}, R_{\mathcal{H}}) \).

Perceptually, it has been shown that precision has a higher relevance than recall (Nieto et al., 2014), i.e., there seems to be a cognitive preference to estimate fewer but correct boundaries than more but less accurate ones. To address this potential problem, one may use the weighted hit rate \( F_\alpha \) measure, as follows:

\[
F_\alpha(P, R) = \left(1 + \alpha^2\right)^{-1} \frac{P \cdot R}{\alpha^2 P + R}
\]

where \( \alpha < 1 \) emphasizes precision and whose parameters are \( P_{\mathcal{H}} \) and \( R_{\mathcal{H}} \).

The Median Deviation is another previously introduced technique, but it often overlooks boundary outliers when several boundaries have to be assessed (Smith and Chew, 2013). Thus, the hit rate measure tends to be preferred, especially when employing small tolerance parameters such as 0.5 seconds. We refer the reader to Turnbull et al. (2007) for more information about the median deviation scores.

2.4.2 Segment Labeling
The evaluation of label agreement tends to operate at a frame level, similar to clustering metrics. One of the most standard techniques is the so-called Pairwise Clustering (Levy and Sandler, 2008), where the set \( A \) of equally labeled time frame pairs \((i, j)\) for a given label assignment \( L \) is computed as follows:
\[ A(S) = \{(i,j) \mid L(i) = L(j)\} \] (4)

From the intersection \( I_P = |A(L^P) \cap A(L^S)|\), we can compute two scores: precision \( P_P = I_P / |A(L^P)| \) representing the proportion of correct label estimations, and recall \( R_P = I_P / |A(L^S)| \) quantifying the reference labels successfully found in the estimated ones. Analogous to the hit rate boundary evaluation, these values can be further combined by using the harmonic mean: \( F_1(P_P, R_P) \). This metric tends to be overly sensitive to exact boundary placement between reference and estimation (Nieto and Bello, 2014).

The Normalized Conditional Entropy scores (Lukashevich, 2008) address this by taking a probabilistic approach. The first conditional entropy \( \mathbb{H}(P^S \mid P^P) \) indicates the amount of information required to explain the estimated label distribution \( P^S \) given the reference one \( P^P \). By swapping the marginal distributions, we obtain \( \mathbb{H}(P^P \mid P^S) \), which can be explained analogously. Intuitively, the more similar these distributions are, the closer to zero the conditional entropy will be.

These conditional entropies can be further normalized by dividing by the maximum entropy annotation \( \log|Y| \) and \( \log|Y| \), respectively, yielding the over- and under-segmentation scores:

\[
S_o = 1 - \frac{\mathbb{H}(P^P \mid P^S)}{\log|Y|} \quad \text{(5)}
\]

\[
S_u = 1 - \frac{\mathbb{H}(P^S \mid P^P)}{\log|Y|} \quad \text{(6)}
\]

such that they reside in a \([0, 1]\) range where 1 becomes the highest degree of similarity.

These scores can be artificially inflated due to the potential lack of uniformity in the marginal distributions \( P^\ell \). Thus, it has been proposed\(^6\) to normalize over the marginal entropies \( \mathbb{H}(P) \) instead, resulting in the following V-measures (Rosenberg and Hirschberg, 2007) that should allow a fairer comparison across multiple tracks:

\[
V_o^\ell = 1 - \frac{\mathbb{H}(P^P \mid P^S)}{\mathbb{H}(P^\ell)} \quad \text{(7)}
\]

\[
V_u^\ell = 1 - \frac{\mathbb{H}(P^S \mid P^P)}{\mathbb{H}(P^\ell)} \quad \text{(8)}
\]

Intuitively, when the over-segmentation metrics \( S_o \) and \( V_o^\ell \) are small, the reference labels are not able to convey the estimated ones, i.e., \( \mathbb{H}(P^P \mid P^S) \) is large. On the other hand, small under-segmentation metrics \( S_u \) and \( V_u^\ell \) translate into a substantial amount of information needed to explain the reference labels from the estimated ones, i.e., \( \mathbb{H}(P^S \mid P^P) \) is large. The over- and under-segmentation metrics can be merged into a single score with the \( F_1 \) measure, resulting in \( F_1(S_o, S_u) \) for the original normalized conditional entropies and \( F_1(V_o^\ell, V_u^\ell) \) for the V-measures.

### 2.4.3 Hierarchy evaluation

The boundary and label evaluation metrics described above are designed to compare flat segmentations, and implicitly assume that the two segmentations in question operate at the same scale. However, when asked to produce a flat annotation, human annotators often produce results of varying specificity, sometimes corresponding to differences in attention to musical cues or time scales. Hierarchical segmentation evaluation seeks to remedy this by combining multiple segmentations into a unified structure:

\[ H = (S_1, S_2, S_3, \ldots) \] (9)

where \( S_\ell \) is an implicit “null” segment mapping the entire time-series to a single label, and subsequent segmentations \( S_2, S_3, \ldots \) provide progressively finer detail.

McFee et al. (2017) defined the L-measure as a generalization of the pairwise classification metric described above to the hierarchical case. Rather than seeking pairs of frames \((i, j)\) which receive the same label, the L-measure seeks triples of frames \((i, j, k)\) where the pair \((i, j)\) receive the same label deeper in the hierarchy than the pair \((i, k)\).

\[
M(i, j) = \arg \max_{\ell} \{S_\ell(i) = S_\ell(j)\} \quad \text{(10)}
\]

\[
A(H) = \{i, j, k \mid M(i, j) > M(i, k)\}. \quad \text{(11)}
\]

This intuitively relaxes the strict equality test of (4) to a relative comparison, facilitating comparison between hierarchies \( H^P \) and \( H^F \) of arbitrary (and differing) depths. Precision, recall, and \( F_1 \) scores are computed analogously to the pairwise metrics by comparing the triplet sets \( A(H^P) \) and \( A(H^F) \).

One must rely on any music structure analysis metric with certain skepticism given the degree of subjectivity and ambiguity in this task (discussed in Sections 3.1 and 3.2, respectively) and the perceptual preferences on different types of segmentations (Nieto et al., 2014). Nevertheless, these metrics are used to assess the music segmentation task of the yearly MIR evaluation exchange (MIREX).\(^7\) This MIREX task focuses on the flat segmentation problem exclusively, and is evaluated against several datasets ranging from pop to classical music. The most relevant and openly available datasets are described next. The best MIREX performances are reported in Section 2.6.

### 2.5 Datasets

Several human-labeled datasets are publicly available to train and/or assess automatic music structure analysis approaches. In this section, we enumerate them and describe their specific peculiarities. While studies on how computational MSA performs on different musical genres are available (Tian and Sandler, 2016), Western popular music tends to outnumber other genres in most of these sets. All of these datasets are conveniently available at a single URL\(^8\) under the same JAMS format (Humphrey et al., 2014). Moreover, a subset of the discussed datasets are also contained in the recently published mirdata...
2.5.1 SALAMI
The Structural Annotations for Large Amounts of Music Information (SALAMI) dataset is the largest publicly available set, containing hierarchical annotations for 1,359 tracks (Smith et al., 2011). These tracks are reasonably diverse and can be divided into five different classes of music: classical, jazz, popular, world, and live music. References are available from at least two distinct annotators for 884 of the tracks. A total of 10 music experts annotated the entire dataset at three different hierarchical levels: (i) the fine level, corresponding to short motives or phrases, (ii) the coarse level, representing larger fragments, and (iii) the functional level, which adds semantic labels (e.g., verse, bridge) to these larger fragments, hence containing sets of boundaries that typically overlap with those from the coarse level (see Figure 4). The annotators were asked to listen to the given track twice: first, to mark the timestamps whenever a boundary is identified and second, to adjust the boundaries and to label the different levels of the segments in the hierarchy. 171 annotations were recently corrected (McFee et al., 2017) and can be found online. A subset of 253 SALAMI tracks are freely available in the Internet Archive.

2.5.2 The Harmonix Set
The largest publicly available dataset including beats, downbeats, and flat musical segment human annotations is The Harmonix Set (Nieto et al., 2019). This dataset is mainly focused on Western popular music such as hip hop, dance, rock, and metal, and it contains annotations for 912 tracks. Given that beat and downbeat annotations are also available, this set can help develop systems that might combine several MIR tasks to yield potentially superior results (as discussed in Section 4.6). The available segmentation data contains flat boundaries and functional labels, and was annotated and revised by musical experts. A single segment annotation is available per track. The annotations were logged as follows: first, a tempo track was created for each song in a Digital Audio Workstation software (e.g., Logic Pro). Then, beats, downbeats, and segments were added into the tempo track. Therefore, the segment boundaries in this collection always fall on an annotated beat.

2.5.3 SPAM
The Structural Poly-Annotations of Music (SPAM) dataset is composed of 50 tracks automatically chosen such that 45 of them are meant to be difficult to segment, while the rest should be fairly simple for this task (Nieto and Bello, 2016). Besides the sampling of the songs, the most interesting feature of this set is its high number of annotators: at least five different hierarchical annotations are provided per track. The five annotators were music students (four graduates and one undergraduate) from the Steinhardt School at New York University, with an average number of years in musical training of $15.3 \pm 4.9$, and with at least 10 years of experience as players of a musical instrument. These annotations are available for the coarse and fine levels, following the same guidelines as in SALAMI. Moreover, 32 of the tracks available in SPAM overlap with those in the SALAMI set, therefore these tracks contain an extra annotation for each level, i.e., those originally contained in SALAMI.

2.5.4 RWC
The Real World Computing (RWC) dataset, which is also known as the AIST Annotations (Goto, 2006a), contains 300 tracks annotated with beats, melody lines, and flat structural segments. The music style ranges from pop to classical, including a large jazz subset. One particularity of this set is that all of its music is copyright-cleared, such that researchers can freely obtain the exact audio content.

Figure 4: Example of a hierarchical structure annotation from annotator 4 of track 10 in SALAMI. The functional level is plotted on top. In the middle, the coarse level is shown, with notable differences from those segmentations plotted in Figure 1 due to annotators disagreements. In the bottom, the fine level is displayed.
used to produce these annotations. Boundaries are always placed on beat positions, and the annotations were gathered by a single music college graduate with perfect pitch using an undisclosed multipurpose music-scene labelling editor.

2.5.5 Isophonics
This dataset was originally gathered by the Centre for Digital Music (C4DM) of Queen Mary University of London (Mauch et al., 2009a). It is composed of 300 singly-annotated tracks of Western popular music with flat, coarse segmentation information. The type of music is mostly pop-rock, including the entire Beatles catalog, the greatest hits by Michael Jackson and Queen, and two additional albums by Carole King and Zweieck. Furthermore, beat and downbeat annotations are also available for the Beatles subset, which can be exploited by algorithms that operate at a beat level. The Beatles annotations were initially collected by Alan Pollack, and were later revised and enriched by music experts at C4DM. The rest of the annotations were collected by experts at C4DM.

2.5.6 TUT Beatles
This is a refined version of 174 annotations of The Beatles catalog, originally published in the Isophonics dataset described above, and further corrected and published by members of the Tampere University of Technology.12

2.5.7 INRIA
Using the semiotic description scheme described by Bimbot et al. (2014), INRIA released annotations for music from three sources: a set of 124 songs from the Eurovision contest from 2008 to 2010; 159 pieces selected for the QUAERO project; and new annotations for the 100 songs in the RWC popular music database.13

2.5.8 Sargon
Finally, a small set of 30 minutes of heavy metal containing all the tracks from a release of the band Sargon annotated by a single music expert at a flat, coarse level, is also publicly available (Nieto and Bello, 2016). Its main singularity is that its tracks are released under a Creative Commons license, thus freely available.14

2.6 Performance
In this section we discuss the state-of-the-art performances reported by MIREX during the years 2012 to 2017.15 This MIREX task is centered on flat segmentations exclusively, and we focus on this publicly available evaluation exchange due to the challenges that originate with independently reported results mostly due to operating on different versions of audio content and annotations. MIREX runs submitted algorithms against their private audio collection, and therefore, while these comparisons are not exhaustive across algorithms (since not all authors submit to MIREX), these comparisons should be the most transparent. In Table 1, we report the top scores for the Hit Rate boundary retrieval measures at 0.5 and 3 seconds ($F_{1}(P_{r}, R_{k})_{0.5}$ and $F_{1}(P_{r}, R_{k})_{3}$, respectively) and the Pairwise Clustering metrics for segment similarity ($F_{1}(P_{r}, R_{k})$). While the Normalized Conditional Entropies are also reported in MIREX, it is known that MIREX uses a problematic implementation (Raffel et al., 2014) and they can be misleading (as discussed in Section 2.4), therefore we do not include them in the table.

The datasets employed by MIREX are subgroups of the sets described above. More specifically:

- **MIREX 2009** contains 297 tracks from the Isophonics and TUT Beatles datasets.
- **MIREX 2010** is composed of 100 Japanese Pop tracks from the RWC set. There are two versions available: (1) Annotations by INRIA, which do not contain segment labels; and (2) AIST annotations, including segment labels.
- **MIREX 2012** is a subset of 859 tracks from the SALAMI dataset.

The best performing results follow a clear trend: the best boundary retrieval algorithms submitted to MIREX are those based on deep convolutional architectures (Grill and Schlüter, 2015a; Ullrich et al., 2014). Additionally, the best performing algorithm in terms of segment similarity is the one presented by Serrà et al. (2014). There does not seem to exist a method that performs best on all metrics for a single dataset, which also exposes the potential limitations of the current metrics. Some datasets seem more challenging than others, with SALAMI obtaining the worst scores in most metrics, and MIREX 2010 having the best boundary (1) and segment similarity (2) scores. The differences in boundary retrieval between the two versions of the MIREX 2010 dataset warrant discussion. While such differences might originate due to the ambiguity and subjectivity of the task (the tracks on the two versions of the dataset are the same, but annotated by different experts), the boundary retrieval algorithms

| Dataset            | $F_{1}(P_{r}, R_{k})_{0.5}$ | $F_{1}(P_{r}, R_{k})_{3}$ | $F_{1}(P_{r}, R_{k})$ |
|--------------------|-----------------------------|-----------------------------|------------------------|
| MIREX 2009         | 56.42 ± 17.04†             | 70.35 ± 14.87†             | 65.28 ± 15.11†         |
| MIREX 2010 (1)     | 69.70 ± 13.59†             | 79.34 ± 9.43†              | –                     |
| MIREX 2010 (2)     | 52.37 ± 17.54†             | 73.80 ± 11.68§             | 68.83 ± 11.91§         |
| SALAMI*            | 54.09 ± 18.50†             | 68.94 ± 17.51§             | 58.09 ± 15.77†         |

Table 1: Best performing evaluation metrics (percentages) for the MSA task in MIREX for the years 2012 to 2017.

*: Smaller subset of SALAMI; †: 2015 submission by Grill and Schlüter (2015a); §: 2012 submission by Serrà et al. (2014); §: 2014 submission by Ullrich et al. (2014).
have been trained on a subset of such datasets, therefore overfitting effects might be occurring. Thus, we warn that the boundary retrieval results might be artificially inflated. Regardless, none of the reported metrics are close to human performance, which is thought to be around 90% for \(F(p, R_{p})\) and \(F(P, R_{p})\) (Serrà et al., 2014), which contrasts with other MIR tasks such as onset detection, where a much more mature level of performance has been reached.6 Overall, these results not only expose the ambiguity and subjectivity problems inherent in MSA, but they also illustrate that this task is far from being a solved problem. These limitations and current challenges are discussed next.

3. Current Trends and Open Challenges

The effectiveness of music structure analysis algorithms has increased greatly over the past two decades. At the same time, these advances have widened our understanding of how the MSA task is ambiguous and has required refinement, and they have further exposed the open challenges we still face. Here, we identify three key challenges in structure analysis that remain unsolved, and highlight recent advances toward addressing them: A) subjectivity, the fact that different people may disagree about a particular song’s structure; B) ambiguity, where the same person may reasonably agree with multiple interpretations depending on which musical attributes they attend to; and C) hierarchy, the fact that structure exists simultaneously at multiple timescales.

3.1 Subjectivity

The evaluation measures described in Section 2.4 compare an algorithm’s prediction to a single reference annotation. However, reference annotations are provided by human listeners, who sometimes disagree about how a piece is structured; this is true for Western popular music (Bruderer et al., 2009; Serrà et al., 2014) and even more for music not bound to a notated score (Kliën et al., 2012). A notable study on subjectivity for segment boundary retrieval was performed by Wang et al. (2017), where they crowdsourced the problem to a large set of annotators and identified significant differences between strong vs weak boundaries, gradual vs sudden boundaries, and perceptual differences based on the musicianship of the annotators.

To address the subjectivity of annotations, some dataset curators have collected multiple annotations per piece (e.g., SALAMI, SPAM), with the view that each annotation, and the discrepancies between them, are important for evaluation and further study. While there is no consensus about the minimum number of annotators to properly deal with this problem, thanks to having access to a set of these annotations, an estimated structure could be evaluated by comparing it to each reference annotation and taking the average—or, generously, the maximum—score. (Faced with a similar challenge in image boundary detection, Martin et al. (2004, p. 7) devised a variant of computing the hit rate (Section 2.4) against multiple annotations: only predicted boundaries that match none of the boundaries by any annotator are counted as false positives, while the recall is averaged over all annotators.) Alternatively, the multiple annotations could be merged into a single ‘gold standard’, as did Nieto (2015, Sec. VI.3).

Having multiple annotations is also important because inter-rater agreement provides a performance ceiling for algorithms, as noted by Flexer and Grill (2016). Although algorithmic approaches still fall short of this ceiling in general, they have approached it for certain genres, such as the classical and non-Western music categories, which were perhaps annotated less consistently than for jazz and pop. As another example, McFee et al. (2017, Sec. 6) paired human annotations both with algorithmic predictions and other human annotations, and carried out a two-sample Kolmogorov-Smirnov test to determine how close human-algorithm agreement is to inter-rater agreement. Finally, conflicting annotations stemming from subjective decisions may also be exploited by learning algorithms: by training a model on two sets of annotations per music piece, boundary hit rates are improved over arbitrarily selecting a single set of annotations (Grill and Schlüter, 2015b).

While these discussed techniques aim at addressing the subjectivity problem, there are no standard methodologies to deal with this issue, and we hope to see more refined data acquisition or evaluation metrics to fully address it in the upcoming years.

3.2 Ambiguity

Related to subjectivity, but an entire problem on its own, is the ambiguity of a given structure. This is due to the fact that there are many dimensions to music similarity and novelty, but most annotations of structure are the outcome of holistic judgements. That is, even given a single listener's annotation of structure, the meaning of the annotation can be ambiguous. For example, in the annotation in Figure 1, we do not know whether the B segments are all given the same label because they are homogeneous, or because they are sequential repetitions, and we know that segments C and C' are similar but different, but not whether that difference relates to harmony, melody, instrumentation, or some other factor. And despite the song having at least three segment types—A, B, and C—there could be some parts, such as a drum pattern or ostinato, that are consistent across all segments. Many factors go into making similarity judgments, but they are conflated into a single task: whether two segments are the same or not. In short: because there are many dimensions to music similarity, musical structure is also multi-dimensional. Next, we discuss the current attempts at addressing ambiguity in MSA, which still remains an open challenge for this task.

3.2.1 Multi-dimensional structure

To reduce ambiguity, dataset curators have created detailed annotation guidelines (Bimbot et al., 2010; Smith et al., 2011) to isolate certain dimensions of similarity. For example, Peeters and Deruty (2009) recognized that typical labeling systems tended to conflate three separate notions of similarity based on ‘musical role’ (e.g., introductory vs transitional), ‘acoustic similarity’ (e.g., chord progression or melody), or ‘instrument role’ (e.g., whether the lead is
sung or played on a guitar), and this insight inspired the design of the SALAMI annotation format (Smith et al., 2011). The annotation scheme of Bimbot et al. (2014) also distinguishes function from musical similarity, and furthermore provides a rich set of symbols to transcribe internal and between-segment relationships such as extension, insertion, increased intensity, and hybridization.

3.2.2 Novelty vs Repetition

Pieces could be annotated according to one structuring principle at a time. A novelty-only segmentation would consist only of boundaries, with no segment labels. Several music cognition studies effectively collect novelty-based annotations when they ask listeners to indicate whenever they perceive a boundary (e.g., Bruderer et al., 2009). In contrast, a repetition-only analysis would indicate all segments in a piece that are repeated; these typically short segments could potentially overlap. This resembles the definition of a related task, music pattern discovery, proposed by Collins et al. (2013).

3.2.3 Single-feature descriptions

Structure is also ambiguous because different musical attributes, such as harmony, rhythm or timbre, could be important at different points in the piece. Thus, another approach that would reduce ambiguity is to have listeners annotate pieces of music multiple times while focusing on different musical attributes, such as melody, harmony, or rhythm (Befus, 2010). These are the kinds of factors that listeners tend to use to justify their analyses, and there is evidence that paying attention to different features could influence the perception of structure (Smith, 2014).

3.2.4 Multi-part descriptions

Another dimension to musical structure is the number of instrument parts within a single piece, since these parts may repeat and vary independently from each other. Smith and Goto (2017) argued that structure could be much less ambiguous if it were annotated part by part, even to the extent that annotations could be produced automatically from MIDI files.

These efforts to reduce ambiguity by isolating dimensions of musical similarity are mirrored by efforts to model structure more accurately by merging the contributions of multiple dimensions. For example, Hargreaves et al. (2012) showed the advantage of using multi-track recordings to estimate structure. Kaiser and Peeters (2013) modeled homogeneity and repetitions individually before fusing the results, while Grill and Schlüter (2015b) improved a CNN, which had mostly modeled novelty, by incorporating information about repetition from a time-lag matrix. And lastly, many approaches collect audio features related to multiple musical attributes, such as chromatograms, MFCCs, and rhythmograms (see McFee et al. (2015) for a description and implementation of these and other music features).

3.3 Hierarchy

Although many music styles exhibit structure at different timescales—segments, bars, beats, notes—the majority of work in music structure analysis operates at a single level of granularity at a time. Moreover, while multi-level datasets are available (as discussed in Section 2.5), relatively few methods exist to take full advantage of the depth dimension of structure. Broadening the applicability of MSA to hierarchical notions of musical structure is currently an exciting, active, and relatively unexplored research area.

Concretely, the hierarchical structural analysis task consists of producing a sequence of (labeled) segmentations arranged from coarse to fine. At the extremes of the sequence, the coarsest segmentation consists of a single segment (the entire recording), while the finest segmentation encodes individual notes. So far, there have been relatively few data-driven methods for multi-level MSA, but we highlight a few approaches here. McFee and Ellis (2014b) proposed an algorithm for multi-level MSA that encodes multi-level structure in the eigenvectors of a graph Laplacian derived from audio features. Grill and Schlüter (2015a) developed a joint model of segment boundaries on SALAMI at both the coarse and fine levels using convolutional neural networks. Kinnaird (2016) developed aligned hierarchies for detecting nested repetition structures in SSMs, which produce a natural encoding of hierarchical structure. Seetharaman and Pardo (2016) use the activations of increasing subsets of NMF bases as segmentation cues, which exploits depth of polyphony to produce multi-level analyses. Finally, Traile and McFee (2019) propose a method to enhance the spectral clustering method described in Section 2.3.3 by using similarity network fusion to combine several frame-level features into clean affinity matrices.

Evaluation of multi-level MSA has also been historically difficult, and many authors have reduced the problem to existing flat segmentation metrics. The L-measures (described in Section 2.4.3) account for hierarchical depth in annotations, and are relatively less sensitive to alignment errors, but more sensitive to truly incompatible annotations (McFee et al., 2017). Similarly, Kinnaird’s aligned hierarchy representation naturally lends itself to a distance function which can support comparisons between hierarchical decompositions of tracks with differing lengths. This distance metric is not normalized and therefore cannot be used directly for evaluation, but it does have applications to cover detection, where structural similarity can be an informative cue (Kinnaird, 2018). Finally, McFee and Kinnaird (2019) recently presented a novel method to automatically expand hierarchical annotations to facilitate their assessment.

Hierarchical MSA has only been superficially explored so far, and it is our hope to see further advances in such methods and their evaluation in the near future, potentially in upcoming MIREX competitions.

3.4 Richer Annotations

We have discussed three major areas within MSA that are not only unsolved, but expose its inherent difficulty. Together, they point to the main open challenge for MSA: to obtain richer descriptions of musical structure. Researchers should aim beyond obtaining flat, one-dimensional descriptions. They should estimate hierarchical descriptions and the salience of each boundary; they should
specify which structuring principles (homogeneity/novelty, repetition, and/or regularity) justify the segment labels, as well as what musical attributes are homogeneous, repeated, or regular within the audio signal.

Given the recent major advances in transfer learning (Raffel et al., 2020), where unsupervised learning is performed on a large unlabelled corpus and then the model is fine-tuned with a subset of annotated data (similarly to the work by McCallum (2019) discussed above), even if these richer structural data are provided in a not substantially large dataset, the benefits for the research community could be significant. Moreover, and as discussed next, rich descriptions like these may be better exploited by applications.

4. Applications
Computational MSA has a number of applications for different groups of users including music creators, consumers, researchers, and musicologists. Nevertheless, successful and popular applications employing MSA are surprisingly scarce, especially when one considers its long term promise of delivering relevant music-related products (Goto, 2006b). Rather than blaming a lack of interest in having access to such applications, we hypothesize that this might be due to the difficulty of having accurate computational approaches to MSA. In this section, we highlight a few major application areas.

4.1 Music Creation and Production
Typical music creation and production software packages (Pro Tools, Adobe Audition, Audacity, Ableton Live, Logic Pro, Cakewalk Sonar, etc.) provide limited semantic information by using the waveform as the main representation. However, such information could be particularly useful for remixing music, looping, and applying different forms of processing to different segments. For example, accurate segmentation boundary markers provide efficient navigation time stamps during recording and mixing sessions. Furthermore, these boundary markers allow efficient synchronizations between music and other media, such as video or graphics. Segmentation labels can be used to manage audio effects efficiently into different groups based on semantic context within a song. For example, the audio effects and their parameters during chorus segments might be the same across a song, but might be different during verses. With accurate segmentation labels it is possible to re-use tuned audio effects more efficiently. These software packages typically allow users to provide markers to highlight specific points in a song. Users sometimes manually perform MSA and mark the change points between segments with such markers. Computational MSA can ameliorate this laborious process. Moreover, different levels of hierarchical MSA could provide new insights and control in different time scopes that could spark creative pursuits.

4.2 Automatic Music Generation
Recent advances in machine learning (especially with the advent of generative adversarial networks (Goodfellow et al., 2014) and flow-based generative models (Dinh et al., 2017)) have resulted in significant contributions in the field of computer image generation (Guérin et al., 2017; Kingma and Dhariwal, 2018). The field of automatic music generation, which originates back in the middle of the 20th century and is currently an active research area of MIR, has notably advanced with these novel machine learning techniques (Dong et al., 2018; Roberts et al., 2018; Dieleman et al., 2018; Dhariwal et al., 2020). One of the key aspects when automatically generating music is to produce a meaningful long-term music structure such that the final piece is coherent and appealing. This is particularly challenging due to the difficulty in capturing long term structures by most sequential models used for this task, such as long short-term memory networks (Manzelli et al., 2018), generative adversarial networks (Engel et al., 2019), or other recurrent models (Thickstun et al., 2019).

To this end, we believe computational MSA may play a significant role when synthesizing music, especially when aiming to produce cohesive tracks with recurring phrases and motives (Jhamtani and Berg-Kirkpatrick, 2019). Moreover, systems that generate music may be able to provide personalized results, in that a potential listener could adjust, e.g., the type of form, segment length, or degree of repetition that a generated song would ultimately contain.

4.3 Music Recommendation
The field of music recommendation has also been impacted by the drastic development of deep architectures (van den Oord et al., 2013; Pons et al., 2018). Given that the actual audio content is generally available in any music recommendation service, more sophisticated recommendations could be produced if computational MSA would be applied to all their music collection. For instance, having a segmented catalog could yield recommendations where certain parts of a track contain the desired musical attributes that a given listener might have identified in a track. By recommending items at a segment level, music recommenders would potentially yield more fine-grained recommendations where the listener could query pieces with specific types of segments (e.g., loud electric guitar solo). Another example of the benefits of applying computational MSA in such recommender systems is when previewing a set of recommendations that the user can chose from. In such cases, short music summaries (Logan and Chu, 2000; Peeters et al., 2002; Levy et al., 2006; Nieto et al., 2012) produced by identifying the most prominent segments of a piece (thus producing short audio thumbnails) could help the final listener to choose the next song/album to play/purchase.

4.4 Live Performances, Video Games, and Recordings
In recent years, several MIR projects have been designed to enhance the experience of live musical concerts (Liem et al., 2015). To this end, computational MSA may provide tools where the light and/or video projections of the live performance may adapt according to the segment of the song currently being played, thus providing the audience with more in-depth and likely enjoyable experiences. Such implementations for live music require MSA techniques that can operate at small windows of time to identify segment boundaries, such as spectrogram-based CNNs.
(as opposed to SSMs, which need the full song to produce results). It remains to be seen how similarly identified segments could be labeled in real-time, causal (i.e., no access to future samples) scenarios.

Video games that are directly related to computational MSA are those in which the user has to play or dance along to songs, following specific scores on the screen (e.g., Rock Band, Rocksmith, Dance Dance Revolution). Such scores could potentially be automatically generated, while still being consistent with the structure of the song to follow. Furthermore, MSA-related techniques have been applied to non-musically centered video games (e.g., Final Fantasy VII Remake, The Secret of Monkey Island 2 Special Edition), where music transitions between scenarios take place seamlessly by employing segment-based anchors.

Moreover, live recordings or long broadcasts could also benefit from computational MSA by identifying those large-scale segmentation points, e.g., for easier navigation by the final user. This could further be exploited by allowing the placement of potentially non-invasive ads in those automatically located key points in such long audio signals.

### 4.5 Visualization

Visualizing the structure of a song can be useful for musicians, musicologists, and consumers to understand a song in more depth, or to get a quick sense of it. For example, the web-based service Songle\(^7\) (Goto et al., 2011) provides users with a timeline of the main repeating segments, with the predicted choruses (detected using the RefraiD (Goto, 2006b) algorithm) highlighted; clicking on a segment quickly directs playback to that segment. The service also displays beats, downbeats, melody and chord estimations. Within Sonic Visualiser\(^8\) (Cannam et al., 2010), a general tool for audio visualization, the VAMP plug-in Segmentino (Mauch et al., 2009b) estimates and displays segment boundaries and labels. Other visualization approaches include Paul Lamere’s Infinite Jukebox,\(^9\) Martin Wattenber’s The Shape of Song,\(^10\) McFee’s circular hierarchies (McFee and Ellis, 2014a), and the scape plot representation (Müller and Jiang, 2012).

### 4.6 Tools for Researchers

MSA is often useful for MIR researchers as a first step towards other applications. For example, Mauch et al. (2009b) use segmentation labeling as part of the chord recognition process. The intuition is that chord progressions within segments that have the same labels are more likely to be consistent with each other than the chord progressions in segments with different labels.

Another example is MSA for source separation. REPET-SIM (Rafii et al., 2014) uses the repetitive nature of background music to help separate background music from vocals (or the lead instrument). The repetitive structure of certain songs is constant within a segment but changes in different segments. Modeling these repetitions differently in each segment tends to yield a higher performance than a global repetition model across the whole song. Using MSA as a pre-processing step allows this local modeling of repetition. Furthermore, a method that uses NMF to simultaneously estimate segmentation and voice separation of audio signals has been proposed (Seetharaman and Pardo, 2016). Moreover, it has recently been shown that music structure can help at identifying downbeats (Fuentes et al., 2019b). This is particularly interesting since it is a clear example where segmentation can inform other areas of MIR (and vice-versa) to obtain more coherent results.

The capacity of automated techniques to analyze a corpus of millions of songs—far more than a single listener could hope to analyze manually—enables digital musicologists to seriously investigate questions such as whether pop songs became more repetitive over the 20th century, or to seek new evidence for well-known subjects, such as how the hierarchical structure of sonatas evolved in the classical period.

For musicological research, in the CHARM Mazurka Project,\(^11\) though not directly conducting computational MSA on the Mazurkas, the scape plots are used to show hierarchical harmonic relations throughout each performed Mazurka at different time scales (Müller and Jiang, 2012).

A number of open source libraries such as Librosa\(^2\) and MSAF\(^3\) in Python and Essentia\(^4\) in C++ support MSA, which allows it to be easily incorporated into the algorithm development process.

### 5. Conclusions

Audio-based MSA is a compelling and active area of research within the broader field of MIR. In this article we have reviewed its current state of the art, including its most relevant methods, principles, evaluation metrics, datasets, and current performance. Furthermore, we have discussed the main challenges that this task is currently facing, placing a strong emphasis on subjectivity, ambiguity, and hierarchy; all of which may be alleviated by collecting richer human labels in upcoming MSA datasets. Finally, a set of applications that could exploit computational MSA have been exposed, thus showing the potential of this task in future musical experiences.

This timely topic is facing rapid changes, and we hope this work helps motivating novel and experienced researchers in the field to focus on the major open challenges and potential applications to bring this task forward to an even more mature state.

### Notes

1. Strictly speaking, the latter is not a hierarchy, and more correctly referred to as a multi-level analysis. In this work we use both terms indistinguishably.
2. How Beautiful You Are by the band The Cure.
3. For clarity: \(L(1500) = L(150 \text{ seconds}) = L(2:30)\), near the end of the second B section.
4. https://youtu.be/s08jcd3E6Mpg.
5. The technique applied to graphical models, not to be confused with the spectral representation of an audio signal.
6. https://github.com/craffel/mir_eval/issues/226.
7. http://www.music-ir.org/mirex/wiki/2017:MIREX2017_Results.
8. https://github.com/marl/jams-data.
Competing Interests
The authors have no competing interests to declare.

References
Befus, C. (2010). Design and evaluation of dynamic feature-based segmentation on music. Master’s thesis, University of Lethbridge, Lethbridge, Alberta, Canada.

Bimbot, F., Blouch, O. L., Sargent, G., & Vincent, E. (2010). Decomposition into autonomous and comparable blocks: A structural description of music pieces. In Proc. of the 11th International Society for Music Information Retrieval Conference, pages 189–194. Utrecht, The Netherlands.

Bimbot, F., Sargent, G., Deruty, E., Guichaoua, C., & Vincent, E. (2014). Semiotic description of music structure: An introduction to the Quaero/Metiss structural annotations. In Proc. of the AES 53rd Conference on Semantic Audio.

Bittner, R., Fuentes, M., Rubinstein, D., Jansson, A., Choi, K., & Kell, T. (2019). mirdata: Software for reproducible usage of datasets. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 99–106. Delft, The Netherlands.

Bruderer, M. J. (2008). Perception and Modeling of Segment Boundaries in Popular Music. PhD thesis, Technische Universität Eindhoven.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2009). The perception of structural boundaries in melody lines of Western popular music. Musicae Scientiae, 13(2), 273–313. DOI: https://doi.org/10.1177/102986490901300204

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In Proc. of the International Computer Music Conference, pages 17–22. La Havana, Cuba.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In Proc. of the 18th ACM International Conference on Multimedia, pages 1467–1468. ACM. DOI: https://doi.org/10.1145/183951.1874248

Chen, T.-P., & Su, L. (2019). Harmony Transformer: Incorporating chord segmentation into harmony recognition. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 259–267. Delft, The Netherlands.

Cheng, T., Smith, J. B. L., & Goto, M. (2018). Music structure boundary detection and labelling by a deconvolution of path-enhanced self-similarity matrix. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 106–110. Calgary, Alberta, Canada. DOI: https://doi.org/10.1109/ICASSP.2018.8461319

Collins, T., Arzt, A., Flossman, S., & Widmer, G. (2013). SIARCT-CFP: Improving precision and the discovery of inexact musical patterns in point-set representations. In Proc. of the 14th International Society for Music Information Retrieval Conference, pages 549–554. Curitiba, Brazil.

Dannenberg, R. B., & Goto, M. (2008). Music structure analysis from acoustic signals. In Havelock, D., Kuwano, S., & Vorländer, M., editors, Handbook of Signal Processing in Acoustics, pages 305–331. Springer, New York, NY. DOI: https://doi.org/10.1007/978-0-387-30441-0_21

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. arXiv preprint 2005.00341.

Dieleman, S., van den Oord, A., & Simonyan, K. (2018). The challenge of realistic music generation: Modelling raw audio at scale. In Advances in Neural Information Processing Systems 31, pages 7989–7999. Curran Associates, Inc.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real NVP. In 5th International Conference on Learning Representations (ICLR). Toulon, France.

Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2018). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proc. of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, LA, USA.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). GANSynth: Adversarial neural audio synthesis. In 7th International Conference on Learning Representations (ICLR). New Orleans, LA, USA.

Flexer, A., & Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. Journal of New Music Research, 45(3), 239–251. DOI: https://doi.org/10.1080/09298215.2016.1200631

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In Proc. of the IEEE International Conference of Multimedia and Expo.
(ICME), pages 452–455. New York City, NY, USA. DOI: https://doi.org/10.1109/ICME.2000.869637

Fuentes, M., Maia, L. S., & Biscainho, L. W. P. (2019a). Tracking beats and microtiming in Afro-Latin American music using conditional random fields and deep learning. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 251–258. Delft, The Netherlands.

Fuentes, M., McFee, B., Crayencour, H., Essid, S., & Bello, J. (2019b). A music structure informed downbeat tracking system using skip-connection conditional random fields and deep learning. In Proc. of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 481–485. DOI: https://doi.org/10.1109/ICASSP.2019.8682870

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. Proc. of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 776–780. DOI: https://doi.org/10.1109/ICASSP.2017.7952261

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems 27, pages 2672–2680.

Goto, M. (2003). A chorus-section detecting method for musical audio signals. In Proc. of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 437–440. Hong Kong, China.

Goto, M. (2006a). AIST annotation for the RWC Music Database. In Proc. of the 7th International Conference on Music Information Retrieval, pages 359–360. Victoria, BC, Canada.

Goto, M. (2006b). A chorus section detection method for musical audio signals and its application to a music listening station. IEEE Transactions on Audio, Speech, and Language Processing, 14(5), 1783–1794. DOI: https://doi.org/10.1109/TASL.2005.863204

Goto, M., Yoshi, K., Fujihara, H., Mauch, M., & Nakano, T. (2011). Songle: A web service for active music listening improved by user contributions. In Proc. of the 12th International Society for Music Information Retrieval Conference, pages 311–316. Miami, FL, USA.

Grill, T., & Schlüter, J. (2015a). Music boundary detection using neural networks on combined features and two-level annotations. In Proc. of the 16th International Society for Music Information Retrieval Conference. Málaga, Spain. DOI: https://doi.org/10.1109/EUSIPCO.2015.7362593

Grill, T., & Schlüter, J. (2015b). Music boundary detection using neural networks on spectrograms and self-similarity lag matrices. In Proc. of the 23rd European Signal Processing Conference (EUSIPCO). Nice, France. DOI: https://doi.org/10.1109/EUSIPCO.2015.7362593

Groves, R. (2016). Automatic melodic reduction using a supervised probabilistic context-free grammar. In Proc. of the 17th International Society for Music Information Retrieval Conference, pages 775–781. New York, NY, USA.

Guérin, É., Digne, J., Galin, É., Peytavie, A., Wolf, C., Benes, B., & Martinez, B. (2017). Interactive example-based terrain authoring with conditional generative adversarial networks. ACM Transactions on Graphics, 36(6), 228:1–228:13. DOI: https://doi.org/10.1145/3130800.3130804

Hamanaka, M., Hirata, K., & Tojo, S. (2006). Implementing “A Generative Theory of Tonal Music”. Journal of New Music Research, 35(4), 249–277. DOI: https://doi.org/10.1080/09298210701563238

Hargreaves, S., Klapuri, A., & Sandler, M. (2012). Structural segmentation of multitrack audio. IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2637–2647. DOI: https://doi.org/10.1109/TASL.2012.2209419

Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). Moving beyond feature design: Deep architecture and automatic feature learning in music informatics. In Proc. of the 13th International Society for Music Information Retrieval Conference, pages 403–408. Porto, Portugal.

Humphrey, E. J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R. M., & Bello, J. P. (2014). JAMS: A JSON annotated music specification for reproducible MIR research. In Proc. of the 15th International Society for Music Information Retrieval Conference, pages 591–596. Taipei, Taiwan.

Janssen, B., de Haas, W., Volk, A., & Van Kranenburg, P. (2013). Discovering repeated patterns in music: State of knowledge, challenges, perspectives. In Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR), pages 225–240. Marseille, France.

Jhamtani, H., & Berg-Kirkpatrick, T. (2019). Modeling self-repetition in music generation using generative adversarial networks. In Machine Learning for Music Discovery Workshop, ICML. Long Beach, USA.

Kaiser, F., & Peeters, G. (2013). A simple fusion method of state and sequence segmentation for music structure discovery. In Proc. of the 14th International Society for Music Information Retrieval Conference. Curitiba, Brazil.

Kaiser, F., & Sikora, T. (2010). Music structure discovery in popular music using non-negative matrix factorization. In Proc. of the 11th International Society for Music Information Retrieval Conference, pages 429–434. Utrecht, Netherlands.

Kim, J. W., & Bello, J. P. (2019). Adversarial learning for improved onsets and frames music transcription. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 670–677. Delft, The Netherlands.

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1 × 1 convolutions. In Advances in Neural Information Processing Systems 31, pages 10215–10224. Montreal, Canada.

Kinnaird, K. M. (2016). Aligned hierarchies: A multiscale structure-based representation for music-based data
stream. In Proc. of the 17th International Society for Music Information Retrieval Conference, pages 337–343. New York City, NY, USA.

Kinnaird, K. M. (2018). Aligned sub-hierarchies: A structure-based approach to the cover song task. In Proc. of the 19th International Society for Music Information Retrieval Conference, pages 585–591. Paris, France.

Klien, V., Grill, T., & Flexer, A. (2012). On automated annotation of acoustical music. Journal of New Music Research, 41(2), 153–173. DOI: https://doi.org/10.1080/09298215.2011.618226

Lerdahl, F., & Jackendoff, R. (1983). A Generative Theory of Tonal Music. MIT Press.

Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 318–326. DOI: https://doi.org/10.1109/TASL.2007.910781

Levy, M., Sandler, M., & Casey, M. (2006). Extraction of high-level musical structure from audio data and its application to thumbnail generation. In Proc. of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 5. DOI: https://doi.org/10.1109/ICASSP.2006.1661200

Liem, C. C., Gomez, E., & Schedl, M. (2015). PHENIX: Innovating the classical music experience. In Proc. of the 2015 IEEE International Conference on Multimedia and Expo Workshops, pages 3–6. Torino, Italy. DOI: https://doi.org/10.1109/ICMEW.2015.7169835

Logan, B., & Chu, S. (2000). Music summarization using key phrases. In Proc. of the 25th IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 749–752. Istanbul, Turkey. DOI: https://doi.org/10.1109/ICASSP.2000.859068

Lukashevich, H. (2008). Towards quantitative measures of evaluating song segmentation. In Proc. of the 9th International Conference on Music Information Retrieval, pages 375–380. Philadelphia, PA, USA.

Maezawa, A. (2019). Music boundary detection based on a hybrid deep model of novelty, homogeneity, repetition and duration. In Proc. of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 206–210. Brighton, United Kingdom. DOI: https://doi.org/10.1109/ICASSP.2019.8683249

Manzelli, R., Thakkar, V., Siahkamari, A., & Kulis, B. (2018). Conditioning deep generative raw audio models for structured automatic music. In Proc. of the 19th International Society for Music Information Retrieval Conference, pages 182–189. Paris, France.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(5), 530–549. DOI: https://doi.org/10.1109/TPAMI.2004.1273918

Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., Tidhar, D., & Sandler, M. (2009a). OMRAS2 Metadata Project 2009. In Late Breaking/Demo at the 10th International Society for Music Information Retrieval Conference. Kobe, Japan.

Mauch, M., Noland, K., & Dixon, S. (2009b). Using musical structure to enhance automatic chord transcription. In Proc. of the 10th International Society for Music Information Retrieval Conference, pages 231–236. Kobe, Japan.

McCallum, M. (2019). Unsupervised learning of deep features for music segmentation. In Proc. of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 346–350. Brighton, United Kingdom. DOI: https://doi.org/10.1109/ICASSP.2019.8683407

McFee, B., & Ellis, D. P. W. (2014a). Analyzing song structure with spectral clustering. In Proc. of the 15th International Society for Music Information Retrieval Conference, pages 405–410. Taipei, Taiwan.

McFee, B., & Ellis, D. P. W. (2014b). Learning to segment songs with ordinal linear discriminant analysis. In Proc. of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing, pages 5197–5201. Florence, Italy. DOI: https://doi.org/10.1109/ICASSP.2014.6854594

McFee, B., & Kinnaird, K. (2019). Improving structure evaluation through automatic hierarchy expansion. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 152–158. Delft, The Netherlands.

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating hierarchical structure in music annotations. Frontiers in Psychology, 8: 1337. DOI: https://doi.org/10.3389/fpsyg.2017.01337

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In Proc. of the 14th Python in Science Conference (SciPy), pages 18–25. Austin, TX, USA. DOI: https://doi.org/10.25080/Majora-7b98e3ed-003

Müller, M., & Jiang, N. (2012). A scope plot representation for visualizing repetitive structures of music recordings. In Proc. of the 13th International Society for Music Information Retrieval Conference, pages 97–102. Porto, Portugal.

Murthy, V. V. S., & Koolagudi, S. G. (2018). Content-based music information retrieval (CB-MIR) and its applications toward the music industry: A review. ACM Computing Surveys, 51(3). DOI: https://doi.org/10.1145/3177849

Nieto, O. (2015). Discovering Structure in Music: Automatic Approaches and Perceptual Evaluations. PhD thesis, New York University.

Nieto, O., & Bello, J. P. (2014). Music segment similarity using 2D-Fourier magnitude coefficients. In Proc. of the 39th IEEE International Conference on Acoustics Speech and Signal Processing, pages 664–668. Florence, Italy. DOI: https://doi.org/10.1109/ICASSP.2014.6853679

Nieto, O., & Bello, J. P. (2016). Systematic exploration of computational music structure research. In Proc. of
the 17th International Society for Music Information Retrieval Conference, pages 547–553. New York City, NY, USA.

Nieto, O., Faroord, M. M., Jehan, T., & Bello, J. P. (2014). Perceptual analysis of the F-measure for evaluating section boundaries in music. In Proc. of the 15th International Society for Music Information Retrieval Conference, pages 265–270. Taipei, Taiwan.

Nieto, O., Humphrey, E. J., & Bello, J. P. (2012). Compressing music recordings into audio summaries. In Proc. of the 13th International Society for Music Information Retrieval Conference, pages 313–318. Porto, Portugal.

Nieto, O., & Jehan, T. (2013). Convex non-negative matrix factorization for automatic music structure identification. In Proc. of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing. DOI: https://doi.org/10.1109/ICASSP.2013.6637644

Nieto, O., McCallum, M., Davies, M., Robertson, A., Stark, A., & Egozy, E. (2019). The Harmonix Set: Beats, downbeats, and functional segment annotations of Western popular music. In Proc. of the 20th International Society for Music Information Retrieval Conference, pages 565–572. Delft, The Netherlands.

Panagakis, Y., & Kotropoulos, C. (2012). Music structure analysis by ridge regression of beatsynchronous audio features. In Proc. of the 13th International Society for Music Information Retrieval Conference, pages 271–276. Porto, Portugal.

Paulus, J., & Klapuri, A. (2009). Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. IEEE Transactions on Audio, Speech, and Language Processing, 17(6), 1159–1170. DOI: https://doi.org/10.1109/TASL.2009.2020533

Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. In Proc. of the 11th International Society for Music Information Retrieval Conference, pages 625–636. Utrecht, Netherlands.

Pauwels, J., Kaiser, F., & Peeters, G. (2013). Combining harmony-based and novelty-based approaches for structural segmentation. In Proc. of the 14th International Society for Music Information Retrieval Conference. Curitiba, Brazil.

Peeters, G., & Bisot, V. (2014). Improving music structure segmentation using lag-priors. In Proc. of the 15th International Society for Music Information Retrieval Conference, pages 337–342. Taipei, Taiwan.

Peeters, G., Burthe, A. L., & Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In Proc. of the 3rd International Conference on Music Information Retrieval. Paris, France. ISMIR.

Peeters, G., & Deruty, E. (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In Proc. of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS), pages 75–90. Graz, Austria.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., & Serra, X. (2018). End-to-end learning for music audio tagging at scale. In Proc. of the 19th International Society for Music Information Retrieval Conference, pages 637–644. Paris, France.

Raffel, C., Mcfee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common MIR metrics. In Proc. of the 15th International Society for Music Information Retrieval Conference, pages 367–372. Taipei, Taiwan.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Rafi, Z., Liutkus, A., & Pardo, B. (2014). REPET for background/foreground separation in audio. In Naik, G. R., & Wang, W., editors, Blind Source Separation, pages 395–411. Springer. DOI: https://doi.org/10.1007/978-3-642-55016-4_14

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In Proc. of the 35th International Conference on Machine Learning, volume 80 of Proc. of Machine Learning Research, pages 4364–4373. Stockholm, Sweden.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLPCoNLL). pages 410–420.

Sargent, G., Bimbot, F., & Vincent, E. (2011). A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs. In Proceedings of the International Conference on Music Information Retrieval, pages 483–488. Miami, United States.

Sargent, G., Bimbot, F., & Vincent, E. (2017). Estimating the structural segmentation of popular music pieces under regularity constraints. IEEE/ACM Transactions on Audio Speech and Language Processing, 25(2), 344–358. DOI: https://doi.org/10.1109/TASLP.2016.2635031

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. Foundations and Trends in Information Retrieval, 8(2–3), 127–261. DOI: https://doi.org/10.1561/1500000042

Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. (2011). Using mutual proximity to improve content-based audio similarity. In Proc. of the 12th International Society for Music Information Retrieval Conference, pages 79–84. Miami, FL, USA.

Seetharaman, P., & Pardo, B. (2016). Simultaneous separation and segmentation in layered music. In Proc. of the 17th International Society for Music Information Retrieval Conference. New York City, NY, USA.

Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2014). Unsupervised music structure annotation by time
series structure features and segment similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, 16(5), 1229–1240. DOI: https://doi.org/10.1109/TMM.2014.2310701

Serrà, J., Serra, X., & Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9), 1138–1151. DOI: https://doi.org/10.1088/1367-2630/11/9/093017

Shibata, G., Nishikimi, R., Nakamura, E., & Yoshii, K. (2019). Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-Markov model. In *Proc. of the 20th International Society for Music Information Retrieval Conference*, pages 268–275. Delft, The Netherlands.

Smith, J. B. L. (2014). *Explaining listener differences in the perception of musical structure*. PhD thesis, Queen Mary University of London.

Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proc. of the 12th International Society for Music Information Retrieval Conference*, pages 555–560. Miami, FL, USA.

Smith, J. B. L., & Chew, E. (2013). A meta-analysis of the MIREX structure segmentation task. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, pages 251–256. Curitiba, Brazil.

Smith, J. B. L., & Goto, M. (2016). Using priors to improve estimates of music structure. In *Proc. of the 17th International Society for Music Information Retrieval Conference*, pages 554–560. New York City, NY, USA.

Smith, J. B. L., & Goto, M. (2017). Multi-part pattern analysis: Combining structure analysis and source separation to discover intra-part repeated sequences. In *Proc. of the 18th International Society for Music Information Retrieval Conference*, pages 716–723. Suzhou, China.

Thickstun, J., Harchaoui, Z., Foster, D., & Kakade, S. (2019). Coupled recurrent models for polyphonic music composition. In *Proc. of the 20th International Society for Music Information Retrieval Conference*, pages 311–318. Delft, The Netherlands.

Tian, M., & Sandler, M. B. (2016). Towards music structural segmentation across genres. *ACM Transactions on Intelligent Systems and Technology*, 8(2), 1–19. DOI: https://doi.org/10.1145/2950066

Tralie, C. J., & McFee, B. (2019). Enhanced hierarchical music structure annotations via feature level similarity fusion. In *Proc. of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 201–205. Brighton, United Kingdom. DOI: https://doi.org/10.1109/ICASSP.2019.8683492

Turnbull, D., Lancikret, G., Pampalk, E., & Goto, M. (2007). A supervised approach for detecting boundaries in music using difference features and boosting. In *Proc. of the 8th International Conference on Music Information Retrieval*, pages 42–49. Vienna, Austria.

Ullrich, K., Schlüter, J., & Grill, T. (2014). Boundary detection in music structure analysis using convolutional neural networks. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, pages 417–422. Taipei, Taiwan.

van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems* 26, pages 2643–2651.

Wang, C.-l., Mysore, G. J., & Dubnov, S. (2017). Re-visiting the music segmentation problem with crowdsourcing. In *Proc. of the 18th International Society for Music Information Retrieval Conference*, pages 738–744. Suzhou, China.

Wang, J.-C., Lee, H.-S., Wang, H.-M., & Jeng, S. K. (2011). Learning the similarity of audio music in bag-of-frames representation from tagged music data. In *Proc. of the 12th International Society for Music Information Retrieval Conference*, pages 85–90. Miami, FL, USA.

Weiss, R., & Bello, J. P. (2011). Unsupervised discovery of temporal structure in music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6), 1240–1251. DOI: https://doi.org/10.1109/JSTSP.2011.2145356