Coffee Bean (and other) Models about the Metric Structure of Music

Anja Volk (Fleischer)
University of Southern California
avolk@imsc.usc.edu

Abstract

This article discusses the question of how to evaluate computational models for metric structures in music. The great amount of computational models developed in this field in recent years has provoked the request to compare the results of these models. Starting from a suggestion to count the number of correct results produced by the model, the article points to some of the difficulties associated with this criterion for a model in music. We give some examples of how the comparison to other models or theories nevertheless can contribute to new insights into the phenomenon of meter in music.

1 Introduction

Florestan and Eusebius[1] study a manuscript about metric theory in music while sitting in a coffeehouse (described by Harald Krebs in [6]). After reading the manuscript's definition of *metrical dissonance* Florestan asks the waiter to bring a handful of coffee beans: "I thought we might find it easier to understand the author's descriptions if we constructed some little models. Why don't you build a coffee bean model of the dissonance G3/2 while I build G5/3?"[2]

Hence they arrange streams of coffee beans on the table in order to illustrate the different metric layers consisting of equally spaced coffee beans

[1] Two fictional characters Robert Schumann invented as incarnations of his various personae.

[2] [6], p. 3

(representing notes in music) which are responsible for creating the effect of metrical dissonance. They count bravely until Florestan smashes his fist on the table and shouts: "Is this manuscript about music or about mathematics? I am sick to death of these numbers and x's and y's!"[3]

Indeed the proximity of the music theoretic problem to a mathematical description seems to have inspired many researchers of the 20th century to ease Florestan's (and other's) life to let the computer automatically count the coffee beans or notes (or beats) of entire pieces:

"In the young and rapidly expanding field of music artificial intelligence, one particularly active area of research has been metrical analysis ... the problem of extracting metrical information from music. Indeed, it would probably be fair to say that no problem in the field has attracted as much attention and energy as this one."[4]

The impressive number of computational models for metric analysis causes Temperley to search for criteria of how to evaluate and to compare these different models:

"Whatever the goals and assumptions of a metrical model, an important and obvious question to ask is, *How good is it*?"[5]

He specifies the last question into the following criteria:

"That is, what percentage of the time does it actually produce the correct result?"[6]

While this is a legitimate measurement for the evaluation of a model (especially in the case of a computational model) I want to discuss in the following some problems evoked by this questions which will lead to a different direction that could be considered in the search for an answer of the question: How good is it?

2 The evaluation problem

A model represents a theoretical construct of the "real" world and usually reduces or simplifies the complexity of the addressed processes in order to

[3] [6], p. 43
[4] [10], p. 28
[5] [10], p. 28
[6] [10], p. 28

allow reasoning within its framework. In the case of a metric model for music we face the nontrivial problem: what actually is the metric structure of a piece of music the model should address? And how can we justify that our model produces the correct result? In a field such as meteorology it is quite straightforward to count the percentage of days the weather forecast gave the right predictions. In the research field of music study we face a more delicate situation. The bar lines of the score might serve as a correct answer, but Florian and Eusebius are counting coffee beans and not bar lines in order to learn how to describe the metric structure of the piece that might differ from the notation of the bar lines. Indeed they discover numerous examples in Schumann's music where the bar lines give us not the ultimate answer.

Another way to evaluate a metric model often suggested is to ask listeners about the metric structure of a specific piece. After admitting that rhythm is difficult to define Scheirer suggests in [9]: "The only ground truth is what human listeners agree to be the rhythmic aspects of the musical content of that signal".[7] But what happens if the listeners do not agree? Hence before even comparing different computational models we face the problem of how to evaluate a single one. Even if there might be a considerable high amount of pieces where the listeners do agree we cannot assume this to be always the case. And perhaps those pieces where they do not agree are exactly the most interesting examples? Gouyon and Dixon in [5] contradict Scheirer's optimistic suggestion by stating that we face the problem of a missing "ground truth" in meter or rhythm description.

Another strategy for the evaluation of computational models for music is to compare the results to existing analyses in music theory. By doing so we test whether the model agrees with the results produced by experts - even though also experts don't necessarily agree in all cases. In other words we test whether the results of the computational model fit to existing knowledge about a piece. But would it not be especially fascinating if a model did not only replicate phenomena we did already know before but tell us something new? Would a disagreement to existing knowledge then always mean a failure of the model?

[7] [9], p. 81

3 The comparison problem

The existence of different computational models for meter in music provokes the question of how to compare them. But the different models cope with different features of the music and produce different forms of output which makes this task difficult, as has been addressed, for instance in [3], [10] or [5]. Some of these models result in a simple beat induction, whereas other address the hierarchic complexity of meter. Even more delicate is the situation that music is located within a communication chain that does not map one to one the production of the sounds to the perception of music. In the case of computational models this fact is reflected by the different data structures that are chosen as input for the models. Does the data represent the events of the symbolic score (such as in MIDI) in an exact (quantized) way or the events as played by a performer (unquantized) or does the data contain the acoustic signal? Gouyon and Dixon consider this as another important argument to undermine the difficulties for the creation of a general evaluation and comparison system:

"However there are several reasons why such an evaluation is not possible for rhythm description. ... since rhythm description systems have been built for diverse applications using diverse data sets".[8]

Facing this problem, Temperley suggests in [10] an evaluation system that only works for models that are based on symbolic score representation. Nevertheless the comparison between models that are based on audio data and those that are based on the score remains especially interesting because they might tell us something about this communication process from the producer of the signal to the receiver. On the other hand the evaluation of a model that is based on audio data calls for the use of methods that work on symbolic data:

"The chief goal in automatic rhythm description is the parsing of acoustic events that occur in time into the more abstract notions of metrical structure, tempo and timing ... "[9]

These more abstract notions contain features lists (such as onsets) that are then mapped to a metrical hierarchy (see figure 2, p. 5 in [5]). Hence the step from the feature list to the metrical structure is similar to a metric model that works on symbolic data such that the search for a combination

[8] [5], p. 49
[9] [5], p. 4

of methods based on audio and symbolic data might be an even more productive approach then the question of how to compare them.

One could argue that the most straightforward way to compare different models is to suggest a common database that all models should be applied to. Such a database is suggested by Temperley who on the other hand admits in [10]:

"Thus, I certainly do not claim that the corpus proposed below would be a suitable or fair one for all of the models listed ... ."[10]

Hence any given database might not be fair for all models but nevertheless it would allow to investigate which models are successful or not in these specific cases. In addition one could also include for each model the most "spectacular" pieces they give good answers to. Both understanding when the model gives a plausible result and when not helps to evaluate the model and to understand something about the phenomenon of meter in music.

What happens if a model gives an unexpected result? Could we modify the criterion "what percentage of the time does it actually produce the correct result"? It implies that we just have to compare the result to something we know otherwise already. If a computational model mostly produces unexpected results we might argue that the model does not describe the phenomenon properly. In cases where a model did produce very often plausible or expected results but gives for a certain piece something that looks "wrong" at a first sight - might this be the case where the model might tell us something new about the phenomenon? Are these not the interesting pieces we should look at instead of just counting the result as wrong? This might lead us to examples where we focus the attention more on the question as to how the comparison of different models help us to understand something new about pieces we applied the models to.

An interesting example in this respect is the discussion of the analysis of the Kyrie I in Bach's Mass in B-Minor in [4], p. 148. The application of the model of *Inner Metric Analysis*[11] to the choral parts of this piece results in an off-beat accentuation in the first half of the piece - the greatest metric accents are located on the weak beats of the bars. In contrast to this the result for the second half of the piece is very different, although this segment is almost a repetition of the first half of the piece. The model

[10] [10], p. 36

[11] a mathematical model that describes the metric hierarchy of a piece based on notes' onsets

assigns in the second half the greatest accents to the main beats of the bars. Since both parts are quite similar these different results are not plausible at first sight. A more detailed analysis reveals that in the first half of the piece the different choral parts create a constant busy flow on the off beats being responsible for creating the off beat pattern in the metric analysis. In contrast to this the second half of the piece contains a cesura in bar 111 that interrupts this busy flow and is responsible for the very different result of the model. This cesura is hence responsible for causing a different computational result that reflects an important compositional effect.

4 A refined strategy

The evaluation of computational models based on the percentage of correct answers is desirable in order to give objective measurements about the performance rate of a model. The decision about how to define the correct answer is highly dependent on the musical test set. In some cases most people might agree as to what the correct answer is. But in other situations this agreement is less likely. Hence such a measurement might be difficult to achieve for metric models in music. In the following we want to discuss some examples where the comparison to other models is not based on such a quantitative criteria but might help us to learn something about the research subject - about the metric structure of music.

This author applies in [4] the model of *Inner Metric Analysis* to a test set proposed by Povel and Essens in [8]. Surprisingly the results of Inner Metric Analysis are very similar to the results of the algorithm in [8] when applied to these pieces. Whereas Inner Metric Analysis is based on metric pulses that consist of equally spaced onsets, Povel's and Essen's model is based on a rhythmic accentuation that arises from certain grouping rules. Because of the very different features these two models take into account we cannot expect in general similar results. But the consistency of the results between the two models when applied to the test set in [8] points towards a special solidarity between rhythmic grouping and meter in this data set. An example is constructed in [4], p. 171 where both models would not agree on the result that demonstrates that this relation cannot be generalized but is specific to the chosen set of examples.

Another example that demonstrates how the comparison of different models might bring new perspectives to our knowledge about music is Brahms'

Intermezzo Op. 76 No. 8. Lewin and Cohn state in [7] and [2] a deep affinity between metric and harmonic processes in this piece. In [11] the computational models of Inner Metric Analysis and Chew's Spiral Array for tonal analysis (see [1]) are applied to the piece. Both models are based on very different assumptions in comparison to the arguments given by Lewin and Cohn. Nevertheless the results of the computational models agree with those by the music theorists to a great extend. Hence the models reconfirm Lewin's and Cohn's observation from a different perspective suggesting that the affinity between metric and harmonic processes in this piece is indeed an intriguing phenomenon. The music theoretic findings by Lewin and Cohn thereby did not serve as "the" correct answer concerning the metric structure in this piece which should be replicated by any other model in order to test it. This piece reaches a complexity both concerning tonal and time structures that makes the search for "the" correct answer very difficult. Hence the comparison between the results of different models is part of the process to explore and characterize this complexity.

Florestan and Eusebius apply the metric model suggested in the manuscript they are reading in the coffeehouse to a large amount of examples in Schumann's Œuvre in [6]. After changing their seats at the coffeehouse with comfortable easy chairs at home they extend their observations to composers who had been a major influence on Schumann's compositions. As one examples for metric dissonance they mention the last bars of the first movement of Beethoven's Eroica (see Figure 1).

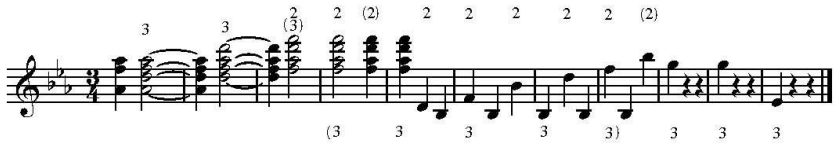

Figure 1: Harald Krebs' analysis of mm. 681-690 of Beethoven's *Eroica*, 1st mvmt

This example contains pulses of competing periods of three and two which overlap each other causing metrical dissonance. Earlier in the manuscript about metric theory Krebs states the following:

"I define the meter of a work as the union of all layers of motion (i.e., series of regularly recurring pulses) active within it" ([6], p. 23).

This rises the question whether the competing layers in Figure 1 can be understood to a certain extend as independent metric structures which would imply the existence of competing metric hierarchies. The model of Inner Metric Analysis calculates from the superposition of all existing pulses in a piece a metric accent for each note. It also allows the exclusion of certain pulses from the calculation which may help answer the question whether competing layers can be understood as competing metrical hierarchies. Figure 2 shows the result of the analysis of the example from the Eroica in figure 1 after excluding the pulses of period 2 (left picture) and excluding the pulses of period 3 (right picture). Thereby the lines in the foreground indicate the metric accent or weight of the corresponding time point in the score (the higher the line the higher the weight) whereas the background marks the bar lines of the score.

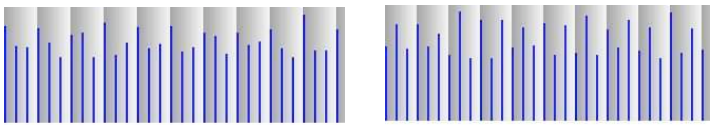

Figure 2: Metric analysis of mm. 681-690 of Beethoven's *Eroica*, 1st mvmt. without pulses of period 2 (left) and without pulses of period 3 (right)

Excluding all pulses of period 2 induces a metric hierarchy with the greatest weights located on the beginning of the bar lines and a lower level induced by the weights of the second and third beats of the bars. This reflects the typical metric hierarchy associated with the notated 3/4 time signature. On the other hand excluding all pulses of period 3 induces two different layers which can be interpreted as the metric hierarchy associated with a 2/4 time signature. Hence we can indeed argue that the model of Inner Metric Analysis shows how these competing layers form independent metric hierarchies within the piece - a discovery which was new to the author[12] of [6].

We might have assumed that a computational model about meter induction would just save Florestan and Eusebius the time to lay and count the beans for entire pieces - but in some cases we might end up detecting something new we did not know already before. These cases could be considered as the most exciting discoveries a computational model might

[12] According to a personal conservation with Harald Krebs on November 12, 2004

reveal that should be part of answering the question, of *how good* the model is. Hence the percentage of correct results, which implies the comparison to something we did already know, should not be the only criteria for evaluating a computational model for meter induction.

References

[1] Chew, E. *Towards a Mathematical Model of Tonality*, Ph.D. dissertation. Operations Research Center, MIT. Cambridge, MA. 2000.

[2] Cohn, R. *Complex Hemiolas, Ski-Hill Graphs and Metric Spaces*, in: *Music Analysis*, 20:iii, 2001, 295-326.

[3] Eck, D. *Finding Downbeats with a Relaxation Oscillator*, in: *Psychological Research*, 66:1, 2002, 18-25.

[4] Fleischer, A. *Die analytische Interpretation. Schritte zur Erschließung eines Forschungsfeldes am Beispiel der Metrik*, dissertation.de - Verlag im Internet Gmbh, Berlin 2003.

[5] Gouyon, F. and Dixon, S. *A review of automatic rhythm description systems*, in *Computer Music Journal*, 29:1, 2005, 34-54.

[6] Krebs, H. *Fantasy Pieces*, Oxford University Press 1999.

[7] Lewin, D. *On Harmony and Meter in Brahms's Op. 76, No. 8*, in: *19th-Century Music*, 4:3, 1981, 261-265.

[8] Povel, Dirk-Jan & Peter Essens. *Perception of Temporal Patterns*. In: *Music Perception*, Summer 1985, 2:4, 411-440.

[9] Scheirer, E. *Music-Listening System*, Ph.D. dissertation. MIT Media Laboratory. Cambridge, MA. 2000.

[10] Temperley, D. *An Evaluation System for Metrical Models*, in: *Computer Music Journal* 28:3, 2004: 28-44.

[11] Volk, A. and Chew, E. *Re-Considering the Affinity between Metric and Tonal Structures in Brahms' Op.76 No.8*, Annual Meeting of the Music Theory Society of New York State, April 2005.