POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

A Study on the Effect of Dynamic Accents on the Perception and Measurement of Rhythm Complexity

Tesi di Laurea Magistrale in Music & Acoustic Engineering

Gabriele Maucione, 952290

Advisor:
Prof. Massimiliano Zanoni

Co-advisors:
Alessandro Ilic Mezza
Dr. Luca Comanducci

Academic year:
2022-2023

Abstract: Rhythm complexity perception has been quite extensively studied in the last decades for its interest in the Music Information Retrieval field. However, many existing studies focus solely on the duration aspect of rhythm, and ignore the influence of other important factors. This thesis explores the impact of dynamic accents on the perception and measurement of rhythm complexity. Two listening tests were designed, respectively investigating the perception of monophonic and polyphonic patterns. In the tests participants were presented with rhythms featuring various types of accents - including patterns with constant intensity and actual dynamic drums performances - and were asked to assess their rhythm complexity. Additionally, six monophonic rhythm complexity metrics were presented and a polyphonic version of each was proposed. The quality of these metrics was analyzed using the results from the listening tests as a reference and computing the Pearson Correlation Coefficient between them. This analysis shed light on the relationship between the subjective experience of perceiving rhythm complexity and its possible quantitative description, and provides insights into the nuanced role of dynamic accents in shaping rhythm perception, in both monophonic and polyphonic contexts.

Key-words: Rhythm complexity, dynamic accents, perception, Music Information Retrieval

1. Introduction

Rhythm is one of the main key concepts involved in music. It can be defined as the placement of sounds in time [44]. When a series of notes and rests repeats, it forms a rhythmic pattern. In addition to indicating when notes are played, musical rhythm also stipulates how long they are played and with what intensity [2]. The importance of rhythm lies in the fact that music listening is a temporal experience, and the temporal structures by which music unfolds are critical to listeners' aesthetic, emotional, and behavioral responses [8]. In terms of cognitive perception, rhythm is the musical feature that triggers our inherent ability of synchronization to periodic stimuli, known as entrainment process [32], and it is strictly related to perceived emotions. Thus, from the rhythmic structure of a song meaningful information about more complex high level features of the song itself can be derived. The discipline whose technologies are focused on the link between low and high level information of an audio track is known as Music Information Retrieval, which is a specific branch of Multimedia Information Retrieval.

Multimedia Information Retrieval (MMIR) is a modern research discipline which aims at extracting and providing reliable information on the content of multimedia data, in such a way to better understand it and use it for scientific applications. With such an information, indeed, it is possible to realize systems that can classify data, modify it according to many tasks, or even generate it. However, the kind of information which usually describes multimedia data as image, video and audio is usually associated with a subjective and abstract language. That's why MMIR results to be an interdisciplinary task, including computer science sub-fields (pattern analysis, machine learning, signal processing, interaction design) as well as psychological areas (perception theory, cognitive neuroscience, psychoacoustics, sociology).

Music Information Retrieval (MIR) is a specific field of MMIR, which pertains to musical data, and is based on the combination of knowledge from both informatics and musicology. As a field, its goal is the development of computational systems able to retrieve high level attributes which are meaningful for a listener, in order to use these for complex applications, from music classification and recommender systems to source separation and music generation. All these technologies are based on the computation of some audio features to primarily analyze the audio content. Among these features, a wide number is based on the rhythm. Rhythm features include tempo, syncopation, onset rate, swing ratio, and these ones can even be used to compute higher level and more complex features, such as danceability.

This work focuses on the study of rhythm complexity. This topic was studied and discussed by many researchers over the decades [7, 15, 17, 30, 38, 45, 66, 68, 74, 76, 77], as its interest in the MIR perspective is undeniable. Also, according to Berlyne [6], rhythm complexity is a very relevant semantic descriptor of the music content. Complexity, together with novelty, is one of the main variables affecting neural activities as attention, reactivity and excitement. It is also involved in the preference process associated with the presentation of an acoustical stimuli.

Many models to describe rhythm complexity were proposed in the literature [11, 15, 21, 22, 33, 38, 45–48, 55, 60, 68, 70–72]; they all adopt some algorithm that return a numeric score associated with the magnitude of the quantity. However, such models only refer to rhythms in terms of binary sequences, which result to be a too poor approximation. Indeed, in a drummer's real performance there are many more factors beyond the onsets positions, like the dynamics. Thus, the application of these models on real music is not validated. In this work we focus on the analysis of the impact of dynamics accents on rhythmic perception, as well as on the application of some of the most widely used metrics in the literature on rhythmic patterns that include dynamics information.

This thesis is organized as follows. In Section 2, the fundamental definitions useful to fluidly understand the conducted study are introduced. In Section 3, the rhythm complexity measures which are object of discussion in the next chapters are presented. In Section 4, we illustrate the design principles adopted for the realization of two subjective listening tests, which have the aim to collect information about the human perception of rhythm complexity in different dynamics scenarios. In Section 5, the results of the listening tests are discussed: meaningful considerations about the influence of dynamic accents in rhythm are derived, and the previously presented measures are evaluated by means of the complexity scores assigned by the users. Finally, Section 6 concludes this work.

2. Preliminaries

In this Section, the fundamental definitions of the music concepts discussed in this thesis are presented. These are *rhythm, pulse, duration, meter, beat, onset, accent, measure, time signature, BPM, tempo, rhythmic pattern* and finally *rhythm complexity.*

The definition of *rhythm* is not obvious. It is a core concept involved in music listening; however, music listening is a subjective experience and it is hard to formally define its attributes, especially when considering one of them independently from the others. Indeed, no melody or harmonic structure can exist without a rhythm; also, when an instrument reproduces a rhythmic pattern, the envelope, the pitch and the timbre of the instrument contribute to the finally perceived rhythm. Here the definition given by Schulkind [36] is adopted, which also was used by Levitin and London in [31]. According to this definition, *rhythm* is "the serial pattern of variable note durations in a melody".

In this definition, the term *duration* is presented. A duration is the number of time units between musical notes. Following what done in similar studies [66, 68], the time unit here adopted is the *pulse.* As defined by Cooper & Meyers, a pulse is "one of a series of regularly recurring, precisely equivalent stimuli" [29]. This regular recurrence happens within a musical *measure* or *bar*. A measure is namely a defined segment of time within a piece of music. As stated by London [31], a *meter* is an equal subdivision of the *pulse*, marked by strongly accented pulses which are defined as *beats*. Thus, a meter defines the structure of a measure. London even says that "meter is *how* you count time, while rhythm is *what* you count - or what you play while you are counting" [32]. Also, the meter and the beats are indicated with a notation called *time signature*. As an example, the most common time signature is the one of 4/4, which indicates four beats on 16 pulses. *Tempo* refers to the pace of music, or the rate at which musical events unfold over time [35]. Tempo is strongly associated with beat rate, which is also referred as beats per minute (*BPM*). An *onset* is a pulse associated with the attack of a note (i.e. the point in time where the note starts). Indeed, an onset marks the beginning of the duration of a note.

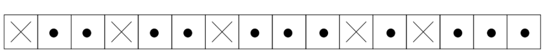
In general, an onset can be associated with an intensity, i.e. when a musician plays a rhythmic pattern, not all the notes are necessarily associated with the same loudness. An *accent* is an onset that is underlined by a higher intensity, and somehow stands out in the rhythm. To be precise, it is possible to distinguish among many kind of accents. *Dynamic accents* (or *rhythmic* accents, or *phenomenal* accents [32]) are sounds that are marked for attention because they louder than their surroundings. *Agogic accents* underline a note by associating it with a longer duration. *Tonic accents* indicate when a note is emphasized with a higher pitch then the near ones in the musical phrase. Finally, *metric accents* are regularly occurring points of peak attention [32] in our metrical experience. These points are psychologically salient, because we use them to synchronize our expectations to what the music is doing (*entrainment* [32]); these accents do not refer to any modification executed by the musician, they only exist in the listener's perception. These kind of psychological phenomena is fundamentally related to our perception of a metrical beat in a whatever rhythmic pattern.

When dynamic accents occur at pulses which do not correspond to the stressed beats of the prevailing meter, patterns are said to be syncopated [39]. *Syncopation* refers to the deliberate shifting of accents onto the weak beats or off-beats in music, creating a rhythmic tension and a feeling of unexpectedness [65].

As the definition of rhythm brought forward, a *rhythmic pattern* is directly associated to a rhythm, and these two names are often interchangeably used to refer to the same thing. However, in the context of this thesis, we discriminate between a *pattern*, which only refers to the positional and description of a rhythm (i.e. no accents or intensity modulations are included in this definition), and a *rhythm*, which is what is derived by applying dynamics to a binary pattern. In other words, a pattern is an abstraction of a rhythm represented by a sequence of 0s and 1s, and a rhythm refers to the performance resulting from an actual drummer playing a pattern. This distinction was adopted to both align to the literature that only studies binary sequences calling them rhythmic patterns [66, 71], and then introduce a further step by transporting the same discussion to a more realistic scenario.

The possible representations of a rhythmic pattern are shown in Figure 1. The most common one is the *music notation*, which is reported in Figure 1a. Another used representation is the *Time Unit Box System* (TUBS), also referred as *box notation*, which was introduced by ethnomusicologists Phillip Harland and James Koetting [28]. The example with the same pattern is shown in Figure 1b. Similarly, a pattern can also be described as a binary string with the *binary notation* (Figure 1c). And finally, the *geometric representation* is presented with an example in Figure 1d.

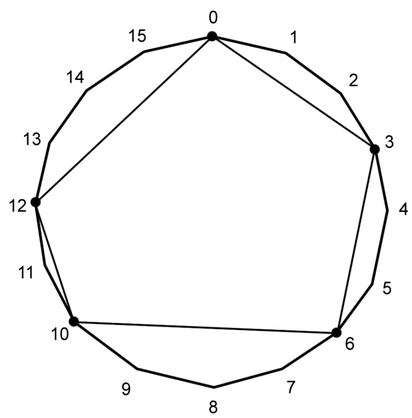
(a) Music notation.



(b) Box notation.

`1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0`

(c) Binary notation.



(d) Geometric representation.

Figure 1: Example of the possible representations of the *clave son* rhythmic pattern.

Finally, the concept of *rhythm complexity* is introduced, which generally refers to how complex a rhythm result to be for a listener. A large literature exists on this subject, but still not a proper definition of it was established. Some studies defined rhythm complexity as the difficulty for a listener to recognize the meter in a rhythmic pattern [74], or sometimes to be able to play it or tap along to its tempo [7]. Some other based their perspective about rhythm complexity on the idea of syncopation [22, 33, 68, 72], which is namely an element of surprise for the listener. Also, the heterogeneity of the rhythm was considered as a descriptor of its complexity [11, 37, 56]. Many of these studies were associated to the presentation of some measures that, according to specific algorithms, return a score to describe the complexity of a rhythm. Other researches focused on the analysis of brain activity while listening to rhythms [7, 31], to provide a deeper explanation of this experience. What's sure is that the perception of the complexity of a rhythm is a cognition phenomena that can neither be explained with few words, nor be objectively classified, as it varies from person to person. In this work, we investigate the subject of rhythm complexity by means of two subjective surveys, together with the application of some well known metrics, which are going to be presented in the next section.

3. Rhythm Complexity Measures

In the previous section, rhythm complexity has been introduced. Rhythm complexity measures are algorithms that aim at describing the complexity of rhythms. Still, this topic is based on subjective music perception, making it hard to establish which is the right method to measure it. In this section, different families of rhythm complexity measures are going to be presented, and some particular measure's algorithms are going to be explained in detail. The presented measures are the same ones which are used in this study.

Several rhythm complexity measures have been discussed in the literature [21, 22, 41, 47, 56, 66–68, 72]. These are groupable into six main categories: *rhythmic syncopation*, *pattern matching*, *distance*, *information entropy*, *inter-onset intervals*, *mathematical irregularity.* Each of these look at the rhythm complexity from a different perspective: *rhythmic syncopation* measures are based on the definition of a metrical hierarchy where each pulse is associated with a weight; *pattern matching* measures search for sub-rhythms in the pattern, *distance* measures compute the difference between some reference pattern and the analyzed one, *information entropy* are based on purely mathematical formulas, *inter-onset intervals* measures consider the duration of the pattern's onsets, and finally *mathematical irregularity* measures exploit the geometric representation of rhythms to get irregularity information as an indicator of complexity.

In the context of this research, we choose six among the most widespread measures available in the literature [66]. Specifically, the chosen measures are Toussaint's Metrical Complexity, Longuet-Higgins & Lee Complexity, Pressing's Cognitive Complexity, Weighted Note to Beat Distance, Toussaint's Off-Beatness, and Inter-Onset Intervals' Information Entropy. Furthermore, Grouped Voice Polyphonic Measure is presented and utilized for the realization of the polyphonic version of these algorithms. Following, all these measures are presented in detail.

3.1. Toussaint's Metrical Complexity

Toussaint's Metrical Complexity measure is by far the most widely recognized and extensively documented in the literature [3, 14, 38, 66]. It belongs to the *rhythmic syncopation* category. This group aims at understanding of syncopated a pattern is by means of a weighting system. This particular measure is based on the metrical hierarchy from Lerdahl and Jackendoff [30], which assigns a weight to each pulse of a N-length pattern, according to their metrical strength. A pulse's weight is indicative of how much that position in the pattern is relevant for a listener to recognize the meter. As a consequence, onsets in metrical positions with low weights are associated with a stronger sense of syncopation, and thus with a higher rhythm complexity.

An example of this hierarchy for a 4/4 time signature pattern of 16 pulses is reported in Figure 2. On the x-axis we have the pulses, on the y-axis the corresponding weights.

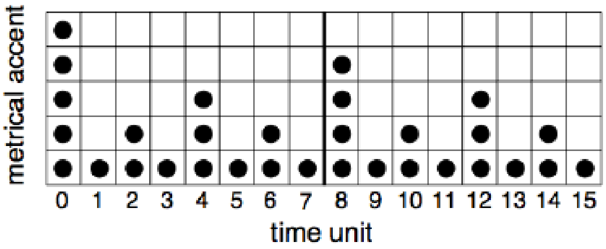


Figure 2: Lerdahl and Jackendoff's metrical hierarchy [22].

To obtain such a hierarchy, units of weights have been iteratively added to pulses spaced according to regular subdivisions at different metrical levels. A general algorithmic procedure to derive this hierarchy for any kind of pattern is below described.

1. Let $\mathbf{r} \in \{0,1\}^N$ be a rhythmic pattern of length N.
2. Compute the prime factorization of N. For each unique permutation of the list of prime factors a different metrical hierarchy can be obtained.
3. Chose one permutation P, and initialize the weights vector $\mathbf{w}$ as a sequence of zeros.
4. For each metrical level $i = 0, ..., len(P)$, increment by 1 the elements of $\mathbf{w}$ which are spaced every N/l steps, where l is a level associated value. For the first level, consider $l = 1$; for the other levels set l equal to $l \cdot p_i$, which is the *i-th* element of P.

For a pattern of length 16, this procedure leads to the following weight vector, which corresponds to the hierarchy shown in Figure 2:

$$\mathbf{w} = [5, 1, 2, 1, 3, 1, 2, 1, 4, 1, 2, 1, 3, 1, 2, 1]. \tag{1}$$

Once the metrical hierarchy is derived, given a pattern $\mathbf{r}$, a quantity known as *metricity* is computed as:

$$m(\mathbf{r}) \coloneqq \mathbf{w}^\top \cdot \mathbf{r}. \tag{2}$$

This corresponds to the sum of the weights in the hierarchy which are indexed by the positions of the onsets in the rhythmic pattern $\mathbf{r}$. The higher the weights of the onsets in $\mathbf{r}$, the higher will be its metricity. However, since high weights are associated with less syncopation, metricity is noticeably inversely proportional to metric complexity. Hence, the final metric complexity score has to be inversely proportional to metricity.

The maximum metricity $M_\mathbf{r}$ for a pattern with the number of onsets of $\mathbf{r}$ is computed. Operatively, it is defined as the sum of the k largest weights in $\mathbf{w}$, where k is the number of onsets in $\mathbf{r}$:

$$M_\mathbf{r} = \sum_{i=1}^{k} \mathbf{w}_{sort}, \tag{3}$$

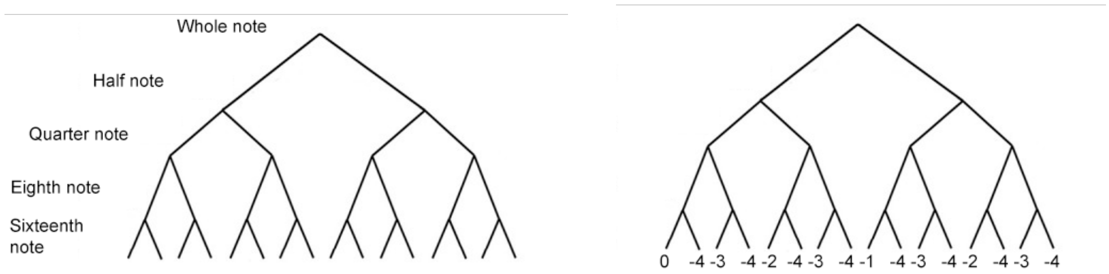
where $\mathbf{w}_{sort}$ is the descending order sorted version of $\mathbf{w}$. Lastly, Toussaint's Metrical Complexity is finally obtained as:

$$T(\mathbf{r}) = M_\mathbf{r} - m(\mathbf{r}) \tag{4}$$

This step, which is also called *Onset Normalization* in [66], is important for two reasons: first, the final result is finally expressive of the metrical complexity of the rhythmic pattern; and second, this passage also makes the measure relative, taking into account the number of onsets in the rhythmic pattern.

3.2. Longuet-Higgins & Lee Complexity

Longuet-Higgins & Lee's (LHL) Complexity Measure [33] also belongs to the *rhythmic syncopation* category. Indeed, it is based on a weighting scheme as well. This scheme is not dissimilar from the Toussaint's one seen in Section 3.1, but it's formulated as a tree structure, where the root represents the whole bar and the leaves are associated with each pulse (Figure 3a).



(a) Tree structure LHL metrical hierarchy [16].

(b) Weighted LHL hierarchy [66].

Figure 3: Longuet-Higgins & Lee metrical hierarchy

This structure is obtainable as follows. Let N be the number of pulses in a rhythmic pattern. As done in 3.1, the first step is to compute the prime factorization of N. Among all the possible permutations of prime factors, one is chosen. Let $l = 1$ be a starting metrical level, which corresponds to the whole bar. For each prime factor p in the list, a new level is added to the tree by creating p children from each node in the current level. For a 16-pulse pattern, the prime factors list will be $2 \cdot 2 \cdot 2 \cdot 2$. In that case, only one hierarchy is obtainable, and that is the one with 5 levels represented in Fig 3a.

In the last level the hierarchy is defined, as each node is associated with a metrical weight (Figure 3b). Let $\mathbf{w}$ be the weights vector, of length N. All the weights are initialized to 0. Let l be a value associated with the metrical level. Starting from $l = 1$, do the following:

1. Subtract 1 to all weights $w_k \in \mathbf{w}$, with $0 \leq k < N$, unless $k = 0$ or k equal to an integer multiple of N/l.
2. Update the value of l by multiplying it for the next prime factor in the list.

Stop when all prime factors in the list have been considered. For a 16-pulse pattern, this algorithm will lead to the hierarchy represented in Figure 3b.

Once the hierarchy has been generated, syncopations in the pattern are searched. Syncopation is here defined as the occurrence of a silence that has a higher weight of the one of a preceding onset [33].

1. Let $\mathbf{r}$ be a rhythmic pattern of length N, and with a weight vector $\mathbf{w}$.
2. For each pulse i in the pattern, with $0 \leq i < N$, check if the i_{th} position is a silence.
3. If the i_{th} pulse in $\mathbf{r}$ is a silence, then search backwards for the nearest onset in $\mathbf{r}$, assuming it to be a cyclic pattern [16].
4. Let the j_{th} pulse be the nearest preceding onset. If the value of the difference:

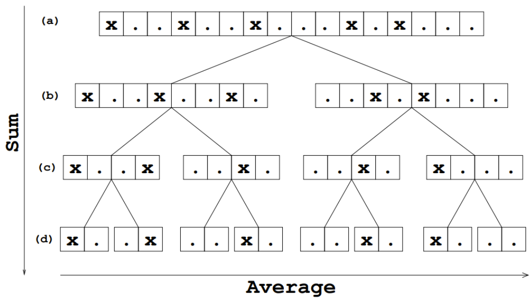
$$s_i = w_i - w_j \tag{5}$$

   is greater than 0, Longuet-Higgins & Lee say that a syncopation has occurred [33], and the value of s_i is the syncopation degree.
5. When a syncopation is found, all the silences after the i_{th} pulse are skipped until the weight of a following silence is greater than w_i. This is because silences which follow a silence with a higher metrical weight are considered to be an undivided metrical unit, according to the rule stated by Longuet-Higgins & Lee [33].
6. When all syncopations have been found, the final metric complexity score is obtained by summing all syncopation degrees.

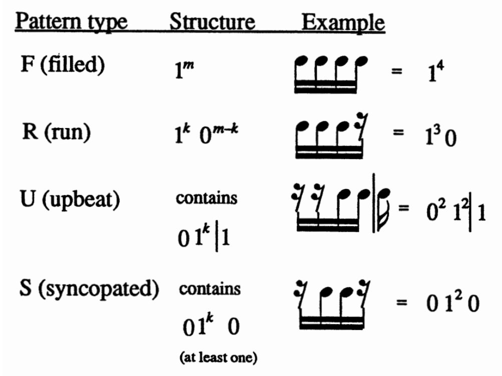
3.3. Pressing's Cognitive Complexity

Pressing's measure falls within the *pattern matching* category. Indeed, this measure's main idea is to divide the rhythmic pattern into sub-patterns and match them to some archetypes associated with known degrees of complexity. Pressing [48] describes that these archetypes have a cognitive cost, and that's why this measure is called *Pressing's Cognitive Complexity*.

The first operative step is to derive a hierarchical subdivision of the rhythmic pattern, in the same way described in Section 3.2, resulting in a tree representation. In Figure 4a an example of this subdivision is shown, applied to the *clave son* pattern.



(a) Example of metrical division of *clave son* pattern [66].



(b) Pressing's classes [47].

Figure 4: Pressing's multi-resolution matching analysis: in (a) an example of a pattern's subdivision in sub-rhythms across different metrical levels and in (b) the stated classes the sub-rhythms could match.

Once this structure is derived, Pressing [47, 48] defines six classes of patterns, with an associated weight according to their complexity degree. For every sub-rhythm in the tree structure, the following classes are considered to see if there is a match:

- *Filled*: All the sub-sub-rhythms below the considered sub-rhythm start with an onset.
- *Run*: The considered sub-ryhthm starts with an onset followed by a sequence of onsets.
- *Upbeat*: The considered sub-rhythm ends with an onset and the following sub-rhythm starts with an onset as well.
- *Subbeat*: This kind of pattern has been mentioned by Pressing [48], but no additional information were given; for this reason this class was not considered in this study.
- *Syncopated*: A sub-sub-rhythm of the considered sub-rhythm does not start with an onset, meaning that in the sub-rhythm there is an onset that starts and ends 'off beat'.
- *Null*: The considered sub-rhythm has no onsets or it only has one onset on its first pulse.

Each of these classes is associated with a complexity weight, as indicated in Table 1.

	Filled	**Run**	**Upbeat**	**Syncopated**	**Null**
Weight	1	2	3	5	0

Table 1: Complexity weights for each Pressing class.

The idea behind this measure is to get the rhythm complexity with a multi-resolution analysis of the tree, that is made up of three passages. First, all sub-rhythms in each metrical level are matched to a cognitive class, and thus to a complexity weight. Second, these weights are averaged horizontally, resulting in a unique value for each level of the tree. Finally, the level average values are summed vertically: this sum is defined as Pressing's Cognitive Complexity.

3.4. Weighted Note to Beat Distance

Weighted Note to Beat Distance [21, 22] is a complexity measure that belongs to the *distance* measures category. This kind of measures uses the mathematical notion of distance to determine how complex rhythms are, essentially looking at how far these are from meters with even beat distributions [66]. In particular, Weighted Note to Beat Distance (WNBD) assigns a weighted score to each onset in a rhythmic pattern depending on that onset distance from the nearest beat. The sum of the weighted scores of all the onsets in the rhythmic pattern determines WNBD complexity.

More formally, let $\mathbf{r}$ be a rhytmic pattern, of length N, with k onsets. Let $\mathbf{e}$ be the reference meter, defined as a vector of length N with m equally spaced beats. For each of the k onsets in $\mathbf{r}$:

1. Let x be the considered onset of $\mathbf{r}$, and consequently select the beats e_i, e_{i+1}, e_{i+2} from $\mathbf{e}$. These are defined respectively as the nearest beat before x, the nearest beat after x, and the next beat after this

latter.
2. Compute the distances between the onset and the its nearest beats $d(x, e_i)$ and $d(x, e_{i+1})$. Let:

$$T(x) \coloneqq min(d(x, e_i), d(x, e_{i+1})) \tag{6}$$

be the minimum distance among these two.
3. Then a complexity weight *D(x)* is assigned to the onset x, based on these rules:

$$D(x) = \begin{cases} 1/T(x), & \text{if } x \text{ ends before or on } e_{i+1} \\ 2/T(x), & \text{if } x \text{ ends after or on } e_{i+1}\text{, but before or on } e_{i+2} \\ 1/T(x), & \text{if } x \text{ ends after } e_{i+2} \\ 0, & \text{if } T(x) = 0 \end{cases} . \tag{7}$$

Once a weight *D(x)* has been obtained for each of the k onsets, the WNBD complexity score is computed as:

$$WNBD(\mathbf{r}) = \frac{1}{k}\sum_{i=0}^{k-1} D(x_i) \tag{8}$$

Equation (7) precisely reflects two musicological intuitions this measure is based on: first, the closer a note is to a strong beat, the more syncopated it sounds; and second, the syncopation effect is stronger when a note crosses over only one strong beat of the meter [22].
Indeed, a smaller weight is assigned in the first and third case of (7), that is when the note occurs exactly in the middle of two strong beats; also, as the onset approaches a strong beat, its weight increases. Besides, if an onset crosses over a strong beat or just ends on the next one, the syncopation degree is even higher, as beat e_{i+1} is unsounded: this case is associated with the highest weight in (7).

Finally, some examples of this measure's application are reported in Figure 5.

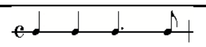
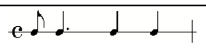
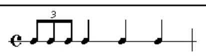
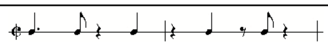
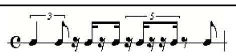
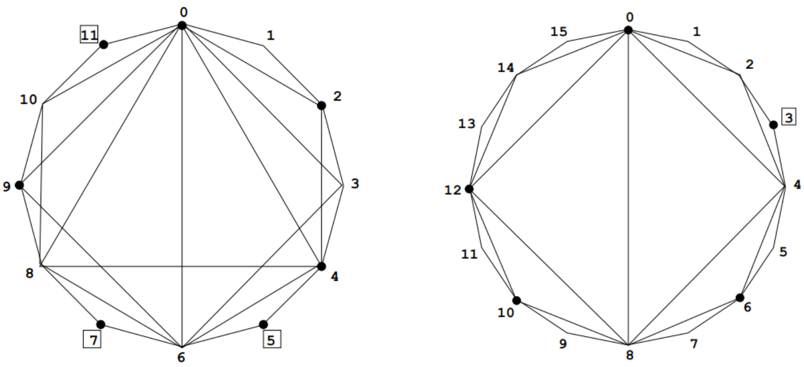
Rhythm	Musical Scores	WNBD
Hesitation		1/2
Anticipation		1/2
Syncopation		1.2
Triplet		0.857
Bembé		3
Bossa-Nova		4
Irregular		5

Figure 5: WNBD scores of example patterns from [21].

3.5. Toussaint's Off-Beatness

Toussaint has also proposed other measures, including the Off-Beatness, [71], which can be classified as a *mathematical irregularity* measure. This family of measures describes complexity as a form of irregularity, often underlined with a geometric visualization of the pattern. Toussaint's Off-Beatness's core idea is to obtain the rhythm complexity of a pattern by looking at how many of the onsets in the pattern are placed off-beat.

Let $\mathbf{r}$ be a cyclic rhytmic pattern, with N pulses and k onsets.
1. Place it on a circle, where each pulse is evenly distributed around the circumference, resulting in a geometric representation of the pattern, Fig (6).
2. Find all the integer values i, with $0 \leq i < N$, such that $\mathbf{r}$ is evenly divided into i partitions of equal length. For each of the found values, inscribe a regular polygon with i vertices in the circle, superposing the vertices to the pulses starting from pulse 0.
3. Each pulse that is included in a polygon is marked as a *beat*, while the remaining pulses outside of all polygons are defined as *off-beats*.
4. The final score of Toussaint's Off-Beatness measure is simply the count of all the onsets in the pattern that are off-beat.



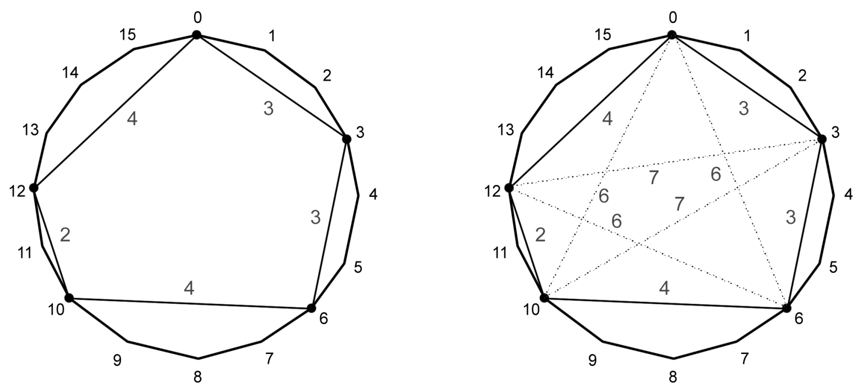
(a) *Bembé* pattern. (b) *Clave son* pattern.

Figure 6: Examples of geometric visualization with inscribed polygons and marked off-beats, with bembè rhythm (a) and clave son rhythm (b) [66].

This measure is said to be a measure of *irregularity* as it looks at how many onsets do not belong to *regular* polygons. Off-Beatness also has group and number theoretic implications, which are not covered here. For a thorough explanation of of these concepts, we refer the interest reader to [21].

3.6. Inter-Onset Intervals Measures

This family of measures is based on the analysis of Inter-Onset Intervals (IOI), which are defined as the intervals between onsets. Precisely, an IOI is computed by counting the number of pulses between two onsets.
Actually, it is possible to define two kinds of IOIs: *local* and *global* ones. *Local* IOIs are the ones obtained by only counting the distances between consecutive pairs of onsets, while *global* IOIs are derived by considering all unique pairs of onsets in a rhythmic pattern, i.e. also including non-adjacent onsets [75]. Given a cyclic rhythm with n pulses and k onsets, we have $\binom{k}{2}$ unique interval pairs, and thus the same number of *global* IOIs. In Figure (7) an example of *local* (7a) and *global* (7b) is presented, with reference to the same *clave son* pattern already seen in the previous sections.

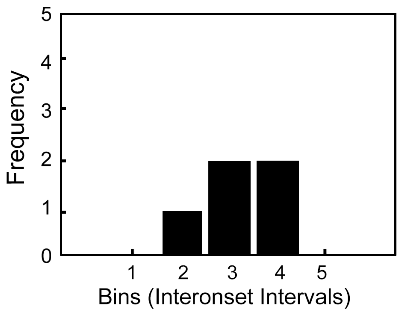


(a) *Local* IOIs of *clave son* pattern. (b) *Global* IOIs of*clave son* pattern.

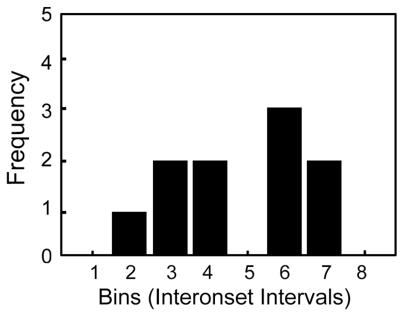
Figure 7: Examples of *local* (a) and *global* (b) IOIs visualization on geometric representation of *clave son* pattern [69].

IOIs are a relevant subject when it comes to rhythm complexity, and have namely been chosen as an analysis approach in many studies [11, 20, 56, 69, 70]. The main idea behind IOIs is that by looking at their distribution by means of a histogram, a meaningful information about the considered pattern's complexity can be obtained. In general, histograms have a large application in the statistics field as a useful tool to estimate the density of an unknown probability distribution [57], and can be usefully applied also to the rhythm complexity analysis field. For example, when dealing with rhythmic patterns, a IOI histogram showing a uniform distribution is indicative of many IOIs occurring with the same probability in the same pattern, making it unpredictable and thus more rhythmically complex.

In Figure 8, the histograms associated to the IOIs reported in Figure (7) are shown, separating the *local* case from the *global* one.



(a) *Local* IOIs histogram of *clave son* pattern.



(b) *Global* IOIs histogram of*clave son* pattern.

Figure 8: Examples of IOIs histograms for a *clave son* pattern [69]: (a) shows the distribution of *local* IOIs reported in Figure (7a) and (b) the one of *global* IOIs shown in Figure (7b).

More than a complexity measure has been presented in literature in the context of IOI histograms [37, 50, 66]; they are all based on the analysis of the IOIs. Following, the IOI Information Entropy measure is presented.

3.6.1 Information Entropy

Once derived the IOI histogram of a pattern, one of the most effective methods to measure its complexity from the histogram is to compute its entropy. Indeed, the entropy of a discrete random variable is maximum when the variable has a uniform probability distribution. By associating a rhythmic pattern to a discrete variable which takes values on the IOI of each bin of its IOI histogram, an information about this latter's flatness can be obtained. This can be used to derive considerations about its rhythm complexity, as a pattern involving a large variety of onset durations is a complex pattern.

Let $\mathbf{r}$ be a rhythmic pattern, whose IOI histogram has been derived. Let's normalize the bins in the histogram such that the sum of the frequencies is equal to 1, then such a normalized histogram may be considered a discrete probability distribution. Let X be the set of IOIs, where each value in X represents a bin label of the histogram, and χ be a discrete random variable taking on the IOI of each bin with probability corresponding to the normalized frequency. The probability mass function p of χ can be derived as follows. Define a function $f : X \to \mathbb{Z}$, which returns the frequency count for a given IOI based on the histogram. Let c be the sum of all values $f(x)$ where $x \in X$. The p.m.f. p of χ is defined as:

$$p(x) = \begin{cases} f(x)/c, & \text{if } x \in X \\ 0, & \text{otherwise} \end{cases} . \tag{9}$$

The Shannon information entropy [53] of χ can be obtained with the following formula:

$$H(\chi) = -\sum_{x \in X} p(x) \log_2 p(x), \tag{10}$$

which represents the uncertainty that χ takes on IOI values, based on probability p.

In [50, 73] this is referred as uncertainty, as it tells how predictable is the variable associated with the onset duration. The less predictable this variable, the more complex the considered pattern. The information entropy measure can be applied both to the *global IOIs* and to the *local IOIs*. In the context of this study, the version considering only *local IOIs* is adopted.

3.7. Grouped Voice Polyphonic Measure

All the measures presented so far are designed for monophonic patterns, i.e. sequences of single events identified by a 0 or a 1. If such a pattern was played by a drummer, no more than a sound at a time would be heard. Additionally, only that specific sound would be heard at every onset, as the pattern refers to a single instrument. This is typically not the case of rhythmic patterns that is actually common to hear in real music.

Mezza et al. proposed the Grouped Voice Polyphonic Measure [38], which represents the most recent work in this sense. Their metric is capable to analyze the rhythm complexity of polyphonic patterns, capturing the interaction between multiple voices. The main idea behind this metric is to expand and adapt Toussaint's Metrical Complexity measure, presented in Section (3.1) to the polyphonic case. Previous works by Sioros [58, 59] tried to do the same, but did not consider the interaction between different instruments of a drumkit.

Let $\mathbf{r}$ be a polyphonic pattern with M voices: $\mathbf{r} = [r_1[n], ..., r_M[n]]^\top$. Let N be the pattern's length: then, the pattern can be thought as a binary matrix of dimensions $M \times N$.

A set of K monophonic groups is defined, where each group represents a musicologically fundamental element that constitutes the polyphonic pattern. Given a subset of L voices $r_1[n], ..., r_M[n]$, the k-th group can be defined as:

$$g_k[n] = \bigvee_{l=1}^{L} r_l[n]. \tag{11}$$

In this way the pattern can be represented with a new matrix of dimensions $K \times N$, where possibly $K \gg M$. The 9 groups proposed in [38] are reported in Table 2.

Group	Voices
k = 1	Bass drum (0); Snare drum (1)
k = 2	Closed Hi-Hat (2)
k = 3	Open Hi-Hat (3)
k = 4	Closed Hi-Hat (2); Open Hi-Hat (3)
k = 5	High Floor Tom (4)
k = 6	Low-Mid Tom (5); High Tom (6)
k = 7	Ride Cymbal (7)
k = 8	Crash Cymbal (8)
k = 9	Ride Cymbal (7); Crash Cymbal (8)

Table 2: Grouped Voice Polyphonic Measure's grouping [38].

Let's indicate with $T(p)$ Toussaint's monophonic measure of a generic monophonic pattern p. Then, the final polyphonic measure $T_p(\mathbf{r})$ score is obtained as:

$$T_p(\mathbf{r}) \coloneqq \sum_{k=1}^{K} w_k T(g_k[n]), \tag{12}$$

where $T(g_k[n])$ is the score of the pattern representing the k-th group and K is the total number of groups. In Mezza et al.' formulation [38], $K = 9$. The 9 selected groups are shown in Table 2. The weights w_k in (12) can be selected as uniform, or set as $1/K$, or even determined via linear regression against the results of a large-scale listening test. The version considered in this study is with $w_k = 1$ for $k = 1, ..., K$.

Lastly, even if Mezza [38] presents this metric using as a reference Toussaint's measure (Section 3.1), a polyphonic version of all the seen monophonic measures can be realized in the same way. Let f be any one monophonic measure presented in Section 3. Its polyphonic version f_p, with reference to pattern $\mathbf{r}$ is:

$$f_p(\mathbf{r}) \coloneqq \sum_{k=1}^{K} w_k f(g_k[n]). \tag{13}$$

4. Subjective Listening Test

In this section, we present the design of two listening tests where rhythms with different dynamics are compared, which represents a new subject of study compared to the existing literature. Indeed, the metrics presented in

Section 3 have been developed to deal with and were only tested on binary patterns, which are taken as an approximated representation of real rhythmic patterns. Binary patterns consist of only 0s and 1s, meaning that if such a notation was used by a drummer to read what to play, they would only understand when to hit the drum, but not how strongly to do so. No information about the intensity of the onset is carried in binary patterns. Instead, in commonly used music notation, dynamics information is reported (Figure 9).

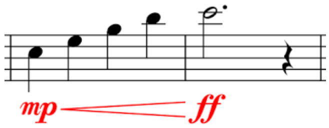


Figure 9: Example of music staff with the intensity notation of a *crescendo* from *mezzopiano* to *fortissimo* [26].

The main goal of this research is to make some significant considerations about how onsets' intensity influences the rhythm complexity perception, and at the same time to analyze how reliable the discussed metrics are when used on rhythmic patterns involving onsets with different intensities.

In real music performances, the sound intensity is a crucial aspect of a musician's feel and of a listener's perception. It is one of the main tools the musician has to communicate emotion and personality while playing. Also, it is nearly impossible for a musician to play with no intensity variations. That's why patterns with constant intensity are perceived as algorithmic, artificial and unnatural for a listener. These considerations state the importance of intensity in music.

In the MIDI protocol, *velocity* is one the most relevant parameters as it maps the intensity of the considered note, as well as other aspects such as timbre, envelope, pitch - that also in real music performance are related to intensity. Given a MIDI representation of a rhythmic pattern, it is sufficient to change the velocity of its notes to impact on the intensity of the onsets of the resulting rhythm.

In order to fulfill the purpose of this study, which is to investigate the sound intensity on the rhythm complexity perception and to evaluate the metrics presented in Section 3, two subjective listening tests have been designed. In these tests, the same rhythmic patterns with different sound intensities are proposed. The user who is presented with the test is asked to rate the perceived rhythm complexity for each of these. The implicit aim of the tests is to operate a comparison among rhythms with different sound intensities. Starting from well known datasets [16, 34, 49], different versions of the same patterns are derived by changing the velocities of the onsets in their MIDI representations. Collected data from the responses of the users are used to analyze how the objective metrics correlate with various sets of subjective opinions that can represent reliable judgements on rhythm complexity in each specific intensity scenario.

Two experiments have been conducted: a first one applied to monophonic patterns, and a second one about polyphonic patterns. These two subjective listening tests are presented in this section, and for each of these, used data and design choices are explained in detail.

4.0.1 Test Question

In the two presented subjective tests, users are presented with rhythms and asked to rate the perceived complexity for each of them. Some other studies tried different approaches in terms of test question [7, 8, 74]. For example some chose to ask to the listener to tap along to the rhythmic pattern, or to reproduce it. These are valid ways to evaluate the complexity of a rhythm, but still they are limited to the measurement of how hard it is to reproduce it and not how hard it is to conceptualize it, making it a less general question, and somehow changing the cognition perspective we look at the problem.

Furthermore, tapping and playing a rhythmic pattern may be easy in the case it is monophonic, but the story is different for polyphonic rhythms. Polyphonic rhythmic playing is a specific skill mastered by drummers, thus such a capability would've required prior knowledge about the act of playing drums. This not only automatically restricts the pool of users who the tests can refer to, but especially introduces a difference in the psychological problem addressed by these two tests, which instead are intended to be comparable. By simply asking to rate the complexity perceived in a rhythm, thus, the considerations deriving by the two tests can be compared even if referring to two different cases, as one test is the extension of the other.

4.1. Monophonic Patterns

The first experiment investigates the perception of monophonic patterns through a subjective listening test. These kind of patterns is not the most referable to real music, as drum musical tracks are in general the superposition of many monophonic elements, but it still represents a good starting point in a *divide et impera* perspective. Besides, the vast majority of the literature focuses on these kind of patterns [7, 16, 22, 45, 46, 66, 68], thus a likewise experiment can align well with the previous studies.

4.1.1 Experimental Data

Several sets of experimental data have been used in listening tests that, similarly to this one, were designed to investigate the perception of rhythm complexity [7, 15, 16, 46, 66]. All of these datasets are composed of binary monophonic rhythmic patterns of length 16 (with reference to a single measure in 4/4 with the sixteenth note as resolution), generated or selected according to some rule, which is usually connected to the presentation of some complexity measure. Thul's work [66] is the most exhaustive one about complexity metrics among the literature. In his study, Thul mentions and uses three datasets from three different studies: the one from Povel & Essens in 1985 [46] (which is also used by Shmulevich & Povel [55]), the one from Essens in 1995 [15] and lastly the one from Fitch & Rosenfeld in 2007 [16]. The one from Povel & Essens [46] consisted of 35 sequences, with fixed onset intervals and 830 Hz square wave tones of duration 50 ms as sound associated to the onsets, and the one from Essens [15] is a set of 20 sequences with same tone onsets but different onset intervals. The set from Fitch & Rosenfeld [16] is constituted by 30 patterns with four or five onsets, associated with a non-tonal click sound of duration 20 ms. In their work, Fitch & Rosenfeld also proposed an alternative version of the *Longuet-Higgins & Lee* complexity measure, validated by means of this set of stimuli.

In order to obtain results comparable with those of the literature anyone of these three datasets, or even more than one, would've been a good starting candidate for this experiment. However, an important design need was to limit the duration of the listening test, as it can have a significant influence on the performance of the listeners and consequently on the validity of the results [52]. For this reason, only one of them could've been used. The choice fell on the most recent dataset, from Fitch & Rosenfeld [16], because it is the only one among these three whose stimuli were manually composed. Indeed, Fitch & Rosenfeld created 30 rhythms by ranging the syncopation index from 0 to 15, with reference to their measure. Instead, the other two sets are mathematically derived by applying all the possible permutations of the same set of inter-onset intervals. Thus, Fitch & Rosenfeld's set result to be more realistic, especially with the introduction of dynamic accents.

The original dataset is made up of 30 monophonic binary patterns in 4/4 time signature. The patterns all last one bar, adopt a 16-pulses resolution and contain from 4 to 5 onsets. The stimuli were manually composed to guarantee a good syncopation range, according to Fitch & Rosenfeld's study [16].

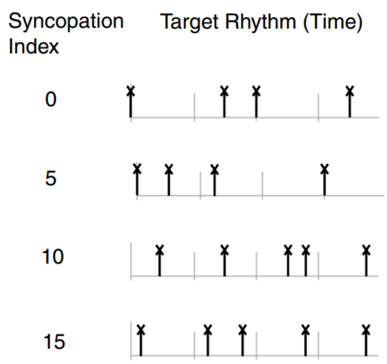
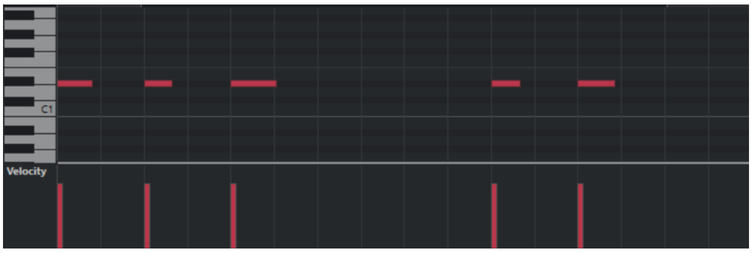
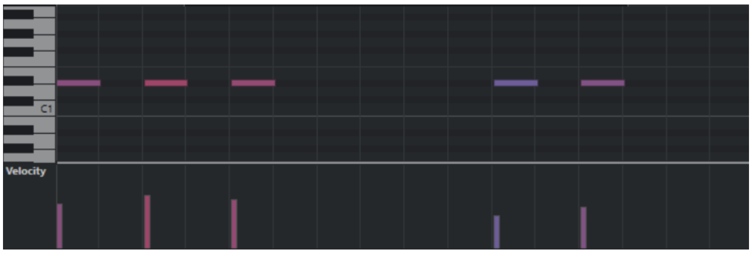


Figure 10: 4 example rhythms with different syncopation index taken from the original dataset of 30 rhythms by Fitch & Rosenfeld [16].

Since the focus of the study is to analyze how the intensity of the onsets in rhythmic patterns impacts on complexity perception, different versions of each pattern of the original dataset were obtained, where each version has different onsets intensities. In particular, for each of the 30 patterns, 4 different rhythms were derived, resulting in a total of 120 different rhythms. The new versions were derived by applying 4 different *velocity modes* to the MIDI representation of the original patterns, meaning that onsets' velocity values were set according



(a) *Constant velocity mode*



(b) *Random velocity mode*

Figure 11: Example of MIDI representation of the same pattern in two different velocity modes

to 4 different procedures. The first class is composed by rhythms with no intensity variations and represents the patterns from the original dataset with no modifications. As the original dataset was composed by binary rhythms and no intensity was thus associated to any of the note in the patterns, in the MIDI representation, all notes have been associated with a constant velocity value equal to 100. This class is understandably named *"Constant"*. A second class, which is named *"Hierarchy"*, was obtained by taking Lerdahl and Jackendoff's perceptive study seen in 3.1 as a reference: velocity values were assigned according to the weight of the pulses in the hierarchy. This means that when onsets occurred in pulses associated with low weights, low velocity values were assigned to those notes; viceversa, for accents in high weights positions, high velocity values were proportionally attributed. The third class was derived by assigning to all the notes in the patterns a random velocity value in the range between 35 and 120. This class was named *"Random"*. The fourth and last class was obtained in a rather less analytic way, as the velocity values of the patterns' onsets were not assigned according to any specific law. Instead, in this case they were expression of a real human musical performance, making this class the closest one to a real music analysis scenario. The performer listened to and subsequently played each of the pattern in the original dataset, and his performance was recorded through MIDI devices. Successively, onsets were quantized to a 16-th note resolution in a professional DAW, in order to make the patterns comparable with the ones from the other class for everything but the velocity. This is the *"Performed"* class. In Figure 11 an example of the difference between two velocity modes is shown by comparing the *piano roll* representations of the two inside a DAW [61].

The procedure just described to get the 4 classes *Constant, Hierarchy, Random* and *Performed* is only relative to the MIDI environment, and results in symbolic representations of the the considered patterns. In order to get the final audio files included in the test, for each of the derived 120 MIDI rhythms the following steps were performed. Each pattern was repeated 2 times, obtaining a set of 2-bars stimuli. This was done to point out the cyclic nature of the rhythmic patterns. Furthermore, each pattern in the test was anticipated by 2 measures of metronome, which fulfill the function of declaring the meter, providing the listener a reference structure to contextualize the rhythmic patterns and their syncopations. All the patterns were reproduced with 120 BPM, in order to avoid the scenario where different BPM could invalidate the complexity comparison between different patterns, making a faster rhythm seem more complex of a slower one independently of their symbolic patterns. Hence, each audio file associated to a rhythmic pattern in the test lasts 8 seconds in total.

Many similar studies [7, 16, 66] have chosen a click sound, similar to the one of a metronome, as the sample used for the stimuli creation. However, these studies did not consider the intensity's influence on rhythm complexity perception, thus some specific considerations about this aspect are needed. When a musician plays a real acoustic instrument with varying intensity, many aspects of the resulting musical performance change with it. Intensity, intended as volume, is the first one of course, but not the only one. Indeed, a different intensity hit also makes the instrument resonate differently, having an impact on the timbre of the associated sound, as well as on the pitch, even if in smaller portion. This is true for drum and percussion instruments, and for plucked and string instruments as well. Besides, even in digital emulations of acoustic instruments the same thing happens: velocity does not only map volume, but also timbre and other factors. That's why for such a research, instead of a clicking metronome sound, whose velocity variations only would've been associated with a volume modification, a drum snare sound was preferred. In particular, the pattern's MIDI tracks have been reproduced with a realistic acoustic drum snare sample from the high quality plugin *Addictive Drums 2* [5], which ensures a good expression in terms of range of possible intensities' simulation.

4.1.2 Methodology

Once the rhythms were derived as explained in the previous section, they were used to develop the listening test. The test design aimed to balance the need for an engaging comparison across rhythms from different intensity modes, with the constraint of a short test duration. Indeed, according to recent studies about the influence of duration on listening tests validity [52], users have a relatively short amount of time before their attention drops and their perception gets unreliable. The adopted solution involved dividing the 120 rhythms into three subsets of 40, creating three test versions in an "A/B testing" fashion. Each subset, obtained by randomly selecting 10 of the 30 original patterns by Fitch & Rosenfeld, included all four possible versions for each pattern. The subsets division is reported in Table 3. This approach ensured both reliable and meaningful results, reducing the overall test duration from 20 to approximately 5 minutes, thereby improving user attention and meaningfulness of the results.

Subset 1		**Subset 2**		**Subset 3**	
N°	**Pattern**	**N°**	**Pattern**	**N°**	**Pattern**
7	x . . . . . x . . x . . . . x .	4	x . . . x x x . . . . . x . . .	1	x . . . . . x . x . . . . . x .
14	. . x . . x . . . . . x . . . x	6	. . x . . . x . x . . . x . x .	2	x . . . x . . . x . . . x . x .
19	x . x . . . x . x . . x	8	x . . x . . . x . . x . x . . .	3	x . x . x . . . . x . x .
20	. . x . . . x . x . . . x . . .	10	x . . . . x . . . x . . . x . .	5	. . x . . . x . x . x . x . . .
23	x x . . . x . . . . x . . . x .	11	. . x . . . x . . . x x . x . .	9	x . . . . x . . . x . . . . x x
24	. x . . . x . . . x x . x . . .	12	. x . . . x . . . x . . x . x .	13	. . x . . . x x . . . x . x . .
25	. . x . . . x . . . x . . x . .	16	x . . . x . x x x	15	. . . x . x . . . . x x . . . x
26	. . x . . . x . . . x x . . . x	18	. . x . x . . . x . x	17	x . x . x . . . . . x . x . . .
28	. . x x . . . x . . . x . x . .	21	x . . . . x . x . . x . x . . .	22	. . x . . . x . x . x . . . x .
30	. x . x . . . x . . . x . . . x	29	. x . x . x . x . . . x	27	. . x . . x . . . . x . . x . x

Table 3: Adopted subdivision of the Fitch & Rosenfeld's dataset in three groups for the first listening test.

The online test page has been developed by means of the webMUSHRA web audio API based experiment software, developed by AudioLabs [51]. Also, the web page was hosted and made accessible through the Firebase platform [19]. Three different web pages were deployed with the same structure but different audio files loaded in it. All the pages were linked to the same url[1], and the user was redirected to the version of the test which had less visits.

The test consists in a likert scale questionnaire [25], where the 40 rhythms of each version were presented to the users, that were asked to express a personal rating about how rhythmically complex each rhythm is. The rhythms were divided in 4 pages with 10 rhythms each, not to overload to listener's attention with too many comparison all at once. Also, the rhythms were presented in a random order to the user, in such a way to minimize order biases [52]. For each rhythm, the users were asked to provide a rating from 1 to 5 about the rhythm complexity they perceived. No labelled examples were presented to the user in the introduction of the test, in order to avoid to condition its perception [13, 40, 78].

Together with their ratings about the rhythms, users were also asked to provide personal information as age, country and years of musical training. Besides, a specific slot for eventual feedbacks about the test was present.

[1]The test is available at this link: `https://maelisteningtest.web.app/`.

4.2. Polyphonic Patterns

The second experiment is an extension of the first one to the polyphonic case. Thus, it represents a further step in the perspective of the aim to apply rhythm complexity measures to real music. Indeed, when listening to music in everyday life, it is very improbable it does not involve the superposition of many instruments, and at a deeper level of many monophonic sources. Additionally, the most logical application of the metrics presented in Section 3 is to measure the performance of a drummer. And even only considering the drum track in a song, it is really unlikely to come across a pattern where a single piece of the drum is playing, as this specific instrument's full sense is found in the clever combination of all its parts. The drum elements include kick, snare, hi-hat, cymbals as ride or crash, low, mid and high toms, and all of these play a different role in the rhythmic narrative of a song, interacting with each other. Given the nature of this instrument, it is clear why it's important to extend the so far collected research about the monophonic patterns to the polyphonic case. The test described in this section is designed on this purpose.

From a top level perspective, the test's structure is very similar to the one presented in Section 4.1. Naturally, a different dataset was needed for this second experiment. Also, the metrics presented in Section 3 only refer to monophonic binary patterns, so a new approach to use them on polyphonic patterns was necessary.

Most of the available studies do not focus on polyphonic patterns. An exception is constituted by the works of [38, 58, 59], which in their researches deal with polyphonic patterns in the context of a multi-instrument drum. However, [58, 59] only computed a first step in this direction, as their work lacks of an important consideration needed when considering a real acoustic drum. Indeed, they did not take into account that the single elements of the drums do not play individually, but each of them has a role in the total pattern that the drummer is globally playing. This means that the patterns of the drum parts such as kick and snare interplay and complement each other. Hence, they should be measured considering the way they contribute to the feel of the total polyphonic pattern, and not merely by isolating them from the other monophonic patterns. Instead, Mezza et al. [38] included in their analysis the interplay between different drum elements, by means of the *k-grouping* presented in Table 2, in Section 3.7. The polyphonic metric presented in their paper, already discussed in Section 3.7, was indeed chosen as a reference and a starting point for the development and the analysis of this experiment.

4.2.1 Experimental Data

Following the example of [38], Groove MIDI Dataset (GMD) was used to obtain the audio files included in the test. This dataset was released by the authors of [18], and contains "13.6 hours of aligned MIDI and (synthesized) audio of human-performed, tempo-aligned expressive drumming" [34], and more than 22000 single measure drum patterns. It was obtained by recording the musical performance of both professional and amateur drummers (80% of the total was composed by professional) on a Roland TD-11 electronic drum kit, and multiple styles and genres are included. The dataset is annotated with metadata such as tempo, BPM, and anonymous drummer ID, and the MIDI files include notes and control changes for the hi-hat pedal position. Also, both the MIDI transcription of the drummers performance and the associated synthesized audio files are available in the dataset. However, for this experiment, only the MIDI files were used, and new audio files have been synthesized.

The original recordings include multiple tempos, but the analysis of this research is limited to 4/4 time signature drum patterns. There are two reasons justifying this choice: first, the studies in the literature are basically all also referred to 4/4 patterns; second, in the perspective of an application of the metrics presented in Section 3 to contemporary music, the 4/4 time signature is adopted by the vast majority of modern musical pieces. Indeed, this is also reflected in the GMD tempos distribution, which shows a 97% percentage of 4/4 time signature patterns.

GMD includes drummer performances of variable length; nonetheless, in this as well as in other studies [27, 41, 64, 66, 72], the definition of pattern is implicitly assumed to be limited to a single measure. Whereby, each of the recordings was splitted in single measures. A total number of 19920 measures in 4/4 time signature was obtained.

Next, for each of the obtained single-measure patterns, the reduction mapping proposed from [18] was applied. The General MIDI (GM) standard for drum kits assigns a number between 1 and 255 to each drum instrument, but this mapping is actually different from the one of the Roland TD-11, used to record the drum performances of GMD. Hence, the authors of [18] applied to the MIDI files of their dataset the mapping shown in Table 4, and the same is done here. In this way, a twofold advantage is accomplished: on the one hand, the pitches are handled more easily when synthesizing new audio files with a drum library; on the other hand, the representations of the drum performances are simplified avoiding unnecessary redundancies of similar instruments. As an example, many kinds of crash cymbals are included in GM and in Roland mapping, but in this case it would've been uselessly specific to consider them all individually, as they all carry out the same role in a drum architecture. Thus, it is possible to merge them into a single channel, without giving up the polyphony of the drum pattern. After the mapping, the number of instruments included in the drum patterns has gone from 22 to 9.

Pitch	Roland Mapping	GM Mapping	GMD Reduction Mapping
36	Kick	Bass Drum 1	Bass (36)
38	Snare (Head)	Acoustic Snare	Snare (38)
40	Snare (Rim)	Electric Snare	Snare (38)
37	Snare X-Stick	Side Stick	Snare (38)
48	Tom 1	Hi-Mid Tom	High Tom (50)
50	Tom 1 (Rim)	High Tom	High Tom (50)
45	Tom 2	Low Tom	Low-Mid Tom (47)
47	Tom 2 (Rim)	Low-Mid Tom	Low-Mid Tom (47)
43	Tom 3 (Head)	High Floor Tom	High Floor Tom (43)
58	Tom 3 (Rim)	Vibraslap	High Floor Tom (43)
46	HH Open (Bow)	Open Hi-Hat	Open Hi-Hat (46)
26	HH Open (Edge)	N/A	Open Hi-Hat (46)
42	HH Closed (Bow)	Closed Hi-Hat	Closed Hi-Hat (42)
22	HH Closed (Edge)	N/A	Closed Hi-Hat (42)
44	HH Pedal	Pedal Hi-Hat	Closed Hi-Hat (42)
49	Crash 1 (Bow)	Crash Cymbal 1	Crash Cymbal (49)
55	Crash 1 (Edge)	Splash Cymbal	Crash Cymbal (49)
57	Crash 2 (Bow)	Crash Cymbal 2	Crash Cymbal (49)
52	Crash 2 (Edge)	Chinese Cymbal	Crash Cymbal (49)
51	Ride (Bow)	Ride Cymbal 1	Ride Cymbal (51)
59	Ride (Edge)	Ride Cymbal 2	Ride Cymbal (51)

Table 4: Reduction mapping proposed from the authors of [18], which is adopted also for this experiment.

Next, each of the obtained patterns has been quantized to the 16-th note resolution, which is also here chosen as a reference. No modification about the original BPM of the rhythms was computed, as in a polyphonic case there are many instruments influencing the complexity and the speed plays a minor role.

GMD includes both long sequences of continuous playing in musical genres such as rock, funk and shuffle, and short beats and fills. Due to this reason, some measures within the dataset could potentially include some silences, or too overly sparse temporal patterns, which are unwanted for such a research. Also, sometimes in a long sequence there may be some variation measures where only few drum instruments are playing, e.g. when only the hi-hat is involved. In order to handle these problems, Mezza et al. [38] adopted a filtering approach. They discarded all the measures with less than three voices playing (to ensure to evaluate only proper polyphonic patterns), and with less than eight pulses where at least one onset is present in two measures. The same approach is used here, but referring to a single measure pattern.

Among the single measure patterns obtained in this way, 10 polyphonic rhythmic patterns were selected for the subjective listening test. The selection procedure was set in such a way to get a set of patterns with increasing expected complexities. For each of the candidate patterns, the complexity score according to Grouped Voice Polyphonic version of Toussaint's Metrical Complexity (presented respectively in Section 3.7 and 3.1) was computed. Also, for each pattern the variance of the velocity values of all the onsets was taken into account. The sorted array with all the complexity values was split in 10 groups, and in each group the first 100 candidates with the highest velocity variance was kept. Among these, a random pattern for each group was chosen. This procedure ensures multiple advantages. The split in groups and the random choice in a sub-set of patterns allow both to take into account the test design need to propose to the user patterns with different complexity degrees, and not to introduce an experimental bias by specifically selecting the samples. At the same time, using the velocity variance as a selection parameter keeps the focus of the experiment on the analysis of patterns where the dynamics have a considerable range, and thus to derive observations about the influence of the intensity on the perceived complexity.

Once selected the rhythmic patterns, two different versions for each of them were synthesized by changing the velocity values of the onsets in their MIDI representations, in a similar fashion with respect to what done in Section 4.1. As the patterns are from a dataset of human performance recordings, the first class is the *"Performed"* one and it is composed by patterns with the original velocity values. A second class was obtained by simply setting all the velocity values equal to the constant value of 100; this class is named *"Constant"*.

No other classes such as *"Random"* and *"Hierarchy"* (both present in the first experiment) were included in this comparison, as they would've uselessly burden the discussion with considerations that are not that interesting in a polyphonic case. This is due to the fact that the analysis that this test set is referred to realistic drum patterns, played by actual musicians and not algorithmically or digitally synthesized. Hence, the application of some dynamics trend obtained with different rules other than the human feeling would've lead to far-fetched scenarios, and thereby to unnecessary comments. The so-designed test thus effectively proposes a simple comparison about rhythmic patterns when human intensity is present and when not, in order to scope its influence on rhythm complexity perception.

Finally, each rhythm was repeated 4 times. No metronome measures were added to the beginning of the rhythms, as, with respect to the first test, in the obtained rhythms there was not need to stress the underlying beat. Also, the addition of a metronome would've resulted in a more boring listening experience throughout the whole test, risking to invalidate the users' clarity. The audio files included in the test was obtained by synthesizing the MIDI patterns with a professional drum library from the Groove Agent software by Steinberg [62].

4.2.2 Methodology

This experiment is based on a subjective listening test, where 20 audio files representing different rhythms are proposed to a listener. The listener is asked to provide a complexity rating for each of them according to their opinion.

The 20 audio files include 2 versions for each of 10 original patterns. An implicit comparison about the two versions is proposed to the user. The 10 original patterns include different level of expected perceived rhythm complexity. The rhythms were randomly split in 2 pages of 10, and this division is always the same for all the users. In each page the rhythms are presented to the user in a random order, in order to avoid biases associated to this aspect.

As the first experiment, also this test is based on a likert scale questionnaire [25]. The user is indeed asked to express a personal judgement about the rhythm complexity perceived in each of the audio files in the test, on a scale from 1 to 5. The ratings associated to this scale are *Very Low, Low, Medium, High* and *Very High*. The test structure is in common with the one explained in 4.1.2: no labelled examples are presented to the user, only some rhythmic patterns with varying complexity degree, taken from the GMD dataset.

Also personal information as age, gender, country and years of formal music training were collected from the users. Finally, the duration of the test is expected to be less then 5 minutes.

The web page of the test was realised using the WebMUSHRA software by AudioLabs [51], and published[2] and managed by means of the Firebase platform [19]. 82 users completed the test. They were all from Italy, the minimum recorded age is 20 and the max is 29. More or less, the majority of the users had some knowledge of music theory concepts. All the results are analyzed and discussed in Section 5.

[2]The test is available at this link: `https://polilisteningtest.web.app/`.

5. Results

In this section the discussion of the results of the experiments presented in sections 4.1 and 4.2 is presented. Each experiment aimed at collecting reliable data about rhythm complexity perception by means of a subjective listening test. These data were used as ground-truth reference to evaluate the quality of the metrics presented in Section 3. Pearson's Correlation Coefficient [9, 23] is here used to give a measure of the similarity between the complexity scores from the subjective tests and the ones from the measures.

5.1. Monophonic Patterns

The first experiment investigates the influence of the onsets intensity on human rhythm complexity perception in the context of monophonic patterns. This is done by means of a subjective listening test where versions with different intensities of the same rhythmic patterns are compared.

Three different versions of the same subjective listening test are associated to this experiment. In particular, each test version presents to the user 10 different rhythmic patterns, each of which was presented in four different variants, obtained by applying specific "velocity modes" to their MIDI representation. We called these modes *Constant* (*C.*), *Hierarchy* (*H.*), *Random* (*R.*) and *Performed* (*P.*) (Section 4.1.1). One test thus contained 40 rhythms; and totally, 120 rhythms were proposed to the users for this experiment.

Each test version was completed from 24 users, which means that in total, 72 users participated the experiment. Each of the three sets of users rated the perceived complexity of a specific subset of the total collection of rhythms. All the users who participated are from the Music & Acoustic Engineering MSc at Politecnico di Milano, or musicians, or researchers in the field of MIR. Therefore, we can anticipate a certain level of familiarity with the concepts of music theory from each of them.

5.1.1 Scores Distribution

The experiment focuses on the comparison between different velocity modes, so a first, top-view analysis tool can be found in the distributions of all the scores collected by each velocity mode, once merged together the results of all three the test versions. A total of 2880 were collected. The test is based on a likert-scale questionnaire [25], and only admits five possible ratings, which are *Very low, Low, Medium, High* and *Very high* complexity. The scores distribution differentiated for velocity mode is represented in Figure 12.

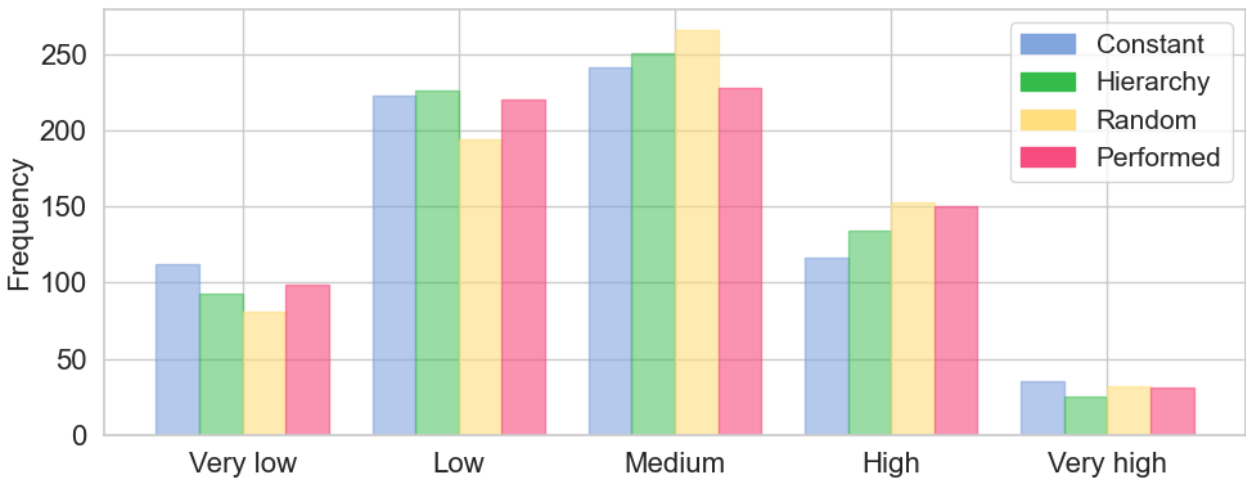


Figure 12: Scores distribution in the first subjective listening test, differentiated for velocity modes.

This histogram shows with which frequency each rating was indicated by the users, and grants a fast access to the general perception of each of these velocity modes. The first noticeable aspect is that the proposed rhythms were generally perceived as not that complex, as all the modes have their peak in the *Medium* score, and at the same time *Low* and *Very low* collected more votes than *High* and *Very high* (Figure 13). This latter in particular reached less than 40 votes in all the velocity modes, and by summing the votes obtained by all the modes it only counts 123 votes out of 2880. This percentage is too small to include this category in the discussion about the comparison among different velocity modes, as it could get to misleading conclusions.

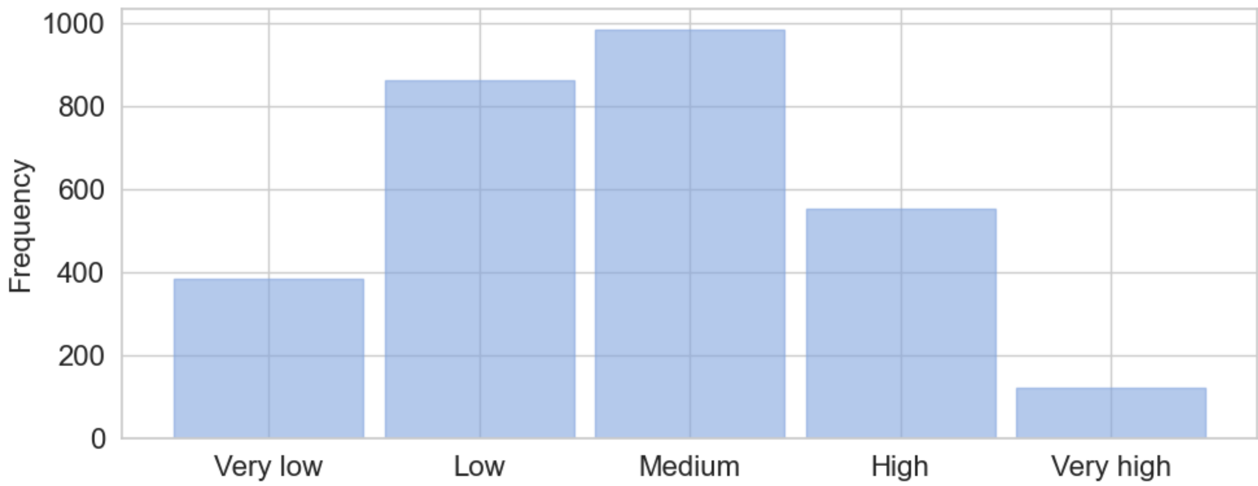


Figure 13: Scores distribution in the first subjective listening test.

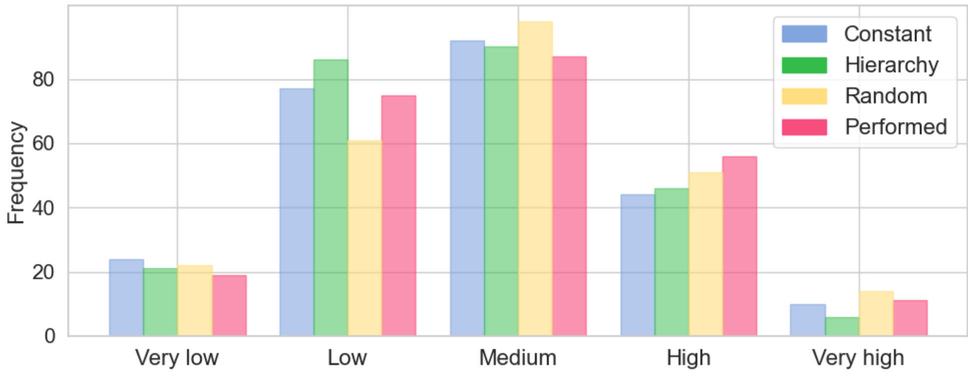
Once excluded the *Very high* rating, it is possible to derive some meaningful considerations about the velocity modes comparison, by looking at the modes distribution of Figure 12. The distribution of the *R.* mode shows that it is generally perceived as more complex than the others. Indeed, it is the mode that collected the highest number of votes in the *High* and *Medium* rating categories, and at the same time the lowest number of scores in the *Low* and *Very low* categories. On the other hand, *C.* mode has the opposite behaviour, as the histogram shows that in the high ratings it was voted less frequently than the other modes, and the opposite happens for the lower scores. *H.* and *P.* are somewhere midway between the first two modes. *P.* is the second one in the *High* category, but also the last one in *Medium*; and *H.* has a similar behaviour to *C.*, but with a slightly higher medium score (Table 5).

Constant	**Hierarchy**	**Random**	**Performed**
2.6410	2.6872	2.8085	2.7170

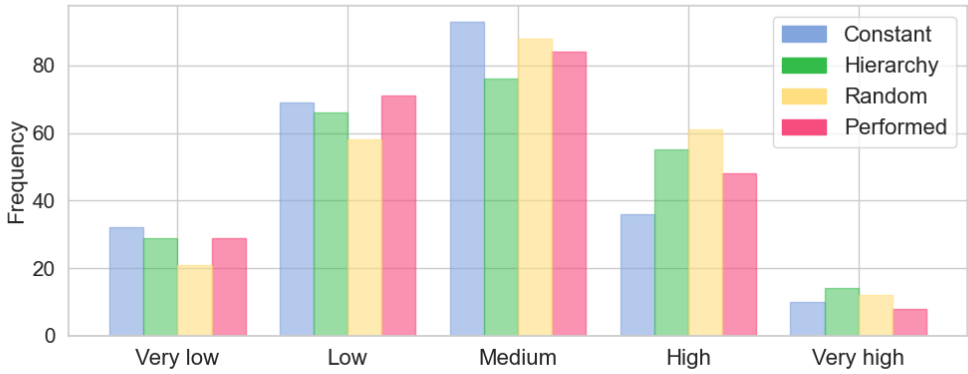
Table 5: Mean score for each velocity mode.

The behaviour of the distributions of Figure 12 is complex and deserves a deeper analysis to be comprehended. One thing to consider is the fact that the strategy of using three test versions each with a specific subset of data may have introduced different complexity degrees among the three tests. In other words, it is possible that the mean complexities of the three tests are not equal, and that this reflects in unbalanced distributions of different velocity modes in different tests.

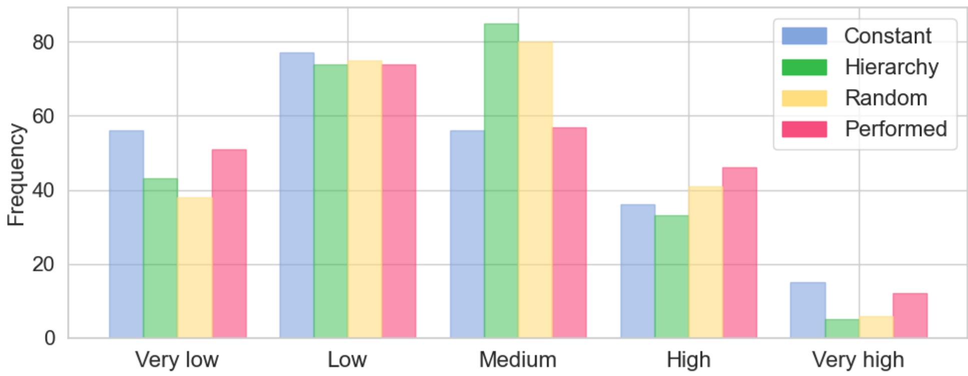
As an example, fig. 14 shows that in the *High* and *Medium* ratings, *R.* is not always the mode with the higher frequency. Indeed, in the test version associated to group of rhythms n° 1 *P.* obtained a larger number of *High* votes. Instead, in the second group of rhythms, *R.* mode gets a lower number of *Medium* scores than *C.* mode. Finally, in the third group of rhythms *R* has both a lower *Medium* rating frequency then *H.* and a lower *High* rating frequency than *P.* mode. It is also possible to observe that *C.* mode was not always the one which collected more *Low* scores, as in the first group the higher frequency of this rating category belongs to *H.* mode. The superposition of all three these histograms naturally results in the one of Figure 12.



(a) Group 1 subset scores distribution.



(b) Group 2 subset scores distribution.



(c) Group 3 subset scores distribution.

Figure 14: Scores distribution in the first subjective listening test for each group of rhythms.

5.1.2 Metrics Evaluation

In this section the evaluation of the complexity metrics presented in Section 3 is discussed. This assessment is performed by taking the subjective listening test results as a reliable reference of quality in terms of psychological rhythmic perception. In particular, following the example of other studies [38, 66], it is based on the computation of the correlation between the metric scores and the mean scores obtained from the test responses. In particular, *Pearson's Correlation Coefficient (PCC)* [42, 43, 63] was used for this purpose.

As a correlation coefficient, PCC measures the linear correlation between two sets of data, and can range between -1 and 1, where -1 indicates a perfect negative relationship, 1 indicates perfect positive linear correlation, and 0 indicates that the two observed variables are not correlated at all.

We here recall that six metrics are presented in this study: *Toussaint, Longuet-Higgins & Lee, Pressing, Weighted Note to Beat Distance, Inter-Onset Intervals Information Entropy* and *Toussaint's Off-Beatness*. For each of these a PCC was computed by coupling it to the average complexity scores obtained by the test. In order to highlight the comparison among the perception of rhythms with different dynamic accents, the test scores were split in groups according to the intensity mode of each rhythm. The combination of all the six metrics with all four the intensity modes leads to the computation of 24 different PCCs, which are presented in Table 6.

	Toussaint	**LHL**	**Pressing**	**WNBD**	**IOI Entropy**	**Off-Beatness**
Constant	0.569570	0.584780	0.703710	0.664680	0.551650	0.552740
Hierarchy	0.442090	0.485310	0.615350	0.560720	0.543100	0.445330
Random	0.668360	0.707180	0.747770	0.771720	0.480080	0.611540
Performed	0.679810	0.700700	0.759780	0.732970	0.478920	0.664490

Table 6: Pearson Correlation Coefficients between mean test scores and metrics' scores in each velocity mode.

This table shows how well the different metrics correlate with the subjective mean scores in each velocity mode. The Pearson coefficients range from a minimum of 0.44209, which is the correlation between *Toussaint*'s scores and the mean scores assigned by the users to the rhythms of the *Hierarchy* set, to a maximum of 0.77172, achieved by the *Weighted Note to Beat Distance* measure when compared to the users' average scores on the *Random* mode. By generally comparing row-wise the correlations in the table, *WNBD* and *Pressing* result to be the ones with the majority of high correlation scores. However, in specific velocity scenarios also other measures as *LHL* and sometimes *Toussaint* and *Off-Beatness* reach coefficients that are near to the highest of the considered row.

In general, there are some pretty good correlation performances shown in the table, as well as some others which are barely sufficient in the perspective a real Music Information Retrieval application. As a starting consideration it is important to stress that the object proposed by this experiment is constituted by a set of monophonic rhythmic patterns. This is not something that common listeners, or even professional musicians, are used to deal with. Indeed, in contemporary music the main carrier of the rhythmic aspect is constituted by the drums, or sometimes some percussion. But even in the simplest cases it is rare to observe a monophonic pattern, where a single instrument is playing and a unique sound is associated to the pattern. This is due to the fact that the monophonic patterns used here as well as in other studies are chosen as an abstraction of the rhythm itself, and should allow to concentrate only on the pattern characteristic, isolating other elements that in real music influence the rhythm. However, it is hard for a listener to listen to such a pattern and find it natural. Thus, it is dutiful to remember that the users who participated to the test may have found themselves more or less consciously displaced, with a subsequent alteration of their judgement.

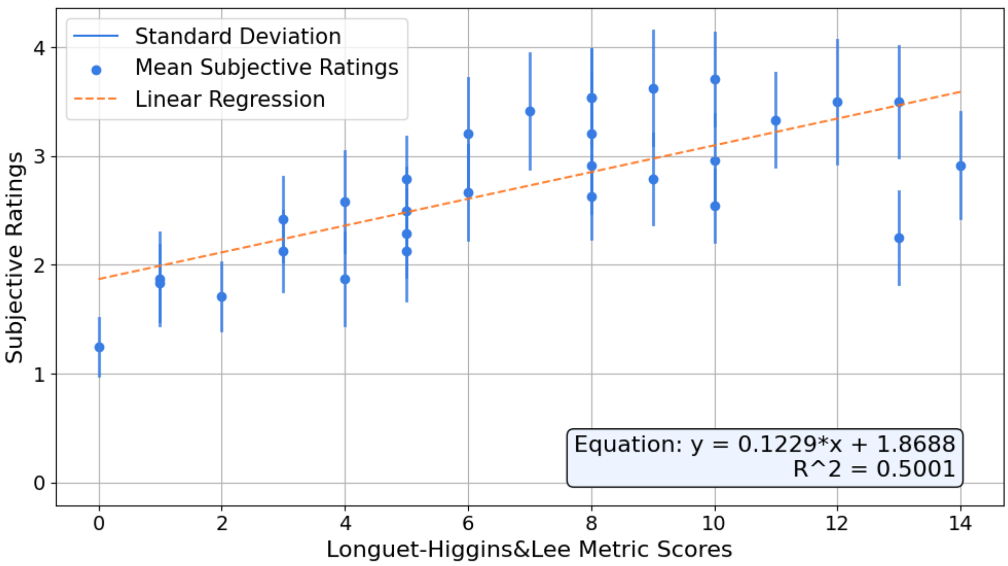


Figure 15: Linear regression between Longuet-Higgins & Lee's complexity metric and users' responses.

As regards the very low correlation scores in the table, one of the reason for that is that the users' responses showed a pretty high diversity, in addition to the fact that some metrics provide a restricted number of complexity scores on this dataset. This can be seen in Figure 15, 16, where a deeper representation of the correlation between metrics and test results is illustrated, taking as examples *LHL* and *WNBD* measures. In these charts, the orange dashed line represents the linear regression, the blue dots are the mean scores provided by the users, and the vertical blue lines are the scores' standard deviations. Undoubtedly, the audition of a monophonic rhythmic pattern does not represent what people are used to think to when it comes to listening rhythms, which is why the responses of the test have a high variance. Indeed, it makes sense to assume that people do not react in the same way to something that is unexpected. The perception of rhythm complexity is somehow a subconscious phenomena, and pointing out explicitly to evaluate that aspect to a listener may be misleading. On the other hand, no examples about a theoretic idea of rhythm complexity was provided with the tests. Thus, no biases related to this conditioned knowledge are expected in the results. In this perspective, a high variance is almost comfortable, as it is natural that each of the users that completed the test had their own idea of rhythm complexity.

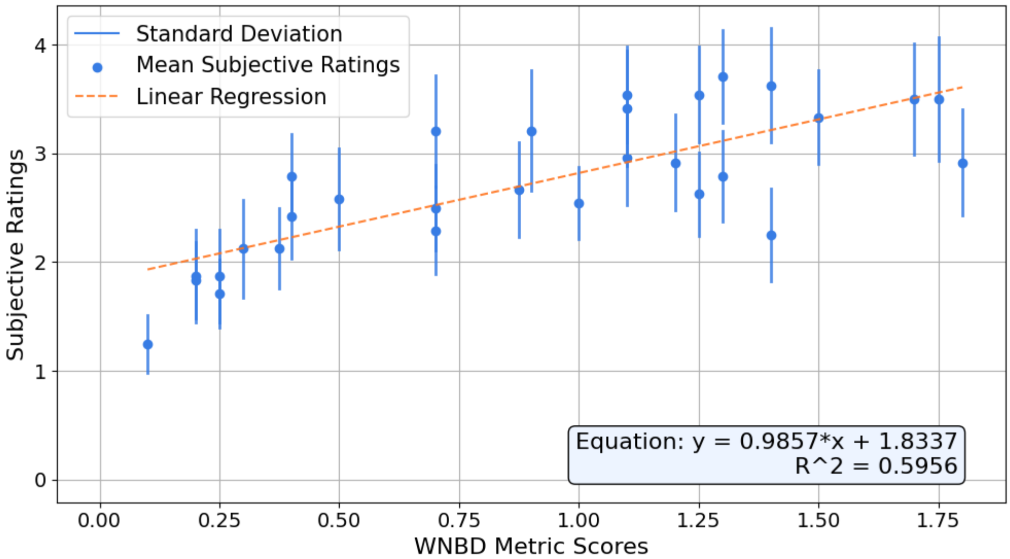
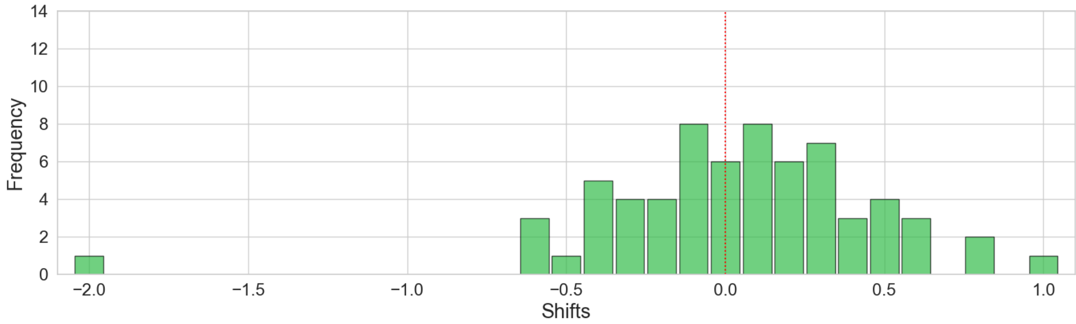


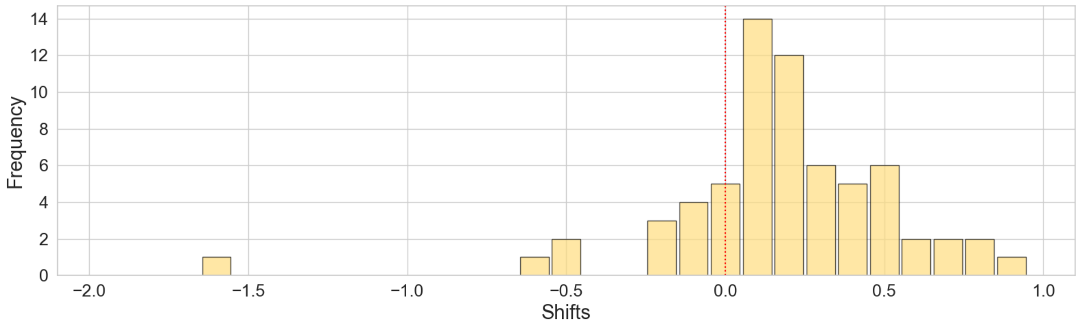
Figure 16: Linear regression between Weighted Note to Beat Distance's complexity metric and users' responses.

The representation advantage constituted by Table 6 is not limited to the possibility of investigating which metrics perform better than the others, which is done by examining the table horizontally. Indeed, we can even monitor if there is one or more velocity modes that are consistently associated to a better performance of the metric algorithms, by comparing the correlation values vertically. Actually, by comparing all the columns, a common trend is detectable. In particular, the metrics *Toussaint, LHL, Pressing, WNBD* and *Off-Beatness* share the same behaviour. In the columns related to these measures, namely, *Hierarchy*'s row always has the lowest score. This means that the metrics correlate worse with rhythms which follow that particular intensity behaviour than with other classes of rhythms. *Constant* mode, instead, is always the row that reaches the second lowest correlation score. The two classes which correlate fairly well with all the considered metrics are *Random* and *Performed*. Let's point out that this was not an expected result, as these measures do not include in their calculation any element which is associated with intensity variations of the rhythmic patterns. For this reason, the intuitive expectation was to find a good correlation for *C.* mode and something lower for all the others. However, there are a few considerations which may help to understand these numbers.

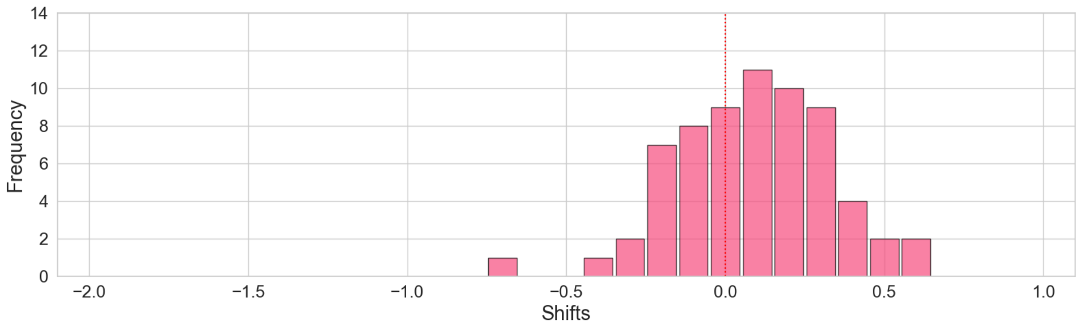
In order to explain the low correlation scores systematically reached by *H.* mode, a deeper analysis was conducted to gain an insight about how each mode was perceived in relation to each other by the users. Figure 17 shows the distributions of the users' mean perceived complexity shifts between different velocity modes. In particular, *C.* mode was taken as a reference, to whom the other three modes were compared. The purpose of this comparison was to look at how the introduction of intensity variations affects the perception of rhythm complexity for a person. To generate the histograms, the following procedure was employed: the scores given by each user for rhythms in each velocity mode were isolated. Then, the perceived complexity shifts for each user on the original 30 patterns were obtained by computing the pattern-wise subtraction between the scores for the *C.* mode and each of the *H., R.* and *P.* modes. The mean perceived complexity shift for each velocity mode was then calculated by averaging these shifts. Consequently, each user was associated with a mean shift perceived from *C.* to *H.*, one from *C.* to *R.*, and one from *C.* to *P.* mode. The histograms in Figure 17 represent the distribution of the shifts of all the users.



(a) Difference between *Hierarchy* and *Constant* modes.



(b) Difference between *Random* and *Constant* modes.



(c) Difference between *Performed* and *Constant* modes.

Figure 17: Distributions of the users' mean shifts perceived among different velocity modes.

The comparison among the three histograms shed some light on the correlation coefficients' relationship. Indeed, even if a good variance is observable in all three the distributions, and sometimes there even are some anomalous outliers to discard, it is noticeable that both the *Random* and the *Performed* classes were mainly perceived as more complex than the *Constant* class, while the users exhibited more indecision on the *Hierarchy* one. In particular, both the histograms of Figures 17b and 17c present a statistic mode on a shift value greater than zero; in addition, they also can boast a majority of positive shifts over the negative ones, meaning that there were more users that on average perceived these rhythms as more complex than the ones from the *C.* class. The same does not hold anymore when observing the histogram of Figure 17a, as a single statistic mode cannot be individuated, and the distribution of the mean shifts is 0 centered. Two peaks can be identified, one in the positive region and one in the negative one. These observations allow to understand that this particular class of rhythms was psychologically divisive, as it was not perceived in a univocal way by the users who completed the test. Actually, their behaviours split in two nearly equally distributed groups: half of the users found this class less complex than the *C.* one, the other half thought the opposite. Thus, it is reasonable to assume that one single number, which represents any of the considered metrics' score, cannot capture the divisive nature of this family of rhythms, and thereby none of these metrics can accurately correlate with the *H.* class.

It makes sense to see similar correlation values and similar mean shifts distributions for the classes *R.* and *P.*, as they are probable to share similar treats. This is because the intensity variations on a human performance are never systemic and can be fairly well approximated by random changes. Indeed, in the context of music production, one of the most common strategies adopted by producers to increase the realism of a digital instrument track is to apply random velocity variations on its MIDI notes. This is known to be a method that makes the sound more natural and human, and less digital, and there even exist many audio plugins that in the context of a DAW can automatically fulfill this task [1].

Behind this practice there is a psychological explanation: we find pleasure in surprise [24]. In every aspect of music perception, it is proven [10, 54] that a controlled amount of surprise increases the enjoyment of music listening: the brain finds boring what is totally predictable. The surprise can go from a conspicuous harmonic modulation to a subtle dynamic variation in a percussion. In this context, these variations are one of the key elements that separate a realistic human performance from an artificial, lifeless imitation. Everyone who plays an instrument, indeed, will introduce even unintentionally some intensity alternations, and the brain innately knows that if this kind of alterations are perceived, he's dealing with a realistic performance. On the basis of these considerations, it is understandable how the rhythms from mode *C.* would be perceived by the users. Such an unnatural performance's perception would lead to a disorienting sensation. Furthermore, this aspect is even accentuated by the fact we're dealing with a monophonic pattern, which already constitutes a relatively natural element on its own. As a consequence, it is not easy to predict how users may on average react to this class of rhythms.

Still, apart from sounding unnatural and artificial, these rhythms should still resemble the most appropriate candidates for an evaluation of the discussed metrics. In the literature, namely, similar examples with no intensity variations were used as stimuli, and moreover, the so far seen measures do not include dynamics related algorithms. Hence, one could ask themselves why some metrics which do not account for dynamic variations show a better correlation with the scores of rhythms that present them than with the scores of constant intensity rhythms. On this purpose, one further consideration on the users' perception is needed here. The point of view of who designed the test surely has some differences from the one of the listeners. Indeed, our a priori knowledge coming from the study of the literature implicitly made us set the *Constant* mode as a reference (the original dataset from Fitch & Rosenfeld [16] did not include intensity changes), and think to the other modes as alterations with respect to this one. Actually, when a user with no a priori information is presented with the test, they listen to 40 rhythms, whereof 75% of these contains some dynamic variations (each of the *H., R.* and *P.* classes). From the listener's perspective, the vast majority of the stimuli in the test presents intensity modulations and only a pretty small percentage of them do not. Thus it is reasonable to assume that a listener would infer as a kind of reference the rhythms with intensity alterations and then perceive the constant intensity ones as outliers, adopting a different criterion for their judgement. This example is important to remember that the results shown in this section always have to be interpreted as relative to the other test items, and not absolutely. Based on the previous assumptions, the relatively low correlation scores of the *C.* mode should not suggest that the absence of velocity decreases the quality of the metrics performance; instead, those coefficients simply are associated with the fact that in that rhythms' category people expectations were not met.

In the last few considerations we were not referring to all the metrics included in this analysis, indeed only *Toussaint, LHL, Pressing, WNBD* and *Off-Beatness*' measures share the fact they have a low correlation coefficient for *H.* mode, a mid one for *C.* and their highest ones for *R.* and *P.* modes. Thus, let's now comment *IOI Information Entropy* measure. In the case of this measure, something almost opposite happens: *R.* and *P.* classes reach the lowest correlation scores, while *C.* achieves the highest one, with *H.* placed midway. Why does this metric constitute an exception with respect to the other ones?

This is the only measure in this set that simply considers the distance between onsets, and not their absolute position. This is a deep difference from all the other considered measures, which for a reason or another, all deal with onsets' positions in the bar. Indeed, *Toussaint* and *Longuet-Higgins & Lee* are based on assigning a metrical weight to each pulse depending on its location, *Pressing* aims at detecting specific sub-patterns based on the onsets' position, *Weighted Note to Beat Distance* associates a weight to each onset based on its distance from the nearest beats, and *Off-Beatness* even only admits some specific pulses to be associated with a complexity contribute. All these measures share the underlying vision of thinking that the rhythm complexity is based on syncopation, which is related to the onsets' positions. When musicians play syncopated patterns, they will always - even unconsciously - accentuate that syncopation with intensity variations that can emphasize the syncopation feeling. This is due to the fact that rhythm and sound are hard to separate [12], as they both contribute to the total musician's feel. Thus, even from the listener's perspective, syncopation is psychologically linked to dynamic variations. On the other hand, *IOI Information Entropy* looks at the problem from a totally different perspective: the pattern is described by a discrete probability distribution, and the flatter this latter, the more complex the pattern. In this scenario there is no space left for a listener's innate association of the location complexity with an unconscious feeling of dynamics. Also in [76], a different kind of entropy measure (computed on the wave audio file samples) was compared to syncopation as predictor of pleasure and desire to move, and subsequently of rhythm complexity. That comparison already showed that entropy and syncopation achieve different results in terms of correlation with the the listeners' responses because of their different nature. Thus, it makes sense to see such a difference in the columns of Table 6.

5.2. Polyphonic Patterns

In this section the results of the second subjective listening test are discussed. This experiment focuses on the analysis of polyphonic rhythmic patterns, where by polyphonic we mean a pattern where multiple elements are simultaneously playing. With reference to the scenario of a drummer playing, a monophonic pattern would be constituted by them only playing a single drum element, e.g. the snare, while a polyphonic pattern would include more than one instrument (for example a pattern where kick, snare, toms and hi-hat are played together). In this sense, it is interesting to observe subjective listeners' responses in the case of polyphonic patterns, as this is the the kind of rhythms one would extrapolate from a real music track. Also, in the literature there are still not so many studies about this specific category [38, 58].

In this test ten original rhythmic patterns are present, with an expected varying degree of complexity. For each of these patterns, two versions are included, where each version follow a different intensity profile. The two compared intensity classes are *Performed* and *Constant*. Thus, a total number of 20 rhythms was listened by each user. For each of them, the listener was asked to assign a complexity rating from 1 to 5 (from *Very low* to *Very high*) according to their opinion, in a Likert-questionnaire fashion.

The test was completed from 82 people. All the participants were musicians, or students from the course of Music and Acoustic Engineering at Politecnico di Milano. Thus, they are all expected to have some music theory familiarity. Anyway, the years of formal music training was one of the personal information collected in the test, together with age, gender and country, as well as eventual feedbacks on the test.

5.2.1 Scores Distribution

Following the same analysis approach used for the first experiment, the comparison between the general perceptions of each of the different intensity modes can be observed by means of the distribution of the scores given by the users who completed the test. Indeed, by separating the scores assigned to the rhythms of each of the analyzed classes of intensity, one can rapidly infer if one of these was perceived on average as more complex than the other one. Two possible realizations of this distribution are shown in Figures 18 and 19. These figures show the distribution of 1640 collected ratings. Figure 18 shows the distribution in relation to the five scores included in the test, meaning that no pre-processing was performed on the test data. Figure 19 is still a representation of the scores distribution, but the scores were first standardized for each user. This standardization is useful to identify and avoid the eventual influence of some subjective judgement bias, e.g. some users may adopt a standard of medium complexity lower or higher than some others, based on their personal taste. This can invalidate the interpretation of such a distribution, as it would be distorted by incomparable ratings.

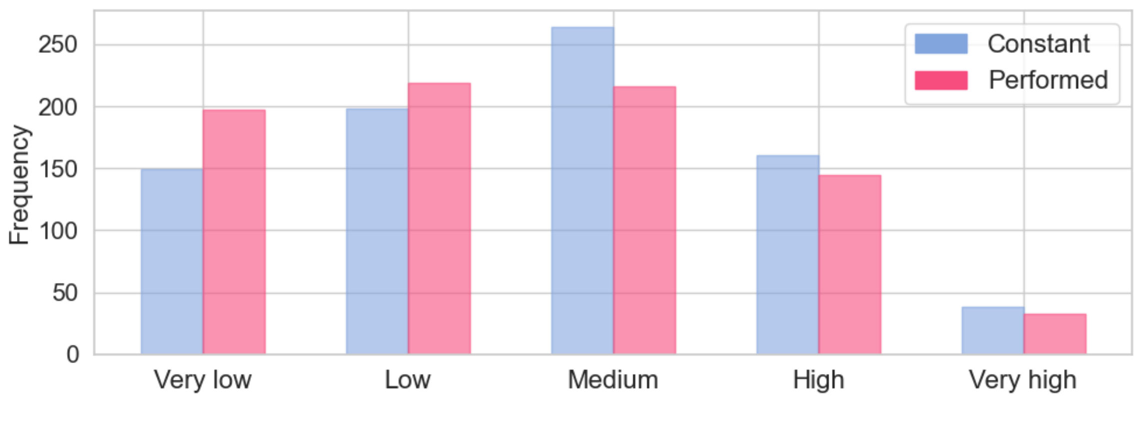


Figure 18: Second subjective listening test responses' distribution (raw data).

However, in this case, no particular differences were tracked. Indeed, both the charts highlight how the *Performed* class' scores are more densely distributed in in the *Medium* - *Low* portion of the ratings spectrum, while

Constant had a higher tendency to receive higher scores. In particular, it is easy to see that *P.* mode collected more votes than *C.* on *Low* and *Very Low* scores, and instead *C.* obtained a higher number of votes with respect to *P.* on *Medium, High* and *Very high* scores.

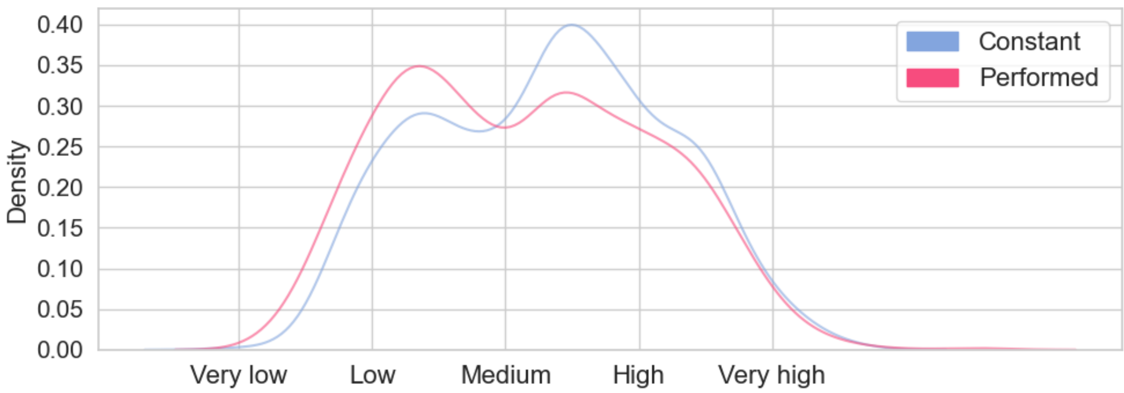
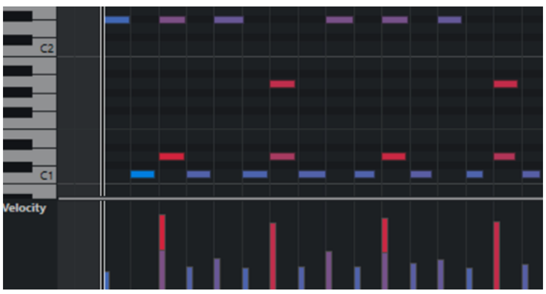


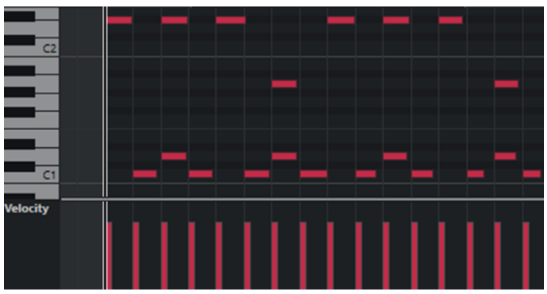
Figure 19: Second subjective listening test responses' distribution (standardized scores for each user).

On the highlighted trend of the *P.* mode being perceived as generally less complex than the *C.* mode, it is possible to consider how the rhythms of the two modes were synthesized. As explained in Section 4.2.1, the stimuli used in the test were picked from a pretty wide dataset of polyphonic drum MIDI transcriptions, from where single measures were isolated and used as candidates. Out of this vast pool, ten rhythmic patterns were selected in such a way to grant a varying expected complexity degree and a generally high dynamic range. Then, 20 rhythms were synthesized by applying two different velocity profiles on the MIDI representation of each pattern, which are *Performed* and *Constant* themselves. The first one simply leaves intact the velocity values of the MIDI drum transcriptions from GMD, while the second one applies a constant value of 100 to the velocity of each onset in the patterns.

In the case of a real drum performance, the musician does not play every note with the same intensity. This is both due to the fact that different instruments in the drum are played with different intensities (e.g. the hi-hat is played at a really low volume compared to the kick drum), and due to the presence of ghost notes, which is one of the key elements giving the rhythm a human feel and a deeper dimension in terms of groove. Ghost notes - sometimes named also as dead notes, silent notes, or mute notes - are defined as musical notes with a rhythmic value but no discernible pitch when played [4]. These are often played on the snare, which is one of the main characters in a drum rhythm performance. The main notes played on the snare are usually played quite loudly; the ghost notes, instead, can be described as notes played very quietly between the main notes. Although they are played softly, they are really important as they fill out the beat and add depth and space to the music, in a subtle yet meaningful way. The drum MIDI transcriptions of GMD [34] are obtained from professional drummers' performances, which of course included ghost notes in their playing style. Also, in the selection of the ten patterns a pretty extended dynamic range was one of the requisites.



(a) Pattern 5 in *P.* mode.



(b) Pattern 5 in *C.* mode.

Figure 20: Comparison between the same pattern in *P.* (a) and *C.* (b) velocity modes.

Ghost notes have a high responsibility in the distinction between the *Performed* and *Constant* classes. Indeed, while in the *P.* one they are associated with a velocity value directly derived from the drummer's performance, which is very low, in the *C.* mode they are brought to a relatively really high velocity value. As a consequence, *C.* mode rhythms result to have a higher number of main notes than their counterpart of the *P.* mode. This can significantly modify the perception and the feeling of a rhythm. Indeed, rhythms from *C.* mode will be perceived as denser in terms of onsets, sometimes even more confusing and intuitively more complex than the rhythms belonging to the *P.* class. A visual representation of this alteration is observable in Figure 20, where using the pattern n° 5 as an example, the difference between the two versions is illustrated. In those images, the velocity values of the notes are represented by the height of the vertical lines and from the color of the bar (red indicates high values and blue indicates low ones).

Figure 20 clearly shows that the number of main notes goes from four in *P.* mode to 16 in *C.* mode. Also here it is important to stress that these results cannot be observed on their own, but they always have to be put in relation with each other. In this case, the comparison between *P.* and *C.* modes' scores distributions of Figure 18 does not necessarily mean that a constant intensity pattern is in general more complex than a realistic drum pattern. Instead, it simply tells that in the scenario of this test, the increase of the velocity of the ghost notes comported a shift in the listeners' perception.

5.2.2 Metrics Evaluation

The listeners' responses collected with the second test were used to evaluate the proposed metrics, similarly as done for the first test. However, in this case the object of the experiment is constituted by polyphonic rhythmic patterns, and the original metrics presented in Section 3 are described for monophonic patterns. Thus, a polyphonic version of the discussed metrics was computed by applying to each of them the Grouped Voice Polyphonic measure algorithm by the authors of [38], presented in 3.7.

Thus, the measures which are going to be analyzed and compared in this section are the polyphonic versions of *Toussaint, Longuet-Higgins & Lee, Pressing, Weighted Note to Beat Distance, IOI Entropy* and *Off-Beatness*. Each of these was used to obtain a complexity score for each of the ten patterns included in the experiment. As done for the first test results, these measures are evaluated by means of the Pearson's Correlation Coefficient. In order to compute these correlations, the subjective listening test's scores assigned by the users were averaged for each rhythm, and then split according to each rhythm's category. Finally, PCC was computed for each couple of polyphonic measure and test mean scores relatively to each particular rhythm class. The obtained coefficients are reported in Table 7.

	Toussaint	**LHL**	**Pressing**	**WNBD**	**IOI Entropy**	**Off-Beatness**
Constant	0.836070	0.600332	0.812444	0.722569	0.679872	0.592801
Performed	0.967446	0.828665	0.923137	0.811797	0.916034	0.547831

Table 7: Pearson Correlation Coefficients between mean test scores and polyphonic metrics in each velocity mode.

At first sight, the correlation coefficients are in generally way higher than the ones shown in Table 6, which refers to the first experiment. This fact represents a comforting result, as it means that the measures presented in this study are pretty reliable when it comes to a realistic drum pattern's analysis scenario. Also, it proves that it is possible to use the polyphonic versions of the discussed measures with no loss of generality (on the contrary, the quality of the measures has increased), extending the work of [38] to the realization of five more polyphonic measures with the same approach. Moreover, the fact that here higher correlations are shown with respect to the first experiment can be simply motivated by considering that in this context the nature of the analyzed objects is way more similar to what the users expect by thinking to a rhythm. Indeed, a drum track, and thus a polyphonic pattern, is the most diffused example of rhythmic element in contemporary music. Hence, the listeners of this test faced something they were more used to and found familiar, with the consequence of a more consistent judgement.

From the numbers showed in Table 7, it can be seen that *Toussaint*'s complexity measure was always the one with the highest correlation coefficient in relation to the mean judgements collected from the test. This is a quite reasonable result, considering that the set of stimuli included in the test was selected by using *Toussaint*'s metric as a reference to compute the expected complexity. At the same time, it was not obvious that the users' responses were in agreement with the scores attributed by the *Toussaint*, so this still constitutes an encouraging result.

The absolute highest correlation score in the table was reached by *Toussaint*, and corresponds to 0.967446. Such a number indicates a strong linear relationship between this metric and the test scores, which is a highly informative quality measurement of the metric. Aside from *Toussaint*, also *Pressing* achieved highly positive correlation scores, collocating as the second best option in every row of the table. *WNBD* and *LHL* generally placed themselves at half of the ranking with respect to the other measures in each row, but *WNBD* reached slightly higher absolute correlation values than the ones of *LHL*. *IOI Entropy* performed better with respect to the first test, but in the *C.* class still got one of the lowest scores. However, the lowest correlation score obtained by a metric in the table was scored by *Off-Beatness*, and equals to 0.547831. This measure got the lowest position in both the rows, indicating that its algorithm is not up to the other discussed metrics, when dealing with polyphonic patterns including dynamic accents. For this reason, a modified version of this measure is presented in Appendix A.

Also, with respect to the first experiment, higher coefficients of determination are obtainable with simple linear models, showing a higher correlation among measures' scores and human's scores. Indeed, the simple linear models represented in Figure 21 and 22 can fit the average scores assigned by the users with a coefficient of determination of respectively $R^2 = 0.852$ (Figure 21) and $R^2 = 0.936$ (Figure 22).

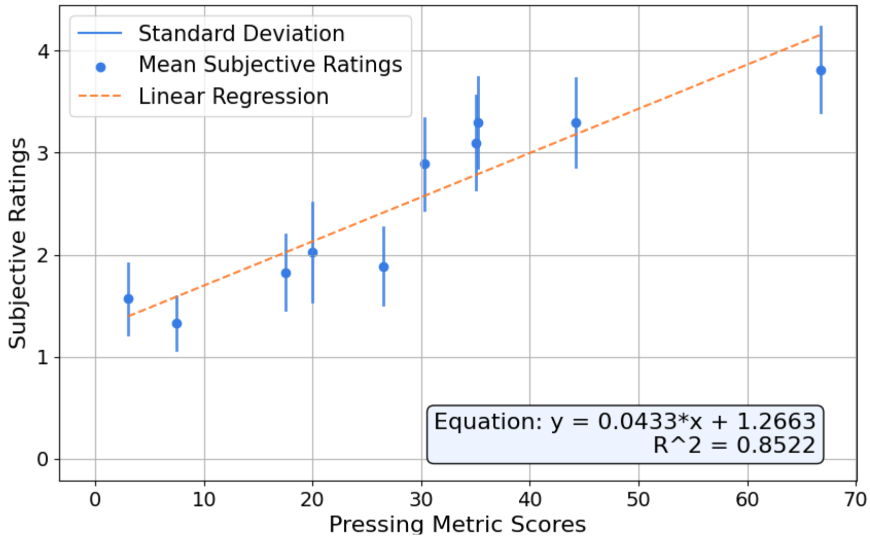


Figure 21: Linear model which shows the correlation between *Pressing*'s polyphonic measure and the average scores assigned by listeners.

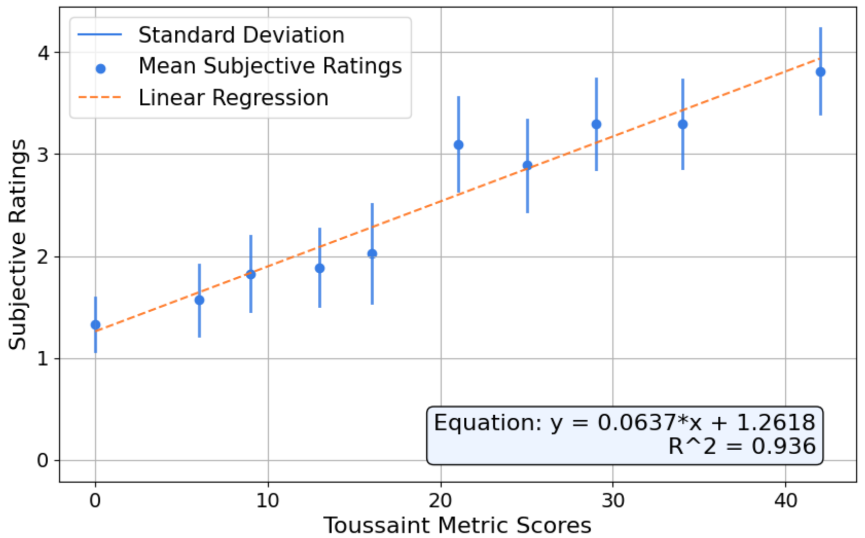


Figure 22: Linear model which shows the correlation between *Toussaint*'s polyphonic measure and the average scores assigned by listeners.

In Table 7, the comparison among the *Constant* and *Performed* velocity modes highlights quite clearly that in almost each column the correlation score with the $P.$ mode is higher than the one of $C.$ row. The only exception is found in *Off-Beatness* measure. Thus, for all the other measures the interpretation of the results seems to tell that the quality of the measurement performance is better on rhythms with realistic intensity profiles, as the ones played by actual musicians. What is unexpected is that these metrics result to be more reliable when applied to rhythmic patterns which include intensity variations, although they are not sensible with respect to this aspect, as they shallowly look at the analyzed patterns as binary sequences. However, this may be a superficial interpretation.

It is important to remember the nature of the values shown in Table 7: these are correlation coefficients between the scores assigned by the measures and the mean of the scores assigned by the users. In particular, what we are stressing here is that the measures return a unique value that indicates the complexity of the analyzed object, but the users are many and each of them is associated with a judgement. By averaging across all the users, we are combining yet blurring information. The intuition behind this consideration is that a low correlation may be justified by a too high variance in the users' responses. On this purpose a further analysis was conducted, and the following charts represent evidence to support this argument.

Of course, the internal representation of a generic rhythm for an average listener includes intensity variations, as it is the kind of rhythm they are used to listen in contemporary music. This means that for a person that both listens to a $P.$ rhythm, which includes realistic human performance intensity variations, and to a $C.$ rhythm, which only admits one level of intensity for all its notes, the first one will be perceived as more familiar. Thus, in a comparison among the two, the constant rhythm is felt as unnatural, and this may lead to internal confusion for the listener. Such a confusion, in the context of many listeners could lead to disagreement among them. In other words, people's ratings are inconsistent when dealing with unfamiliar objects.

Figures 23 and 24 show how the users changed their ratings while listening rhythms from $C.$ mode to $P.$ mode. In particular, Figure 23 was obtained by computing for each user the mean score given to the rhythms of $C.$ mode and the one given to the rhythms of $P.$ mode and subtracting these two numbers. The result of this operation represents the mean displacement that the user perceived between the two modes, and Figure 23 illustrates the distribution of these mean displacements. On the other hand, Figure 24 looks at the same question from another perspective: it is indeed the distribution of the displacements perceived for each pattern. Thus, for each rhythm the average score given by all the users was computed for each of the two modes, and their difference shows if the same pattern in $P.$ mode was perceived as more or less complex than its version in $C.$ mode.

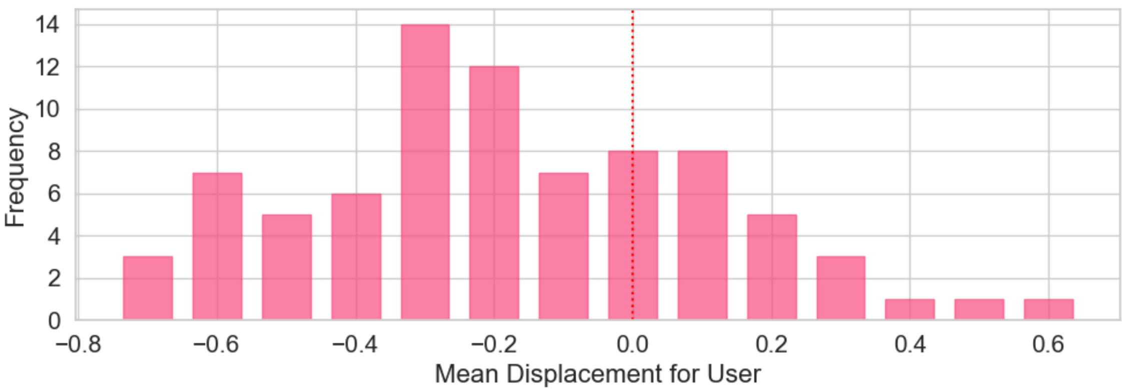


Figure 23: Distribution of the displacements perceived by each user from $P.$ mode to $C.$ mode, averaged for all patterns.

Both the charts suggest that the class *Performed* was almost unequivocally perceived as less complex than *Constant*. This result is in line with the scores distributions shown in Figures 18 and 19, and can be motivated with the same considerations done for those, but it was still not obvious that the majority of the users was in agreement on finding less rhythm complexity in $P.$ mode than in $C.$, for the same patterns. Indeed, in Figure 23, it can be observed that the mode of the distribution is in the negative portion of the x-axis, at -0.3, and that there are many more users which perceived a negative shift with respect to the ones who felt the opposite. The histogram of Figure 24, instead, can be interpreted even more precisely: seven patterns out of ten scored a lower complexity when associated to the $P.$ class with respect to their counterpart in $C.$ class. In particular, the most frequent shift is -0.1, but the lowest one is -1.2, which on a scale from 1 to 5 is pretty high.

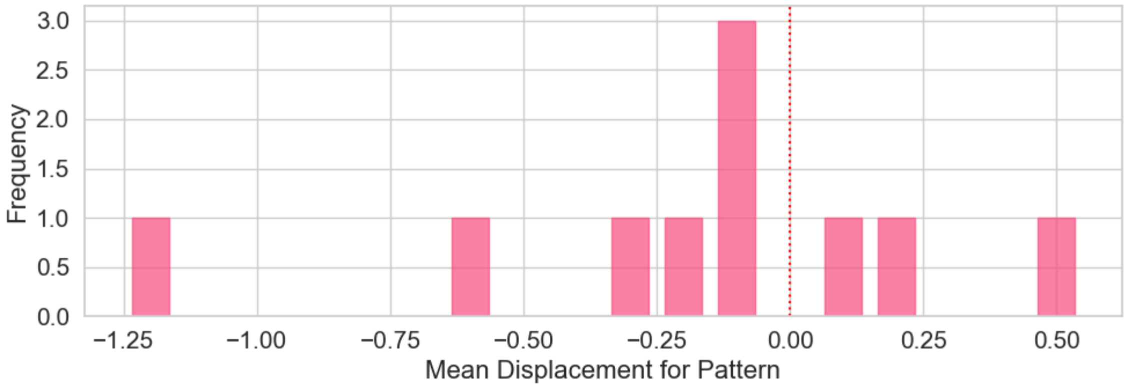


Figure 24: Distribution of the displacements perceived for each pattern from its version in *P.* mode to its version in *C.* mode, averaged for all the users.

What the last two figures suggest is that the users were quite in agreement on rating the rhythms from *Performed* class. At least, no distinction between two different populations were detected, as the statistical modes were unique for each chart. This fact represents a first confirmation of the previous consideration according to the correlation coefficients between measures and mean users' scores are higher for the *P.* class because this latter is more familiar for the users, with the consequence that their judgement is more consistent in that scenario.
A second confirmation is needed in the analysis of the variance of the users ratings in the *Constant* mode, which is shown in Figure 25. This graph shows the variance by means of a violin plot, where the distributions (Kernel Density Estimators (KDE)) of the scores obtained by each of the patterns are reported one next to each other. This representation allows to see how the users distributed in rating the rhythms, and in particular, if one or more statistical modes can be observed.

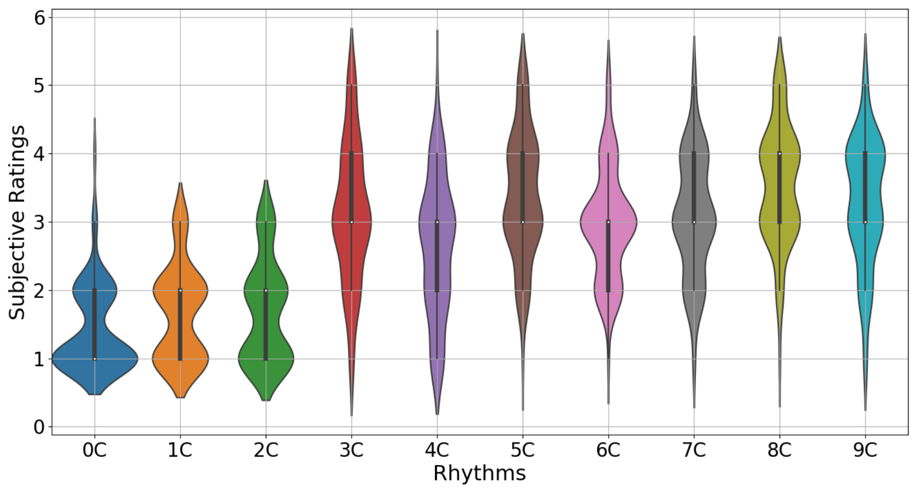


Figure 25: Violin plot showing the KDE distribution of the scores obtained by each rhythm in the *Constant* mode.

It is clear by looking at the violin plot that many rhythms were divisive with respect to the opinions of the users, especially those which were at the extremes of the proposed complexity scale. In particular, the first three and the last two rhythms are the ones showing higher indecision from the totality of the users; indeed, in the case of these rhythms two distinct populations of ratings are observable. Also a couple of the mid complexity rhythms have a scores distribution with a similar behaviour, but the a main statistical mode is still recognizable. In general, the rhythms associated with the mid complexity positions are the ones who collected a distinguishable majority for a unique complexity score. However, the same rhythms are also the ones who can boast the highest variance, as their mid complexity brought to discordant opinions, collecting all the five possible ratings. Both the distinction of two populations and the high variance are two factors that can invalidate the correlation with a measure which is associated with a unique number. Indeed, the points that in the linear regression charts like the one of Figures 21 and 22 are outlier with respect to the linear model are motivated by one of these

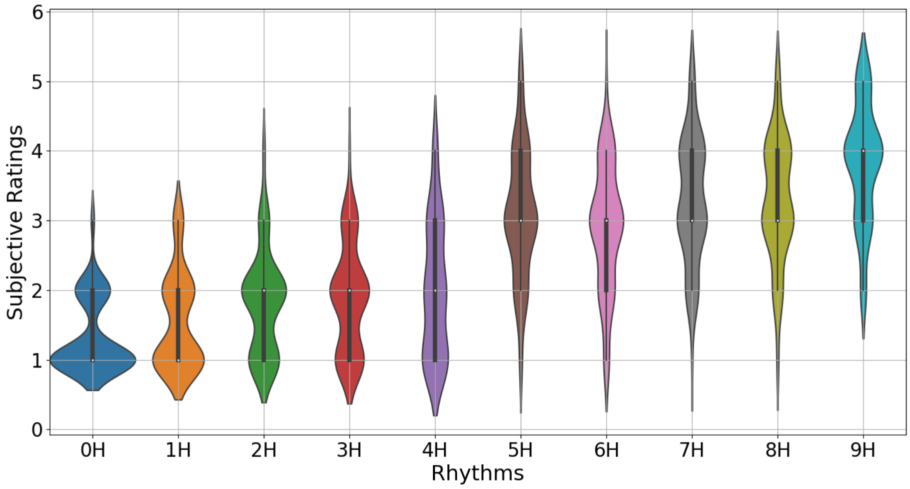


Figure 26: Violin plot showing the KDE distribution of the scores obtained by each rhythm in the *Human* mode.

two causes. It can be seen that the same violin plot realized for the rhythms of *P.* mode (Figure 26) shows less sparse distributions. Indeed, the KDEs are less extended - because the users were less undecided - , show fewer examples of population distinction - because their judgement was more consistent - , and the different complexity ranks are more recognizable.

The last charts made clear that the keeping in mind the users' perspective is fundamental when doing considerations on the correlation coefficients of Table 7. Indeed, the simple reason for the higher correlations with the mean scores of *P.* mode's rhythms is that people's judgements were too separated in the rhythms with constant intensities. This is due to the unfamiliarity that is associated to these kind of rhythms: when the listener has to rate something they perceive as unnatural, their internal judgement scale is compromised and they evaluate that object differently with respect to how they usually do. However, this difference is not predictable. That's why many users' opinions split in the test, when dealing with *C.* mode rhythms. As a consequence, the comparison allowed by Table 7 can be understood with a mental experiment. Let a whatever listener imagine a spectrum of rhythms with continuously increasing rhythm complexity. If we asked them to sample ten rhythms on this spectrum, the resulting set of samples would be more similar to the *P.* class set of rhythms than to the *C.* class one. The listener internal representation of a rhythm innately includes the intensity information. Which is why the discussed measures correlate better with the average scores assigned by the users when they listened to rhythms which include intensity variations.

6. Conclusions

In this study the impact of the dynamic accents on a rhythm's perceived complexity was investigated. This investigation was realized by means of two experiments based on subjective listening tests. In particular, the first experiment focused on the analysis of monophonic rhythmic patterns and the second one referred to polyphonic rhythms. In both these scenarios, different versions of a common set of rhythmic patterns, realized by applying varying dynamics, were compared. The data collected from the two tests was used to derive meaningful considerations about the human rhythmic perception and how it is influenced by the presence of rhythmic accents. Also, six rhythm complexity measures from the literature were evaluated taking the tests' responses data as a reference.

The definitions of the fundamental concepts on which the study is based were given in Section 2. In Section 3, rhythm complexity measures of Toussaint, Longuet-Higgins & Lee, Pressing, Weighted Note to Beat Distance, Inter-Onset Intervals Information Entropy, Off-Beatness, and Grouped Voice were presented. Section 4 explains in detail the design of both the monophonic and polyphonic rhythms experiments. Finally, the results of these tests were discussed in Section 5.

Although the rhythm complexity represents an interesting topic for musicologists and computer scientists which operate in the Music Information Retrieval field, almost no studies were conducted on the dynamics influence on rhythm complexity perception. Also, the majority of research was only referred to monophonic patterns, so a properly conducted analysis on the more realistic circumstance of polyphonic pattern was missing, so was and some appropriate complexity measures for the same scenario. Thus, this study's main contribution can be found in the careful design of the two subjective listening tests on which this analysis is based on. Dealing with subjective perception poses several challenges, as it can be influenced by many factors even unintentionally, which is why the planning of a test is not a trivial task. The tests were largely participated, indeed 72 and 82 listeners respectively completed them. Furthermore, the experiment on polyphonic rhythms revealed generally high correlation between the complexity measures' scores and the results of the test. Given that we are talking about Person's correlation coefficient, a strong linear relationship was found among objective metrics and mean subjective opinions, meaning that these measures can be reliably adopted even for real music performances and MIR applications.

An additional contribution of this study was the extension of the work conducted by Mezza et al. [38]: their Grouped Voice procedure was applied to develop the polyphonic version of six rhythm complexity measures which have been extensively studied in literature but only refer to monophonic rhythmic patterns.

It is hard to summarize all the considerations that were derived from the results of the tests. In general, the context of analysis revealed to significantly impact the perception of the listeners. Indeed, in the first test rhythms with dynamic accents were perceived as more complex than the ones with constant intensity, while in the second test, the opposite happened. On this regard, it is worth to notice that the first test refers to a more abstracted conceptualization of what we usually conceive as rhythm. Also, rhythms with dynamics have gathered less conflicting opinions with respect to their constant intensity counterparts. This is because the memory and familiarity play an important role in the cognitive process, and people who generally listen contemporary music, are used to rhythms which involve dynamics modulations.

Just as there is a difference between listening to a pattern with dynamics and one without, dealing with monophonic or polyphonic rhythmic patterns, from a perception perspective, makes a huge difference. It is possible to conclude that a too general consideration about rhythm complexity perception is not feasible. It is a subjective experience, and involves many factors that should be considered. Each of these deserves a deep and dedicated analysis. Indeed, it would be interesting to consider them in future works. In this sense, this study has been a pioneering work, paving the way for many other questions.

The impact of dynamic accents was central in this thesis, but was addressed a bit generally, given that it is, nonetheless, the first piece of research to tackle it. Indeed, coming from a literature that doesn't address the issue at all, the main comparison here proposed was between the presence and the absence of dynamic accents. In a future work it would be interesting to focus more on this aspect and to compare cases with different amounts of intensity for the same accents, or to compare cases where the same amount of intensity is introduced in different metrical positions. In addition to the influence of onset intensity, many other aspects could be investigated to deepen the discussion about rhythm complexity. Indeed, typical terms that occur when discussing about complex rhythms are *polyrhythms*, *polymeters* and *tercet rhythms*: it would be interesting to introduce these aspects in the analysis, given that the current analysis of rhythmic complexity only limits to binary rhythms, and some of these are very diffused as drumming techniques.

References

[1] Randomvelo. MindThePressure.

[2] Rhythm, definition.

[3] Douglas Abrams. Rhythmic complexity in jazz: An information theory approach, 2023.

[4] Academic Accelerator. Ghost note.

[5] XLN Audio. Addictive drums 2.

[6] D. E. Berlyne. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, 31(4):553–553, 1973.

[7] Fleur L. Bouwer, John Ashley Burgoyne, Daan Odijk, Henkjan Honing, and Jessica A. Grahn. What makes a rhythm complex? the influence of musical training and accent type on beat perception. *PLoS ONE*, 13, 2018.

[8] Daniel Cameron and Jessica Grahn. *Perception of Rhythm*, pages 20–38. 09 2020.

[9] Peter Y. Chen and Paula M. Popovich. Correlation: Parametric and nonparametric measures. 2002.

[10] Vincent K.M. Cheung, Peter M.C. Harrison, Lars Meyer, Marcus T. Pearce, John-Dylan Haynes, and Stefan Koelsch. Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, 29(23):4084–4092.e4, 2019.

[11] Edward Coyle and Ilya Shmulevich. A system for machine recognition of music patterns. volume 6, pages 3597 – 3600 vol.6, 06 1998.

[12] Anne Danielsen. Musical rhythm in the age of digital reproduction. 2010.

[13] Isabelle Dautriche, Hugh Rabagliati, and Kenny Smith. Subjective confidence influences word learning in a cross-situational statistical learning task. *Journal of Memory and Language*, 121:104277, 2021.

[14] Remi de Fleurian, Tim Blackwell, Oded Ben-Tal, and Daniel Müllensiefen. Information-theoretic measures predict the human judgment of rhythm complexity. *Cognitive Science*, 41(3):800–813, 2017.

[15] Peter Essens. Structuring temporal sequences: Comparison of models and factors of complexity. *Perception & Psychophysics*, 57:519–532, 1995.

[16] W. Tecumseh Fitch and Andrew J. Rosenfeld. Perception and production of syncopated rhythms. *Music Perception: An Interdisciplinary Journal*, 25:43–58, 2007.

[17] Jacopo Foglietti. Rhythm complexity analysis for odd time signatures, 2016.

[18] Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to groove with inverse sequence transformations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2269–2279. PMLR, 09–15 Jun 2019.

[19] Google. Firebase.

[20] Esther Grabe. Durational variability in speech and the rhythm class hypothesis. 2005.

[21] Francisco Gómez, Andrew Melvin, David Rappaport, and G Toussaint. Mathematical measures of syncopation. *Proceedings of the BRIDGES: Mathematical Connections in Art, Music and Science*, 01 2005.

[22] Francisco Gómez, Eric Thul, and G Toussaint. An experimental comparison of formal measures of rhythmic syncopation. pages 101–104, 01 2007.

[23] Jan Hauke and Tomasz M. Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. 2011.

[24] David Huron. Musical aesthetics: Uncertainty and surprise enhance our enjoyment of music. *Current Biology*, 29(23):R1238–R1240, 2019.

[25] Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science and Technology*, 7:396–403, 2015.

[26] Don Rath Jr. Dynamics – music theory – level 2 – part 21b.

[27] M. Keith. *From Polychords to Pólya: Adventures in Musical Combinatorics.* Vinculum Press, 1991.

[28] J. Koetting. *Analysis and notation of west African drum ensemble music.* Selected Reports in Ethnomusicology, 1970.

[29] G.W. Cooper L. B. Meyer. *The Rhythmic Structure of Music.* The University of Chicago Press, 1960.

[30] Fred Lerdahl and Ray Jackendoff. *A generative theory of tonal music.* The MIT Press, Cambridge. MA, 1983.

[31] Daniel J. Levitin, Jessica A. Grahn, and Justin London. The psychology of music: Rhythm and movement. *Annual Review of Psychology*, 69(1):51–75, 2018. PMID: 29035690.

[32] Justin London. *Hearing in Time: Psychological Aspects of Musical Meter.* Oxford University Press, 05 2012.

[33] H. C. Longuet-Higgins and C. S. Lee. The rhythmic interpretation of monophonic music. *Music Perception: An Interdisciplinary Journal*, 1(4):424–441, 1984.

[34] Magenta. Groove midi dataset.

[35] J. Mcauley. *Tempo and Rhythm*, pages 165–199. 08 2010.

[36] Schulkind MD. Long-term memory for temporal structure: evidence form the identification of well-known and novel songs. *Mem Cognit*, pages 896–906, 1999.

[37] Well-A.D. Lorch Jr R.F. Meyers, J.L. *Research Design and Statistical Analysis: Third Edition (3rd ed.)*. Routledge, 2010.

[38] Alessandro Ilic Mezza, Massimiliano Zanoni, and Augusto Sarti. A latent rhythm complexity model for attribute-controlled drum pattern generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023.

[39] Randel Don Michael. *The Harvard Dictionary of Music. 4th ed.* Belknap Press of Harvard University Press, 2003.

[40] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175 – 220, 1998.

[41] Caroline Palmer and Carol Krumhansl. Mental representations for musical meter. *Journal of experimental psychology. Human perception and performance*, 16 4:728–41, 1990.

[42] Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, 1896.

[43] Karl Pearson. *VII. Note on regression and inheritance in the case of two parents.* Proceedings of the Royal Society of London, 1985.

[44] Encyclopedia Britannica Peter Crossley-Holland. Understanding rhythm in music.

[45] Dirk-Jan Povel. A theoretical framework for rhythm perception. *Psychological Research*, 45:315–337, 1984.

[46] Dirk-Jan Povel and Peter Essens. Perception of temporal patterns. *Music Perception*, 2:411–440, 1985.

[47] Jeff Pressing. *Transcribe: A comprehensive autotranscription program.* Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 1993.

[48] Jeff Pressing. Cognitive complexity and the structure of musical patterns. 1999.

[49] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4364–4373. PMLR, 10–15 Jul 2018.

[50] Thomas Schneider. Information theory primer with an appendix on logarithms pdf version. 07 2013.

[51] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. webmushra — a comprehensive framework for web-based listening tests. *Journal of open research software*, 6, 2018.

[52] Diemo Schwarz, Guillaume Lemaitre, Mitsuko Aramaki, and Richard Kronland-Martinet. Effects of test duration in subjective listening tests. In *International Conference on Mathematics and Computing*, 2016.

[53] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, 1948.

[54] Ofir Shany, Neomi Singer, Benjamin Paul Gold, Nori Jacoby, Ricardo Tarrasch, Talma Hendler, and Roni Granot. Surprise-related activation in the nucleus accumbens interacts with music-induced pleasantness. *Social Cognitive and Affective Neuroscience*, 14(4):459–470, 03 2019.

[55] Ilya Shmulevich and Dirk-Jan Povel. Measures of temporal pattern complexity. *Journal of New Music Research*, 29:61 – 69, 2000.

[56] Ilya Shmulevich, Olli Yli-Harja, Edward Coyle, Dirk-Jan Povel, and Kjell Lemström. Perceptual issues in music pattern recognition: Complexity of rhythm and key finding. *Computers and the Humanities*, 35:23–35, 02 2001.

[57] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[58] George Sioros and Carlos Guedes. Complexity driven recombination of midi loops. In *International Society for Music Information Retrieval Conference*, 2011.

[59] George Sioros, André Holzapfel, and Carlos Guedes. On measuring syncopation to drive an interactive music system. In *International Society for Music Information Retrieval Conference*, 2012.

[60] Leigh M. Smith and Henkjan Honing. Evaluating and extending computational models of rhythmic syncopation in music. *IEEE Transactions on Information Theory*, 2006.

[61] Steinberg. Cubase.

[62] Steinberg. Groove agent.

[63] Stephen M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.

[64] Andranick Tanguiane. Artificial perception and music recognition. In *Lecture Notes in Computer Science*, 1993.

[65] David Temperley. The origins of syncopation in american popular music. *Popular Music*, 40(1):18–41, 2021.

[66] Eric Thul. Measuring the complexity of musical rhythm. 2008.

[67] Eric Thul and Godfried T. Toussaint. Rhythm complexity measures: A comparison of mathematical models of human perception and performance. In *International Society for Music Information Retrieval Conference*, 2008.

[68] Godfried Toussaint. A mathematical analysis of african, brazilian, and cuban clave rhythms. In Reza Sarhangi, editor, *Bridges: Mathematical Connections in Art, Music, and Science*, pages 157–168, Southwestern College, Winfield, Kansas, 2002. Bridges Conference.

[69] Godfried Toussaint. The geometry of musical rhythm. In Jin Akiyama, Mikio Kano, and Xuehou Tan, editors, *Discrete and Computational Geometry*, pages 198–212, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[70] Godfried Toussaint. A comparison of rhythmic dissimilarity measures. 21:129–149, 01 2006.

[71] Godfried T. Toussaint. Classification and phylogenetic analysis of african ternary rhythm timelines. 2003.

[72] Godfried T. Toussaint and Konstantinos Trochidis. On measuring the complexity of musical rhythm. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 753–757, 2018.

[73] Paul Vitz. Information, run structure and binary pattern complexity. *Perception Psychophysics*, 3:275–280, 07 1968.

[74] Peter Vuust and Maria A. G. Witek. Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Frontiers in Psychology*, 5, 2014.

[75] E. W. Weisstein. Polygon diagonal. from MathWorld–a Wolfram web resource.

[76] Maria A. G. Witek, Eric Clarke, Mikkel Wallentin, Morten L. Kringelbach, and Peter Vuust. Syncopation, body-movement and pleasure in groove music. *PLoS ONE*, 9, 2014.

[77] Maria A. G. Witek, Jingyi Liu, John Kuubertzie, Appiah Poku Yankyera, Senyo Adzei, and Peter Vuust. A critical cross-cultural study of sensorimotor and groove responses to syncopation among ghanaian and american university students and staff. *Music Perception*, 2020.

[78] Sławomir K. Zieliński. On some biases encountered in modern listening tests. 2006.

A. Dynamic Toussaint's Off-Beatness

In this appendix, a modified version of the Off-Beatness complexity measure, introduced in Section 3.5, is presented. The reason for such a modification is that the Pearson Correlation Coefficient of this measure, reported in Table 7 in Section 5.2.2, resulted to be the worst among all the other ones. In particular, the PCC was significantly lower in the case of rhythms which included intensity modulations. Such a poor correlation performance risks to invalidate the usage of this measure in real drum performance scenarios.

The novel version is very similar to the original one, except that it takes into account the intensities of the pulses in the rhythmic pattern. Indeed, the analyzed pattern is no longer seen as a binary sequence, where, regardless of the dynamic component of the rhythm, a 0 is associated to a rest and a 1 is associated to an onset. Instead, onsets are associated with a value between 0 and 1, depending on their intensity. As a consequence, the complexity score no longer is the count of the off-beat onsets, but the sum of their normalized intensities.

Formally, let $\mathbf{r*}$ be a rhythmic pattern with dynamic modulations, represented with a similar notation to binary notation (Figure 1c), except that onsets are indicated by a value $0 < x \leq 1$. Let N be the number of pulses and k be the number of onsets in $\mathbf{r*}$.

1. Place the pattern on a circle, where each pulse is evenly distributed around the circumference, resulting in a geometric representation of the pattern, which is shown in Figures 1d, 6, in Sections 2 and 3.5.
2. Find all the integer values i, with $0 \leq i < N$, such that $\mathbf{r*}$ is evenly divided into i partitions of equal length. For each of the found values, inscribe a regular polygon with i vertices in the circle, superposing the vertices to the pulses starting from pulse 0.
3. Each pulse that is included in a polygon is marked as a *beat*, while the remaining pulses outside of all polygons are defined as *off-beats*.
4. The final score of Toussaint's Off-Beatness measure is simply the sum of the values of all the onsets that are off-beat.

As a check, the same correlation coefficient computed also for the other measures in Table 7 was calculated. The comparison between the original version of the measure and the one just described is presented in Table 8.

	Original Off-Beatness	**Dynamic Off-Beatness**
Constant	0.592801	0.592801
Performed	0.547831	**0.808119**

Table 8: Comparison between the Pearson Correlation Coefficients of the original and modified versions of Toussaint's Off-Beatness measure with the test results.

The proposed modification represents an extension of the original measure to the dynamic case, as it behaves in the same way when dealing with constant intensity rhythmic patterns. Furthermore, as the comparison underlines, considering the dynamics in the rhythmic pattern significantly improved the performance of this measure in a dynamic pattern case.

Abstract in lingua italiana

La percezione della complessità ritmica è stata oggetto di un'ampia analisi nelle ultime decadi, soprattutto per il suo interesse nel campo del Music Information Retrieval. Tuttavia, molti studi esistenti si concentrano esclusivamente sull'aspetto posizionale del ritmo, ignorando l'influenza di altri fattori importanti. Questa tesi esplora l'impatto degli accenti dinamici sulla percezione e sulla misurazione della complessità ritmica. Sono stati realizzati due listening tests, che indagano rispettivamente sulla percezione di pattern monofonici e polifonici. Ai partecipanti sono stati presentati ritmi con vari tipi di accenti - inclusi pattern con intensità costante e vere performance di batteristi - ed è stato chiesto loro di valutarne la complessità ritmica. Inoltre, sono state discusse e valutate sei metriche di complessità ritmica monofoniche, ciascuna delle quali è stata adattata anche per il caso polifonico. La bontà di queste metriche è stata determinata calcolando la correlazione tra i loro punteggi e quelli degli utenti che hanno completato i test, ed è stata così indagata la relazione tra il fenomeno soggettivo della percezione della complessità ritmica e una possibile descrizione quantitativa. I risultati offrono approfondimenti sul ruolo sfumato degli accenti dinamici nella modellazione della percezione ritmica e portano contributi significativi per la comprensione di questo tema sia in contesti monofonici che polifonici.

Parole chiave: Complessità ritmica, dinamica, accenti , percezione, Music Information Retrieval