Real-time Qualification of Percussive Sounds Based on Correspondences Between Schaeffer's *Solfège* and Low-level Audio Descriptors

Sérgio Freire[1], **José Henrique Padovani**[1], **Caio Campos**[1]

[1]LaPIS – School of Music/Federal University of Minas Gerais
Av. Antônio Carlos, 6627 – 31270-901 Belo Horizonte, MG

sfreire@musica.ufmg.br, jhp@ufmg.br, costacaiocampos@gmail.com

Abstract. *Pierre Schaeffer's typomorphology (1966) proposes seven criteria of musical perception for the qualification of sound objects, which form the basis of his musical theory. This* Solfège *fits well into contexts where pitch is not the dominant dimension. Relying on similarities between the practice of reduced listening and the utilization of low-level audio descriptors, we present the first version of a real-time setup in which these descriptors are applied to qualify percussive sounds. The paper describes the tools and strategies used for addressing different criteria: envelope followers with different window sizes and filtering; detection of transients and amplitude modulations; extraction and counting of spectral components; estimation of intrinsic dissonance and spectral distribution; among others. The extracted data is subjected to simple statistical analysis, producing scalar values associated with each segmented object. Finally, we present a variety of examples.*

1. Introduction

The association of audio descriptors with the Schaefferian theory is not rare in recent literature. One can find it in analytical contexts [1], in the association of acoustic data and subjective labels [2, 3, 4], and automatic sound indexing [5]. Solomon [6], although not using audio descriptors and not explicitly using this theory, presents a conceptual approach for the classification and association of percussive sounds similar to the one intended here.

In the last years, we developed a real-time setup in Max[1] to qualify percussive sounds using the *Solfège*'s perceptual criteria, aiming at its application in interactive situations. Our approach uses time and frequency domain representations to tackle the diversity of criteria and sounds. In this paper, we present its first version, tested with diverse sound types. The intention was to build a general framework upon which we could later refine some specific tools. The background assumption is that the practice of reduced listening has similarities with the application of low-level audio descriptors.

The text organizes as follows. First, we present the criteria for musical perception that constitute the Schaefferian Solfège, followed by the segmentation and preprocessing procedures used in the setup. The description of the implemented tools comes next. Then we discuss the intended correspondences between the criteria and the descriptors. A section with examples and their discussion completes the presentation of the setup.

[1]www.cycling74.com

2. Schaeffer's Criteria

In the Treatise on Musical Objects (TOM])[7, 8], Pierre Schaeffer introduced a theoretical framework and a practical methodology for classifying and manipulating electroacoustic sounds, what he would come to call *Musical Research Program*. By inverting the traditional notion of *solfège* – associated with the practice of singing intervals and scalar excerpts through the solmization of musical notes – Schaeffer proposed a "generalized *solfège*", based on the "art of better hearing".

To furnish means to his proposal of describing perceptual aspects of sound phenomena — despite any referential or causal events such as physical or instrumental factors that may have generated them — Schaeffer proposes, borrowing the Husserlian concept of *epoché*, a *reduced listening*. By this term, the author refers to a conscious attempt to scrutinize the attributes of sound objects by taking into account the perceptual dimensions of pitches, durations, and intensities. To enable a more detailed description of sound objects in these three dimensions, Schaeffer proposes seven typomorphological criteria: *mass*, *dynamic*, *harmonic timbre*, *melodic profile*, *mass profile*, *grain*, and *allure*.

The seven typomorphological criteria proposed by Schaeffer allow him to build a methodological framework to typify, qualify, and evaluate audible characteristics of sound objects with the help of categories like *types*, *classes*, *genres*, and *species*. This comprehensive method is compiled in the TARSOM[2], a Summary Diagram that outlines the analytical concepts developed by the author. Thus, while the first three columns of this table – *types*, *classes*, and *genres* – relate the criteria (7 rows), respectively, to *typology*, *morphology*, and *characterology*, the following six seek to draw connections between these seven morphological criteria and the perceptual dimensions of *pitches*, *intensities*, and *durations*, employing the *site/calibre* binomial (two columns per perceptive field).

As Michel Chion clarifies (section 25, on the perceptual field, in [9]), each criterion has a more evident relationship with one or more of these perceptual fields. In

[2]*Tableau Récapitulatif du Solfège des Objets Musicaux*, presented at the pp. 584-587 of the original edition[7] and on pp. 464-467 of the English version[8] of the TOM.

Table 1, we present the summary description of the morphological criteria – provided by Chion in section 88 of the same work, but here described with fragments from the English translation of TOM [8].

3. Pre-processing and Segmentation

Defining the adequate sound portion to be analyzed in a real-time setup is crucial for working with reliable data. In our program, the inputs are audio streams delivered by microphones, pickups, or mixers, presenting diverse background noise and dynamic ranges. It is not difficult to control background noise by observing the input amplitude during a "silent" period. On the other hand, the dynamic range poses some challenges, like the leakage from other sources (including sounds coming from loudspeakers). We decided to set a maximal dynamic range of 40 dB, which can cover the range of the majority of instruments [10] but may cut off earlier some long resonances. In the development phase, we decided to use a set of pre-recorded sounds, which offers not only variety but also repeatability, two relevant factors for building and refining tools. The sound selection depicted in Table 8 was based on Schaefferian typology. These sounds become real-time inputs to the setup[3], running with a sampling frequency of 48 KHz.

Portions to be analyzed are segmented between onset and offset points. A new segmentation clue may occur before the offset; in these cases, this clue determines the offset of the last event and the onset of the present one, which is qualified as "slurred". The detection of onsets and offsets relies on the comparison of a dynamic envelope with two thresholds, 6 dB and 3 dB, respectively, above the background noise level. This envelope is an RMS curve estimated with a short window size – 256 points – and a hop size of 64. We will refer to this curve as *rms256:4*, and use a similar notation for other envelopes. The implementation uses `[gen~]` routines, which employ native audio signal processing and offer more efficiency and precision. A low-pass filter (a single one-pole filter, with a -6 dB per octave attenuation) smoothes these envelopes, and we use different cutoff frequencies depending on the purpose. The detection of onsets and offsets employs a cutoff frequency of 4 Hz. The same setting applies to the detection of the end of an attack (in this case, the input signal may pass through a filter before the estimation of the envelope). Routines dedicated to attack profiles and iterative grains use the same envelope with a cutoff frequency of 30 Hz. The attack profiles are expressed by a control-rate version of this curve. We also use an *rms2048:4* curve, low-pass filtered at 10Hz, for the global dynamic envelope and the detection of allures. Other processes use spectral peaks values estimated by the `[sigmund~]` object [11], using the same window and hop sizes.

[3]The sound files used in this study are available in the following link: `https://bit.ly/sbcm2021_soundExamples`

4. Implemented Tools

4.1. Time Domain Low-level Descriptors

Duration is a simple attribute, whose value is the time interval (in ms) between onset and offset. The estimation of onsets and offsets was described in Section 3. The dynamic profile is the portion of the *rms2048:4* curve comprised between onset and offset. The dynamic level is a simple feature, represented by the mean value and standard deviation of the same curve. Even if we try to improve the correspondence between dynamic levels and perceptual attributes by observing the total duration and spectral region, it is necessary to remember Schaeffer's words: "For sounds with unremarkable mass and profile this dynamic field is *almost unknown*." (TOM[8], p. 432)

Schaeffer has stressed the importance of the attack for some sound typologies and the perception of durations. In the case of percussive sounds, this is a primordial characteristic. Therefore, we prefer to analyze the entire attack profile, which may surpass 300 ms, instead of stopping at the point usually called the "end of attack" [12]. For the same reason, this point will be called the "attack first plateau". We define the "first plateau" as the moment when the derivative of the low-pass filtered audio-rate *rms256:4* curve from the (possibly filtered) input audio stream comes near (or cross) to zero, just after having surpassed a predetermined positive threshold. We call the ratio between the amplitude difference and the time interval that occur between the onset and the first plateau as **FPSlope**. Since we prefer to consider multiple fast strokes (such as flans, drags, ricochets) as belonging to the same profile, a value of 200 ms stays as the reattack threshold. Depending on the settings (filtering and thresholds), the algorithm may not detect soft attacks[4]. On the other hand, some iterative sustainments are considered single allured sound objects, even when the distance between peaks exceeds the given threshold. Figure 1 depicts the attack profiles of eight percussive sounds with varied characteristics.

4.1.1. *Allures*

As practically any sound event with mechanical origins presents small fluctuations in its envelope, it is necessary to define a minimum threshold for the occurrence of a noticeable allure. In our setup, this threshold is a value expressed in dB (3 dB by default). The detection of allures and related features uses a control-rate version of the *rms2048:4* curve of the input stream, expressed in dBFS. The sign change of the derivative of the signal (crossing a tiny region around zero) is indicative of a possible peak or trough, which may be validated if the difference between the present peak and the last trough (and vice-versa) exceeds the chosen threshold. The estimated descriptors are (1) number of occurrences; (2) mean value and standard deviation of the (a) difference of intensity between peaks and troughs; (b) time interval between successive peaks;

[4]These parameters (reattack time and sharpness) help to redefine the fluids limits between the context ("whether the criteria are artificially put into a structure...") and the contexture ("...or naturally form a structure") of percussive sounds in diverse situations. (TOM[8], p. 402)

Table 1: Schaeffer's descriptions of typomorphological criteria and the related perceptual fields

criterion	description	fields
mass	"...quality through which sound installs itself (in a somewhat a priori fashion) in the pitch field." (p. 412)	pitches
dynamic	"...the variation in intensity of this sound in the course of its duration" (p. 33); "...its *energetic development*" [*évolution énergétique*] (p. 174)	intensities durations
harmonic timbre	"...more or less diffuse halo and more generally the secondary qualities that seem to be associated with mass and enable us to describe it." (p. 412)	pitches
melodic profile	"Neumes, although intended to represent variations in a specific source (the voice), can provide us with a model. (...) Type of sounds deliberately varied in tessitura" (p. 458)	pitches intensities durations
mass profile	"The *mass profile* is made up of all the (perceived) intensities of the various components of the spectrum of a sound." (p. 433)	pitches intensities durations
grain	"...a microstructure, generally due to sustainment from a bow, a reed, or even a drum roll. This property of *sound matter* reminds us of the *grain* of a textile or a mineral. (...) We find ourselves in a zone where two sensations from the same phenomenon [bassoon reed] merge: the perception of pitch from the beats, and the perception of beats from differentiation of the impacts" (p. 437)	pitches intensities durations
allure	"...the more or less regular oscillations that are its hallmark also cause variations in pitch (vibrato in stringed instruments, singers, etc.) and harmonic timbre. We could say that allure is made up of many factors (...), the most important of which are associated with the dynamic and pitch of sounds." (p. 438)	pitches intensities durations

(c) proportion between peak/trough and trough/peak intervals (symmetry); and (d) maximal value of the derivative in each inflection (spikiness). For a rough estimation of the distribution of peaks through the duration of the entire sound, their temporal centroid and spread are also calculated (in a time series filled with zeros, we insert the value 1 for each detected peak).

4.1.2. Grains

We use two different algorithms for the estimation of the presence and quality of grains in the audio stream, one linked to iteration, the other to resonance and friction types. These last two types are treated jointly under the term tiny grains. For iterative grains, we use the derivative of an *rms256:4* audio-rate curve low-pass filtered at 30 Hz. This signal goes through a Schmitt trigger (with thresholds -0.01 and 0.01). When its value goes below the lower limit, there is an indication of a possible grain, which is confirmed if the time interval between two occurrences has a value below 75 ms. The grain amplitude correlates with the difference between the peaks and troughs of the RMS curve in the same time interval. The estimated descriptors are (1) number of iterative grains; (2) mean value and standard deviation of (a) grain amplitudes; (b) time interval between successive peaks.

The estimation of tiny grains is done on the proper audio stream, or better, on its derivative. We assume that a granular signal will change directions more often than a smooth one. Although it may be objected that signals with higher frequencies will also bend more often (and with a larger difference between samples, due to the sampling resolution), our experience has showed that in the realm of percussive sounds the granular influence is stronger than the register[5]. An alternative would be the use of the spectral modeling synthesis (SMS), an algorithm proposed by Serra and Smith [13] to separate between deterministic and residual parts of a signal, using a frequency domain representation. However, as it presupposes the existence of a fundamental frequency, it is not adequately apply to most percussive sounds.

4.2. Frequency Domain Low-level Descriptors

This set of descriptors employs the spectral peaks estimation algorithm used by the `[sigmund~]` object, with the following parameters: analysis size of 2048 points; hop size of 512 points; 20 peaks; outputs (peaks, envelope, and pitch). These outputs are also controlled by the onset and offset descriptors described above and have a refresh rate of 10.67 ms with a sampling frequency of 48 kHz. For each frame analysis, we calculate the following descriptors.

Pct50 and **pct80** (number of peaks for energy percentiles 50% and 80%). According to Parseval's theorem, the energy of a time signal equals the sum of the energy of the absolute values of its frequency components. Applying this principle to `[sigmund~]` outputs (peaks and

[5] A few examples: sine waves with frequency of 100 Hz (2 grains), 1000 Hz (21 grains), 10 kHz (213 grains), pink noise (ca. 300 grains) and white noise (ca. 350 grains).

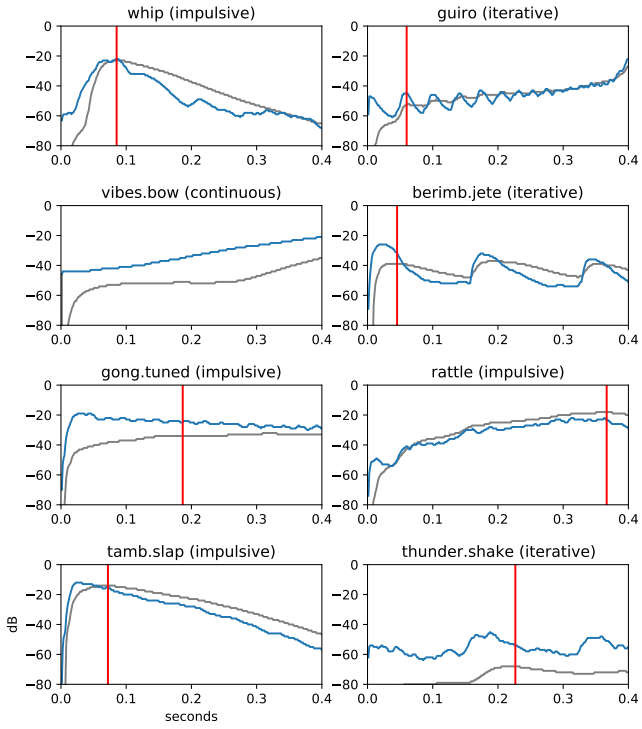

Figure 1: Attack profiles of eight sounds.The blue and gray curves derive from *rms256:4* estimations; the latter is calculated from a filtered audio. The red line indicates the first plateau. The gender of the attack profiles are indicated between parentheses.

envelope), it is possible to estimate the number of peaks needed to obtain the 50% (-3 dB) and the 80% (-1 dB) energy percentiles. The descriptor outputs two curves with the estimated number of peaks. For sounds with a more continuous spectral distribution, 20 peaks may not reach the chosen percentiles. In such cases, the output will be 20 peaks, and further information is obtained through the next descriptor. **20P/total** (percentage of sound energy represented by up to 20 peaks) estimates the amount of energy represented by the set of peaks calculated for each analysis frame and ranges between 0 and 1. The descriptor, along with the last one, helps differentiate sounds between the classes tonic and node. **MPP** (most prominent peak) is a very simple descriptor, represented by the curve formed by the values (expressed in Midicents) of the peaks with the largest amplitude in each analysis frame.

Estimated fundamental frequency. The object `[sigmund~]` outputs a value in Midicents for frames considered to bear a fundamental frequency and the value -1500 for unpitched frames. Our descriptor outputs a scalar (ratio of pitched to total frames– **unpitched/total**), a curve with all numbers, in which -1500 is substituted by 1, another curve with only the pitched values, and a third with 1 for pitched, and 0 for unpitched frames .

Intrinsic dissonance. The estimation of the intrinsic dissonance uses an implementation of the algorithm developed by Sethares [14], using the frequency and amplitude values delivered by the sigmund analysis. As he prescribes the use of SPL pressure values, we used 0.00001 for the minimal audible reference.

SC (spectral centroid). We use a `[gen~]` routine delivered with the Max program since its version 6 for the estimation of the spectral centroid. Instead of using a nominal value in Hz, we use values in Midicents, which define a scale ranging from 15.5 to 155 in the audible range.

Δ peaks (interval between the lowest and highest peaks). For each frame, we calculate the difference between the highest and the lowest estimated peaks, which is also expressed in Midicents.

Spectral region. The starting point for the definition of a region is a rough division of the audible spectrum in three ranges. The first three octaves (20–160 Hz) define the low range, the four intermediate octaves (160–2,560 Hz) the medium range, and the last three octaves (2,560–20,000 Hz) the high range. We estimate the energy carried by the peaks in each analysis frame for each of these ranges. If none of these ranges hold 40% or more of the total energy, the sound frame is classified as wideband, labeled as (7). Otherwise, any range with more than 40% of the total energy contributes to qualifying one of the six spectral combinations: (1) Low, (2) Low/Medium, (3) Medium, (4) Low/High, (5) Medium/High, (6) High.

The flowchart in Figure 2 depicts the main processes involved in the estimation of all low-level descriptors. The time series generated by most descriptors are subjected to simple statistical analysis just after the offset. We adapted the algorithms given in [12], and have chosen the following scalar descriptors: mean value and standard deviation; temporal centroid and spread (normalized by the total duration); skewness; kurtosis; crest; flatness. The correlations with the Schaefferian perceptual attributes rely on these values.

5. Intended Correspondences

As stated in the Introduction, our purpose is to find correlations between low-level descriptors and the criteria of musical perception defined by Schaeffer. In this study, we reduced the seven criteria to six, rearranging them as follows. Since in the realm of percussion (and of everyday sounds) tonic sounds are not the rule, we prefer to unify mass and harmonic timbre under one single category, relying on an observation made by Schaeffer, “considering them rather as connecting vessels, with the exception of certain specific examples...” (TOM, p. 412) Melodic and mass profiles are joined by the same reasons. Although the same set of descriptors supports the qualification of sounds under the four mentioned criteria, we prefer to put the profiles in a dedicated category since they use different statistical values. On the other hand, due to the great variety of attack types, we chose to treat them not as genres belonging to the criterion *dynamic* but as a separated criterion.

Table 2 depicts the intended correlations between Schaefferian criteria (and their attributes expressed in types, classes, genres, or species) and the selected

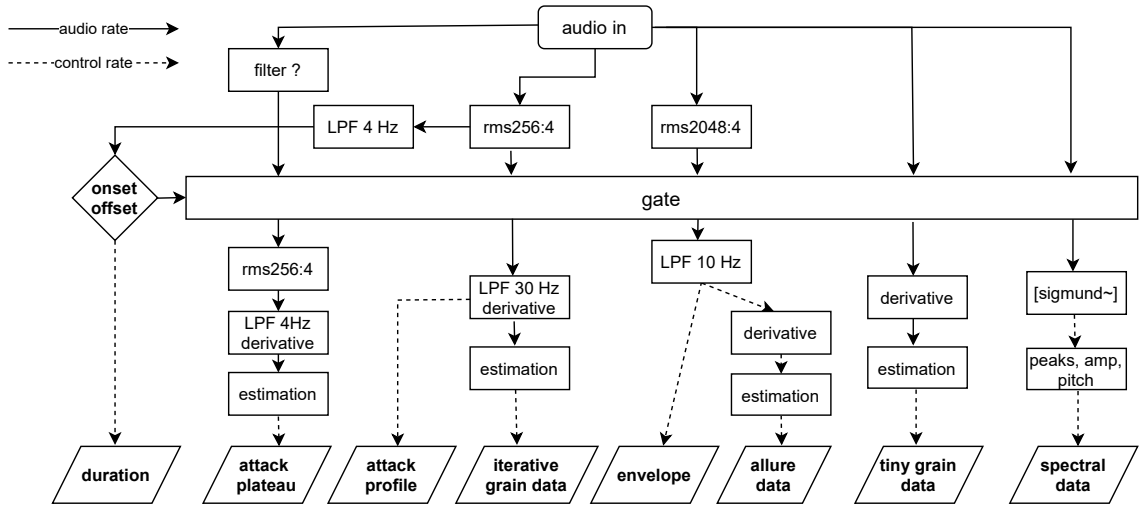

Figure 2: Flowchart showing the main routes and procedures for the estimation of low-level audio descriptors.

Table 2: Intended correspondences between Schaeffer criteria and low-level audio descriptors.

Criterion	Attributes	Audio Descriptors
dynamic	duration: short, formed, long dynamic level: pp to ff dynamic forms: shock, resonance, profiles (5 classes), flat, nil	onset–offset rms curve statistics skewness of spectral centroid spectral region, attack genre
attack profile	genre: abrupt, solid, soft, gentle, stressed, nil	attack profile statistics iterative grain / allure data
mass / harmonic timbre	class: tonic, channeled, nodal group, node region: low, medium, high genre/species: full/hollow/narrow, rich/poor	spectral peaks data / statistics spectral centroid
melodic / mass profiles	density of information: weak, medium, strong type: *fluctuation*, *evolution*, *modulation*[6]	spectral peaks data / statistics spectral centroid allure data
grain	type: iterative, tiny (friction or resonant) density: rough, matt, smooth	iterative / tiny grain statistics
allure	intensity: weak, medium, strong genre: regular, progressive, irregular, etc.	allure data

low-level descriptors. Note that a few estimated attributes function as descriptors for other categories, like the spectral regions for the dynamic levels or allures for the melodic/mass profiles. It is important to note that any sound object enacts at least three perceptual criteria: mass, dynamic (including the attack), and harmonic timbre. The other criteria may be relevant to characterize specific sounds, and their presence may affect the others. For example, allures may affect the melodic profile, grains may affect the mass and the mass profile, and percussion-resonance types condition the harmonic profile. A comprehensive discussion of the use of the profile concept by Schaeffer (and its implications on our work) is beyond the scope of this paper.

[6]Schaeffer proposes a "typology of variations". *Fluctuation* is a "variation that is felt only as an imperfection in a desired stability". *Evolution* stands for a progressive variation. *Modulation* is understood as a variation with "a development in stages — already sketching out a scalar structure." (TOM[8], p. 453).

6. Examples and Discussion

For every input sound, our program generates real-time curves (or markers) for all descriptors and calculates the scalar values described in section 4. These results are relatively numerous and probably present some degree of redundancy, not yet analyzed. For the sake of clarity, we discuss these results separated by criterion (or sub-criterion), using subsets of sounds and descriptors.

Our first example depicts the quantitative results (Table 3) for the different attack profiles illustrated in Figure 1. These shapes depend on the facture of the sound objects (single stroke, iteration, continuous excitation) and dynamic levels. Schaeffer defines seven genres of attack: abrupt, steep, soft, flat, gentle, sforzando, and nil. The whip sound has a short duration, low values for temporal centroid and spread, a positive skewness, and a high crest. All this data corresponds to an abrupt genre. The slap on a tambourine presents the steep profile. It has a short reso-

nance following the attack. Its temporal centroid is similar to the whip, but with a larger spread and less pronounced values for skewness and crest. A reinforcement of the resonance follows a soft profile, such as with the tuned gong. Here, a longer duration, a slightly positive skewness, and a high value for flatness contribute to the characterization of the genre. We perceive the rattle profile as gentle, given the absence of an initial shock. Its sound production combines iterative and continuous energies, which correlates with a negative skewness, a small crest, and medium flatness.

Two sounds have a clear iterative or granular profile, the guiro rub, and the berimbau jeté. This last has a profile between steep and soft, due to the reinforcement of the resonance by repeated strokes. The guiro profile is sforzando, demonstrated by a high temporal centroid, a negative skewness, a high crest. Two sounds bear a nil profile, the bowed vibes, and the thunder shake. This fact is reflected by the high temporal centroid, negative skewness and small crest. There was no estimation of the first plateau during the first 400 ms for the bowed sound. Long sounds will rely less on their attack profile for their qualification. We also believe that the iterative/granular character should be a second-order qualifier for the attack profiles. A similar approach can be made to dynamic classes, analyzing the entire envelope.

Table 4 depicts *allure* data for six sounds, each of them with a different excitation pattern. As already exposed, we have chosen to interpret as allures (and not as new attacks) iterative sustainments with clearly differentiable impulses, as in the cases of the berimbau and sleigh bells. Time intervals with small standard deviation values are related to ordered or regular instances (like the berimbau), while the opposite points to higher irregularity (rainstick). The amplitudes indicate the depth of variation; for example, the tuned gong resonant allures are much softer than the iterative allures of the sleigh bells. Symmetry indicates the regularity of transitions between peaks and troughs (and vice-versa), and spikiness values point to the suddenness of variation.

The estimation of *grains* for eight sounds is depicted in Table 5. As expected, sounds with iterative sustainment present a large number of these grains. The exception is the bass drum, whose resonant grains, due to the slow rate, also fit into this category. The diverse values for size and duration also helps differentiating between iterative grains. Our tiny grain descriptor is dedicated to resonant and friction grains, and their mixture. A closer inspection of the number of tiny grains and their standard deviation furnish information about their temporal behavior. The large standard deviation, along with a high value of temporal centroid in the bass drum, indicates an increase of background noise at the end of the resonance. The bowed cymbal also displays a considerable standard deviation, but a temporal centroid below 0.5; in this case, the granular characteristic is more present at the beginning. Friction grains tend to have higher bend values than resonant ones, as displayed by the ratchet and cymbal data. On the other hand, the resonant characteristic seems to prevail in the bowed cymbal and vibraphone, despite the excitation mode.

Seven classes of *mass* are defined by Schaeffer: pure sound, tonic, tonic group, channeled, nodal group, node, white noise. Pure and tonic sounds bear a clear pitch, and tonic groups indicate chord-like sonorities. Node is a filtered noise (or a dense spectral region), while nodal group is a set of nodes. In the middle of this classification we encounter the ambiguous channeled sound, sharing properties for pitched and unpitched classes. The most common classes in the percussive realm are the tonic, channeled, node, and nodal group, although this latter occurs more in combination than in a single object[7]. As stated above, the *harmonic timbre* is a complementary characteristic of the spectral perception, and Schaeffer points to oppositions like hollow/full, rich/poor, and bright/matt. Data related to mass and harmonic timbre is displayed in Tables 6 and 7. The ratio between the unpitched and total frames estimated for the entire sound object points to its tonic character. Low values indicate tonic sounds, like the bass drum, whistle, and the friction of a pandeiro's skin. Higher values of this descriptor, combined with small values of percentiles 50% and 80%, can qualify channeled sounds, like the snare drum (with no snare) and the tuned gong. A high unpitched/total ratio, along with a low value for the energy carried by the 20 first spectral peaks, characterizes nodes, as in the cases of the ratchet and rattle. Spectral centroid and region values are self-explaining.

The intrinsic dissonance values are more ambiguous. Although they may help differentiating between tonic and non-tonic sounds, this is not a univocal association since inharmonic spectral peaks only increment this value if they are close enough in frequency (see [14]). We try to approximate the perceptive attributes *full/hollow/narrow* with Δ peaks, pct50, pct80, and region, and the opposition *rich/poor* with the values estimated for pct80 and 20P/total. For instance, two sounds classified in the medium region may be contrasted through the attributes hollow (snare drum) and narrow (rattle).

High standard deviation values may indicate the presence of audible profiles. The analysis of profiles deserves a dedicated inquiry for the following reasons: (1) the non-univocal use of the concept in the TOM; (2) their variety (dynamic, melodic, mass, harmonic, allures) and interdependency; (3) the choice of the best fitting parameters for each situation — which may require non-realtime dimensionality reduction techniques (PCA, NMF, etc).

7. Final remarks

After presenting a general overview of the setup and some examples, our next steps are directed to a more comprehensive approach of sound profiles, to its implementation in PureData, to the work with musicians in specific instrumental configurations (when the pandemic permits), to the

[7] It is advisable to increase the number of spectral regions for dealing with nodal groups.

Table 3: Attack parameters for eight selected percussive sounds (the same from Figure 1), plus total duration and dynamic level.

sound	FPSlope (dB/ms)	temp. centroid	temp. spread	skewness	kurtosis	crest	flatness	dur (ms)	DL (dB)
whip	0.42	0.24	0.26	1.25	2.87	6.08	0.30	396	-46
tamb.slap	0.85	0.22	0.74	0.31	0.25	4.07	0.42	461	-33.5
vibes.bow	-	0.71	0.68	-0.35	0.40	2.84	0.70	2177	-26.5
guiro	0.63	0.72	0.39	-0.66	1.36	11.09	0.67	631	-42
rattle	0.13	0.67	0.67	-0.21	0.28	2.18	0.67	1151	-42
berimb.jete	0.81	0.35	0.53	0.33	0.66	4.77	0.60	2019	-55.6
gong.tuned	0.24	0.42	1.13	0.08	0.11	1.96	0.92	9219	-42.6
thunder.shake	0.18	0.53	0.20	-0.19	3.82	2.80	0.86	15541	-33

Table 4: Allure values for six selected percussive sounds.

sound	dur (ms)	number	amp (dB)	Δ t (ms)	symmetry	spikiness
berimb.jete	2019	11	5.1 ± 0.8	162 ± 32	1.99 ± 0.94	4.93 ± 3.2
berimb.vib	4687	12	7.3 ± 2.4	349 ± 68	0.99 ± 0.48	1.91 ± 0.79
gong.tuned	9219	6	4.2 ± 1.2	1476 ± 485	1.05 ± 0.56	0.26 ± 11
sleighbells	10295	30	17.1 ± 4.6	333 ± 55	1.47 ± 1	5.34 ± 3.3
thunder.shake	15541	57	7.2 ± 3.8	261 ± 95	1.46 ± 1.2	2.97 ± 2.2
rainstick	17295	27	5.6 ± 3.1	647 ± 448	1.23 ± 1.1	1.77 ± 1.65

Table 5: Grain values for 10 selected percussive sounds.

sound	dur (ms)	iterative grains			tiny grains		
		number	size (dB)	dur (ms)	number	TC	bend (dB)
guiro	631	14	5 ± 3	22.5 ± 13	140 ± 25	0.49	-60.5
ratchet	753	12	15 ± 5	54.5 ± 8	173 ± 18	0.49	-31.0
rattle	1103	2	4.8	73.3	129 ± 18	0.48	-59
pand.rim.frict	1365	3	5.2 ± 1.5	27	218 ± 17	0.5	-29
whistle	1463	30	4.6 ± 2.7	37.5 ± 12	56 ± 11	0.48	-58
bassdrum	3671	96	3.4 ± 1.4	37.4 ± 17	190 ± 122	0.67	-86
cymbal.bow	4565	3	2.7 ± 0.4	34.7 ± 8	65 ± 34	0.44	-76
cymbal	4963	-	-	-	116 ± 48	0.57	-71
tamb.tremolo	7884	59	5.6 ± 1.8	49.5 ± 17	171 ± 11	0.49	-41.5
rainstick	17245	110	6.8 ± 3.3	44.3 ± 16.5	191 ± 32	0.49	-54

Table 6: Mass and harmonic timbre parameters (1) for 10 selected percussive sounds.

sound	dur	pct50	pct80	20P/total (ratio)	unpitched/total (ratio)
tabla.gliss	227	1.3 ± 0.5	6.6 ± 8.5	0.85 ± 0.12	0.24
sdrum.nosnare	560	1.1 ± 0.4	2.4 ± 3.7	0.96 ± 0.1	0.32
ratchet	753	19.8 ± 0.6	20 ± 0	0.42 ± 0.1	0.86
rattle	1103	8.4 ± 2.2	19.7 ± 1.3	0.71 ± 0.1	1.0
pand.skin.frict	1915	1 ± 0.2	1.5 ± 2.1	0.98 ± 0	0.07
slidewhistle	641	6.3 ± 7.7	14.4 ± 8.4	0.66 ± 0.5	0.13
chin.opera.gong	1911	2.5 ± 1.9	8.5 ± 5.9	0.9 ± 0.1	0.75
berimb.jete	2019	4.6 ± 5.4	10 ± 6.7	0.82 ± 0.2	0.94
bassdrum	3671	1.1 ± 1	1.2 ± 1.5	0.98 ± 0.1	0.04
gong.tuned	9219	1.3 ± 0.8	2.2 ± 1.6	0.98 ± 0.05	0.42

Table 7: Mass and harmonic timbre parameters (2) for 10 selected percussive sounds.

sound	diss	MPP (mc)	Δ peaks (mc)	SC (mc)	region
tabla.gliss	37 ± 22.5	46.2 ± 10.4	69.9 ± 19.3	54 ± 7.6	1.8 ± 1
sdrum.nosnare	45.3 ± 21.9	61.2 ± 7.7	59.4 ± 15.6	65.5 ± 13.3	3
ratchet	122.5 ± 30.9	97.6 ± 9.5	30.7 ± 4.5	110.2 ± 3.6	6.5 ± 1.3
rattle	138.8 ± 42.6	95.5 ± 1.9	17.5 ± 9.1	101.2 ± 2.7	3
pand.skin.frict	43.2 ± 30.5	48 ± 1.6	70.2 ± 6.7	51.7 ± 6.7	1
slidewhistle	141.1 ± 66.8	89.8 ± 14.8	79.8 ± 20.5	95.3 ± 8.1	4.4 ± 1.8
chin.opera.gong	69.1 ± 15.1	73.2 ± 5.7	40.9 ± 12.6	82.8 ± 6.8	3
berimb.jete	24.8 ± 18.6	65.5 ± 13.9	80.8 ± 13.2	86.5 ± 8.5	3.5 ± 1.3
bassdrum	22.6 ± 23	27.8 ± 3.8	99.3 ± 21.6	30.5 ± 8.7	1 ± 0.3
gong.tuned	24.2 ± 17.8	61 ± 2.9	76.4 ± 31	68.2 ± 7.6	3 ± 0.2

Table 8: Selected sounds.

sound	description
tabla.gliss	single tabla stroke with glissando
whip	single whip attack
tamb.slap	single tambourine hand slap
sdrum.nosnare	single snare drum stroke, without snare
chin.opera.gong	single chinese opera gong stroke
bassdrum	single bassdrum stroke
cymbal	single cymbal stroke
gong.tuned	single tuned gong stroke
guiro	single directional guiro rub
ratchet	single ratchet swing
rattle	single directional rattle shake
tamb.tremolo	tambourine tremolo
berimb.jete	berimabau jete, multiple strokes
berimb.vib	single berimbau stroke, with vibrato
pand.rim.frict	pandeiro tremolo-like rim friction
sleighbells	multiple sleighbells shakes
thunder.shake	multiple thunder sheet shakes
rainstick	rainstick tip
slidewhistle	slide whistle blow with glissando
pand.skin.frict	single pandeiro skin friction
vibes.bow	single vibraphone key bow
cymbal.bow	single cymbal bow

use of machine learning with larger sets of sounds, and creative applications.

We would like to finish with a quotation from Schaeffer, for us a standing source of caution and stimulus: "It is perhaps disconcerting to see us, after so many warnings, recommending the use of the bathygraph and the Sonagraph to describe a piece of music." (TOM [8], pp. 556–567)

Aknowledgments

This work is supported by the Brazilian research agency CNPq (National Council for Scientific and Technological Development).

References

[1] Andrea Valle. Schaeffer Reconsidered: a Typological Space and its Analytical Applications. *Analitica*, 8(1):1–15, 2015.

[2] Rolf Inge Godøy. Perceiving Sound Objects in the Musique Concrète. *Frontiers in Psychology*, 12, 2021.

[3] G. Bernardes, M. Davies, and C. Guedes. A Pure Data Spectro-Morphological Analysis Toolkit for Sound-Based Composition. pages 31–38, Aveiro, 2015. Proceedings of the eaw2015.

[4] Julien Ricard. *Towards Computational Morphological Description of Sound*. Doctorate, Universitat Pompeu Fabra, Barcelona, September 2004.

[5] Geoffroy Peeters and Emmanuel Deruty. Sound Indexing Using Morphological Description. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):675–687, March 2010.

[6] Samuel Z. Solomon. *How to Write for Percussion: a Comprehensive Guide to Percussion Composition*. Oxford University Press, New York, 2nd edition, 2016.

[7] Pierre Schaeffer. *Traité des Objets Musicaux*. Éditions du Seuil, Paris, 1966.

[8] Pierre Schaeffer. *Treatise on Musical Objects: Essays Across Disciplines*. University of California Press, Oakland, California, 2017.

[9] Michel Chion. *Guide des Objets Sonores*. Buchet/Chastel, Paris, 1983.

[10] W. Gieseler, L. Lombardi, and R. Weyer. *Instrumentation in der Musik des 20. Jahrhunderts: Akustik, Instrumente, Zusammenwirken*. Moeck Verlag, Celle, 1985.

[11] Miller S Puckette, Theodore Apel, and David D Zicarelli. Real-time Audio Analysis Tools for Pd and MSP. In *Proceedings of the International Computer Music Conference*, pages 109–112, San Francisco, 1998.

[12] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, November 2011.

[13] Xavier Serra and Julius Smith. Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition. *Computer Music Journal*, 14(4):12–24, 1990. Publisher: The MIT Press.

[14] William A Sethares. *Tuning, Timbre, Spectrum, Scale*. Springer, London, 2005.