Journal of Mathematics and Music, 2023
https://doi.org/10.1080/17459737.2023.2213471

Taylor & Francis
Taylor & Francis Group

Check for updates

Minimum description length for selection of models of musical rhythm

Verónica Rumbo[a*], Ernesto Mordecki[b], and Martín Rocamora [c]

[a]*Instituto de Matemática y Estadística "Rafael Laguardia", Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay;* [b]*Centro de Matemática, Facultad de Ciencias, Montevideo, Uruguay;* [c]*Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Montevideo, Uruguay*

(*Received 19 March 2021; accepted 26 April 2023*)

We apply the minimum description length (MDL) methodology from information theory in order to select mathematical models for musical rhythm. We consider six models proposed by David Temperley and mathematically formalize them, which allows for an MDL analysis. As a consequence, we find that two of the models are not suitable to apply this methodology, as their codification strategy does not represent all possible rhythm sequences. A slight modification of Temperley's hierarchical model provides an improvement in codification performance, robust across all the four corpora considered in the experiments: three classical corpora commonly used in music studies and one of tango songs recently released. Our study confirms the usefulness of the MDL approach to solve the classical trade-off between model complexity and its ability to fit the data.

Keywords: Musical rhythm; probabilistic modelling; information theory; minimum description length; model selection

2010 Mathematics Subject Classification: 00A65; 68Q30; 62B10
2012 Computing Classification Scheme: Applied computing

1. Introduction

The phenomenon of rhythm is of paramount relevance in music. In a general sense, it refers to the placement of sounds in time, organized and grouped, forming different structures and patterns. Several research fields, such as music theory, cognitive science, and musicology, have devoted increased attention to musical rhythm during the past few decades. This gave rise to novel theories of rhythmic organization (Lerdahl and Jackendoff 1985) and music perception (London 2004), as well as the development of new computational models of rhythm (Temperley 2001, Mavromatis 2012, Toussaint 2013).

The concept of rhythm involves patterns of organized durations that are phenomenally present in the music (London 2004). One of the most important contributions to the new conception of rhythmic organization is that of Lerdahl and Jackendoff (1985), in which rhythm perception is thought to be structured by an underlying metric framework. This metrical structure is a mental construct that consists of a regular pattern of points in time (called beats) that are hierarchically

*Corresponding author. Email: verorumb@yahoo.com, vrumbo@fing.edu.uy

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2023 Informa UK Limited, trading as Taylor & Francis Group

organized into levels. It is not directly heard, yet, it can be inferred by the listener from the actual events present in the musical piece (even if they do not coincide in time with the beats). The organized pulsation of the metrical structure provides specific anchor points to structure the rhythm (London 2004).

In Western tonal music, the metre of a piece is usually specified at the beginning of the musical score through the time signature, such as $\frac{4}{4}$, $\frac{2}{4}$, or $\frac{3}{4}$. The upper number indicates the number of pulsations within the bar, and the lower one indicates their durational value. The metre hierarchy is constructed by recursively dividing (or grouping) the main pulse into lower (or higher) metrical levels. Figure 1 shows an example of a $\frac{4}{4}$ metre hierarchy. This time signature defines a metre of four quarter-notes in each bar. Each metrical level, from the whole-note level to the eight-note level, is subdivided into two equal parts. The lowest metrical level, the eight-note level in this case, defines a grid of metrical positions within the bar. The elements of each metrical level at each position are indicated with $*$ symbols below the score of Figure 1. A metrical strength for each position can be derived from this hierarchy. The more elements of the metrical structure present at a given position, the higher its metrical strength.

One of the most appealing features of computational models is that they allow for the analysis of large bodies of data capturing regularities and patterns at a scale not feasible by traditional means. In this way, the models can be used to assess the validity of certain theoretical hypotheses about music based on large music collections. In addition, models for musical rhythm have several practical applications in music information retrieval (Gómez et al. 2016), such as music search and recommendation, or automatic music transcription and generation. For these reasons, the scientific community has undertaken a long-standing effort at the development of computational models for rhythm and metre (see for instance Desain and Honing 1999, Dixon 2001, Ellis 2007, Volk 2008, Naveda et al. 2011, Srinivasamurthy, Holzapfel, and Serra 2014, Bello et al. 2015.). This paper deals with such models for musical rhythm, in particular with the approach of Temperley (2004, 2007, 2009, 2010), which has largely influenced subsequent work (see for instance Vuust and Witek 2014, Sioros, Davies, and Guedes 2018, Pesek, Leonardis, and Marolt 2020). His approach is based on a generative process similar to that proposed by Lerdahl and
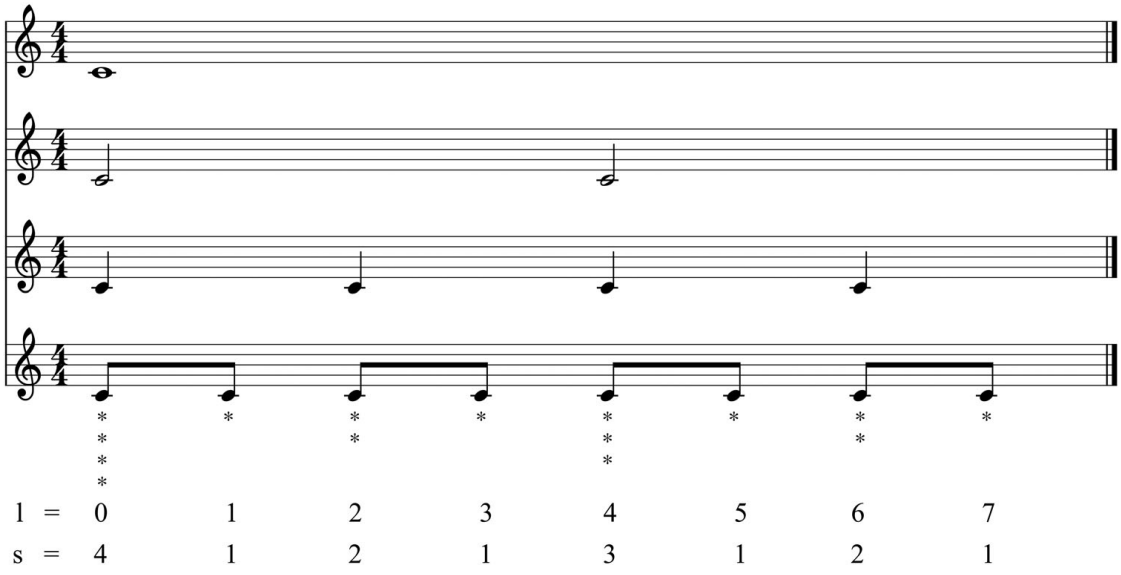


Figure 1. Example of the metrical grid considered for a time signature of $\frac{4}{4}$, consisting of eighth-note, quarter-note, half-note, and whole-note levels. Note that ℓ indicates the metrical position, the $*$ symbols denote the elements of each metrical level at each position, and s shows their corresponding metrical strength.

Jackendoff (1985), and postulates that metre is inferred from the probabilities of different patterns of regularity generated by a given rhythmic input.

Our main sources of inspiration are the works by Temperley (2010) and Mavromatis (2009). In the present paper we follow Temperley (2010), analyzing and slightly modifying the rhythm models he proposed, while from Mavromatis (2009) we borrow the methodology applied for model selection. Mavromatis (2009) studies Greek church chant using the MDL principle to find the best fitting parameters in a hidden Markov model.

In Temperley (2010), the author considers six probabilistic models of musical rhythm and compares them according to their likelihood on two different data sets of common–practice music, the Essen Folksong Collection and a corpus of Mozart and Haydn string quartets. The set of models ranges from simple independent and identically distributed Bernoulli trials (*uniform position model*) to more sophisticated conditional probability models based on the metrical structure (such as the *hierarchical position model*).

Faced to the fact that models with more number of parameters give better fits, Temperley argues that not only the likelihood of the data should be taken into account but also the complexity of the model that produces this likelihood. Thus, he chooses the hierarchical position model, which is not as good as the best performing one but has fewer parameters. This criterion is a version of the *law of parsimony*, also known as *Occam's razor* principle.

The law of parsimony is a classical issue in statistical modelling: more complex models tend to produce better likelihood results than simpler ones; though simpler models are pursued and reasonably preferred than complex ones. As noticed by Temperley (2010), this compromise arises in the comparison of the six proposed models. The complexity of a (parametric) model can be measured as the complexity of the model's parameter set. Different approaches have been proposed to deal with this trade–off. Traditional ones include the Akaike information criterion (AIC) introduced by Akaike (1973) and the Bayesian information criterion (BIC) proposed by Schwarz (1978). The AIC compares the log-likelihood given by the data with the number of parameters of the model, whereas in the BIC the comparison is with respect to the number of parameters multiplied by the logarithm of the sample size.

Within other possible existing alternatives for model selection we choose the *minimum description length* (MDL) principle, a method based on information theory introduced by Rissanen (1978). The aim is to provide a measure of the amount of information conveyed by the data, which can be interpreted as the length of its most compact description. In this setting, the messages to be encoded are supposed to be outcomes of a known random source, whose characteristics determine the encoding. Therefore, given a random source of known characteristics, we are interested in the minimum expected number of bits per symbol to transmit a message from the source through an error-free channel. This is the data compression paradigm that was introduced in the foundational work on information theory by Shannon (1948). His source coding theorem establishes the limits to possible data compression stating that it is not possible to losslessly compress the data using an average number of bits per symbol (i.e. coding rate) smaller than the entropy of the source. In this context, entropy is the expected value of the information contained in each message, where information is defined as the cologarithm of the symbol probabilities. The MDL approach allows the quantification of the complexity of both the data and the model under the same basis, providing a reasonable strategy for model selection. We refer the interested reader to Section 2 of Mavromatis (2009), which provides a detailed and gentle introduction to this approach. In particular, Mavromatis discusses the fundamental idea of leveraging regularity for data compression, the problem of overfitting, and the two-part code approach in MDL (used in the present work), with some illustrative examples. A comprehensive treatment of MDL can be found in Grunwald (2007).

1.1. *Our contributions*

The main motivation of our work is to provide more comprehensive quantitative arguments to carry out the model selection problem posed by Temperley (2010). To do that, we rely on an information theoretic basis, applying the MDL principle. Our main contributions can be summarized in two points: (i) a detailed analysis of the models proposed by Temperley (2010), and their mathematical formalization; (ii) an implementation of MDL for model selection that seeks for the optimal value of the precision used for describing the parameters of each model. Additionally, the formal analysis based on information theory allows us to clarify that two of the models proposed in Temperley (2010) actually ignore some feasible rhythm sequences. For this reason, we exclude these two models from the comparison (namely the *zero-th order duration model* and the *first-order metrical duration model*). Regarding the data used in the computational experiments, we depart from the classical corpora commonly used in these music studies and introduce a novel music corpus of tango songs recently compiled by Satorre (2018). This broadens the musical genres available for computational modelling of music. We released a public repository with the software that implements the models and experiments. This allows the reproduction of the reported results and the analysis of other music corpora.

The rest of the paper is organized as follows. The next section describes how the rhythmic information is represented and introduces the probabilistic models considered. Section 3 presents our approach for model selection using MDL. The experiments and results are described in Section 4. The paper ends with a critical discussion of the work.

2. Models and data

2.1. *The data*

The generic input of our models is the rhythmic information of a certain music corpus. Given a musical piece in the corpus, its rhythmic information is encoded as a pattern of note onsets, disregarding note durations. To do this, following Temperley (2010), we consider a metrical grid consisting of eighth-note, quarter-note, half-note, and whole-note levels (see Figure 1). For the quantization of onset locations we use the most refined of these metrical levels, which is the eighth-note level in this case[1]. Therefore, we build a binary sequence at the eighth-note level indicating the presence of a note onset at each position – i.e. a one for an onset and a zero otherwise.

The encoded music corpus $\mathcal{D}$, namely the data set, is the set of encoded music pieces, written as

$$\mathcal{D} = \{x^0, x^1, \ldots, x^{C-1}\},$$

where x^j is a piece and C denotes the total number of pieces in the music corpus. Then, the encoded piece x^j is written as

$$x^j = x_0^j x_1^j x_2^j \cdots x_{n^j-1}^j,$$

where $x_i^j \in \{0, 1\}$ indicates the presence or absence of an onset at the i-th beat and n^j is the total number of beats in the piece. Note that the total number of beats in the data set is $N = \sum_{j=0}^{C-1} n^j$. For the sake of brevity, the super-index j is omitted when referring to a concrete piece.

[1] Note that other metrical levels could be used as reference, such as the sixteenth-note level.

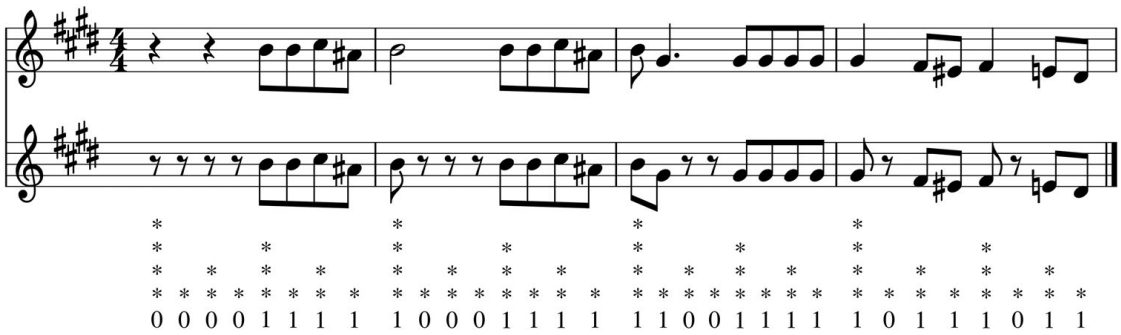


Figure 2. Example of a melody (fragment of *Madame Ivonne* from Satorre's tango corpus) encoded to the eighth-note level. Its corresponding metrical structure is also depicted with $*$ symbols.

In some cases, it is useful to divide the complete piece into patterns of one bar in length (i.e. a whole-note duration). In this way, a piece x is written as the sequence

$$x = m_0 m_1 m_2 \cdots m_{M-1},$$

where m_k represents a one-bar pattern and M is the total number of bars in the piece.

If the number of eighth-note beats in a bar is denoted as L, we encode each one-bar pattern $\mathrm{m_k}$ as

$$m_k = x_0 x_1 x_2 \cdots x_{L-1},$$

where $x_\ell \in \{0, 1\}$ and ℓ is the index of the beat within the k-th bar.

The process is depicted in Figure 2 for the first four bars of a popular tune. The upper part of the figure shows the original score. The lower part illustrates the mapping of the notes to the eighth-note level and shows the metrical levels depicted with $*$ symbols (as in Figure 1). To represent the mapping, the notes are re-written with eighth-note duration at the location of the onsets, and eighth-note rest are added whenever needed. This clearly conveys the binary sequence that results from the encoding and that is depicted at the bottom.

As shown in Figure 2, a musical piece with a time signature of $\frac{4}{4}$ encoded at the eighth-note level yields one-bar patterns of 8 beats. Other situations can be treated analogously. If the time signature of the piece is $\frac{2}{4}$, then the eighth-note level encoding gives patterns of 4 beats. For this reason, the musical pieces of a corpus are grouped by time signature and each group is treated separately, resulting in different data sets.[2]

2.2. *Probabilistic modelling*

By a *model* we mean a probability defined on the sample space of all binary sequences of a given finite length.[3] Each element of this space represents a possible rhythmic sequence corresponding to a musical piece. We furthermore assume that each (observed) musical piece of a certain corpus (the sample) is a realization of a random vector, distributed according the aforementioned probability, that we pursue to estimate. In our work we consider the six models introduced by Temperley (2010) and propose a seventh one, being a minor modification of one of them. As our goal is to understand the underlying rhythmic structure of music, we apply a model selection methodology based on the minimum description length approach to decide which of this models better fits the data, i.e. which of the models better describes the rhythmical nature of the considered corpora.

[2] This is not the only reason. The models discussed in the following also treat different time signatures differently.

[3] Mavromatis (2012) uses the term *probabilistic musical grammar.*

2.3. *Temperley's models*

In the following we formally revisit the six models presented in Temperley (2010), keeping their original naming. We also give the rationale for excluding two of the models from the MDL comparison.

2.3.1. *Uniform position model* (M_1)

We begin by presenting the simplest of the models. It is assumed that in each beat an onset occurs with probability $\theta \in (0,1)$, independently of the other ones. The resulting random vector has Bernoulli independent coordinates with success probability equal to θ. The log-likelihood of a piece x^j in this model is[4]

$$\log \mathbf{P}_1(x^j \mid \theta) = \left(\sum_{i=0}^{n^j-1} x_i^j\right) \log\theta + \left(n^j - \sum_{i=0}^{n^j-1} x_i^j\right) \log(1-\theta).$$

Then, the likelihood of the whole data set is

$$\mathbf{P}_1(D \mid \theta) = \prod_{j=0}^{C-1} \mathbf{P}_1(x^j \mid \theta).$$

It follows that the the maximum likelihood estimator is

$$\hat{\theta} = \frac{1}{N} \sum_{j=0}^{C-1} \sum_{i=1}^{n_j} x_i^j,$$

which is simply the rate of 1s in the data set.

2.3.2. *Zero-th order duration model* (M_2)

The description of this model, as presented by Temperley (2010), goes as follows. We start by sampling the position of the first onset from a certain distribution π_0. Then, we sequentially sample the number of consecutive zeros up the next onset (i.e. the inter–onset interval, IOI) from another distribution π_1, independently of the initial position and independently from each other. The process is repeated until we fill n beats.

Criticism

To construct this model it is necessary to specify the range $\{0,1,\ldots,n_0\}$ of π_0 and $\{0,1,\ldots,n_1\}$ of π_1. But, given that n_0 and n_1 are finite, not all possible sequences are included. In particular, the number of consecutive zeros in the encoded piece is always smaller than $\max(n_0,n_1)$. For instance, if $n_0 = n_1 = 16$, as in Temperley (2010), there is always an onset within 16 beat positions, hence, pieces with longer silences can not be encoded. Therefore, this model is not suitable for the MDL comparison.

[4] As usual in information theory we use $\log x = \log_2 x$.

2.3.3. *Metrical position model* (M_3)

This model is based on the metrical position of beats, in particular, its formulation uses the notion of *metrical strength*. Given a certain position ℓ within a bar of length L, its metrical strength s_ℓ is the number of metrical levels that correspond to it. For instance, given a musical piece with a time signature of $\frac{4}{4}$ encoded at the eighth-note level, as in the example of Figure 1, the first beat position $\ell = 0$ has a metrical strength of $s_\ell = 4$, while the position $\ell = 4$ has $s_\ell = 3$.

In this model, the positions with the same metrical strength are grouped together. In the example of Figure 1, there are only four groups: (i) position $\ell = 0$; (ii) position $\ell = 4$; (iii) positions $\ell = 2$ and $\ell = 6$; (iv) positions $\ell = 1, 3, 5, 7$. Then, the probability of the occurrence of an onset at a certain position depends on the metrical strength of that position, independently of each other. The model has as many parameters as groups resulting from the metrical levels considered. In the case of the example these are

$$\begin{aligned} \theta_4 &= \mathbf{P}(X_\ell = 1), \text{ for } \ell = 0, \qquad \theta_3 = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 4, \\ \theta_2 &= \mathbf{P}(X_\ell = 1), \text{ for } \ell = 2, 6, \quad \theta_1 = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 1, 3, 5, 7. \end{aligned}$$

where X_ℓ is the random variable that models the occurrence of an onset at the position ℓ within the bar. The likelihood of a bar m given these parameters is

$$\begin{aligned} \mathbf{P}_3(m \mid \theta) = &\left(\prod_{\ell=0} \theta_4^{x_\ell} (1-\theta_4)^{1-x_\ell} \right) \left(\prod_{\ell=4} \theta_3^{x_\ell} (1-\theta_3)^{1-x_\ell} \right) \\ &\left(\prod_{\ell=2,6} \theta_2^{x_\ell} (1-\theta_2)^{1-x_\ell} \right) \left(\prod_{\ell=1,3,5,7} \theta_1^{x_\ell} (1-\theta_1)^{1-x_\ell} \right). \end{aligned}$$

Then, the likelihood of a piece x in this model is

$$\mathbf{P}_3(x \mid \theta) = \prod_{k=0}^{M-1} \mathbf{P}_3(m_k \mid \theta). \tag{1}$$

Finally, the likelihood of the whole data set is

$$\mathbf{P}_3(D \mid \theta) = \prod_{j=0}^{C-1} \mathbf{P}_3(x^j \mid \theta). \tag{2}$$

The corresponding estimators are

$$\begin{aligned} \hat{\theta}_4 &= \tfrac{8}{N} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} \sum_{\ell=0} x_\ell, \quad \hat{\theta}_3 = \tfrac{8}{N} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} \sum_{\ell=4} x_\ell, \\ \hat{\theta}_2 &= \tfrac{4}{N} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} \sum_{\ell=2,6} x_\ell, \quad \hat{\theta}_1 = \tfrac{2}{N} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} \sum_{\ell=1,3,5,7} x_\ell. \end{aligned}$$

This corresponds to just counting the number of 1s for each metrical strength along the whole data set. In general, the number of parameters depends on the number of metrical strengths given the time signature and the metrical level used as reference.

2.3.4. *Fine-grained position model (*M_4*)*

As its name suggests, this model can be regarded as a refined version of M_3, in which the grouping by metrical strength is not done. Therefore, the occurrence of an onset in a certain beat is independent of the others and only depends on its metrical position within the bar.

The model has then as many parameters as beats per bar. For instance, in the example of Figure 1 ($\binom{4}{4}$ at eighth-note level) there are eight parameters, defined as:

$$\theta_\ell = \mathbf{P}(X_\ell = 1), \text{ for } \ell \in \{0, \ldots, 7\}.$$

Then, the likelihood of a bar m is

$$\mathbf{P}_4(m \mid \theta) = \prod_{\ell=0}^{7} \theta_\ell^{x_\ell} (1 - \theta_\ell)^{1-x_\ell}.$$

As in the previous model, the likelihood of a piece and the likelihood of the whole data set are computed according to Equations 1 and 2, respectively.

The corresponding estimators are

$$\hat{\theta}_\ell = \frac{8}{N} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} x_\ell, \text{ for } \ell \in \{0, \ldots, 7\}.$$

Note that in both models, $\mathbf{M_3}$ and $\mathbf{M_4}$, the sequences x are independent Bernoulli random variables with different success probabilities.

2.3.5. *Hierarchical position model (*M_5*)*

In this model, not only the metrical strength of each position is used (as in $\mathbf{M_3}$), but also the hierarchical relation among the metrical levels. The occurrence of an onset in a certain beat has a probability that depends on its metrical strength, but is also conditioned by the presence (or absence) of an onset at the neighbouring positions that correspond to higher metrical levels. In order to do that, the notion of *anchor type* is defined.

The following *anchor types* can be distinguished for a certain beat position

un-anchored if none of the two neighbouring higher-level beats have an onset;
pre-anchored or **post-anchored** if only the previous or the next of the two neighbouring higher-level beats has an onset;
both-anchored if both of the two neighbouring higher-level beats have an onset.

Note that the neighbouring positions used to define the anchor type do not necessarily correspond to the closest ones.

Then, the probability of an onset at a certain position within the bar depends on the metrical strength of the position but is also conditioned by its anchor type. This model is similar to the metrical position model ($\mathbf{M_3}$), in the sense that the positions with the same metrical strength are grouped together. However, considering the anchor type of the position gives rise to four different cases. Thus, instead of a single parameter for each metrical strength, as in the $\mathbf{M_3}$ model, four parameters are needed to take into account the anchor type. In the case of the example of Figure 1,

these parameters are,

$$\theta_4^\eta = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 0, \qquad \theta_3^\eta = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 4,$$
$$\theta_2^\eta = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 2, 6, \quad \theta_1^\eta = \mathbf{P}(X_\ell = 1), \text{ for } \ell = 1, 3, 5, 7.$$

where ℓ is the position within the bar, and $\eta \in$ {0-0,0-1,1-0,1-1} represent the un-, post-, pre-, and both-anchored conditions, respectively.

Note that the highest metrical level, position $\ell = 0$, is actually not anchored, since it does not have a higher metrical level to be anchored to. Therefore, the group of parameters corresponding to the highest metrical strength has in fact a single parameter, $\theta_4^\eta = \theta_4$. Moreover, note that this is exactly the same θ_4 parameter defined for the $\mathbf{M_3}$ model, and also corresponds to the θ_0 parameter of the $\mathbf{M_4}$ model.

The resulting probability vector θ is more complex than in the previous models. However, we can concisely write the likelihood of a bar m for the example as

$$\mathbf{P}_5(m \mid \theta) = \prod_{\ell=0}^{7} (\theta_s^\eta)^{x_\ell} (1 - \theta_s^\eta)^{1-x_\ell}.$$

where s∈{0,1,2,3} is the metrical strength group, η represents the current anchor type, and ℓ is the position within the bar. Note that the values of s and η actually change according to the value of ℓ. Once again, the likelihood of a piece and the likelihood of the whole data set are computed according to Equations 1 and 2, respectively.

It is important to note that for determining the anchor type for some positions ($\ell = 4, 6, 7$) it is necessary to check for the presence of an onset at the beginning ($\ell = 0$) of the next bar, i.e. at the downbeat. This has some implications on how to deal with the last bar of the piece. One could consider the piece only up to the last downbeat, inclusive, thus omitting most of the last bar, as done in Temperley (2010). However, for the experiments reported in Section 4 we opted for encoding the last bar by adding an extra downbeat with an onset after the end of the piece.

The corresponding estimators are

$$\hat{\theta}_s^\eta = \frac{1}{N_s^\eta} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} \sum_{\ell \in C_s^\eta} x_\ell.$$

where C_s^η is the set of positions within the bar that share the same metrical strength s and anchor type η, and N_s^η is the total number of such positions in the whole data set.

2.3.6. *First-Order metrical duration model* (M_6)

This model is a modification of the *Zero-th Order Duration Model* ($\mathbf{M_2}$). The first onset is determined with a distribution π_0 with support $\{0, 1, \ldots, n_0\}$. After this, a metrical position is chosen for the next onset, conditional on the metrical position of the previous onset. This can be regarded as a first-order Markov model with finite state space $\{0, 1, \ldots, n_1\}$. The process is repeated until we fill n beats.

Criticism

Similar issues to those discussed for the second model ($\mathbf{M_2}$) apply in this case. First, the value of n_0 arbitrarily restricts the position of the first onset of the piece. Secondly, chosen the metrical position for the next onset actually assumes that an onsets has to occur within the current or the next bar, limiting the number of consecutive zeros in the encoded piece. Therefore, this model is also not suitable for the MDL comparison.
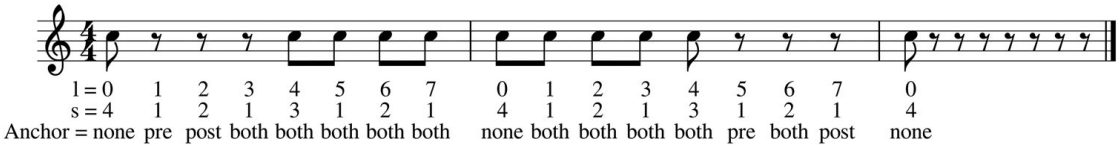


Figure 3. The first two bars have the same probability according to the $\mathbf{M_5}$ model but different probabilities according to the $\mathbf{M_7}$ model. The third bar is given to set the anchor type of the positions of the second bar.

2.4. *A new model proposal*

In this section we propose a new model that is a slight variation on the *Hierarchical Position Model* ($\mathbf{M_5}$) proposed by Temperley (2010). The motivation arises from the fact that in the $\mathbf{M_5}$ model the conditional probabilities of positions with the same metrical strength coincide. For instance, in a $\frac{4}{4}$ bar at the eighth-note level, the conditional probabilities of the positions 2 and 6 are equal. As a consequence of this fact, the model assigns the same probability to the first two bars of the example of Figure 3. In the search of a model that better fits the data, we propose to consider different conditional probabilities for different positions in the bar. This proposal is related with the concept of *indispensability* in music theory, introduced by Barlow (1987), that consists in ranking the positions within a bar according to a given metrical relevance, assigning a different integer value to each position.

2.4.1. *Fine-grained hierarchical position model* (M_7)

This model is a refined version of $\mathbf{M_5}$ but without the grouping by metrical strength. Thus, the occurrence of an onset in a certain beat only depends on its metrical position within the bar and on its anchor type. The model has a group of four parameters for each position within the bar, corresponding to the four different anchor types, except for the downbeat ($\ell = 0$) which has no anchor type at all. In the case of the example of Figure 1, these parameters are,

$$\theta_\ell^\eta = \mathbf{P}(X_\ell = 1), \text{ for } \ell \in \{0, \ldots, 7\} \text{ and } \eta \in \{\text{0-0}, \text{0-1}, \text{1-0}, \text{1-1}\}.$$

The likelihood of a bar m for the example is

$$\mathbf{P}_7(m \mid \theta) = \prod_{\ell=0}^{7} (\theta_\ell^\eta)^{x_\ell} (1 - \theta_\ell^\eta)^{1-x_\ell}.$$

where η represents the current anchor type, and ℓ is the position within the bar. As before, the likelihood of a piece and the likelihood of the whole data set are computed according to Equations 1 and 2, respectively.

The corresponding estimators are

$$\hat{\theta}_\ell^\eta = \frac{1}{N_\ell^\eta} \sum_{j=0}^{C-1} \sum_{k=0}^{M^j-1} x_\ell.$$

where N_ℓ^η is the total number of positions ℓ that share the same anchor type η in the whole data set. Note this is the same refinement applied to model $\mathbf{M_3}$ to obtain $\mathbf{M_4}$.

2.5. *Model's complexity as number of parameters*

It is worth noting that the presented models have different number of parameters, which reflects their different complexity. The number of parameters depends on the time signature and the

Table 1. Number of parameters of the different models for some time signatures and encoding levels.

Model	Time signature					
	8th-note encoding			16th-note encoding		
	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{4}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{4}{4}$
Uniform position (M_1)	1	1	1	1	1	1
Metrical position (M_3)	3	3	4	4	4	5
Fine-grained position (M_4)	4	6	8	8	12	16
Hierarchical position (M_5)	9	9	13	13	13	17
Fine-grained hierarchical position (M_7)	13	21	29	29	45	61

metrical level used for encoding. As an example, Table 1 shows the number of parameters of the models for some time signatures and encoding levels. To take into account this complexity in the model selection, we apply the MDL principle as described in the following section.

3. MDL model selection

We briefly review some elements of the minimum description length (MDL) methodology applied to select the best fitting model. The method is based on information theory, and we refer to Cover and Thomas (2006) in this respect, where our main reference regarding MDL is Grunwald (2007). The departure point is the encoded data set $\mathcal{D}$ that we want to compress, in which the rhythmic information of each musical piece is expressed as a sequence of zeros and ones. To do that, we choose a probabilistic model M and we compute the (expected) length of the compressed data and model. When comparing different models through MDL we select the one that yields the minimum description length. The expected length of $\mathcal{D}$ through a model M can be computed as

$$L(\mathcal{D}, M) = L_{C_1}(M) + L_{C_2}(\mathcal{D} \mid M). \tag{3}$$

Following Grunwald (2007), we call this strategy the *crude two-part MDL*. Here the first term on the r.h.s is the length of the model description, that depends on a code C_1, and the second one is the length of $\mathcal{D}$ coded through C_2 in model M. If the model M has h parameters, each one belonging to $[0, 1]$, and that can take d values, equally spaced (see Grunwald 2007), we choose a parameter in the finite set

$$\Theta_h^d := \left\{ \theta = \left(\frac{k_1}{d-1}, \ldots, \frac{k_h}{d-1} \right) : k_i \in \{0, \ldots, d-1\} \text{ for i} = 1, \ldots, \text{h} \right\} \subset [0,1]^h.$$

The corresponding length of the model description through its parameters is then[5]

$$L_{C_1}(M) = h \log d.$$

With regards to the second term in (3), the existence of a quasi optimal code gives that this quantity is the likelihood of the sample in the probabilistic model, evaluated with the maximum

[5] To encode one element within a set of $d = 2^\ell$ a binary stings, we need to use $\ell = \log_2 d = \log d$ bits.

likelihood estimator. The notation is

$$\hat{\theta} = \arg\min_{\theta \in \Theta_k^d} \{-\log \mathbf{P}_\theta(\mathcal{D})\},$$

and

$$L_{C_2}(\mathcal{D} \mid M) = -\log \mathbf{P}_{\hat{\theta}}(\mathcal{D}).$$

In conclusion, the description length, assuming that d is fixed, is

$$L(\mathcal{D}, M) = h \log d - \log \mathbf{P}_{\hat{\theta}}(\mathcal{D}).$$

The selection of the accuracy of the model parameters, i.e. the determination of the truncation level of the real numbers that the parameters can take as values is a critical decision of the MDL method. It can be considered equivalent to the selection of the ridge regression parameter or the regularization level in the LASSO estimation procedure. An initial proposal by Rissanen (1978) was to take $d = O(\sqrt{N})$. This is a consequence of the minimization of the second order Taylor expansion of the likelihood function assuming that the estimation errors are uniformly distributed, when considering a linear model (see the details in Section 4 of Rissanen 1978 and also the discussion in Rissanen 1983). If for instance we consider $d = \sqrt{N}$, the description length becomes

$$L(\mathcal{D}, M) = \frac{h}{2} \log N - \log \mathbf{P}_{\hat{\theta}}(\mathcal{D}). \tag{4}$$

Note that this expression is close[6] to the optimization function proposed by the Bayesian information criterion (BIC, see Schwarz 1978), given by

$$BIC(M, \mathcal{D}) = \log \mathbf{P}_{\hat{\theta}}(\mathcal{D}) - 2h \log N.$$

Our approach to the selection of d is to choose it by global minimization of (3). As d becomes a parameter necessary to specify the model M, this approach adds the cost of coding d itself to the minimization, this cost being $\log d$. The obtained MDL equation is

$$L(D, M, d) = (h + 1) \log d - \log \mathbf{P}_{\hat{\theta}}(\mathcal{D}). \tag{5}$$

Note that this expression represents a compromise between the description length of the data set and the description length of the model. As the value of d increases, the description length of the model increases while the description length of the data set decreases. The experiments presented below show this global behaviour.

4. Experiments and results

4.1. *Methodological considerations*

4.1.1. *Encoding the musical pieces*

As in Temperley (2010), we are interested in common-practice rhythm, as represented by European classical music and folk music. In particular, the experiments presented in this section focus on music corpora consisting of popular songs or folk tunes. In such cases, there is usually a single main melody of the piece, which is the encoded one.

[6] In both cases the description length is proportional to $\log N$. This fact was observed by Rissanen (1978).

The starting point of the encoding is a corpus of musical pieces written in some symbolic music notation digital format, such as MusicXML, ABC, or **kern. Additionally, one has to select a time signature and an metrical level to do the encoding (e.g. $\frac{4}{4}$ at the eighth-note level). Only the pieces from the corpus with the given time signature are selected. It is assumed that each piece has a single time signature; pieces with time signature changes are ignored. The encoding also ignores a negligible number of note onsets out of the given metrical grid.

4.1.2. *Software implementation*

All the models described in Section 2 are implemented in the Python programming language as an open–source software package available on GitHub.[7] The software encodes a data set of music pieces and allows to find the optimal parameters of a model given the encoded data set. It also computes the minimum description length for the given model and data set. To do that, the optimal value for the precision parameter d is found.

We use the `music21` library by Cuthbert and Ariza (2010) for reading various symbolic music formats and representing the music information. In this way, music corpora from different sources can be encoded and analysed straightforwardly using the implemented software. In particular, some data sets are readily available within the `music21` package.

4.2. *Corpora*

We adopt three standard corpora commonly used in music studies for the experiments. The Essen Folksong Collection, that comprises more than 6000 folk songs mainly from Europe, and is used in Temperley (2010), Aird's Airs, which is a selections of 1180 Scotch, English, and Irish airs, and O'Neill's Music of Ireland, a collection of 1850 fiddle tunes. We obtained all of these corpora from the `music21` package.

In addition, we used another corpus consisting of the vocal melody of 204 tango songs, mainly from the 1940s, which Satorre (2018) recently compiled. Although this corpus also contains popular songs, their cultural origin is different from the previous ones, which motivates their inclusion in the experiments.

The musical pieces in each corpus were grouped by their key signature. Only musical pieces in $\frac{2}{4}$ and $\frac{4}{4}$ were considered in the experiments. There are just a few pieces in $\frac{2}{4}$ in the tango data set, so they were ignored. Table 2 summarizes the number of pieces for each corpus and time signature that were used in the experiments.

4.3. *Results*

The adopted approach to compute the minimum description length involves the selection of the optimal value for the precision parameter d by global optimization. Figure 4 shows the description length obtained when varying the precision value d for the different models considering the $\frac{4}{4}$ data sets and the eighth-note level encoding. Note that the reference value of 8 bits per bar corresponds to the original representation and implies no compression. The minimum value of the description length achieved for each model is depicted with a greyed circle.

It is noteworthy that for a given data set the optimal value of the parameter d is different for the different models. Hence, it becomes clear that the selection of $d = O(\sqrt{N})$ as proposed by Rissanen (1978) would produce sub-optimal solutions.

[7] https://github.com/mrocamora/mdlfit

Table 2. Number of pieces and bars for each of the data sets used in the experiments.

Time signature	$\frac{4}{4}$				$\frac{2}{4}$		
Corpus	Essen	Aird's	O'Neill's	Tango	Essen	Aird's	O'Neill's
Number of pieces	1897	368	225	173	2201	239	184
Number of bars	25056	5687	3875	6188	35830	5097	3788

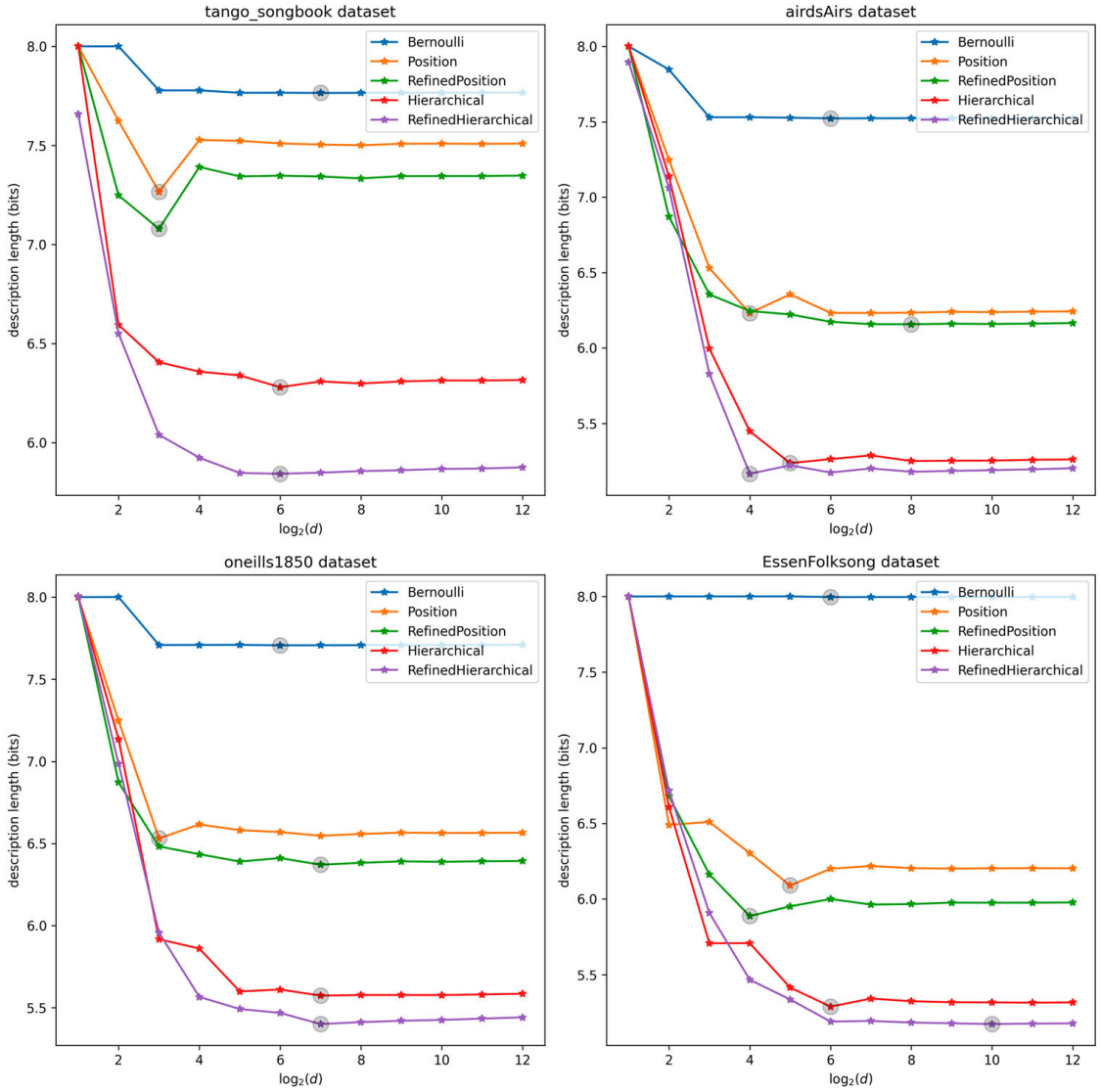


Figure 4. Description length (in bits per bar) for the $\frac{4}{4}$ data sets at the eighth-note level encoding obtained with the different models when varying the precision parameter d.

Despite the differences in the description values attained in each case, the experiments of Figure 4 consistently show the same ordering of the models across the different data sets. Not surprisingly, the models are grouped into three categories according to their behaviour. The simplest one, the uniform position model, yields the highest description length values. Then, the models based on the metrical position, the metrical position model and the fine-grained position model, provide a clear improvement over the uniform position model. Finally, the models
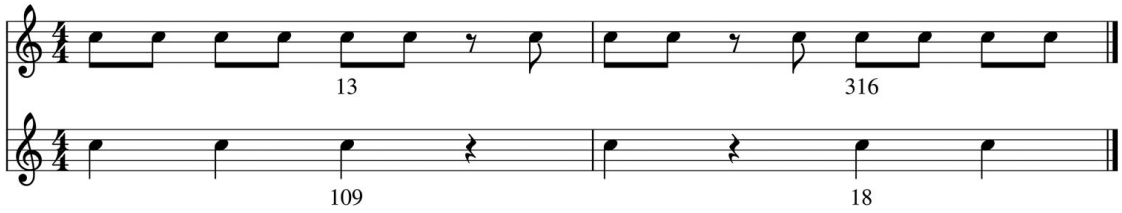


Figure 5. Two examples of symmetric one-bar rhythmic patterns for which each variant exhibit a very different number of occurrences (indicated bellow) in the tango data set.

that exploit the metrical hierarchy, the hierarchical position model and the fine-grained hierarchical position model, further improve the minimum description length values. Note that the fine-grained versions of the models always improve over the original ones.

The improvement provided by the fine-grained model with respect to the hierarchical position model deserves some analysis. We may presume that some rhythmic patterns are considered equivalent by the hierarchical position model, despite the fact that they actually have significant differences in their frequency of occurrence. In this way, the fine-grained version of the model could take advantage of these differences, assigning different probabilities of occurrence to them. In order to get better insight into this phenomenon we compare the number of occurrences of two different rhythmic patterns that are treated differently by the fine-grained version of the model compared to the original one. Figure 5 shows two examples of rhythmic patterns of this kind and their corresponding number of occurrences in the tango data set, which is the one that exhibits the most notable difference between the two models.

The results obtained for the $\frac{4}{4}$ data sets using the sixteenth-note level encoding show a very similar qualitative behaviour to the one reported in Figure 4. The ordering of the models is the same, but the minimum description length is attained for different values of the precision parameter d. Although these results are not included herein, the software package provides the code needed to reproduce all these experiments.

We conducted the same kind of experiments on the $\frac{2}{4}$ data sets of Table 2. Figure 6 shows the results obtained for the Aird's Airs and Essen Folksong $\frac{2}{4}$ data sets using the eighth-note level encoding. The behaviour of the models is similar to the one obtained for the $\frac{4}{4}$ case. The ordering of the models across the different data sets is essentially the same as before. It is worth noting that the refined versions of the position and the hierarchical models show virtually no difference compared to the basic ones. This is because given the metrical grid there are only two positions in the bar that are grouped together, namely $\ell = 1$ and $\ell = 3$. The advantage of considering them separately, i.e. by refining the model, does not compensate the cost of introducing new parameters. Furthermore, for the Aird's Airs data set the hierarchical model slightly surpasses the performance of the refined version. This shows the convenience of properly measuring the cost of describing the model, as in the proposed MDL approach.

Table 3 shows the compression rate obtained for each model and data set, encoded at the eighth-note level. The compression rate is the ratio between the minimum description length value (in bits per bar) and the number of bits in an uncompressed bar.

5. Discussion and conclusion

We tackled the problem of model selection in the context of musical rhythm. Our main contribution is to go a step further in the model selection problem considered by Temperley (2010) through the mathematical formalization of the models and the successful application of the minimum description length (MDL) approach. To do that, we revisited the six models proposed

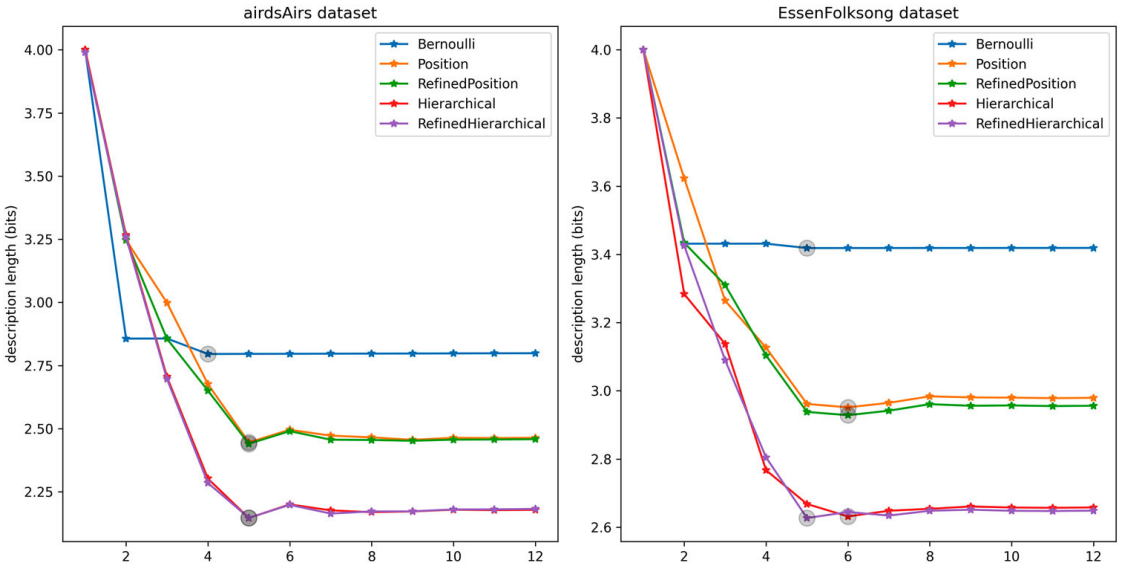


Figure 6. Description length (in bits per bar) for two $\frac{2}{4}$ data sets at the eighth-note level encoding obtained with the different models when varying the precision parameter d.

Table 3. Compression ratio for each model and data set encoded at the eighth-note level.

Time signature	$\frac{4}{4}$				$\frac{2}{4}$		
Corpus	Essen	Aird's	O'Neill's	Tango	Essen	Aird's	O'Neill's
Bernoulli	0.9995	0.940	0.963	0.971	0.855	0.699	0.640
Position	0.761	0.779	0.816	0.908	0.738	0.611	0.599
Refined Position	0.736	0.770	0.796	0.885	0.732	0.610	0.599
Hierarchical	0.661	0.655	0.697	0.785	0.658	**0.537**	**0.454**
Refined Hierarchical	**0.647**	**0.646**	**0.675**	**0.730**	**0.657**	**0.537**	0.459

by Temperley (2010) seeking to provide a detailed mathematical formulation of each of them. As a result, we argued that two of the models are inappropriate for this purpose, because they can not represent all possible rhythmic patterns. Consequently, these two models were discarded in the experiments conducted. We also proposed a seventh model, that is a refined version of the hierarchical position model.

Then, we revised the model selection approach in Temperley (2010), considering in the same framework the cost of describing both the data and the model itself by using the MDL principle. To do that, we adopted an approach similar to the one proposed in Barron, Rissanen, and Yu (1998), in which the incidence of the precision parameter d on the cost of describing the data is taken into account. However, instead of considering a worst case scenario, we explicitly computed the description length when varying the value of the precision parameter d and then applied global optimization.

The different models as well as the model selection approach were implemented in the Python programming language and were released as an open-source software package. Apart from the reproduction of the results reported in this paper, the package allows for the realization of further experiments, for instance, by including new data sets.

We conducted a series of experiments using different corpora of popular music melodies: three data sets commonly used in music studies (Essen Folksong, Oneill's 1850, Aird's Airs) and a novel one of tango songs. The results of the experiments show that despite the differences between the data sets (e.g. the number of pieces) the performance of the models according to the selection approach is quite consistent, indicating that the more complex models – among the ones

considered – tend to produce smaller values of the minimum description length. In particular, the proposed refined hierarchical position model achieves the best results in most cases. However, for the $\frac{2}{4}$ data sets the refined versions of the position and hierarchical position models provide virtually no difference with regards to the original ones. It is even the case that for the Aird's Airs data set the plain hierarchical model barely beats the refined one. This highlights the usefulness of the proposed MDL approach – that accounts for the cost of describing the model – in discerning the situations in which it is worth refining the model and those in which it is not. Regarding the similar performance of the models for the different data sets, note that they all correspond essentially to popular song melodies. However, the tango data set exhibits the most noticeable difference between the hierarchical and refined hierarchical models, which could be related to its different cultural origins.[8] Future work should extend the experiments by considering other kinds of music corpora.

A general consequence of our findings is to stress the fact that integrating music theory knowledge is beneficial for musical rhythm modelling, given that the gradual incorporation of musical structure in the models produces better-fitting results. The hierarchical models stand out as a good starting point, whereas the possible refinements would adjust to the nature and complexity of the corpus under consideration.

Regarding the attractive possibility of considering even more refined hierarchical position models, this would imply at first counting frequencies on the sixteenth-note level. Although an initial exploration of the corpora considered in the present work shows a low rate of occurrence for these events, this would remain as a possible way of further exploration. Other possible future work includes the consideration of hidden Markov models, as in previous work (Mavromatis 2009, 2012). The question here is whether their modelling flexibility can learn the information provided a priori in hierarchical position models.

A possible application of this work could be classifying musical pieces into different music genres according to rhythmic differences. Given a particular type of model, for example, the proposed refined hierarchical position model, one could fit its parameters producing a model instance adjusted to each one of the data sets. Then, it could be feasible to identify to which music corpus a given piece is most similar by comparing the goodness of fit of the previously adjusted models.

Acknowledgments

The authors would like to thank Dr. Ignacio Ramírez, Dr. Marcelo Fiori and Dr. Paola Bermolen for fruitful discussion about this work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by funding agencies Agencia Nacional de Investigación e Innovación (ANII) and Comisión Sectorial de Investigación Científica (CSIC), UdelaR.

Supplemental data

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/17459737.2023.2213471.

[8] Aesthetic preferences can result in a different rate of occurrence of different rhythmic patterns, see Figure 5.

ORCID

Martín Rocamora http://orcid.org/0000-0003-3183-9717

References

Akaike, H. 1973. "Information Theory and An Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory (Tsahkadsor, 1971),* 267–281. Akadémiai Kiadó, Budapest.

Barlow, C. 1987. "Two Essays on Theory." *Computer Music Journal* 11 (1): 44–60. doi:10.2307/3680177.

Barron, A. R., J. Rissanen, and B. Yu. 1998. "The Minimum Description Length Principle in Coding and Modeling." *IEEE Trans. Inf. Theory* 44 (6): 2743–2760. doi:10.1109/18.720554.

Bello, J. P., R. Rowe, C. Guedes, and G. Toussaint. 2015. "Five Perspectives on Musical Rhythm." *Journal of New Music Research* 44 (1): 1–2. doi:10.1080/09298215.2014.996572.

Cover, T. M., and J. A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY: Wiley-Interscience.

Cuthbert, M. S., and C. Ariza. 2010. "Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data." In *Proc. of the 11th International Society for Music Information Retrieval Conference* Utrecht, The Netherlands, August 9-13, 637–642.

Desain, P., and H. Honing. 1999. "Computational Models of Beat Induction: The Rule-Based Approach." *Journal of New Music Research* 28 (1): 29–42. doi:10.1076/jnmr.28.1.29.3123.

Dixon, S. 2001. "Automatic Extraction of Tempo and Beat From Expressive Performances." *Journal of New Music Research* 30 (1): 39–58. doi:10.1076/jnmr.30.1.39.7119.

Ellis, D. P. W. 2007. "Beat Tracking by Dynamic Programming." *Journal of New Music Research* 36 (1): 51–60. doi:10.1080/09298210701653344.

Grunwald, P. D. 2007. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. London, England: The MIT Press.

Gómez, E., M. Schedl, X. Serra, and X. Hu. 2016. "Music Information Retrieval: Overview, Recent Developments and Future Challenges." In *17th International Society for Music Information Retrieval (ISMIR) Conference*, New York City, 07/08/2016, 11–12.

Lerdahl, F., and R. Jackendoff. 1985. *A Generative Theory of Tonal Music*. Cambridge: MIT Press.

London, J. 2004. *Hearing in Time: Psychological Aspects of Musical Meter*. New York: Oxford University Press.

Mavromatis, P. 2009. "Minimum Description Length Modelling of Musical Structure." *Journal of Mathematics and Music* 3 (3): 117–136. doi:10.1080/17459730903313122.

Mavromatis, P. 2012. "Exploring the Rhythm of the Palestrina Style: A Case Study in Probabilistic Grammar Induction." *Journal of Music Theory* 56 (2): 169–223. doi:10.1215/00222909-1650406.http://www.jstor.org/stable/41810053

Naveda, L., F. Gouyon, C. Guedes, and M. Leman. 2011. "Microtiming Patterns and Interactions with Musical Properties in Samba Music." *Journal of New Music Research* 40 (3): 225–238. doi:10.1080/09298215.2011.603833.

Pesek, M., A. Leonardis, and M. Marolt. 2020. "An Analysis of Rhythmic Patterns With Unsupervised Learning." *Applied Sciences* 10 (1): 178. doi:10.3390/app10010178.

Rissanen, J. 1978. "Modeling by Shortest Data Description." *Automatica* 14 (5): 465–471. doi:10.1016/0005-1098(78)90005-5.

Rissanen, J. 1983. "A Universal Prior for Integers and Estimation by Minimum Description Length." *The Annals of Statistics* 11 (2): 416–431. doi:10.1214/aos/1176346150

Satorre, H. 2018. "Cancionero del tango." Universidad Nacional de San Martín.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–464. doi:10.1214/aos/1176344136.http://links.jstor.org/sici?sici=0090-5364(197803)6:2 < 461:ETDOAM > 2.0.CO;2-5&origin=MSN.

Shannon, C. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423. doi:10.1002/bltj.1948.27.issue-3.623–656.

Sioros, G., M. E. P. Davies, and C. Guedes. 2018. "A Generative Model for the Characterization of Musical Rhythms." *Journal of New Music Research* 47 (2): 114–128. doi:10.1080/09298215.2017.1409769.

Srinivasamurthy, A., A. Holzapfel, and X. Serra. 2014. "In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music." *Journal of New Music Research* 43 (1): 94–114. doi:10.1080/09298215.2013.879902.

Temperley, D. 2001. *The Cognition of Basic Musical Structures*. Cambridge: Mit Press.
Temperley, D. 2004. "An Evaluation System for Metrical Models." *Computer Music Journal* 28 (3): 28–44. doi:10.1162/0148926041790621.
Temperley, D 2007. *Music and Probability*. annotated edition ed. London, England: MIT Press.
Temperley, D. 2009. "A Unified Probabilistic Model for Polyphonic Music Analysis." *Journal of New Music Research* 38 (1): 3–18. doi:10.1080/09298210902928495.
Temperley, D. 2010. "Modeling Common-Practice Rhythm." *Music Perception* 27 (5): 355–376. doi:10.1525/mp.2010.27.5.355.
Toussaint, G. T. 2013. *The Geometry of Musical Rhythm: What Makes a "Good" Rhythm Good?* Boca Raton: Chapman and Hall/CRC.
Volk, A. 2008. "The Study of Syncopation Using Inner Metric Analysis: Linking Theoretical and Experimental Analysis of Metre in Music." *Journal of New Music Research* 37 (4): 259–273. doi:10.1080/09298210802680758.
Vuust, P., and M. A. G. Witek. 2014. "Rhythmic Complexity and Predictive Coding: A Novel Approach to Modeling Rhythm and Meter Perception in Music." *Frontiers in Psychology* 5: 1111. doi:10.3389/fpsyg.2014.01111.