Investigating Musical Pattern Ambiguity in a Human Annotated Dataset

Iris Yuping Ren,[1] Oriol Nieto,[2] Hendrik Vincent Koops,[1] Anja Volk,[1] and Wouter Swierstra[1]

[1] *Department of Information and Computing Sciences, Utrecht University, Netherlands*
[2] *Pandora Media, Inc., Oakland, CA, USA*

y.ren@uu.nl

Abstract

Many musical structures, such as musical motifs or patterns, are inherently ambiguous and lead to different equally plausible interpretations by listeners. Accordingly, annotations of these musical structures by different listeners give rise to disagreement between different annotators. In Music Information Retrieval (MIR) tasks such as automatic musical pattern extraction, however, this disagreement poses particular difficulties for evaluation. To provide data for further research on the ambiguity of musical patterns, we present a new annotated patterns dataset. This collection comprises six musical excerpts each annotated by 12 annotators. We observe from the data that disagreement amongst annotators is common. We therefore propose to perform two treatments on annotations to achieve higher pairwise annotator agreement: by using extra rankable metadata on the annotations such as relevance/importance scores, and by adjusting the time resolution of the annotated patterns' time span. We hypothesise that, by using the top-ranked annotations and lowering the time resolution of the annotations, we may obtain more pairwise annotator agreement in the dataset. We perform computational analyses and provide supporting evidence that patterns rated as highly relevant or/and with lower time resolution tend to have more agreement amongst the annotators, in contrast to those rated with lower relevance and higher time resolution. Our analyses could be useful for the development and evaluation of new musical pattern discovery algorithms.

Introduction

In Music Information Retrieval (MIR), automatic musical pattern discovery is an active area of research where algorithms are designed, employed on music data, and evaluated to extract musical patterns computationally for different applications (Collins, 2011). However, ambiguity of musical structures poses difficulties on how one should evaluate the output of such algorithms: if different equally valid interpretations of musical patterns or motifs[1] are possible, what should we take as the ground truth for evaluating algorithms?

Music ambiguity per se is a widely studied topic in music perception and cognition research. Not only musical patterns, but also other dimensions of music can be ambiguous in their interpretation: polyrhythms, for example, can be heard and interpreted with beats at different levels; chords of a harmony sequence can be interpreted with different functions (Randall, 1999, Schoenberg, 1983). Furthermore, creators of music often employ elements of ambiguity to their compositions, and listeners often experience uncertainty where multiple simultaneous interpretations are possible (Bernstein, 1976).

In this paper, we put focus on the ambiguity of monophonic melodic patterns. While musicologists use the term "motif" often intuitively in their analyses of musical compositions, there exists no generally agreed upon definition of what constitutes a motif or pattern: it can be a short musical idea, a salient recurring figure, musical fragment, or succession of notes that has special importance in or is characteristic of a composition. Related concepts also include musical sequence, imitation (Benward, 2014), melody type (Hiley, 1993), musical cell (Nattiez, 1990), phrase (Burkhart, 2005; Sadie, & Tyrrell, 2001), and subject (Scholes, 1970). These different notions of what might constitute a musical pattern pose challenges for creating annotations as reference data for evaluating pattern discovery algorithms.

One attempt to systematically evaluate pattern discovery algorithms is the Music Information Retrieval Evaluation eXchange (MIREX) Discovery of Repeated Themes & Sections task (Collins, Janssen, Ren & Volk, 2017). This task uses a single reference annotation compiled from music theoretic analyses. In the task, a pattern is defined as a set of time-pitch pairs that occurs at least twice in a piece of music. However, in (Ren, Koops, Volk, & Swierstra, 2017) it is shown that pattern discovery algorithms do not agree with each other on what patterns should be extracted from the pieces, and they agree even less with the patterns from the reference annotation. This raises the question on the suitability of one reference annotation for evaluating pattern discovery algorithms. Due to the lack of a clear music theoretic notion of what constitutes a pattern, we therefore propose to take a data-driven approach towards the understanding of the notion of musical patterns, by gathering multiple annotations for the same piece and analysing the disagreement and agreement between different annotators on the discovered patterns.

Therefore, we present a new musical patterns dataset: HEMAN (Human Estimations of Musically Agreeing Notes) where multiple perspectives on six musical excerpts are made available. We asked 12 subjects to annotate patterns in six musical excerpts. HEMAN is a digitised, open source version of the dataset introduced in (Nieto, Farbood, 2012). While we show that there exists considerable disagreement between annotators on what constitutes the patterns of a piece, we demonstrate that the disagreement can be reduced by 1) considering only the patterns that annotators have rated as highly relevant and 2) by lowering the time resolution.

[1] In this paper we use the terms "patterns" and "motifs" interchangeably.

The two main contributions of this paper are:

- Releasing digitised pattern annotation data in the JAMS (JSON Annotated Music Specification) format (Humphrey et al, 2014) and time interval format for facilitating future research.
- Analyses on the HEMAN dataset including two methods for alleviating pairwise annotator disagreement: use annotations that are rated as highly relevant and reduce the time resolution of the annotated intervals.

Experimental setup

The annotation process for obtaining the dataset was conducted at New York University (NYU). Subjects were all graduate students at NYU and had an average of 10 years of formal musical training (Standard Deviation = 2.3). Detailed information on the subject's music experience background can be found in the next section.

The HEMAN collection comprises 6 music excerpts as listed below:

1. Bach – Cantata BWV 1, Movement 6, Horn:
2. Bach –Cantata BWV 2, Movement 6, Soprano:
3. Beethoven –String Quartet, Op. 18, No. 1, Violin I
4. Haydn –String Quartet, Op. 74, No. 1, Violin I
5. Mozart –String Quartet, K. 155, Violin I
6. Mozart –String Quartet, K. 458, Violin I

Some of these excerpts were chosen because they were particularly hard for humans to analyse given the structural ambiguity and creative variations of the musical material. For example, the Bach chorale had very little rhythmic variation or clear grouping cues aside from phrase ending points. The other type of excerpts contains many evident and rigid repeated patterns.

Each piece was annotated by the 12 subjects. Unlimited time was given to the subjects. We did not reveal the name of the pieces on the annotation sheet. The following instructions were given to the annotators:

"Please, analyze the following musical excerpts and mark all the musical motives you can find. A musical motive is defined as a short musical idea, a salient recurring figure, musical fragment, or succession of notes that has some special importance in or is characteristic of a composition. It shouldn't be longer than a musical phrase. If you find a motive that is similar to another (or multiple versions of a motive), choose the one that you think is the most representative. Even though all motives are relevant, please rate each one of them from 1 to 3:

- 1 = Not as relevant
- 2 = Relevant
- 3 = Highly relevant

You can listen to the music excerpts as many times as you like. You can find them here. http://urinieto.com/NYU/Research/MotivesExperiment/"

We deliberately offered several possible interpretations as to what defines a "musical motif" and allow the subjects to both analyse the pieces and use their musical intuitions on what constitutes a musical motif in the process. We did not ask them to laboriously label all occurrences of the same pattern for us. In this way, we obtain at least the prototypes of musical patterns the subjects perceived in the piece.

The HEMAN Dataset

In this section, we present the content of the dataset, how we digitised it, and some difficulties we encountered and decisions we made during the digitising process.

The musical background of the subjects

In Table 1, we report the results of the musical background questionnaire of our subjects.

Table 1. Musical background of annotators. The header Inst and Inst2 denote the main and the secondary instrument the participant plays, followed by how many years of experience. The headers Theory and Overall denote the self-rated musical theory background level and the self-rated musical background level. A professional background has a score of 5 and no background has a score of 1. The header Abs stands for absolute pitch/perfect pitch, with 0 stands for self-identified as no perfect pitch, 1 otherwise.

Inst	Year	Inst2	Year	Theory	Overall	Abs
Piano	10	Violin	4	3	3	0
Oboe	5	Piano	1	2	3	0
Piano	5	None	0	3	3	0
Piano	12	None	0	5	5	1
Piano	14	Trumpet	8	4	4	0
Guitar	13	Bass	0	5	5	0
Piano	13	Violin	4	4	4	0
None	0	None	0	0	0	0
Guitar	8	Piano	4	5	5	0
Guitar	10	Kazoo	4	5	5	0
Violin	8	Guitar	2	3	3	0

As we can see, we have annotators with different levels of musical background. Most of them are trained with western musical instruments. In the following subsections, we detail how the making of the dataset started with paper-based annotation, how we calculate the time intervals from the photos and convert the musical patterns to a numerical format, and finally how we convert the numerical format to the JAMS format.

Digitising paper-based annotations

Figure 1 shows an example of a raw annotation. From this notation, we first calculate the time intervals [start time, end

time] of the pattern annotations. We take the unit of a crochet as one in this time interval format. For example, the first annotated pattern in the time interval format is [0,2], and the second pattern is [8,10].

There are advantages and disadvantages with using a paper-based setting. On the one hand, this could preserve the most natural mindset on perceiving patterns in music; on the other hand, this poses some risks for the correct interpretation of the markings. For example, in Figure 1, it is not immediately apparent whether to include the last quiver in bar 3 into the musical pattern. The same situation applies to the crochet in bar 14 in the patterns. The rule we followed here is to take the midpoint of the gap between the two notes in questions, and depending on which side the annotation start/end with respect to the midpoint, we take include/exclude the note in the patterns. For example, in both bar 3 and bar 14, we include the quiver and the crochet. Furthermore, we noticed that some annotators forgot to mark the relevance even though their time interval annotations look reasonable. In this case, we give all patterns a relevance score of 1.

As mentioned in the last section, since the annotators were not forbidden to mark the occurrences in addition to the prototype patterns, we do find such occurrences in the photo. We take those occurrences into account as long as a relevance score is given.

Finally, although transcribing from the photo format to the time interval format is relatively time consuming and prone to human errors, we do not know of any matured technology which could be used to automatically convert between the two formats. Optical recognition techniques are promising, but could give low accuracies with the imprecise markings.

In the future, it would be ideal to have a digital system which could be as natural a process for the annotators as paper and pen. For a large scale online experiment, it would only be viable with developing such an annotation system.

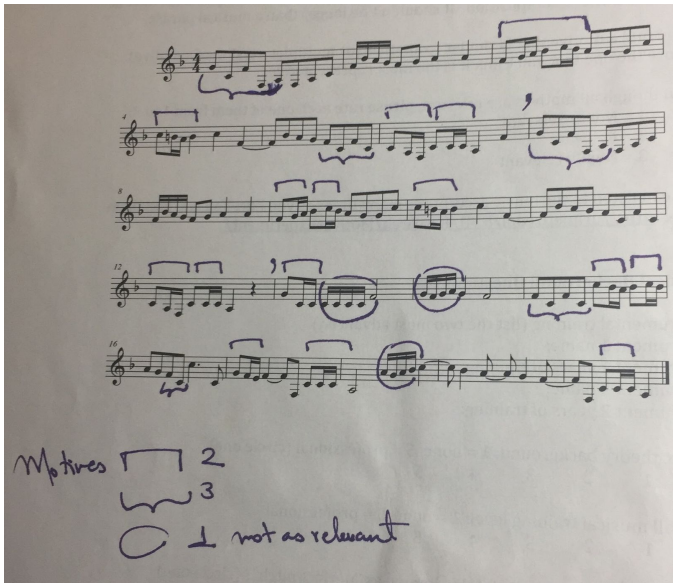

Figure 1. Example of the raw data of annotation

Visualising the time interval data

After converting to the time interval format, we can visualise the data taking the same approach as in (Ren et al 2017) as shown in Figure 2. In this visualisation, we abstract away the actual notes, just preserving the temporal markings in the excerpts. We can see that the disagreement amongst the annotators is prevalent.

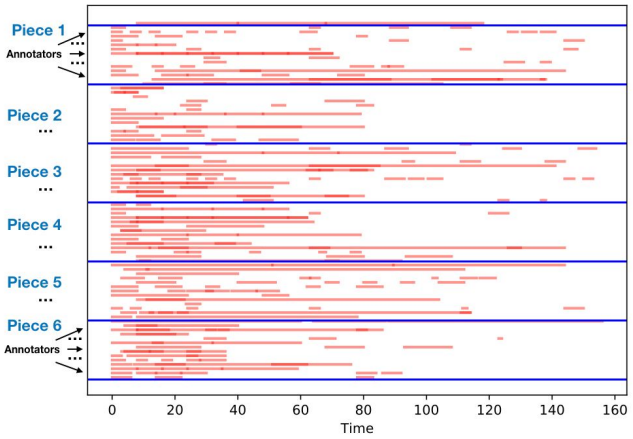

Figure 2. Time intervals in the six musical excerpts from 7 different annotators. Red bars indicate that there exist a pattern, and the absence of red bars indicates the absence of patterns. The blue horizontal lines separate different pieces. The x-axis represents time in the unit of a crochet. The y-axis represents different annotators. Due to limited space, the time intervals of only seven annotators are shown here. We can see different degrees of agreement and disagreement amongst the annotators.

Time intervals to numerical format

After obtaining the time interval data, we took the symbolic music data from a python toolbox, music21. By segmenting the music21 data using the time intervals, the numerical format of a (onset, duration, pitch) triplet was created. Each pattern is thus represented by a succession of triplets and its associated relevance score. We further organise each individual pattern into a "Excerpt -> Annotator -> Relevance -> Pattern" hierarchy in a python dictionary.

Numeric format to JAMS

From the hierarchical structure of the numerical symbolic musical patterns, we further convert the data to the JAMS format. JAMS provides a simple, structured, and sustainable approach to representing rich information in a human-readable, language agnostic format. This format fits our purpose well because it supports multiple types of annotations, multiple annotations for a given task, and rich file level and annotation level metadata. The dictionary format is then converted to the JAMS format using the JAMS python library and a HEMAN parser script, which can be accessed at the official repository of JAMS: https://github.com/marl/jams-data/blob/master/parsers/heman_parser.py. The time-interval representation, the numeric format, and the JAMS files of the dataset can all be accessed online: https://github.com/irisyupingren/HEMANanalysis.

Method

In this section, based on intuition and observation on the dataset and findings about relevance values in the context of segmentation (Bruderer, Mckinney & Kohlrausch, 2009), we investigate our hypothesise that, by taking the most relevant patterns, and lowering the time resolution, we obtain more pairwise annotator agreement in the dataset.

To conduct our computational analysis and verify our hypothesis, we take the time interval representation of the annotations. We do not need to consider the actual notes when analysing pairwise annotator agreement because the only disagreement possible is the starting and/or ending time given a specific piece.

To measure pairwise agreement, we take each individual annotator as the reference and use the standard precision, recall, and F1 score as measurements of agreement (Goutte & Gaussier 2005). A formal definition is given below:

$$\text{Precision} = \frac{\#\ \text{matched annotations}}{\#\ \text{annotations of the referenced annotator}}$$

$$\text{Recall} = \frac{\#\ \text{matched annotations}}{\#\ \text{annotations of the current annotator}}$$

$$\text{F1} = \frac{Precision \times Recall}{Precision + Recall} \times 2$$

The notion of matched annotations is defined on different levels of time resolution: {Annotation1=[begin1, end1], Annotation2=[begin2, end2]} $\in$ {Matched annotations}, if |begin1- begin2| + |end1 - end2| $\leq$ Threshold. The essence of lowering the time resolution of the annotations is taking a larger tolerance on identifying whether two annotations are agreeing (matched) or disagreeing (not matched). In this way, we can see how much disagreement there is on different scales of time resolution.

We use precision, recall, and F1 score instead of the kappa agreement measures like in (Balke et al, 2016) because this approach simplifies the calculation and avoids taking the average across the musical piece.

In addition to numerical methods, we examine the annotations analytically and categorically. In Figure 3, we show an example of disagreement amongst three annotators on the same piece.

We categorise the possible types of disagreement as follows (Annotation1 = [a1,b1], Annotation2 = [a2,b2], symmetrical cases with switched order of the annotators are omitted):

- On the individual pattern level
  - When there is a match
    - a1 < a2, b2 > b1
    - a1 = a2, b2 > b1
  - When there is no match
- On the piece level
  - The number of annotations
  - Whether there are overlaps of patterns
  - Occurrences
    - Exhaustive occurrences
      - with variation
      - without variation
    - Non-exhaustive occurrences
      - Arbitrary
      - Only Prototypes

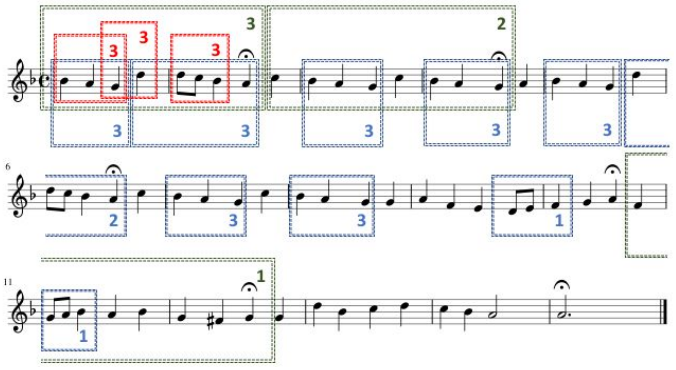

Figure 3. Examples of disagreement. Different colours indicate different annotators. The relevance values are marked with respect to each pattern.

Results

In this section, we show the numerical values of precision, recall, F1 score, and analysis of the annotations. Our initial exploration shows supporting evidence for our hypothesis that by using the top-ranked annotations and lowering the time resolution of the annotations, we may obtain more pairwise annotator agreement in the dataset. We first discuss effects of the relevance score and the time resolution threshold separately, and then see their effects together.

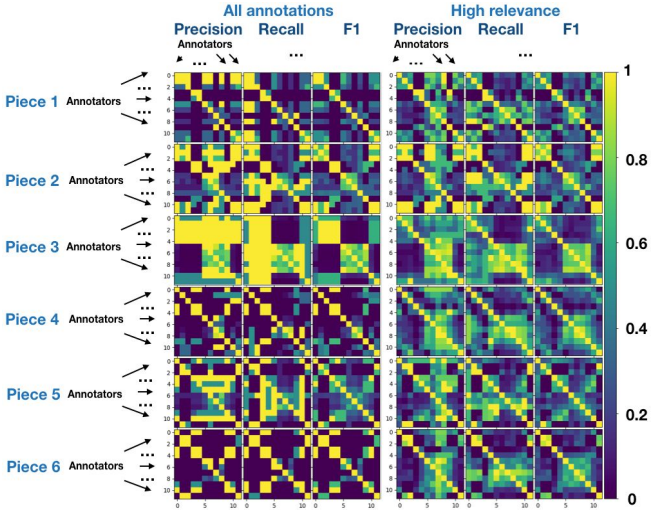

Figure 4. Precision, recall and F1 score across pieces. Different rows denote different pieces and different annotators. Different columns denote different metrics and relevance scores. Within each subfigure, we show the results of 12×12 annotators on one metric $\in$ {Precision, Recall, F1}. Yellow side of spectrum denotes high agreement.

Relevance

In the left subfigure of Figure 4, we show the measurement of agreement using precision, recall and F1 score computed using all annotations; in the right subfigure, we show the precision, recall and F1 score computed using the annotations which are rated as the most relevant.

When considering all annotations, we see a *grouping* phenomenon where subclusters amongst annotators with high

agreement are formed. By considering the high relevance annotations, the grouping phenomenon is reduced and we can see an overall increase of agreement by comparing the results in Figure 4.

Time resolution thresholding

For showing the effects of the time resolution thresholding step, we only show the average across the six excerpts instead of looking at the excerpts individually. The averaged results already show a clear sign of increased agreement with a lower time resolution.

By comparing row by row, we observe in Figure 5 that, with a lower time resolution, that is, a more relaxed threshold, the precision, recall and F1 scores increase. The relaxation of the threshold is effectively loosening the notion of "matched annotations", as mentioned in the last section. For example, if Annotation1 = [a1,b1], Annotation2 = [a1, b1 + 2ε], with threshold = ε, the two annotations are not matched; with threshold = 2ε, the two annotations are matched. With different degrees of loosening, we can examine the pairwise agreement amongst annotators on different scales of time resolution.

In the case where all annotators have the same annotations, taking different time resolution values does not have an effect on the precision, recall and F1 score. In the case of the different annotations, the metrics will reach a stable value as the threshold increases. In our case, we see an increasing trend in the metrics as the threshold increases. Therefore, more pairwise annotator agreement is reached on lower time resolutions.

Relevance ranking and time resolution thresholding combined

In Figure 5, we show the combined effects of relevance ranking and time resolution thresholding. As expected, the agreement increases with the effects from both steps. We therefore conclude that, by using extra rankable metadata on the annotations such as relevance/importance scores, and by adjusting the time resolution of the annotated patterns' time span, we achieve a higher degree of pairwise annotator agreement in the HEMAN dataset.

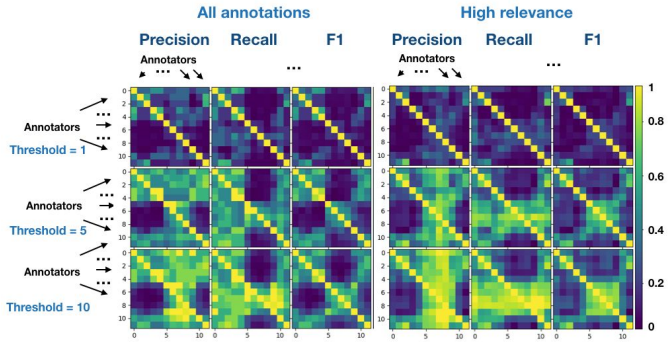

Figure 5. The effects of threshold change and relevance value. The precision, recall and F1 score across pieces are averaged. The threshold values are given as multiples of quarter note length. Other specifications are the same as Figure 4: Different rows denote different pieces and different annotators. Different columns denote results denote different metrics and relevance scores. Within each subfigure, we show the results of 12×12 annotators on one metric $\in$ {Precision, Recall, F1}.

Conclusion

Based on our data, we draw the tentative conclusion that motifs or patterns rated as highly relevant are more trustworthy than those rated with lower relevance values. In addition, by lowering the time resolution of annotations, we gain more agreement amongst the annotators. Therefore, if we choose the patterns of high relevance and threshold the annotations, we could establish an evaluation measure with small irreducible errors for automatically extracted patterns.

In the future, we would like to extend the work by collecting more data from a larger number of subjects using a web interface and verify our current conclusions with more data and further similarity analysis. We could also design other related experiments where we would provide candidate patterns and ask the subjects to choose. Another exploration could be to ask the subjects to annotate and rank the patterns in complete musical pieces rather than excerpts. Finally, for algorithms, we will put the dataset into use for computational pattern extraction evaluation tasks, and see if the current state-of-the-art algorithms can reproduce human annotations with high agreements. The agreement values could also be used as training data for novel computational pattern extraction models, such that the models could give a confidence value when predicting the presence of patterns.

Acknowledgements. We thank the participants of the experiment and everybody who stimulated thought-provoking discussions.

References

Balke, S., Driedger, J., Abeßer, J., Dittmar, C., & Müller, M. (2016). Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings. Proc. of the 15th International Society for Music Information Retrieval Conference (pp. 246-253). New York City, United States.

Benward, B. (2014). Music in Theory and Practice Volume 1. McGraw-Hill Higher Education.

Bernstein, L. (1976). The unanswered question: Six talks at Harvard (Vol. 33). Harvard University Press.

Bruderer, M. J., Mckinney, M. F., & Kohlrausch, A. (2009). The Perception of Structural Boundaries in Melody Lines of Western Popular Music. Musicæ Scientiæ, 13(2), 273–313.

Burkhart, C. (2005). The phrase rhythm of Chopin's A-Flat Major Mazurka, Op. 59, No. 2. Engaging music: Essays in music analysis, 3-12.

Collins, T., Böck, S., Krebs, F., & Widmer, G. (2014, January). Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio. In Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society.

Cross, I. (2005). Music and meaning, ambiguity and evolution. Musical communication, 27-43.

Collins, T. (2011). Improved methods for pattern discovery in music, with applications in automated stylistic composition (Doctoral dissertation, The Open University).

Collins, T., Janssen, B., Ren, I. Y. & Volk, A. (2017). Discovery of Repeated Themes & Sections, 2017. http://www.musicir.org/mirex/wiki/2017:Discovery_of_Repeated_Themes_%26_Sections. Accessed on 20 May 2018.

Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval (pp. 345-359). Springer, Berlin, Heidelberg.

Hiley, D. (1993). Western plainchant: a handbook. Oxford University Press.

Humphrey, J.E., Salamon, J., Nieto, O., Forsyth, J., Bittner, R., Bello, J.P., JAMS: A JSON Annotated Music Specification for Reproducible MIR Research. Proc. of the 15th International Society for Music Information Retrieval Conference (pp. 591-596). Taipei, Taiwan.

Lerdahl, F., & Jackendoff, R. S. (1985). A generative theory of tonal music. MIT press.

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating Hierarchical Structure in Music Annotations. Frontiers in psychology, 8, 1337.

Nattiez, J. J. (1990). Music and discourse: Toward a semiology of music. Princeton University Press.

Nieto, O., & Farbood, M. M. (2012). Perceptual evaluation of automatically extracted musical motives. In Proceedings of the 12th International Conference on Music Perception and Cognition (pp. 723-727).

Randall, D. M. (1999). The Harvard concise dictionary of music and musicians. Harvard University Press.

Ren, I. Y., Koops, V., Volk, A., & Swierstra, W. (2017). In Search Of The Consensus Among Musical Pattern Discovery Algorithms. Proc. of the 18th International Society for Music Information Retrieval (pp. 671-677). Suzhou, China.

Thom, B., Spevak, C., & Höthker, K. (2002, September). Melodic segmentation: evaluating the performance of algorithms and musical experts. International Computer Music Conference.

Sadie, S., & Tyrrell, J. (2001). Dictionary of music and musicians. New York: Oxford University Press. Yónatan Sánchez.

Schoenberg, A. (1983). *Theory of harmony*. Univ of California Press.

Scholes, P. A. (1970). The Oxford Companion to Music. Oxford University Press.