Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description

Ugo Marchand, Geoffroy Peeters

Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description.

Ugo MARCHAND, Geoffroy PEETERS

September 30, 2016

Abstract

In this paper we propose two novel scale and shift-invariant time-frequency representations of the audio content. Scale-invariance is a desired property to describe the rhythm of an audio signal as it will allow to obtain the same representations for same rhythms played at different tempi. This property can be achieved by expressing the time-axis in log-scale, for example using the Scale Transform (ST). Since the frequency locations of the audio content are also important, we previously extended the ST to the Modulation Scale Spectrum (MSS). However, this MSS does not allow to represent the inter-relationship between the audio content existing in various frequency bands. To solve this issue, we propose here two novel representations. The first one is based on the 2D Scale Transform, the second on statistics (inspired by the auditory experiments of McDermott) that represent the inter-relationship between the various frequency bands. We apply both representations to a task of rhythm class recognition and demonstrates their benefits. We show that the introduction of auditory statistics allows a large increase of the recognition results.

2D-Fourier, 2D-Scale, Fourier-Mellin Transform, auditory statistics, rhythm description

1 Introduction

The two cornerstones of automatic music description based on audio analysis are: – extracting meaningful information from the audio signal (audio descriptor extraction) and – performing an efficient mapping between this information and the concept to be estimated (classification algorithm)[1].

In this paper, we propose two novel audio descriptors which aim at representing the time and frequency energy pattern of an audio signal independently of its scale (in the case of music the scale is the tempo). Those are based on

[1]While deep learning methods tend to bring both together, carefully designed audio descriptors are still necessary when a very large amount of training data is not accessible.

the Scale Transform and designed specifically to capture the rhythm characteristics of a music track. Indeed, rhythm, along harmony (melody) and timbre (orchestration) are the three most important perspectives to describe the music content.

The rhythm of a track is usually described by a pattern played at a specific tempo. The pattern includes the characteristics of the meter, the specific accentuation, micro-timing, ... We therefore need a representation which is independent of the tempo (scale-invariant) and of the position of the beginning of the pattern (shift-invariant). There has been several proposals related to shift and scale invariant representation of an audio signal: log-time axis [15], the Scale Transform [13] or our own proposal of Modulation Scale Spectrum (MSS) [18].

However none of them allow taking into account how the different rhythmic events relate to each other in the frequency domain. As an example, let's consider the following pattern [x.o.x.o.], where 'x' is a kick, 'o' a hi-hat, '.' a rest and 'xo' a kick and a hi-hat played simultaneously. In [15, 13], since there is no frequency representation, there will be no difference between [x.o.x.o.] and [x.x.x.x.] or [o.o.o.o.]. In [18], each different frequency has its own representation but there are no inter-relationship modeled between them. Therefore there will be no difference of representation between [x.o.x.o.] and [xo...xo...].

Proposal. In this paper, we propose two novel audio representations that allows modeling this missing inter-relationship. The first one is based on the application of the Scale Transform along the two dimensions of time and frequency. However, while this 2D representation allows representing the inter-relationship between the various frequency bands, it also produces shift-invariance over frequencies (including invariance to circular rotation of the frequency axis) which is not a desired property. Because of this unwanted property, we propose a second representation which uses statistics (inspired by the auditory experiments of McDermott [20, 19]) to represent the inter-relationship between the various frequency bands.

Potential uses. Potential uses of these representations are the search by rhythm pattern (for example looking for identical rhythm patterns without being affected by the tempo) or the automatic classification into rhythm-classes. These representations would also benefit to any genre, mood classification or search by similarity systems that include rhythm description. Applications of these representations outside the music field (i.e. when the scale is not the tempo) also concern generic audio recognition.

Paper organization. The paper is organized as follow. In section 2, we first review related works and highlight the differences with previous proposals. We then introduce in section 3 the Scale Transform along the two dimensions of time and frequency, and propose two methods to build a rhythm-content descriptor that takes into account frequency inter-relationship (section 4). We then evaluate the two novel representations in a task of rhythm class recognition and discuss the results obtained (section 5).

2 Related works

In this section, we briefly review existing methods to represent the rhythm of an audio signal. We also review the two works that inspired our current proposal.

The first set of methods proposed to represent rhythm are based on shift-invariant (but not scale-invariant) representations. For example, Foote [8] and Antonopoulos [1] derive rhythm descriptors from a *Self-Similarity-Matrix*, Dixon [5] samples the onset-energy-function at specific tempo-related intervals to obtain a *Rhythmic patterns*, Wright [25] does the same as Dixon for afro-cuban music clave pattern templates, Tzanetakis [24] proposes the *Beat histogram*, Gouyon [9] proposes a set of 73 descriptors derived from the tempo, the *periodicity histogram* and the *inter-onset-interval histogram*. To obtain tempo invariant descriptors, Peeters [21] proposes to sample a spectral and temporal periodicity representations of the onset-energy-function at specific frequencies related to the tempo. Other approaches use *Dynamic Periodicity Warping* [12] to compare rhythms at different tempi.

A second set of methods uses shift and scale (tempo in the case of music) invariant representations, usually the **Scale Transform** (ST) [4]. Holzapfel et al. [13] were the first to propose the use of the ST for rhythm representation. It should be noted that the method proposed by Jensen [15], while not mentioning the ST, follows a close path. In these works, the ST is applied to the auto-correlation (used to add shift-invariance) of the onset-energy function. This method was also used by Prockup et al. [22] which apply a Discrete Cosine Transform to the ST coefficients to reduce its dimensionality. Since there is only one onset-energy-function for the whole set of frequencies, this method does not allow representing the frequency location of the rhythm events. For this reason, we proposed in Marchand et al. [18] to model the rhythmic information on multiple perceptual frequency bands with the ST. Since this method can be considered as a Modulation Spectrum in the Scale domain it was named Modulation Scale Spectrum (MSS). We showed that for a rhythm class recognition task, the MSS largely increases the recognition performances.

A first work that inspired our current proposal is [3] that proposes a transform to describe a 2D signal in a shift and scale-invariant way in both directions. This transform is a **2D Scale Transform** (scale-invariance) applied to a 2D-Fourier Transform. The latter is used to obtain the shift-invariance property. This transform is often named the Fourier-Mellin Transform. It is extremely useful for, though not limited to, image processing. It has been introduced by the optical research community in the end of the seventies [3, 26] and has been used in many fields since then: radar 2D signal analysis [11, 14], pattern recognition in image [23, 10]. It has never been used however for audio signal description.

The second work that inspired our current proposal is the one of McDermott et al. [20, 19]. They proposed a set of **auditory statistics** to model and generate sound textures. Theses statistics are based on how the auditory system summarizes temporal information, and involves cross-correlations between many frequency bands. Ellis et al. [7] used these statistics for soundtrack

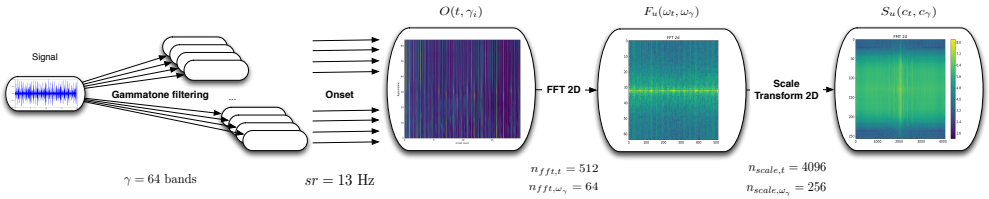

Figure 1: *Flowchart of the computation process of the 2D Modulation Scale Spectrum (2DMSS).*

classification and show a small improvement over the use of Mel Frequency Cepstral Coefficients (MFCC) statistics. It has never been used however for rhythm description.

3 The 1D and 2D Scale Transform

In this section, we introduce the 1D and 2D Scale Transform for audio signal processing, that will be used in the section 4 to build rhythm-content descriptors.

3.1 The 1D Scale Transform

The Scale Transform (ST) is a specific case of the Mellin Transform, which was introduced by Cohen in [4]. For a 1D signal $x(t)$ over time t, the ST at scale c_t is defined as:

$$S(c_t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) t^{-jc_t - \frac{1}{2}} dt \tag{1}$$

Scale-invariance. One of the most important property of the ST is its *scale invariance*. If $S(c_t)$ is the ST of a temporal signal $x(t)$, then the ST of a time-scaled version of this temporal signal $\sqrt{a}\, x(at)$ is $e^{jc_t \ln a} S(c_t)$. Both $x(t)$ and $\sqrt{a}\, x(at)$ have therefore the same modulus of the ST. The ST can viewed as the Fourier Transform of an exponentially re-sampled signal weighted by an exponential window:

$$S(c_t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x(e^t) e^{\frac{1}{2}t} e^{-jc_t t} dt \tag{2}$$

When $x(t)$ represents the audio signal of a music track, the scale correspond to the tempo (the speed at which a rhythm pattern is played). The modulus of the ST is therefore a representation independent of the tempo. This has been used for tempo-invariant rhythm representations by [13, 18, 22].

3.2 The 2D Scale Transform

For a 2D signal $X(t, \omega)$ over time t and frequency ω, the 2D Scale Transform (ST) at scales c_t and c_ω is defined as:

$$S(c_t, c_\omega) = \int_t \left(\int_\omega X(t, \omega) \omega^{-jc_\omega - \frac{1}{2}} d\omega \right) t^{-jc_t - \frac{1}{2}} dt$$

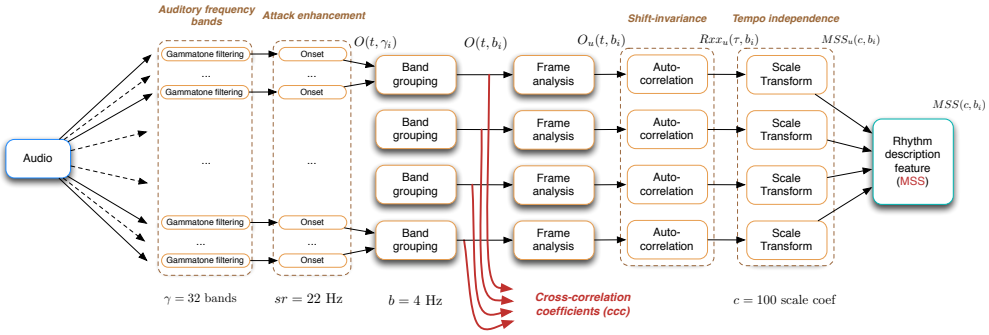

Figure 2: *Flowchart of the computation process of Modulation Scale Spectrum with Auditory Statistics (MASSS).*

As for the 1D ST, the 2D ST can be viewed as the 2D Fourier Transform of an exponentially re-sampled signal on both dimensions $X(e^t, e^\omega)e^{\omega/2}e^{t/2}$.

$$S(c_t, c_\omega) = \int_t \left(\int_\omega \left(X(e^t, e^\omega)e^{\omega/2}e^{t/2} \right) e^{-jc_\omega \omega} d\omega \right) e^{-jc_t t} dt$$

This transform has already been widely used in image processing, to get scale-invariant representations [23, 10], but never in audio signal processing.

3.3 Shift-invariance

It should be noted that neither the 1D nor the 2D Scale Transforms are shift invariant, which means in the 1D case $|S(x(t))| \neq |S(x(t+a))|$. For this reason, when the 1D Scale Transform will be used in part 4.2, we will apply it on a shift-invariant representation (the auto-correlation $Rxx_u(t, b_i)$ of the global-over-frequency onset-energy-function). In the 2D case, when the 2D Scale Transform will be used in part 4.1, it will be applied to a 2D shift-invariant representation (the modulus of the 2D Fourier Transform $F_u(\omega_t, \omega_\gamma)$). In the image processing literature, a 2D Fourier Transform followed by a 2D Scale Transform is named the Fourier-Mellin Transform.

4 Application to rhythm description

We describe here how the Scale Transform (1D or 2D) can be used to build rhythm descriptors. We distinguish 4 descriptors.

Holzapfel. The 1D Scale Transform is applied to the auto-correlation of an onset-energy-function that represents the full frequency range. There is no distinction between the frequency locations of rhythm events. This is the initial method proposed by [13].

MSS. For each frequency band (Gammatone filters) we compute the 1D Scale Transform of the auto-correlation of the onset-energy-function within this band. The frequency locations of rhythm events are represented but independently (no inter-relationships). This is the method we proposed in [18].

2DMSS. In section 4.1, we extend the idea of the MSS but represent the inter-relationship between the frequency bands using the 2D-Scale Transform of the modulus of the 2D-Fourier Transform instead of the independent 1D-Scale Transforms of independent auto-correlation functions.

MASSS. In section 4.2, we also extend the idea of the MSS but simply represent the inter-relationship between the onset-energy-functions using auditory statistics.

4.1 2D Modulation Scale Spectrum (2DMSS)

In this method, we extend the idea of the MSS but represent the inter-relationship between the frequency bands using the 2D-Scale Transform of the modulus of the 2D-Fourier Transform instead of the independent 1D-Scale Transforms of independent auto-correlation functions. The flowchart of the computation process of the 2DMSS is given in Figure 1 and described below.

1. The audio signal $x(t)$ is first separated into 64 Gammatone[2]filters (using 4^{th} order bandpass) centered on a log-space from 26 to 9795 Hz.
2. For each filter output, we calculate an onset-energy function (OEF)[3] using the method of Ellis [6]. This function has a sampling rate of 13 Hz. The OEF of each filter are then stacked into a matrix to form a 2D time/frequency representation $O(t, \gamma_i)$; $i \in [1, 64]$.
3. We then perform a block analysis of $O(t, \gamma_i)$. The block analysis is performed using a 0.5 seconds hop size and a 8 seconds window duration of rectangular shape.
4. For each block u, we compute the modulus of the 2D Fourier Transform of $O_u(t, \gamma_i)$. We denote it by $F_u(\omega_t, \omega_\gamma)$. It has a dimension of $(512, 64)$ (we fixed $n_{fft,t} = 512$ and $n_{fft,\omega_\gamma} = 64$).
5. We then compute the 2D Scale Transform of $F_u(\omega_t, \omega_\gamma)$ denoted by $S_u(c_t, c_\gamma)$. It has a dimension of $n_{scale,t} = 4096$ by $n_{scale,\omega_\gamma} = 256$.
6. We then average $S_u(c_t, c_\gamma)$ over blocks u to obtain $S(c_t, c_\gamma)$.
7. We finally reduce its dimensions by applying Principal Component Analysis (PCA). We only keep the first (with the highest eigenvalues or explained variance) 100 eigenvectors. The final dimension is therefore 100.

4.2 Modulation Scale Spectrum with Auditory Statistics (MASSS)

The previous 2DMSS representation provides a scale and shift invariant representation of the audio content and allows to represent the inter-relationship between the various frequency bands. However, it also produces shift-invariance

[2]These filters model the auditory system. 4th-order Gammatone filters have been shown to be extremely close to the human auditory filters. We used the implementation kindly proposed by Ma [16].

[3]An OEF is a function taking high values when an onset (beginning of a discrete event in the audio signal) is present and low values otherwise.

over frequencies, including shift-invariance when circularly rotating the frequency axis. This property is not desired since it will correspond to consider as equivalent low (kick) and high (hi-hat) patterns. This is the reason why we propose here a second representation which uses statistics (inspired by the auditory experiments of McDermott) to represent the inter-relationship between the various frequency bands.

Auditory statistics. In [20, 19], McDermott and al. show evidence that "the auditory system summarizes temporal details of sounds using time-averaged statistics". They show that, in order to resynthesize sound textures, these statistics should include the statistics of each individual frequency band but should also include the cross-correlations between the temporal energy profiles within each frequency band.

We therefore propose to add to the MSS, the correlations between the onset-energy-functions of the various frequency bands. The flowchart of the computation process of the MASSS descriptor is given in Figure 2 and described below.

1. same as section 4.1 step 1, with 32 Gammatone filters.
2. same as section 4.1 step 2, with a sampling rate of 22 Hz.
3. The number of frequency bands is reduced from 32 to 4 by summing adjacent bands together. The resulting matrix is $O(t, b_i)$ $i \in [1, 4]$.
4. Cross-correlation coefficients $ccc(b_i, b_j)$ are computed between frequency b_i and b_j using $ccc(b_i, b_j) = \sum_k O(t_k, b_i) \cdot O(t_k, b_j)$.
5. For each band b_i, we then perform a frame analysis of $O(t, b_i)$ and compute, for each frame u, its auto-correlation: $Rxx_u(\tau, b_i)$ where τ denotes the time-lag. The frame analysis is performed using a 0.5 seconds hop size and a 8 seconds window duration of rectangular shape.
6. Finally, for each frequency band b_i, we compute the Scale Transform of $Rxx_u(\tau, b_i)$ over τ and average it over frames u. We denote it by $MSS(c, b_i)$ where c is the scale coefficient. We only retain the first 100 coefficient. The dimensionaly of $MSS(c, b_i)$ is therefore $(100, 4)$. We denote by MASSS the concatenation of the $MSS(c, b_i)$ coefficients and the 10 $ccc(b_i, b_j)$ coefficients.

5 Experiments

In this section, we compare the ability to represent rhythm of the proposed descriptors. For this we evaluate their performances for a task of rhythm class recognition.

5.1 Task of rhythm class recognition

The task consists in correctly recognizing the rhythm class of an audio track. For this we use datasets annotated into rhythm classes (see section 5.2). We evaluate the performances of the 2DMSS (section 4.1), the MSS alone, the cross-correlation coefficients ccc alone, and finally both together (MASSS=MSS+ccc) (section 4.2). We compare them to the best results published in [13] and [18].

Table 1: *Results on the 3 different datasets. All the results are in term of mean-over-classes recall (except [13] which is an accuracy). The state-of-the-art results are presented in italic. The results in bold are improving or performing equally with state-of-the-art.*

	Ballroom		Extended Ballroom		Cretan dances	
Method	Result	*parameters*	Result	*parameters*	Result	*parameters*
State-of-the-art	*93,1%* [18]	$\gamma = 12$, $sr = 50$, $c = 100$	-	-	*77.8%* [13]	$sr = 50$, $c = 160$
Proposal: 2DMSS	91.1%	$\gamma = 32$, $n_{fft} = 64; 512$, $sr = 13$, $n_{sc} = 256; 4096$	-	-	63,0%	$\gamma = 32$, $n_{fft} = 64; 512$, $sr = 10$, $n_{sc} = 256; 4096$
Proposal: MSS	**95,1%**	$\gamma = 32$, $b = 4$, $sr = 22$, $c = 100$	94,6%	$\gamma = 32$, $b = 4$, $sr = 22$, $c = 100$	75,6%	$\gamma = 8$, $b = 8$, $sr = 50$, $c = 60$
Proposal: ccc	41,1%	$\gamma = 32$, $b = 4$	31,1%	$\gamma = 32$, $b = 4$	36,6%	$\gamma = 32$, $b = 10$
Proposal: MASSS	**96,0%**	$\gamma = 32$, $b = 4$, $sr = 22$, $c = 100$	**94,9%**	$\gamma = 32$, $b = 4$, $sr = 22$, $c = 100$	**77,2%**	$\gamma = 32$, $b = 10$, $sr = 22$, $c = 40$

For all classification tasks, we use Support Vector Machines (SVM) with a radial basis function kernel. Parameters of the SVM[4] are found using grid-search. The results are presented in term of mean-over-classes recall[5]using 10-fold cross-validation.

5.2 Datasets

Ballroom. The Ballroom dataset contains 698 tracks of 30 seconds divided into 8 music genres ('ChaChaCha', 'Jive', 'QuickStep', 'Rumba', 'Samba', 'Tango', 'VienneseWaltz', 'Waltz'). This dataset was created for the ISMIR 2004 rhythm description contest [2]. It is extracted from the website *www.ballroomdancers.com*. We consider the genre labels as classes of rhythm because in ballroom music, the genre is closely related to the type of rhythm pattern.

Extended Ballroom. Although the Ballroom dataset is relevant for rhythm classification, it suffers from several drawbacks: poor audio quality, small size, presence of duplicates. We therefore decided to update it. Since *www.ballroomdancers.com*

[4]The range of search of gamma is $[10^{-10}; 10^{5}]$ and the range of search of C is $[10^{-10}; 10^{5}]$.

Table 2: *Rhythm repartition of the Ballroom datasets*

Class	Ballroom	Extended Ballroom v1
Chacha	111	455
Foxtrot		507
Jive	60	350
Quickstep	82	497
Rumba	98	470
Samba	86	468
Tango	86	464
Viennesewaltz	65	252
Waltz	110	529
Total	**698**	**3992**

still exists and provides tempo and genre annotations for thousands of 30-second tracks, we extracted all its content again. The new Extended Ballroom dataset is now 6 times larger and has 1 new class 'Foxtrot'. We show in Table 2 the distribution of the number of tracks by rhythm class. It contains 3992 tracks divided into 9 rhythm classes (four additional classes are not displayed since they contained less tracks: Pasodoble (53), Salsa (47), Slowwaltz (65) and Wcswing (23)). The dataset can be found at:

http://anasynth.ircam.fr/home/media/ExtendedBallroom along with the Python scripts used to extract and clean this dataset. The main advantages of the Extended Ballroom dataset over the standard one are: better audio quality, larger size, 1 new rhythm class and repetitions (manual annotations of duplicates, karaoke, repetitions, ...). More details can be found in [17].

Greek dances. The third dataset is the "Greek dances" one. This dataset was kindly provided by the author. It contains 180 excerpts of the following 6 dances commonly encountered in the island of Crete: *Kalamatianos*, *Kontilies*, *Maleviziotis*, *Pentozalis*, *Sousta* and *Kritikos Syrtos*. It was introduced in [13]. This dataset is more challenging than the Ballroom one since: 1) the rhythm classes have a wider tempo distribution hence a good rhythm descriptor will have to be tempo-independent; 2) most rhythm classes share the same meter (All the dances have a $\frac{2}{4}$ meter except *Kalamatianos* which has a $\frac{7}{8}$ meter); hence recognizing the meter will not be sufficient to recognize the classes.

5.3 Discussion of results

In Table 1, we compare the performances obtained by our four rhythm descriptors to the state of the art results of Holzapfel et al. [13] and Marchand et al. [18]. The sign - denotes the fact that the results are not available for the given configuration. All the results are presented in term of mean-over-classes recall, except Holzapfel's which is an accuracy. Along with the mean-recall, we indicate the parameters used for computing our descriptors: γ (number of Gammatone filters), b (final number of frequency band after grouping), sr (sampling rate of the onset-energy function), c (number of scale coefficients kept for each frequency band), n_{fft} and n_{sc} (sizes of the 2D Fourier Transform and 2D ST respectively).

First, it should be noted that the results of [18] on the Ballroom dataset (93.1%) were based on a MSS with many frequency bands. The new results presented as MSS (95.1%) are based on a reduced number of frequency bands which seems beneficial. Over the two proposed new rhythm descriptors (2DMSS and MASSS), only the MASSS succeeded to outperform the MSS descriptor. For the Ballroom dataset, our MASSS descriptor outperforms (96,0%) state-of-the-art methods (93.1%) by 3%. On the Extended Ballroom dataset, our MASSS descriptor scores 94,9%. No comparison with state-of-the-art method is possible since this dataset is new for the research community. While the results obtained on this new dataset are slightly lower than those on the standard

[5]The recall score for a class is the number of correctly detected items over the number of items in this class. The mean-over-class recall is the average of all recall scores of each class.

Ballroom, one should consider that not only the number of files is 5 time larger but also the number of classes is larger (9 over 8). Therefore 94.9% on 9 classes is actually better than 96% on 8 classes. On the Cretan dances dataset, the MASSS descriptor has a mean-recall of 77,2% which is somewhat equivalent to state-of-the-art Holzapfel's accuracy of 77,8%.

6 Conclusion

We proposed two novel audio descriptors (2DMSS and MASSS) that allows representing in a shift and scale invariant way the time and frequency content of an audio signal and differ by the way they model the inter-relationship between the various frequency bands.

The first one, named 2DMSS, is based on the application of the Scale Transform along the two dimensions of time and frequency. This method was not successful and lead to lower scores than our initial results [18]. It can be explained as follow. While this 2D representation allows representing the inter-relationship between the various frequency bands, it also produces shift-invariance over frequency, including invariance when circularly rotating the frequency axis. This means that low and high frequencies can not be distinguished any more, which is not a desired property

For this reason, we proposed a second representation which uses statistics (inspired by the auditory experiments of McDermott) to represent the inter-relationship between the various frequency bands. This second descriptor, named MASSS, provides the new top-results for these datasets. We see that in each of the three experiments, adding the cross-correlation coefficients improves the classification result: 0,9% for the Ballroom, 0,3% for the Extended Ballroom and 1,6% for the Cretan dances dataset. These are promising scores and future works will concentrate on testing MASSS as input to other classification tasks.

References

[1] Iasonas Antonopoulos et al. "Music retrieval by rhythmic similarity applied on greek and african traditional music". In: *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*. 2007.

[2] Pedro Cano et al. "ISMIR 2004 audio description contest". In: *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep* (2006).

[3] David Casasent and Demetri Psaltis. "Scale invariant optical transform". In: *Optical Engineering* 15.3 (1976), pp. 153258–153258.

[4] Leon Cohen. "The scale representation". In: *IEEE Transactions on Signal Processing* 41.12 (1993), pp. 3275–3292.

[5] Simon Dixon, Fabien Gouyon, and Gerhard Widmer. "Towards Characterisation of Music via Rhythmic Patterns". In: *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)*. 2004.

[6] Daniel PW Ellis. "Beat tracking by dynamic programming". In: *Journal of New Music Research* 36.1 (2007), pp. 51–60.

[7] Daniel PW Ellis, Xiaohong Zeng, and Josh H McDermott. "Classifying soundtracks with audio texture features". In: *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, pp. 5880–5883.

[8] Jonathan Foote and Shingo Uchihashi. "The Beat Spectrum: A New Approach To Rhythm Analysis." In: *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME)*. 2001.

[9] Fabien Gouyon et al. "Evaluating rhythmic descriptors for musical genre classification". In: *Proceedings of the25th Audio Engineering Society International Conference (AES)*. 2004, pp. 196–204.

[10] AE Grace and Michael Spann. "A comparison between Fourier-Mellin descriptors and moment based features for invariant object recognition using neural networks". In: *Pattern Recognition Letters* 12.10 (1991), pp. 635–643.

[11] Guo Gui-Rong, Yu Wen-Xian, and Zhang Wei. "An intelligence recognition method of ship targets". In: *Fuzzy Sets and Systems* 36.1 (1990), pp. 27–36.

[12] Andre Holzapfel and Yannis Stylianou. "Rhythmic similarity of music based on dynamic periodicity warping". In: *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2008, pp. 2217–2220.

[13] André Holzapfel and Yannis Stylianou. "Scale transform in rhythmic similarity of music". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.1 (2011), pp. 176–185.

[14] MR Inggs and AR Robinson. "Neural approaches to ship target recognition". In: *Proceedings of th IEEE International Radar Conference*. IEEE. 1995, pp. 386–391.

[15] Jesper Højvang Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. "A tempo-insensitive representation of rhythmic patterns". In: *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*. IEEE. 2009.

[16] Ning Ma et al. "Exploiting correlogram structure for robust speech recognition with multiple speech sources". In: *Speech Communication* 49.12 (2007), pp. 874–891.

[17] Ugo Marchand and Geoffroy Peeters. "The Extended Ballroom Dataset". In: *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR)*. 2016.

[18] Ugo Marchand and Geoffroy Peeters. "The Modulation Scale Spectrum and its Application to Rhythm-Content description". In: *Proceedings of the 17th International Conference on Digital Audio Effects (Dafx)*. Sept. 2014.

[19] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. "Summary statistics in auditory perception". In: *Nature neuroscience* 16.4 (2013), pp. 493–498.

[20] Josh H McDermott and Eero P Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis". In: *Neuron* 71.5 (2011), pp. 926–940.

[21] Geoffroy Peeters. "Spectral and Temporal Periodicity Representation of Rhythm for the Automatic Classification of Music Audio Signal". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.5 (2011), pp. 1242–1252.

[22] Matthew Prockup et al. "Modeling musical rhythm at scale with the music Genome project". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2015, pp. 1–5.

[23] Yunlong Sheng and Henri H Arsenault. "Experiments on pattern recognition using invariant Fourier–Mellin descriptors". In: *JOSA A* 3.6 (1986), pp. 771–776.

[24] George Tzanetakis and Perry Cook. "Musical genre classification of audio signals". In: *IEEE transactions on Speech and Audio Processing* 10.5 (2002), pp. 293–302.

[25] Matthew Wright, W Andrew Schloss, and George Tzanetakis. "Analyzing Afro-Cuban Rhythms using Rotation-Aware Clave Template Matching with Dynamic Programming." In: *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*. 2008, pp. 647–652.

[26] Toyohiko Yatagai, Kazuhiko Choji, and Hiroyoshi Saito. "Pattern classification using optical Mellin transform and circular photodiode array". In: *Optics Communications* 38.3 (1981), pp. 162–165.