THE MODULATION SCALE SPECTRUM AND ITS APPLICATION TO RHYTHM-CONTENT DESCRIPTION.

Ugo Marchand
STMS IRCAM-CNRS-UPMC
1 pl. Igor Stravinsky, 75004 Paris, France
ugo.marchand@ircam.fr

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
1 pl. Igor Stravinsky, 75004 Paris, France
geoffroy.peeters@ircam.fr

ABSTRACT

In this paper, we propose the Modulation Scale Spectrum as an extension of the Modulation Spectrum through the Scale domain. The Modulation Spectrum expresses the evolution over time of the amplitude content of various frequency bands by a second Fourier Transform. While its use has been proven for many applications, it is not scale-invariant. Because of this, we propose the use of the Scale Transform instead of the second Fourier Transform. The Scale Transform is a special case of the Mellin Transform. Among its properties is "scale-invariance". This implies that two time-stretched version of a same music track will have (almost) the same Scale Spectrum. Our proposed Modulation Scale Spectrum therefore inherits from this property while describing frequency content evolution over time. We then propose a specific implementation of the Modulation Scale Spectrum in order to represent rhythm content. This representation is therefore tempo-independent. We evaluate the ability of this representation to catch rhythm characteristics on a classification task. We demonstrate that for this task our proposed representation largely exceeds results obtained so far while being highly tempo-independent.

1. INTRODUCTION

Automatic description of audio content has become over the years an important research field. When the audio content is music, the related research field is named Music Information Retrieval. Its growth over the years is explained by the massive digitalization of music and its accessibility through online-services (on-line music sellers –as iTunes– or music streaming services –as Spotify or Deezer–). This necessitates the organization and tagging of the music data to enhance their accessibility. While manual annotation is still possible, it is time-consuming hence money-consuming. Because of this, tools are developed for automatically organizing, tagging and visualizing the music data. These tools rely on automatic audio content analysis. In the case of tagging, automatic content analysis tools rely on two parts: extracting the right information (named audio features) from the audio signal and matching this information to the required tags.

In this paper, we propose a new representation, named Modulation Scale Spectrum, that allows an efficient description of time/frequency content over time. We apply it to create an audio feature that allows an efficient description of the rhythm content of a music track. Rhythm, along harmony (melody) and timbre (orchestration) are the three perspectives that describe music content.

The direct applications of this feature are the search by rhythm pattern (for example looking for identical rhythm pattens without being affected by the tempo), the automatic classification into rhythm-classes. A better description of rhythm would be also beneficial to genre, mood classification or search by similarity systems. Applications of the Modulation Scale Spectrum outside the music field also concern generic audio recognition.

1.1. Related works

In this part, we only review works related to audio rhythm representation. We also provide an overview on the Modulation Spectrum since our proposal is based on it. For a good overview on other methods to represent time and frequency content see [1].

Works can generally be divided according to the use of temporal, spectral or spectro/temporal representations; according to the matching model (Dynamic Time Warping, or statistical models), according to the necessity to have a preliminary tempo detection.

In [2], Foote introduces the *beat spectrum*. The beat spectrum is computed by parametrizing audio (with STFT amplitude coefficients or with its MFCC). A similarity measure (cosine or euclidian distance) is then taken between each feature vector and embedded in a 2-dimensional representation named Self-Similarity-Matrix (SSM) S. Finally the beat-spectrum is found by looking for periodicities in S with diagonal sums or auto-correlation. In [3], Foote uses his beat spectrum to measure rhythmic similarity between songs. Bartsch extends the beat spectrum to create an audio thumbnail in [4]. Antonopoulos in [5] also uses the SSM approach. He extracts from a song (or its thumbnailed version) a chroma-based MFCC feature. He computes a rhythmic signature by using Dynamic Time Warping (DTW) to calculate the similarity between two feature vectors. He validates his method by measuring rhythmic similarity on a greek and an african music corpus.

Dixon [6] uses *rhythmic patterns* to classify rhythm on the Ballroom Dataset. It achieves 50% correctness, and up to 96% (85.7% without annotated tempo) by combining with other features derived from these patterns, from features of Gouyon [7], and from annotated tempo. However, a limitation of this method is that position of the first bars is needed to extract the rhythmic pattern. It has been estimated with BeatRoot algorithm [8] but have been "corrected manually". Paulus [9] firstly detects tatus, tatum and bar length to segment a song into patterns. Then he computes features on theses patterns (loudness, brightness for example) and measures the similarity of rhythmic patterns with DTW. Wright [10] creates clave pattern templates to analyse afro-cuban music. He uses matched-filtering to enhance claves, and a rotation-aware dynamic programming algorithm to find tempo, beat and downbeat positions. Jensen [11] uses a log-scale autocorrelation to create tempo-independent rhythmic patterns. His method works nicely on the rhythm classification task on the Ballroom Dataset if we ignore nearest neighbours with similar in tempo (50% accuracy

instead of 20% for state-of-the-art method).

Tzanetakis [12] proposes a *Beat histogram* computed by using an enhanced auto-correlation function and peak-picking algorithm. This beat histogram, along with timbre and pitch features is used for music genre classification. Gouyon [7] uses a set of 73 features based on the tempo, on the *periodicity histogram* and on the *inter-onset-interval histogram*. The system achieves a 78.9% accuracy on the Ballroom Dataset classification task, and up to 90.1% using the annotated tempo.

Peeters [13] compares various spectral and temporal periodicity representations to describe the rhythm of a song. He firstly extracts an onset function from the signal then computes three feature vectors (based on the amplitude of the Discrete Fourier Transform, the Auto-Correlation Function, and Hybrid-Axis-Autocorrelation-Fourier-Transform representation). He then makes these vectors tempo independent and compact by sampling them at multiple of the tempo frequency. Theses features are tested for rhythm classification on the ballroom dataset. They achieve 96.1% classification accuracy with the annotated tempo and 88% without.

Holzapfel [14] proposes *Dynamic Periodicity Warping* to classify rhythm. He uses a DTW distance on the periodicity spectrum between each song of the ballroom dataset as input of a k-nearest neighbours classifier. The process achieves 82.1% accuracy on the ballroom dataset, which is a bit less than state-of-the-art methods, but outperforms all other algorithms on a dataset made of Cretan dances (70% against 50%).

Directly related to our method are the following works.

Holzapfel [15] proposes a representation to describe rhythm: the *Scale Transform*. The main property of the scale transform is that the scale transforms of two identical songs played at different speed will be the same. He uses this scale transform in the creation of a tempo-independent descriptor. The performance of this descriptor on the ballroom dataset are good: 85.1% (state of the art is about 88%). In [16], Holzapfel tests his descriptor on two additional datasets (Turkish and Greek traditional music which have wider within-class tempo distribution). It achieves better classes recognition than other methods.

Rodet and Worms [17, 18] proposes the *Modulation Spectrum* (without naming it explicitly Modulation Spectrum) for a task of audio identification by fingerprint. It models the temporal evolution of the energy content in different frequency bands with a Fourier Transform. McKinney [19] computes the Modulation Spectrum on the output of 18 4^{th} order bandpass GammaTone filters, then sums the energy on four frequency bands to get 18x4 descriptors. He uses these features to classify five audio classes (classical, pop music, speech, noise, crowd) and 7 musical genres. Whitman [20] proposes the "Penny" features to create a system of automatic recording reviews that understands and labels songs based on their audio content. Atlas [21] proposes a joint acoustic/modulation frequency model to improve audio coding. This model coded at 32 kbit/s has been found better by listeners than MP3 coding at 56 kbit/s.

1.2. Paper overview and organization

The Modulation Spectrum allows a compact description of the time and frequency content. Its use has been proven for many applications (audio fingerprint [18], generic audio classes or genre recognition [19], auto-tagging [20], speaker separation [21]). However, this representation is not scale-invariant.

In this paper, we propose the Modulation Scale Spectrum as an extension of the Modulation Spectrum to the Scale domain. This makes it scale-invariant. This property allows then to derive audio features for music audio signals which are tempo invariant. As MFCC provides the spectral-envelope information complementary to the fundamental frequency, the Modulation Scale Spectrum provides the information complementary to the music tempo. While the Scale Transform by itself already allows this, the Scale Transform does not allow to represent the spectral-envelope information (only a single energy value is used to represent the whole spectrum). The Modulation Scale Spectrum does represent this spectral-envelope information, as the Modulation Spectrum does, but is tempo-invariant, while the Modulation Spectrum is not.

The paper is organized as follows. In part 2, we introduce the Modulation Scale Spectrum. We first describe the Scale Transform (part 2.1) and the Modulation Spectrum (part 2.2) it is based on and then present in part 2.3 our proposed Modulation Scale Spectrum. We provide a specific implementation of it for rhythm description in part 2.4 and successfully demonstrate its use for a task of rhythm class recognition in part 3.

2. MODULATION SCALE SPECTRUM

In this part we propose the Modulation Scale Spectrum. It is based on both the Scale Transform (ST) and the Modulation Spectrum (MS). We first briefly review those.

2.1. Scale Transform

The Scale Transform is a special case of the Mellin Transform. It has been introduced by Cohen in [22]. Scale, like frequency is a physical attribute of the signal. We can see the scale content of a signal by using the Scale Transform, just like we can see the frequency content by using the Fourier Transform. The scale transform is defined by :

$$D(c) = \frac{1}{\sqrt{2\pi}} \int x(t) t^{-jc-\frac{1}{2}} dt \tag{1}$$

where c is the scale variable, t the time, and x the signal.

It can be viewed as the Fourier transform of an exponentially re-sampled signal $x(e^t)$ weighted by an exponential window $e^{\frac{1}{2}t}$:

$$D(c) = \frac{1}{\sqrt{2\pi}} \int x(e^t) e^{\frac{1}{2}t} e^{-jct} dt \tag{2}$$

One of the most important property of the Scale Transform is *scale invariance*. Scale invariance is expressed as

$$\begin{aligned} x(t) &\rightleftharpoons D(c) \\ \sqrt{a}\, x(at) &\rightleftharpoons e^{jc \ln a} D(c) \end{aligned} \tag{3}$$

This implies that both $x(t)$ and $\sqrt{a}\, x(at)$ share the same modulus of the scale spectrum but differ in their phase.

It should be noted however that the Scale Transform is not shift invariant: $|ST(x(t))| \neq |ST(x(t+a))|$. For this reason, the Scale Transform is often applied to the auto-correlation of $x(t)$ instead of $x(t)$ itself.

If $x(t)$ represents the energy of the audio signal, two audio signals with the same rhythm pattern and starting at the same time but played at different speed will share the same modulus of the scale spectrum (this is the property used by Holzapfel [15, 16]).

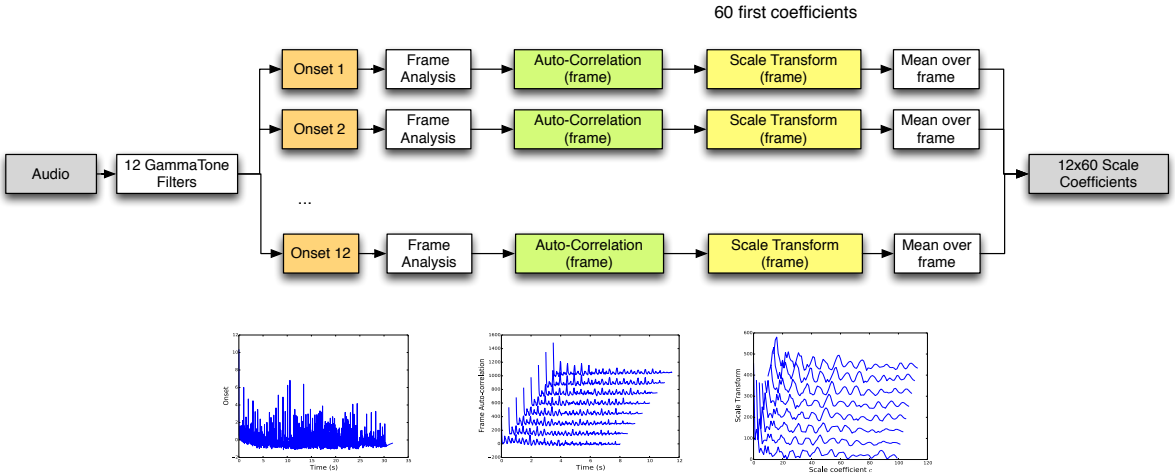

Figure 1: *Flowchart of our proposed Modulation Scale Spectrum for rhythm description.*

Computation: As seen, the ST can be computed in an efficient way from the Fourier Transform of $x(e^t)e^{\frac{1}{2}t}$ with $\omega = c$. In the case of the Discrete Fourier Transform $\omega_k = 2\pi \frac{k}{N} sr$ with $k \in \mathbb{N}$, N the size of the FFT and sr the sampling rate. In this case, c as ω_k is linearly-spaced. In the following we will use the implementation kindly provided by [23][1].

The exponential resampling step is detailled in [23]. We briefly review it here. We create an exponential axe between $t_0 = T_s$ and $t_e = nT_s$ where $T_s = \frac{1}{sr}$ is the sampling period and n the number of samples of the original signal. To fulfill the Nyquist-Shanon resampling condition, the maximum step between two adjacent exponential sample can not exceed T_s. In order not to have too many exponential samples, we choose as the distance between the two last samples the largest allowed distance: T_s. We then use a spline interpolation to create the resampled points. This leads to approximately $n \ln(n)$ exponential samples.

2.2. Modulation Spectrum

The modulation spectrum represents the evolution over time of the amplitude content of the various frequency bands ω of an STFT by a second Fourier Transform. It has been proposed by different authors under various names: dynamic features [18], modulation spectrum [21], auditory filterbank temporal envelopes [19], "penny" features [20]. It is also closely related to the scattering features proposed by [24]. The modulation spectrum $X(\omega, \Omega)$ can be expressed as (using [18] notation)

$$\begin{aligned} x(\omega, \tau) &= \frac{1}{\sqrt{2\pi}} \int_t x(t) h(\tau - t) e^{-j\omega t} dt \\ X(\omega, \Omega) &= \frac{1}{\sqrt{2\pi}} \int_\tau |x(\omega, \tau)| e^{-j\Omega\tau} d\tau \end{aligned} \tag{4}$$

In this ω denotes the frequencies of the STFT (ranging from 0 to half-Nyquist frequencies), t the time, τ the center-time of the STFT windows and Ω the frequencies of the second Fourier Transform (which upper frequency depends on the hop-size of the STFT).

[1] `http://profs.sci.univr.it/~desena/FMT/`

In [19], it is proposed to use 18 GammaTone filters to create $x(\omega, \tau)$ instead of the STFT. The bands of the GammaTone filters are centered on a log-space from 26 to 9795 Hz.

2.3. Modulation Scale Spectrum

We propose here the Modulation Scale Spectrum as an extension of the Scale Transform to allow its application to individual frequency bands. It therefore allows to take the benefits of the Modulation Spectrum while ensuring the *scale invariance* property.

It is expressed as

$$D(\omega, c) = \frac{1}{\sqrt{2\pi}} \int |x(\omega, e^\tau)| e^{\frac{1}{2}t} e^{-jc\tau} d\tau \tag{5}$$

where ω is defined as above and c is the scale (as in eq. (2)).

2.4. Modulation Scale Spectrum for rhythm description

We propose a specific implementation of the Modulation Scale Spectrum to describe the rhythm content of a music signal. In this, $x(\omega, \tau)$ does not represent the signal itself but the auto-correlation of an onset-energy function[2] derived from the signal in a specific frequency band ω.

The flowchart of its computation is indicated into Figure 1 and described below.

1. As in [19], we first separate the audio signal $x(t)$ using 4^{th} order bandpass GammaTone filters centered on a log-space from 26 to 9795 Hz. In our case, we use 12 filters instead of the 18 proposed by [19][3].
2. We then calculate an onset-energy function $o(\omega, \tau)$ on the output of each filter. For this, we used the function proposed by Ellis [26][4]. We parametrized it to have a sampling rate of 50 Hz.

[2] A onset-energy-function is a function taking high values when an onset is present and low values otherwise.

[3] For the GammaTone filters, we used the implementation kindly proposed by Ma [25] whose code is available at `http://staffwww.dcs.shef.ac.uk/people/N.Ma/resources/gammatone/`.

[4] `http://labrosa.ee.columbia.edu/projects/coversongs/`

3. We then perform a frame-analysis with an 8 s. length rectangular window and a 0.5 second step. We denote the frames by u.
4. As in [15], we then compute on each frame the autocorrelation of each $o(\omega, \tau)$. We denote the resulting functions by $R_{xx}(\omega, \tau, u)$.
5. Finally we compute the Modulation Scale Spectra on each autocorrelation functions $R_{xx}(\omega, \tau, u)$. We denote it by $D(\omega, c, u)$.
6. In order to obtain a single representation for the whole audio signal, we compute the average value of $D(\omega, c, u)$ over u. We denote it by $D(\omega, c)$. The resulting dimensionality is $[12, 2000]$.
7. In order to reduce this dimensionality, we could group the values of the various scale-bands c (as McKinney and Whitman did for the FFT frequencies). Another possibility starts from the consideration that most of the information of the Modulation Scale Spectrum is contained in the first scale coefficients c. Using this, Holzapfel in [15] only kept the first 40 coefficients of the Scale-Transform. In our case, we keep the first 60 scale coefficients ($c \in [1, 60]$) of the Modulation Scale Spectrum.

3. APPLICATION OF THE MODULATION SCALE SPECTRUM FOR RHYTHM RECOGNITION

In order to validate our Modulation Scale Spectrum we apply it to the task of rhythm recognition.

3.1. Test-set

For this, we use a test-set annotated into classes of rhythm: the Ballroom test-set. This one contains 698 titles divided into 8 music genres. Because in ballroom music, the genre ("ChaChaCha", "Rumba" ...) are closely related to the type of rhythm pattern, we consider this genre labels as classes of rhythm. This test-set was created for the ISMIR 2004 rhythm description contest [27]. It is extracted from the website `www.ballroomdancers.com`.

3.2. Experimental scenario

In the first experimental scenario, the task consists in recognizing the correct rhythm class.

Ballroom rhythm classes have a predominant tempo (i.e. ChaChaCha tracks are usually around 125bpm, QuickStep around 200bpm ...). Which means that it would be possible to recognize the rhythm classes simply by using the tempo of the track (if tempo==200 then class=QuickStep). This has been shown for example by Gouyon [7] who reported 82.3 % accuracy using only annotated tempo and a k-Nearest Neighbours (k-NN) classifier. Because of this, it is difficult to highlight the benefits of using a tempo-invariant-feature such as the Modulation Scale Spectrum. In order to show this benefit, we created a second experimental scenario. Considering that our classification algorithm is a KNN, in the second scenario we will ignore the tracks that have a tempo within 4 % of the tempo of the target[5].

[5]It should be noted that Viennese Waltz genre has been discarded in this experiment since all its tempi are within a 5 % range.

3.3. Classifier

As explained in part 2.4, each track is represented by its Modulation Scale Spectrum $D_i(c, \omega)$ as described. For the two scenario described above, we use a modified K-Nearest-Neighbor to perform the classification.

As showed in eq. (3), the Scale Transform is scale-invariant ignoring a global multiplication factor $\sqrt{\alpha}$. Because of this factor, we use the one-minus-cosine-distance[6] in the K-NN instead of the usual Euclidean distance.

If we denote by i the target point of the K-NN, and by $j \in [1, K]$ the K nearest-points, we assign to each j the following weight: $w_{i,j} = 1 - \frac{d_{i,j}}{d_{K+1}}$. We then attribute a global score to each class by summing the weights $w_{i,j}$ of the points j that belong to the class. The class with the largest score is then assigned to i. This modified K-NN algorithm has been found superior to the equally weighted K-NN by Holzapfel [14]. The best value of K and C (the number of scale coefficients) have been found by performing a grid-search.

The results presented in the following has been obtained using 10-fold cross-validation.

3.4. Results and discussions

Results are indicated into Table 1. In this table, we compare the performances obtained by our method to the ones obtained by Holzapfel [15] and Peeters [13]. The sign "-" denotes the fact that the result is not available for the given configuration. In order to be able to compare our algorithm to the one of Holzapfel in the second scenario (Exp. 2), we re-implemented Holzapfel method. Surprisingly, this re-implementation provides better results than the ones published by the author[7].

For the first scenario (Exp. 1), our proposed Modulation Scale Spectrum (93.12% accuracy) outperforms state-of-the-art methods by more than 5 %. Holzapfel method obtains 86.9 % and Peeters 87.96 %.

For the second scenario (Exp. 2), which remove tracks closely-related in tempo from the K-NN, our proposed Modulation Scale Spectrum (75.52% accuracy) outperforms state-of-the-art methods by more than 9 %. Jensen [11] reports 48.4% with his "tempo-insensive rhythmic pattern", our re-implementation of Holzapfel reaches 66.48%. This second experiment demonstrates that our Modulation Scale Spectrum is able to capture rhythmic content, without being dependent on the tempo.

On Fig 2, we represent the first 60 coefficients corresponding to the 9-th GammaTone filters[8]of the Modulation Scale Spectrum for the 698 tracks of the Ballroom test-set. Each row of the matrix represents the 60 coefficients of our Modulation Scale Spectrum for a given track. As can be seen, the various tracks are visually clustered according to the rhythm-classes (represented on the y-axis). Each rhythm class has a dominant pattern. Unsurprisingly, Waltz and Viennese Waltz have close patterns where most of their energy is concentrated on the first coefficient.

[6]$d_{i,j} = 1 - \frac{D_i \cdot D_j}{|D_i||D_j|}$

[7]This may be due to a different split of the test-set into ten folds.

[8]It should be noted that we couldn't display all the Modulation Scale Spectrum for the 12 frequency bands, so we selected the 9^{th} GammaTone band as an example. All other bands show similar behavior.

Table 1: *Results and parameters (best values of C and K) for the recognition of the ballroom rhythm classes.*

Method	Exp. 1			Exp. 2		
	Accuracy	C	K	Accuracy.	C	K
Jensen	-	-	-	48.4 %		1
Holzapfel	86.9 %	40	5	-	-	-
Holzapfel (re-implemented)	87.82 %	40	11	66.48 %	20	5
Peeters	87.96 %	-	-	-	-	-
Modulation Scale Transform	**93.12 %**	**60**	**5**	**75.52** %	**20**	**5**

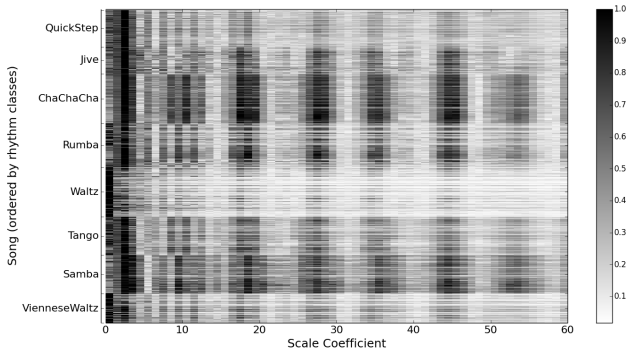

Figure 2: *First 60 coefficients (x-axis) of the Modulation Scale Spectrum for all the songs of the Ballroom test-set. The songs of the test-set are grouped by rhythm-classes on the y-axis. The coefficients are normalized : their amplitudes are represented in gray scale (black for a normalized amplitude of 1, white for 0)*

4. CONCLUSION

In this paper, we proposed the Modulation Scale Spectrum as an extension of both the Modulation Spectrum and the Scale Transform. It takes the benefits from the Modulation Spectrum while ensuring *scale invariance*. We used the Modulation Scale Spectrum to create a rhythm description feature. The whole process includes GammaTone filtering, onset-strength energy estimation, auto-correlation of those and finally Scale Transform. We tested this feature for a rhythm-class recognition task. We demonstrated that for this task our proposed feature largely exceeds results obtained so far. With a second experiment we showed that this feature is indeed tempo-independent. In future works, we will extend our use of the Modulation Scale Spectrum as audio representation for other MIR tasks such as genre classification.

Acknowledgements

This work was founded by the French government Programme Investissements d'Avenir (PIA) through the Bee Music Project.

5. REFERENCES

[1] Cyril Joder, Slim Essid, and Gaël Richard, "Temporal integration for audio classification with application to musical instrument classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 174–186, 2009.

[2] Jonathan Foote and Shingo Uchihashi, "The beat spectrum: A new approach to rhythm analysis.," in *ICME*, 2001.

[3] Jonathan Foote, Matthew L Cooper, and Unjung Nam, "Audio retrieval by rythmic similarity," in *ISMIR*, 2002.

[4] Mark A Bartsch and Gregory H Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 96–104, 2005.

[5] Iasonas Antonopoulos, Aggelos Pikrakis, Sergios Theodoridis, Olmo Cornelis, Dirk Moelants, and Marc Leman, "Music retrieval by rhythmic similarity applied on greek and african traditional music," in *ISMIR*, 2007.

[6] Simon Dixon, Fabien Gouyon, Gerhard Widmer, et al., "Towards characterisation of music via rhythmic patterns.," in *ISMIR*, 2004.

[7] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proceedings of the AES 25th International Conference*. Citeseer, 2004, pp. 196–204.

[8] Simon Dixon, "n interactive beat tracking and visualisation system," in *Proceedings of the international computer music conference*, 2001, pp. 215–218.

[9] Jouni Paulus and Anssi Klapuri, "Measuring the similarity of rhythmic patterns.," in *ISMIR*, 2002.

[10] Matthew Wright, W Andrew Schloss, and George Tzanetakis, "Analyzing afro-cuban rhythms using rotation-aware clave template matching with dynamic programming.," in *ISMIR*, 2008, pp. 647–652.

[11] Jesper Højvang Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen, "A tempo-insensitive representation of rhythmic patterns," in *Proc. of the 17th European Signal Processing Conf.(EUSIPCO-09)*. Citeseer, 2009.

[12] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[13] Geoffroy Peeters, "Spectral and temporal periodicity representation of rhythm for the automatic classification of music audio signal," *IEEE*, 2011.

[14] Andre Holzapfel and Yannis Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2217–2220.

[15] Andre Holzapfel and Yannis Stylianou, "A scale transform based method for rhythmic similarity of music," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 317–320.

[16] André Holzapfel and Yannis Stylianou, "Scale transform in rhythmic similarity of music," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 176–185, 2011.

[17] L. Worms, *Reconnaissance d'extraits sonores dans une large base de donnees*, Practical lessons, Ircam, 1998.

[18] Xavier Rodet, Laurent Worms, and Geoffroy Peeters, "Method for characterinz a sound signal," July 11 2003, WO Patent 2,003,056,455.

[19] Martin F McKinney and Jeroen Breebaart, "Features for audio and music classification.," in *ISMIR*, 2003, vol. 3, pp. 151–158.

[20] Brian Whitman and Daniel PW Ellis, "Automatic record reviews," in *ISMIR 2004: 5th International Conference on Music Information Retrieval: Proceedings: Universitat Pompeu Fabra, October 10-14, 2004*. Audiovisual Institute, Pompeu Fabra University, 2004, pp. 86–93.

[21] Les Atlas and Shihab A Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 668–675, 2003.

[22] Leon Cohen, "The scale representation," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3275–3292, 1993.

[23] Antonio De Sena and Davide Rocchesso, "A fast mellin and scale transform," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[24] Joakim Andén and Stéphane Mallat, "Multiscale scattering for audio classification.," in *ISMIR*, 2011, pp. 657–662.

[25] Ning Ma, Phil Green, Jon Barker, and André Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication*, vol. 49, no. 12, pp. 874–891, 2007.

[26] Daniel PW Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[27] Beesuan Ong, Xavier Serra, Sebastian Streich, Nicolas Wack, Pedro Cano, Emilia Gómez, Fabien Gouyon, and Perfecto Herrera, "Ismir 2004 audio description contest," 2006.