DEEP-RHYTHM FOR TEMPO ESTIMATION AND RHYTHM PATTERN RECOGNITION

Hadrien Foroughmand, Geoffroy Peeters

DEEP-RHYTHM FOR TEMPO ESTIMATION AND RHYTHM PATTERN RECOGNITION

Hadrien Foroughmand
IRCAM Lab - CNRS - Sorbonne Université
hadrien.foroughmand@ircam.fr

Geoffroy Peeters
LTCI - Télécom Paris - Institut Polytechnique de Paris
geoffroy.peeters@telecom-paris.fr

ABSTRACT

It has been shown that the harmonic series at the tempo frequency of the onset-strength-function of an audio signal accurately describes its rhythm pattern and can be used to perform tempo or rhythm pattern estimation. Recently, in the case of multi-pitch estimation, the depth of the input layer of a convolutional network has been used to represent the harmonic series of pitch candidates. We use a similar idea here to represent the harmonic series of tempo candidates. We propose the Harmonic-Constant-Q-Modulation which represents, using a 4D-tensors, the harmonic series of modulation frequencies (considered as tempo frequencies) in several acoustic frequency bands over time. This representation is used as input to a convolutional network which is trained to estimate tempo or rhythm pattern classes. Using a large number of datasets, we evaluate the performance of our approach and compare it with previous approaches. We show that it slightly increases Accuracy-1 for tempo estimation but not the average-mean-Recall for rhythm pattern recognition.

1. INTRODUCTION

Tempo is one of the most important perceptual elements of music. Today numerous applications rely on tempo information (recommendation, playlist generation, synchronization, dj-ing, audio or audio/video editing, beat-synchronous analysis). It is therefore crucial to develop algorithms to correctly estimate it. The automatic estimation of tempo from an audio signal has been one of the first research carried on in Music Information Retrieval (MIR) [11]. 25 years later it is still a very active research subject in MIR. This is due to the fact that tempo estimation is still an unsolved problem (outside the prototypical cases of pop or techno music) and that recent deep-learning approaches [3] [37] bring new perspectives to it.

Tempo is usually defined as (and annotated as) the rate at which people tap their foot or their hand when listening to a music piece. Several people can therefore perceive different tempi for the same piece of music. This is due to the hierarchy of the metrical structure in music (to deal with this ambiguity the research community has proposed to consider octave errors as correct) and due to the fact that without the cultural knowledge of the rhythm pattern(s) being played, it can be difficult to perceive "the" tempo (or even "a" tempo). This last point, of course, opens the door to data-driven approaches, which can learn the specificities of the patterns. In this work, we do not deal with the inherent ambiguity of tempo and consider the values provided by annotated datasets as ground-truth. The method we propose here belong to the data-driven systems in the sense that we learn from the data. It also considers both the tempo and rhythm pattern in interaction by adequately modeling the audio content. The tempo of a track can of course vary along time, but in this work we focus on the estimation of constant (global) tempi and rhythm patterns.

In the following section, we summarize works related to tempo and rhythm pattern estimation from audio. We refer the reader to [12, 32, 44] for more detailed overviews.

1.1 Related works

Tempo estimation. Early MIR systems encoded domain knowledge (audio, auditory perception and musical knowledge) by hand-crafting signal processing and statistical models (hidden Markov, dynamic Bayesian network). Data were at most used to manually tune some parameters (such as filter frequencies or transition probabilities). Early techniques for beat tracking and/or tempo estimation belong to this category. Their overall flowchart is a multi-band separation combined with an onset strength function which periodicity is measured. For example, Scheirer [36] proposes the use of band-pass filters combined with resonating comb filters and peak picking; Klapuri [18] uses the resonating comb filter bank which is driven by the bandwise accent signals, the main extension is the tracking of multiple metrical levels; Gainza and Cole [8] propose a hybrid multiband decomposition where the periodicities of onset functions are tracked in several frequency bands using autocorrelation and then weighted.

Since the pioneer works of [11], many audio datasets have been annotated into tempo. This therefore encourages researchers to develop data-driven systems based on machine learning and deep learning techniques. Such machine-learning models are K-Nearest-Neighbors (KNN) [40], Gaussian Mixture Model (GMM) [33, 43], Support Vector Machine (SVM) [4, 9, 34], bags of classifiers [20], Random Forest [38] or more recently deep learning models. The first use of deep learning for tempo estimation

was proposed by Böck et al. [3] who proposed a deep Recurrent Neural Network (bi-LSTM) to predict the position of the beats inside the signal. This output is then used as the input of a bank of resonating comb filters to detect the periodicity and so the tempo. This technique still achieves the best results today in terms of Accuracy2. Recently, Schreiber and Müller [37] proposed a "single step approach" for tempo estimation using deep convolutional networks. The network design is inspired by the flowchart of handcrafted systems: the first layer is supposed to mimic the extraction of an onset-strength-function. Their system uses as input Mel-spectrograms and the network is trained to classify the tempo of an audio excerpt into 256 tempo classes (from 30 to 285 BPM), it shows very good results in terms of Class-Accuracy and Accuracy1.

Rhythm pattern recognition. While tempo and rhythm pattern are closely interleaved, the recognition of rhythm pattern has received much less attention. This is probably due to the difficulty of creating datasets annotated in such rhythm pattern (defining the similarity between patterns — outside the trivial identity case — remains a difficult task). To create such a dataset, one may consider the equivalence between the rhythm pattern and the related dance (such as Tango): *Ballroom* [12], *Extended-ballroom* [24] and *Greek-dances* [14]. Systems to recognize rhythm pattern from the audio are all very different: Foote [7] defines a beat spectrum computed with a similarity matrix of MFCCs, Tzanetakis [42] defines a beat histogram computed from an autocorrelation function, Peeters [32] proposes a harmonic analysis of the rhythm pattern, Holzapfel [15] proposes the use of the scale transform (which allows to get a tempo invariant representation), Marchand [23, 25] extends the latter by combining it with the modulation spectrum and adding correlation coefficients between frequency bands. For a recognition task on the Extended-ballroom and Greek-dances, Marchand can be considered as the state-of-the-art.

1.2 Paper proposal and organization

In this paper we present a new audio representation, the Harmonic-Constant-Q-Modulation (HCQM), to be used as input to a convolutional network for global tempo and rhythm pattern estimation. We name it **Deep Rhythm** since it uses the **depth** of the input and a **deep** network. The paper is organized as follows. In §2, we describe the motivation for (§2.1) and the computation details of (§2.2) the HCQM. We then describe the architecture of the convolutional neural network we used (§2.3) and the training process (§2.4). In §3, we evaluate our proposal for the task of global tempo estimation (§3.1) and rhythm pattern recognition (§3.2) and discuss the results.

2. PROPOSED METHOD

2.1 Motivations

From Fourier series, it is known that any periodic signal $x(t)$ with period T_0 (of fundamental period $f_0 = 1/T_0$) can be represented as a weighted sum of sinusoidal components which frequencies are the harmonics of f_0: $\hat{x}_{f_0,\underline{a}}(t) = \sum_{h=1}^{H} a_h \sin(2\pi h f_0 t + \phi_h)$.

For the voiced part of speech or pitched musical instrument, this leads to the so-called "harmonic sinusoidal model" [26, 39] that can be used for audio coding or transformation. This model can also be used to estimate the pitch of a signal [21]: estimating the f_0 such that $\hat{x}_{f_0,\underline{a}}(t) \simeq x(t)$. The values a_h can be estimated by sampling the magnitude of the DFT at the corresponding frequencies $a_{h,f_0} = |X(hf_0)|$. The vector $\underline{a}_{f_0} = \{a_{1,f_0} \cdots a_{H,f_0}\}$ represents the spectral envelope of the signal and is closely related to the timbre of the audio signal, hence the instrument playing. For this reason, these values are often used for instrument classification [29].

For audio musical rhythm, Peeters [30] [31] [32] proposes to apply such an harmonic analysis to an onset-strength-function . The period T_0 is then defined as the duration of a beat. a_{1,f_0} then represents the DFT magnitude at the 4^{th}-note level, a_{2,f_0} at the 8^{th}-note level, a_{3,f_0} at the 8^{th}-note-triplet level, while $a_{\frac{1}{2},f_0}$ represent the binary grouping of the beats and $a_{\frac{1}{3},f_0}$ the ternary one. Peeters considers that the vector $\underline{a}$ is representative of the specific rhythm and that therefore $\underline{a}_{f_0}$ represents a specific rhythm played at a specific tempo f_0. He proposes the following harmonic series: $h \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 1.25, 1.33, \ldots 8\}$ Using this, he shows - in [32] that given the tempo f_0, the vector $\underline{a}_{f_0}$ can be used to classify the different rhythm pattern; - in [30], that given manually-fixed prototype vectors $\underline{a}$, it is possible to estimate the tempo f_0 (looking for the f such that $\underline{a}_f \simeq \underline{a}$); - in [31] that the prototype vectors $\underline{a}$ can be learned (using simple machine-learning) to achieve the best tempo estimation f_0.

The method we propose in this paper is in the continuation of this last work (learning the values $\underline{a}$ to estimate the tempo or the rhythm class) but we adapted it to the deep learning formalism recently proposed by Bittner et al. [2] where the depth of the input to a convolutional network is used to represent the harmonic series $\underline{a}_f$. In [2], a constant-Q-transform (time τ and log-frequency f) is expanded to a third dimension which represent the harmonic series $\underline{a}_f$ of each f (with $h \in [\frac{1}{2}, 1, 2, 3, 4, 5]$). When $f = f_0$, $\underline{a}_f$ will represent the specific harmonic series of the musical instrument (plus an extra value at the $\frac{1}{2}f$ position used to avoid octave errors). When $f \neq f_0$, $\underline{a}_f$ will represent (almost) random values. In [2], the goal is to estimate the parameters of a filter such that when multiplied with this third dimension $\underline{a}_f$ it will provide very different values when $f = f_0$ or when $f \neq f_0$. This filter will then be convolved over all log-frequencies f and time τ to estimate the f_0's. This filter is trained using annotated data. In [2], there is actually several of such filter; they constitute the first layer of a convolutional network. In practice, in [2], the $a_{h,f}$ are not obtained as $|X(hf)|$; but by stacking in depth several CQTs each starting at different minimal frequencies $hf_{\min}$. The representation is denoted by Harmonic Constant-Q Transform (HCQT): $X_{hcqt}(f, \tau, h)$.

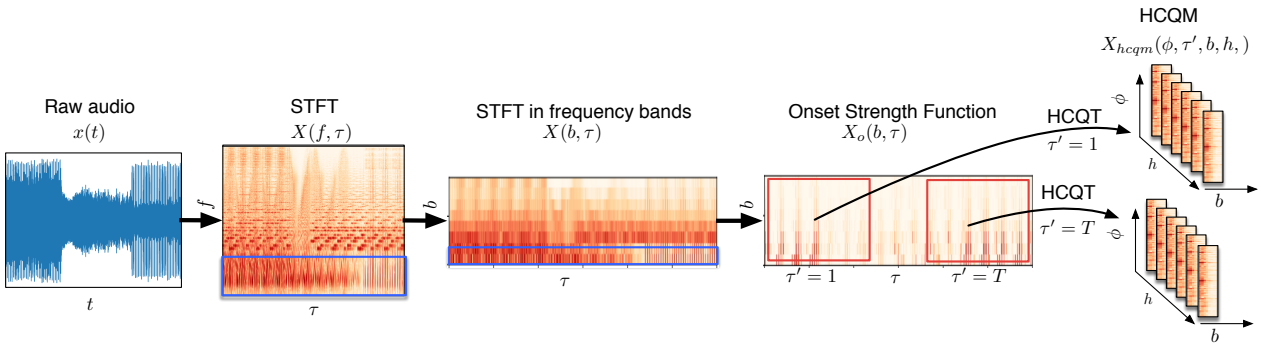

Figure 1. Flowchart of the computation of the Harmonic-Constant-Q-Modulation (HCQM). See text for details.

2.2 Input representation: the HCQM

As mentioned our goal is here to adapt the harmonic representation of the rhythm proposed in [30] [31] [32] to the deep learning formalism proposed in [2]. For this, the HCQT proposed by [2] is not applied to the audio signal, but to a set of Onset-Strength-Function (OSF) which represent the rhythm content in several acoustic frequency bands. The OSFs are low-pass signals which temporal evolution is related to the tempo and the rhythm pattern.

We denote our representation by **Harmonic-Constant-Q-Modulation (HCQM)**. As the Modulation Spectrum (MS) [1] it represents, using a time/frequency (τ'/ϕ) representation, the energy evolution (low-pass signal) within each frequency band b of a first time/frequency (τ/f) representation. However, while the MS uses two interleaved Short-Time-Fourier-Transforms (STFTs) for this, we use a Constant-Q transform for the second time/frequency representation (in order to obtain a better spectral resolution). Finally, as proposed by [2], we add one extra dimension to represent the content at the harmonics of each frequency ϕ. We denote it by $X_{hcqm}(\phi,\tau',b,h)$ where ϕ are the modulation frequencies (which correspond to the tempo frequencies), τ' are the times of the CQT frames, b are the acoustic frequency bands and h the harmonic numbers.

Computation. In Figure 1, we indicate the computation flowchart of the HCQM. Given an audio signal $x(t)$, we first compute its STFT, denoted by $X(f,\tau)$. The acoustic frequency f of the STFT are grouped into logarithmic-spaced acoustic-frequency-bands $b \in [1, B=8]$. We denote it by $X(b,\tau)$. The goal of this is to reduce the dimensionality while preserving the information of the spectral location of the rhythm events (kick patterns tend to be in low frequencies while hit-hat patterns in high frequencies). For each band b, we then compute an Onset-Strength-Function over time τ, denoted by $X_o(b,\tau)$.

For a specific b, we now consider the signal $s_b(\tau) = X_o(b,\tau)$ and perform the analysis of its periodicities over time τ. One possibility would be to compute a time-frequency representation $S_b(\phi,\tau')$ over tempo-frequencies ϕ and time frame τ' and then sample $S_b(\phi,\tau')$ at the positions $h\phi$ with $h \in \{\frac{1}{2},1,2,3,4,5\}$ to obtain $S_b(\phi,\tau',h)$. This is the idea we used in [32]. However, in the present work, we use the idea proposed by [2]: we compute a set of CQTs [1], each one with a different starting frequency $h\phi_{\min}$. We set $\phi_{\min}$=32.7. Each of these CQTs gives us $S_{b,h}(\phi,\tau')$ for one value of h. Stacking them over h therefore provides us with $S_b(\phi,\tau',h)$. The idea proposed by [2] therefore allows to mimic the sampling at the $h\phi$ but provides the correct window length to achieve a correct spectral resolution. We finally stack the $S_b(\phi,\tau',h)$ over b to obtain the 4D-tensors $X_{hcqm}(\phi,\tau',b,h)$. The computation parameters (sample rate, hop size, window length and FFT size) are set such that the CQT modulation frequency ϕ represents the tempo value in BPM.

Illustration. For easiness of visualisation (it is difficult to visualize a 4D-tensor), we illustrate the HCQM $X_{hcqm}(\phi,\tau',b,h)$ for a given τ' (it is then a 3D-tensor). Figure 2 [Left] represent $X_{hcqm}(\phi,b,h)$ on a real audio signal with a tempo of 120 bpm. Each sub-figure represent $X_{hcqm}(\phi,b,h)$ for a different value of $h \in \{\frac{1}{2},1,2,3,4,5\}$. The y-axis and x-axis are the tempo frequency ϕ and the acoustic frequency band b. The dashed rectangle super-imposed to the sub-figures indicates the slice of values $X_{hcqm}(\phi = 120bpm,b,h)$ which corresponds to the ground-truth tempo. It is this specific pattern over b and h that we want to learn using the filters W of the first layer of our convolutional network.

2.3 Architecture of the Convolutional Neural Network

The architecture of our network is both inspired by the one of [2] (since we perform convolutions over an input spectral representation and use its depth) and the one of [37] (since we perform a classification). However, it differs in the definition of the input and output.

Input. In [2], the input is the 3D-tensor $X_{cqt}(f,\tau,h)$ and the convolution is done over f and τ (with filters of depth H). In our case, the input could be the 4D-tensors $X_{hcqm}(\phi,\tau',b,h)$ and the convolution could be done over ϕ, τ' and b (with filters of depth H). However, to simplify the computation, we reduce $X_{hcqm}(\phi,\tau',b,h)$ to a sequence over τ' of 3D-tensors $X_{hcqm}(\phi,b,h)$ [2]. The convolution is then done over ϕ and b with filters of depth H.

Our goal is to learn filters W narrow in ϕ and large in b which represents the specific shape of the harmonic con-

[1] For the STFT and CQT we used the librosa library [27].

[2] Future works will concentrate in performing the convolution directly using the 4D-tensors; which would allow to perform smoothing over time.

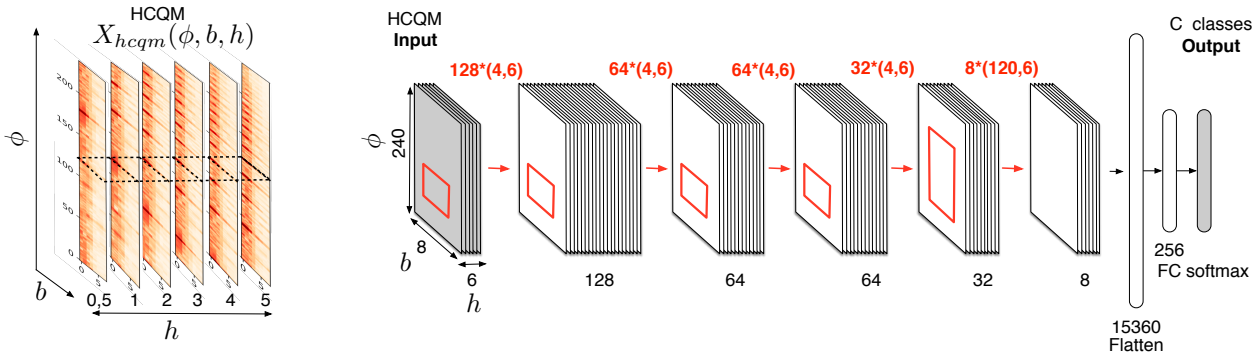

Figure 2. [Left] Example of the HCQM for a real audio with tempo 120bpm. [Right] Architecture of our CNN.

tent of a rhythm pattern. We do the convolution over b because the same rhythm pattern can be played with instrument transposed in acoustic frequencies.

Output. The output of the network proposed by [2] is a 2D representation which represents a saliency map of the harmonic content over time. In our case, the outputs are either the $C = 256$ classes of tempo (as in [37] we consider the tempo estimation problem as a classification problem into 256 tempo classes) or the $C = 13$ (for extended-ballroom) or $C = 6$ (for Greek-dances) classes of rhythm pattern. To do so, we added at the end of the network proposed by [2] two dense layers, the last one with C units and a softmax activation.

Architecture. In Figure 2 [Right], we indicate the architecture of our network. The input is a 3D-tensor $X_{hcqm}(\phi, b, h)$ for a given time τ'. The first layer is a set of 128 convolutional filters of shape $(\phi = 4, b = 6)$ (with depth H). As mentioned, the convolution is done over ϕ and b. The shape of these filters has been chosen such that they are narrow in tempo frequency ϕ (to precisely estimate the tempo) but cover multiple frequency acoustic bands b (because the information relative to the tempo/ rhythm cover several bands). As illustrated in Figure 2 [Left] the goal of the filters is to identify the pattern over b and h specific to $\phi =$ tempo frequency.

The first layer is followed by two convolutional layers of 64 filters of shape $(4, 6)$, one layer of 32 filters of shape $(4, 6)$, one layer of 8 filters of shape $(120, 6)$ (this allows to track down the relationships between the modulation frequencies ϕ). The output of the last convolution layer is then flattened and followed by a dropout with $p = 0.5$ (to avoid over-fitting [41]), a fully-connected layer of 256 units, a last fully-connected layer of C units with a softmax activation to perform the classification into C classes. The softmax activation vector is denoted by $\underline{y}(\tau')$. The Loss function to be minimized is a categorical cross entropy.

All layers are preceded by a batch normalization [16]. We used Rectified Linear Units (ReLU) [28] for all convolutional layers, and Exponential Linear Units (eLU) [5] for the first fully-connected layer.

2.4 Training

The inputs of our network are the 3D tensors HCQM $X_{hcqm}(\phi, b, h)$ computed for all time τ' of the music track. On the other side, the datasets we will use for our experiments only provide **global** tempi [3] or **global** rhythm classes as ground-truths. We therefore have several HCQMs for a given track which are all associated with the same ground-truth (multiple instance learning).

To fix the network hyper-parameters, we split the training set into a train (90%) and a validation part (10%). We used the ADAM [17] optimizer to find the parameters of the network with a constant learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$. We used mini-batches of 256 HCQM with shuffle and a maximum of 20 epochs with early-stopping.

2.5 Aggregating decisions over time

Our network provides an estimation of the tempo or rhythm class at each time τ'. To obtain a **global** value we aggregate the softmax activation vectors $\underline{y}(\tau')$ over time by choosing the maximum of the vector $\underline{y}$ computed as the average over τ' of the $\underline{y}(\tau')$.

3. EVALUATION OF THE SYSTEM

We evaluate our proposed system for two tasks: - **global tempo estimation** (§3.1), - **rhythm pattern recognition** (§3.2). We used the same system (same input representation and same network architecture) for the two tasks. However, considering that the class definitions are different, we performed two independent trainings [4].

3.1 Global tempo estimation

Training and testing sets. To be able to compare our results with previous works, we use the same paradigm (cross-dataset validation [5]) and the same datasets as [37] which we consider here as the state-of-the-art. The **training set** is the union of

[3] It should be noted that this does not always correspond to the reality of the track content since some of them have a tempo varying over time (tempo drift), have a silent introduction or a break in the middle.

[4] Future works, will consider training a single network for the two tasks or applying transfer learning of one task to the other.

[5] Cross-dataset validation uses separate datasets for training and testing; not only splitting a single dataset into a train and a test part.

- *LMD tempo*: it is a subset of the Lack MIDI dataset [35] annotated into tempo by [38]; it contains 3611 items of 30s excerpts of 10 different genres
- *MTG tempo* it is a subset of the GiantSteps MTG key dataset (MTG tempo) [6] annotated using a tapping method by [37]; it contains 1159 items of 2min excerpts of electronic dance music;
- *Extended Ballroom* [24]: it contains 3826 items of 13 genres (it should be noted that we removed from the Ballroom test-set the items existing in the Extended Ballroom training-set).

The total size of the training set is 8596. It covers multiple musical genres to favor generalization.

The **test-sets** are also the same as in [37] (see [38] for their details): - *ACM-Mirum* [33] (1410 items), - *ISMIR04* [12] (464 items), - *Ballroom* [12] (698 items), - *Hainsworth* [13] (222 items), - *GTzan-Rhythm* [22] (1000 items), - *SMC* [14] (217 items), - *Giantsteps Tempo* [19] (664 items). We also added the *RWC-popular* [10] (100 items) for comparison with [3]. As in [37], *Combined* denotes the union of all test-sets (except RWC popular).

Evaluation protocol. We measure the performances using the following metrics:

- **Class-Accuracy:** it measures the ability of our system to predict the correct tempo class (in our system we have 256 tempo classes ranging from 30 to 285bpm);
- **Accuracy1:** it measures if our estimated tempo is within $\pm 4\%$ of the ground-truth tempo
- **Accuracy2:** is the same as Accuracy1 but considering octave errors as correct

Results and discussions. The results are indicated in Tables 1, 2 and 3.

Validation of B and H. We first show that the separation of the signal into multiple acoustic-frequency-bands $b \in [1, B = 8]$ and the use of the harmonic depth H is beneficial. For this, we compare the results obtained using
- $B = 8$ and $h \in \{\frac{1}{2}, 1, 2, 3, 4, 5\}$ (column "new")
- $B = 8$ but $h \in \{1\}$ (column "h=1")
- $B = 1$ and $h \in \{\frac{1}{2}, 1, 2, 3, 4, 5\}$ (column "b=1").

Results are only indicated for the "Combined" test-sets. For all metrics (Class-Accuracy, Accuracy-1 and Accuracy-2), the best results are obtained using $B = 8$ and $h \in \{\frac{1}{2}, 1, 2, 3, 4, 5\}$.

Comparison with state-of-the-art. We now compare our results with the state-of-the-art represented by the 3 following systems: - **sch1** denotes the results published in [38], - **sch2** in [37] and - **böck** in [3].

According to these Tables, we see that our method allows an improvement for two test-sets: - *Ballroom* in terms of Class-Accuracy (73.8) - *Ballroom* and *Giantsteps* in terms of Accuracy1 (92.6 and 83.6) and Accuracy2 (98.7 and 97.9). This can be explained by the fact that the musical genres of these two test-sets are represented in our training set (by the *Extended-Ballroom* and *MTG tempo* respectively). It should be noted however that there is no intersection between the training and test sets. For the *ISMIR04* and *GTZAN* test-sets, our results are very close (but lower) to the one of **sch1**. For the *RWC popular* test sets, our results (73.0 and 98.0) largely outperforms the ones of **böck** in terms of Accuracy-1 and Accuracy-2. Finally, if we consider the *Combined* dataset, our method slightly outperforms the other ones in terms of Accuracy1 (74.4).

The worst results of our method were obtained for the *SMC* data-sets. This can be explained by the fact that the *SMC* data-sets contains rhythm patterns very different from the ones represented in our training-sets.

While our results for Accuracy2 (92.0) are of course higher than our results for Accuracy1 (74.4), they are still lower than the ones of **böck** (93.6). One reason for this is that our network (as the ones of sch1 and sch2) is trained to discriminate BPM classes, therefore it doesn't know anything about octave equivalences.

3.2 Rhythm pattern recognition

Training and testing sets. For the rhythm pattern recognition, it is not possible to perform cross-dataset validation (as we did above) because the definition of the rhythm pattern classes is specific to each dataset. Therefore, as in previous works [25], we perform a 10-fold cross-validation using the following datasets:

- *Extended Ballroom [24]:* it contains 4180 samples of C=13 rhythm classes: 'Foxtrot', 'Salsa', 'Viennesewaltz', 'Tango', 'Rumba', 'Wcswing', 'Quickstep', 'Chacha', 'Slowwaltz', 'Pasodoble', 'Jive', 'Samba', 'Waltz'
- *Greek-dances [15]:* it contains 180 samples of C=6 rhythm classes: 'Kalamatianos', 'Kontilies', 'Maleviziotis', 'Pentozalis', 'Sousta' and 'Kritikos Syrtos'

Evaluation protocol. As in previous works [25], we measure the performances using the average-mean-Recall R[6]. For a given fold f, the mean-Recall R_f is the mean over the classes c of the class-recall $R_{f,c}$. The average mean-Recall R is then the average over f of the mean-Recall R_f. We also indicate the standard deviation over f of the mean-Recall R_f.

Results and discussions. The results are indicated in Table 4. We compare our results with the ones of [25], denoted as **march**, considered here representative of the current state-of-the-art.

Our results (76.4 and 68.3) are largely below the ones of **march** (94.9 and 77.2). This can be explained by the fact that the "Scale and shift invariant time/frequency" representation proposed in [25] takes into account the interrelationships between the frequency bands of the rhythmic events while our HCQM does not.

To better understand these results, we indicate in Figure 3 [Top] the confusion matrices for the *Extended Ballroom*. The diagonal represents the Recall R_c of each class c. We see that our system is actually suitable to detect the majority of the classes: $R_c \geq 88\%$ for 9 classes over 13. ChaCha, Waltz and WcSwing make R completely drop. Waltz is actually estimated 97% of the

[6] The mean-Recall is not sensitive to the distribution of the classes. Its random value is always $1/C$ for a problem with C classes.

Datasets	sch1	sch2	böck	**new**	h=1	b=1
ACMMirum	38.3	**40.6**	29.4	38.2		
ISMIR04	**37.7**	34.1	27.2	29.4		
Ballroom	46.8	67.9	33.8	**73.8**		
Hainsworth	**43.7**	43.2	33.8	23.0		
GTzan	**38.8**	36.9	32.2	27.6		
SMC	14.3	12.4	**17.1**	7.8		
Giantsteps	53.5	**59.8**	37.2	25.5		
RWCpop	X	X	X	66.0		
Combined	40.9	**44.8**	31.2	36.8	29.8	32.4

Table 1. Class-Accuracy

sch1	sch2	böck	**new**	h=1	b=1
72.3	**79.5**	74.0	73.3		
63.4	60.6	55.0	61.2		
64.6	92.0	84.0	**92.6**		
65.8	77.0	**80.6**	73.4		
71.0	69.4	69.7	69.7		
31.8	33.6	**44.7**	30.9		
63.1	73.0	58.9	**83.6**		
X	X	60.0	**73.0**		
66.5	74.2	69.5	**74.4**	64.2	67.9

Table 2. Accuracy1

sch1	sch2	böck	**new**	h=1	b=1
97.3	97.4	**97.7**	96.5		
92.2	92.2	**95.0**	87.1		
97.0	98.4	**98.7**	**98.7**		
85.6	84.2	**89.2**	82.9		
93.3	92.6	**95.0**	89.1		
55.3	50.2	**67.3**	50.7		
88.7	89.3	86.4	**97.9**		
X	X	95.0	**98.0**		
92.2	92.1	**93.6**	92.0	82.0	88.4

Table 3. Accuracy2

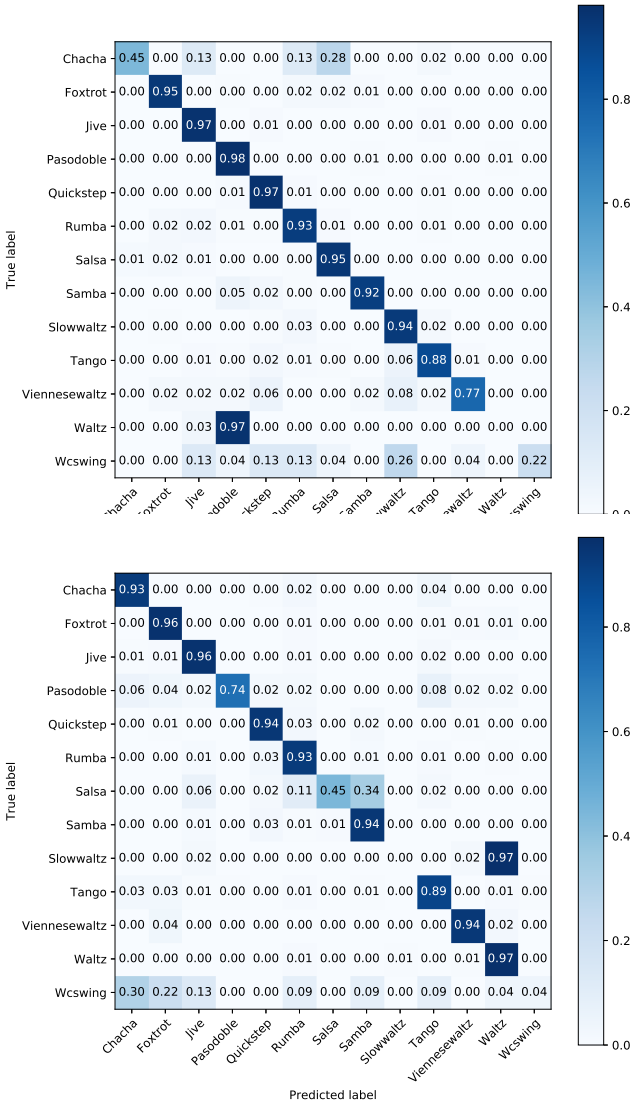

Figure 3. Confusion matrix for the *Extended Ballroom*. [Top] using $h \in \{0.5, 1, 2, 3, 4, 5\}$ [Bottom] using $h \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 1.25, 1.33, \ldots 8\}$

time as Pasadoble. This can be explained by the fact that our current harmonic series $h \in \{\frac{1}{2}, 1, 2, 3, 4, 5\}$ does not represent anything about the 3/4 meter specific to Waltz which would be represented by $h = \frac{1}{3}$. To verify our assumption, we redo the experiment using this time exactly the same harmonic series as proposed in [32]: $h \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 1.25, 1.33, \ldots 8\}$. The corresponding confusion matrix is indicated in 3 [Bottom] where Waltz is now perfectly recognized (97%), however SlowWaltz is now recognized as Waltz in 97% of the cases which makes (while the system is better) the average mean-Recall actually decreases to 74.6%. The low results of Wcswing can be explained by the (too) small number of training examples (23).

Datasets	march	**new**
Extended Ballroom	94.9	76.4 (33.1)
Greek-dances	77.2	68.3 (27.5)

Table 4. Average (std) Recall R for rhythm classification.

4. CONCLUSION

In this paper we have presented a new approach for global tempo and rhythm pattern classification. We have proposed the Harmonic-Constant-Q-Modulation (HCQM) representation, a 4D-tensor which represents the harmonic series at candidate tempo frequencies of a multi-band Onset-Strength-Function. This HCQM is then used as input to a deep convolutional network. The filters of the first layer of this network are supposed to learn the specific characteristic of the various rhythm patterns. We have evaluated our approach for two classification tasks: global tempo and rhythm pattern classification.

For tempo estimation, our method outperforms previous approaches (in terms of Accuracy1 and Accuracy2) for the *Ballroom* (ballroom music) and *Giant-steps tempo* (electronic music) test-sets. Both test-sets represent music genres with a strong focus on rhythm. It seems therefore that our approach works better when rhythm patterns are clearly defined. Our method also performs slightly better (in terms of Accuracy1) for the Combined test-set.

For rhythm classification, our method doesn't work as well as the state of the art [25]. However, the confusion matrices indicate that our recognition is above 90% for the majority of the classes of the *Extended Ballroom*. Moreover, we have shown that adapting the harmonic series h can help improving the performances.

Among future works, we would like to study the use of the HCQM 4D tensors directly, to study other harmonic series and to study the joint training of (or transfer learning between) tempo and rhythm pattern classification.

Acknowledgement: This work was partly supported by European Union's Horizon 2020 research and innovation program under grant agreement No 761634 (Future Pulse project).

5. REFERENCES

[1] Les Atlas and Shihab A Shamma. Joint acoustic and modulation frequency. *Advances in Signal Processing, EURASIP Journal on*, 2003:668–675, 2003.

[2] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan Pablo Bello. Deep salience representations for f0 estimation in polyphonic music. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Suzhou, China, 2017.

[3] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Accurate tempo estimation based on recurrent neural networks and resonating comb filters. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Malaga, Spain, 2015.

[4] Ching-Wei Chen, Markus Cremer, Kyogu Lee, Peter DiMaria, and Ho-Hsiang Wu. Improving perceived tempo estimation by statistical modeling of higher-level musical descriptors. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[5] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[6] Angel Faraldo, Sergi Jorda, and Perfecto Herrera. A multi-profile method for key estimation in edm. In *AES International Conference on Semantic Audio*, Ilmenau, Germany, 2017. Audio Engineering Society.

[7] Jonathan Foote, Matthew L Cooper, and Unjung Nam. Audio retrieval by rhythmic similarity. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Paris, France, 2002.

[8] Mikel Gainza and Eugene Coyle. Tempo detection using a hybrid multiband approach. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(1):57–68, 2011.

[9] Aggelos Gkiokas, Vassilios Katsouros, and George Carayannis. Reducing tempo octave errors by periodicity vector coding and svm learning. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Porto, Portugal, 2012.

[10] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Paris, France, 2002.

[11] Masataka Goto and Yoichi Muraoka. A beat tracking system for acoustic signals of music. In *Proceedings of the second ACM international conference on Multimedia*, San Francisco, California, USA, 1994.

[12] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(5):1832–1844, 2006.

[13] Stephen Webley Hainsworth. *Techniques for the Automated Analysis of Musical Audio*. PhD thesis, University of Cambridge, UK, September 2004.

[14] Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech and Language Processing, IEEE Transactions on*, 20(9):2539–2548, 2012.

[15] André Holzapfel and Yannis Stylianou. Scale transform in rhythmic similarity of music. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(1):176–185, 2011.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Anssi P Klapuri, Antti J Eronen, and Jaakko T Astola. Analysis of the meter of acoustic musical signals. *Audio, Speech and Language Processing, IEEE Transactions on*, 14(1):342–355, 2006.

[19] Peter Knees, Angel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Malaga, Spain, 2015.

[20] Mark Levy. Improving perceptual tempo estimation with crowd-sourced annotations. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Miami, Florida, USA, 2011.

[21] Robert C Maher and James W Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *JASA (Journal of the Acoustical Society of America)*, 95(4):2254–2263, 1994.

[22] Ugo Marchand, Quentin Fresnel, and Geoffroy Peeters. Gtzan-rhythm: Extending the gtzan test-set with

beat, downbeat and swing annotations. In *Late-Breaking/Demo Session of ISMIR (International Society for Music Information Retrieval)*, Malaga, Spain, October 2015. Late-Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference, 2015.

[23] Ugo Marchand and Geoffroy Peeters. The modulation scale spectrum and its application to rhythm-content description. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, Erlangen, Germany, 2014.

[24] Ugo Marchand and Geoffroy Peeters. The extended ballroom dataset. In *Late-Breaking/Demo Session of ISMIR (International Society for Music Information Retrieval)*, New York, USA, August 2016. Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conf.. 2016.

[25] Ugo Marchand and Geoffroy Peeters. Scale and shift invariant time/frequency representation using auditory statistics: Application to rhythm description. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016.

[26] R McAuley and T Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34:744–754, 1986.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Austin, Texas, 2015.

[28] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of ICMC (International Computer Music Conference)*, Haifa, Israel, 2010.

[29] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.

[30] Geoffroy Peeters. Template-based estimation of time-varying tempo. *Advances in Signal Processing, EURASIP Journal on*, 2007(1):067215, 2006.

[31] Geoffroy Peeters. Template-based estimation of tempo: using unsupervised or supervised learning to create better spectral templates. In *Proc. of DAFx (International Conference on Digital Audio Effects)*, pages 209–212, Graz, Austria, September 2010.

[32] Geoffroy Peeters. Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *Audio, Speech and Language Processing, IEEE Transactions on*, 19(5):1242–1252, 2011.

[33] Geoffroy Peeters and Joachim Flocon-Cholet. Perceptual tempo estimation using gmm-regression. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, Nara, Japan, 2012.

[34] Graham Percival and George Tzanetakis. Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):1765–1776, 2014.

[35] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. PhD thesis, Columbia University, 2016.

[36] Eric D Scheirer. Tempo and beat analysis of acoustic musical signals. *JASA (Journal of the Acoustical Society of America)*, 103(1):588–601, 1998.

[37] Hendrik Schreiber and M Müller. A single-step approach to musical tempo estimation using a convolutional neural network. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Paris, France, 2018.

[38] Hendrik Schreiber and Meinard Müller. A post-processing procedure for improving music tempo estimates using supervised learning. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Suzhou, China, 2017.

[39] 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.

[40] Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer. From rhythm patterns to perceived tempo. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Vienna, Austria, 2007.

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[42] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302, 2002.

[43] Linxing Xiao, Aibo Tian, Wen Li, and Jie Zhou. Using statistic model to capture the association between timbre and perceived tempo. In *Proc. of ISMIR (International Society for Music Information Retrieval)*, Philadelphia, PA, USA, 2008.

[44] Jose Zapata and Emilia Gómez. Comparative evaluation and combination of audio tempo estimation approaches. In *Aes 42Nd Conference On Semantic Audio*, Ilmenau, Germany, 2011.