A SCALE TRANSFORM BASED METHOD FOR RHYTHMIC SIMILARITY OF MUSIC

Andre Holzapfel and Yannis Stylianou

Institute of Computer Science, FORTH, Greece,
and Multimedia Informatics Lab, Computer Science Department, University of Crete
{hannover, yannis}@csd.uoc.gr

ABSTRACT

This paper introduces scale transforms to measure rhythmic similarity between two musical pieces. The rhythm of a piece of music is described by the scale transform magnitude, computed by transforming the sample autocorrelation of its onset strength signal to the scale domain. Then, two pieces can be compared without the impact of tempo differences by using simple distances between these descriptors like the cosine distance. A widely used dance music dataset has been chosen for proof of concept. On this data set, the proposed method based on scale transform achieves classification results as high as other state of the art approaches. On a second data set, which is characterized by much larger intra-class tempo variance, the scale transform based measure improves classification compared to previously presented measures by 41%.

Index Terms— Rhythm, similarity, music, information retrieval

1. INTRODUCTION

The detection of rhythmic similarity of musical pieces plays an important role in many applications. In many of these applications similar pieces can be expected to differ widely regarding their tempi. In traditional forms of music, performances of the same style of dance can have very different tempi [1]. Because of that, automatic classification of traditional dances has to be robust to tempo variability. A second example: Composers often search large databases for accompaniments that fit to a chosen drum sample, where the tempo could be adjusted to the drum if an accompaniment is found satisfying. Thus, retrieval of rhythmically similar samples without impact of their tempo differences is desirable.

Previous approaches for measuring rhythmic similarity usually lack the robustness to tempo changes. For example, in [2] a cosine measure between beat spectra is used to measure rhythmic distances. These beat spectra are computed using self similarity between the Fourier transform at different time lags. This measure is shown to work well within a narrow range of tempo variation only. Other approaches make the estimation of tempo necessary, as for example in [3], or demand the estimation of meter organization like in [4], where the tactus is estimated and rhythmic patterns are extracted from the time signal. The patterns are then warped for comparison. The estimation of tempo or meter organization can be considered to be error-prone especially for signals without strong percussive content, as shown in [5].

The difficulty of estimating meter structure or tempo from a piece imposes the usage of descriptors that do not consider the order and positioning of sound events. Such an approach has been presented in [2] by using beat spectra. To improve the robustness to tempo changes when using such descriptors, in [6] periodicity spectra have been computed from onset strength signals [7] and have been used in a method referred to as Dynamic Periodicity Warping (DPW). There, a matrix of point wise distances between periodicity spectra is computed, and a minimum cost warping path through this matrix is found. This path is compared to an ideal warping path to get a distance measure. In [8], warping with different kind of step criteria than in [6] is applied to periodicity representations derived from self similarity measures; thereafter simply the cost of the warping is taken as distance.

In this paper, an approach based on scale transforms [9] is shown to further improve results. From a music signal an onset strength signal is derived similar to [7]. This onset strength signal has high values at moments of large positive changes in the STFT magnitude spectrum of the signal. The sample autocorrelation is then computed from the onset strength signal. It is important to point out that given the same piece of music performed at two different tempi the two sample autocorrelation vectors derived from the onset strength signals mainly differ by a scaling factor. Thus, their scale transform magnitudes (STM) will be very similar [10]. We show that using simple pointwise distance measures (*i.e.* cosine distance) between the STM's the robustness to tempo differences of rhythmic similarity detection is considerably improved. Moreover, the suggested method is much faster than the method suggested in [6]. To our best knowledge, scale transforms have been applied to music signals only for digital audio effects [10].

Two datasets are used for evaluation: The first dataset, D1, has been used in the ISMIR Rhythm Description contest[1] and contains eight classes of ballroom dances. This dataset has also been used in [11]. The second dataset, D2, consists of six different traditional dances commonly encountered in the island of Crete (Greece). This data set is an extension of the data set presented in [6]. Two differences between D1 and D2 are related to instrumentation and tempo: D2 contains almost only string instruments without percussion, which makes the automatic meter analysis a very difficult task as shown in [12]. Furthermore, D2 has a much higher intra-class tempo variability than D1. The validity of the proposed method will be judged based on the classification accuracies on the two data sets using a modified kNN classifier.

The next Section will give a short outline of the scale transform and its characteristics. Section 3 explains the computation of the STM's from audio signals. Section 4 overviews the distance measures that are going to be compared in this paper. Section 5 describes the datasets, pointing out the characteristics of D2. Section 6 gives the classification results for both data sets. Section 7 concludes the paper.

[1] http://www.iua.upf.es/mtg/ismir2004/contest/rhythmContest/

2. SCALE TRANSFORM

As a special case of the Mellin transform, the scale transform has been introduced by Cohen in [9]. This transform is scale-invariant, which means that the magnitude distributions of the scale transform of the signals $x(t)$ and $\sqrt{a}x(at)$ are equal, for a being some number bigger than zero. The scale transform is defined as

$$X(c) = \frac{1}{2\pi} \int_0^\infty x(t) e^{(-jc-1/2)\ln t} dt \tag{1}$$

and as depicted in [10], changes in scale of a signal change only the phase of the scale transform, while the magnitude remains the same. The computation of the scale transform can be done efficiently by using its relation to the Fourier transform, which is referred to as Fast Mellin Transform (FMT). The FMT of $x(t)$ can be computed as [10]

$$X(c) = \int_{-\infty}^{\infty} x(e^t) e^{1/2t} e^{-jct} dt \tag{2}$$

which is the Fourier transform of the exponentially warped signal weighted by an exponential window.

3. SCALE TRANSFORM FEATURE COMPUTATION

The first step in computing scale transform magnitudes (STM) from an audio sample is the computation of an onset strength signal $o(n)$ [7]. In the next step, some representation of the salient periodicities must be found. In [6], STFT of the onset strength signals have been computed, referred to as periodicity spectra. In Figure 1, a periodicity spectrum of a drum beat is shown in bold lines, while the periodicity spectrum of its time scaled version is depicted in dotted lines. It is important to note, that the immediate computation of a point wise distance between these spectra is affected by the time scaling. To cope with this, a warping of periodicity spectra was suggested in [6]. In this paper, we suggest the scale transform to address this problem. A representation of the signal in scale domain is derived from the sample autocorrelation, computed as

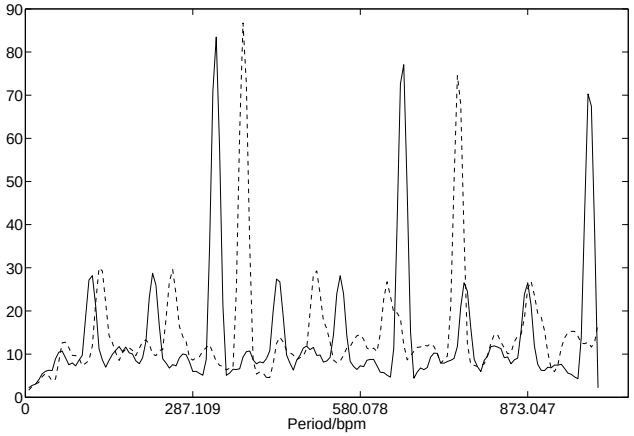

Fig. 1. Periodicity spectra of original (bold) and time scaled (dashed) drum beat

$$\hat{R}_{xx}(m,w) = \sum_{n=0}^{N-m-1} o(n+m+wH)o(n+wH) \tag{3}$$

with N being the number of samples contained in an eight second window, w being the index of the analysis frame, and H the number of samples corresponding to half a second. From each of the autocorrelation frames an FMT is computed as in (2). As the tempo of sound samples is considered not to vary strongly within their duration, a whole sound sample can be efficiently represented by the mean of the derived STM's for sample i, which will be referred to as $X_i(c)$. In Figure 2, the mean STM's derived from the sample autocorrelations of the drum beat examples used in Figure 1 are shown. It is evident that the two time scaled drum patterns have essentially the same STM. Thus, scale representations as depicted in Figure 2 can be compared directly without the necessity of warping. For measuring the distance between two scale representations $X_i(c)$ and $X_j(c)$ of songs i and j the cosine distance

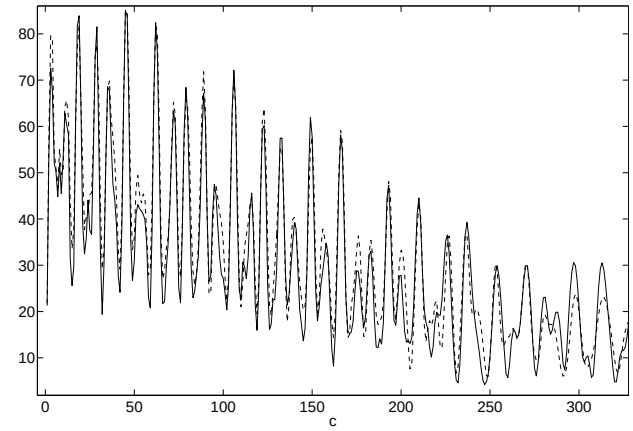

Fig. 2. Scale transform magnitudes of original (bold) and time scaled (dashed) drum beat

$$d_{scale} = 1 - \frac{X_i(c) \cdot X_j(c)}{|X_i(c)||X_j(c)|} \tag{4}$$

has been chosen, because it was found superior to Euclidean distance applied to similar representations in [2] and [6].

Preliminary experiments have been conducted in order to evaluate the number of scale coefficients to be used for the rhythmic similarity measure. For this, different guitar and drum sound samples have been changed in tempo by up to 15% in the audio editor *audacity*, thus forming pairs of an original and a time scaled version. Then, distance matrices have been computed and it was checked, for which number of scale coefficients the pairs can be recognized best from the distance matrices. In general, a number of less than 20 coefficients does not result in a meaningful representation. Using more than 100 coefficients degrades results, because higher scale coefficients tend to have decreasing amplitudes as can be seen in Figure 2. The number of coefficients to use was set to 40. This number of coefficients will be used throughout the following Sections.

The parameters of the onset strength signal computation had to be changed compared to the parameters proposed by Dan Ellis for the following reason: Ellis proposes a hop size of 4ms, which results in an onset strength signal with a sampling frequency f_{ons} of 250Hz. This means that the first $k = 1 : 12$ coefficients of the autocorrelation are related to periodicities of $250Hz/k$. These frequencies are much higher compared to the frequency range that is significant for the rhythmic properties of a sample (below 20 Hz). Because of that, f_{ons} was set to $50Hz$, resulting in only two coefficients related to periodicities larger than $20Hz$, while maintaining sufficient time

accuracy of the Fourier transform applied in the computation of the onset strength signal $o(n)$. This change is important, because scale transform is sensitive to the content of the first values of its input and a misalignment caused by signal properties unrelated to rhythm may corrupt the results.

4. DISTANCE MEASURES

Starting from an onset strength signal sampled at f_{ons} the corresponding autocorrelation vectors and periodicity spectra are computed as representations of the salient periodicities in the signal. They are computed using the same window sizes and hop sizes: eight second rectangular windows with a half second shift, as explained in Section 3. As mentioned in the introduction, a direct comparison of periodicity spectra or sample autocorrelation vectors will be problematic between rhythmically similar pieces with different tempi. In order to shed light on this problem, the mean of periodicity spectra, $\bar{P}_x^i(f)$, and the mean autocorrelation function, $\bar{R}_{xx}^i$, of song i are computed. Then, the corresponding cosine distances, $d_{cos}(P)$ and $d_{cos}(R)$, are used to compare a pair of these representations. Note that the cosine measure has been shown to be optimal in the absence of large tempo variation in [2]. To improve the robustness in presence of tempo changes, the similarity measure d_{DPW} based on Dynamic Periodicity Warping (DPW) is also computed as described in [6].

The aforementioned measures will be compared with the method proposed in this paper: The cosine distance d_{scale} using the scale representations, as denoted in (4).

5. DATA SETS

The first data set, D1, used in this paper has been used in [3, 11, 13], among others. It contains 698 songs from eight different styles of ballroom dances. This data set appears to be easy, because dances vary little in tempo within one class, and classes can be widely distinguished only by using tempo information [13]. In [3], accuracies of about 78% are reported when using the hand annotated bpm values as features for classification.

This is very different for the second data set, D2. It contains short excerpts of six dances commonly encountered in the island of Crete: Kalamatianos, Siganos, Maleviziotis, (fast) Pentozalis, Sousta and Syrtos. In Figure 3, the tempo annotations conducted by the authors have been modelled by Gaussians. The rate of fourth notes has been taken as tempo. It can be seen that the variations of tempi within one class are larger than for the ballroom data (compare with Figure 1 in [13]), and that there are large overlaps between their tempo distributions. When considering that the dance Syrtos is often transcribed in notes of double length, this overlap gets even larger with the tempo distribution of Syrtos moving from the left part of the Figure to the right, creating a distribution overlapping with all dances except of Siganos. These overlaps in tempo make D2 appear a more difficult data set than the collection of Greek music used in [8], where the four classes do not overlap. Furthermore, all traditional Cretan dances have a $\frac{2}{4}$ meter, only Kalamatianos as a dance originating from a different part of Greece has a $\frac{7}{8}$ meter. This makes the separation by considering their meter impossible as well. Also, most of the pieces contain only two kinds of string instruments, while percussive instruments are not contained in most samples, creating a very homogeneous data set considering instrumental timbre. Thus, it can be concluded that a distance metric for D2 has to be robust to tempo changes and cannot rely on tempo or meter characteristics. It has to be able to detect periodicities that are characteristic for the rhythmic pattern of the specific kind of dance. Preliminary listening tests on a randomly chosen subset of D2 which contained 90 songs, support the assumption of the difficulty of the task: Even though all six subjects were able to dance each of the six dances, they labeled only about 76% of the samples correctly.

As detailed in [6], the length of the samples guarantees that they contain at least two repetitions of the music segments characteristic for the dance. Thus, they should be sufficient both for human listeners and for a computational approach, to detect present similarities in their rhythm. Recently, the research presented in [14] shed light on the difficulty of understanding music structured as the samples in D2: it is following a different kind of morphology, the logic of parataxis [14], and a successful method to compute rhythmic similarities in this type of music can provide a useful tool for computational ethnomusicology [15]. Data set D2 will be accessible for interested researchers on request.

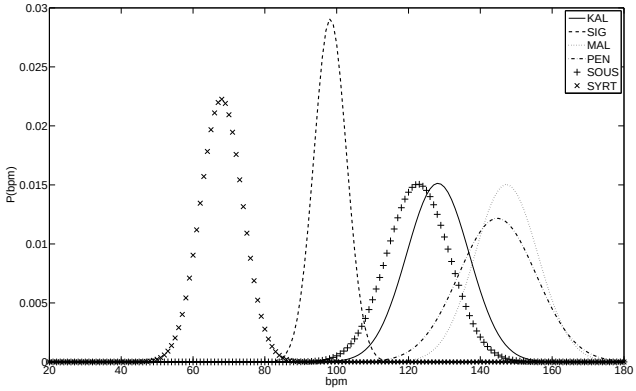

Fig. 3. Tempi of dances in D2, modelled by Gaussian distributions

6. EXPERIMENTS

The accuracies of a modified k-Nearest Neighbor (kNN) classification have been determined by running ten repetitions of a 10-fold cross validation. As shown in [6], a locally weighted kNN was found to improve accuracies on similar data, and therefore it has been used in the experiments. It assigns weight $w_i = 1 - (d_i/d_{k+1})$ to the i-th training sample, where d_{k+1} is the distance of the $k+1$-nearest neighbor to the test sample. Thus, training samples more far away from the test sample, contribute less to its classification.

Table 1. Classification Accuracies on D1

	$d_{cos}(P)$	$d_{cos}(R)$	d_{DPW}	d_{scale}
D1	**86.9**	86.2	84.0	85.1
D2	55.0	44.5	61.7	**77.4**

Table 1 shows the classification accuracies on the two datasets, using the measures as described in Section 4. Similar to the results presented in [6], there is a slight advantage of the cosine measures $d_{cos}(P)$ and $d_{cos}(R)$ on D1. However, compared to the highest accuracy without the usage of the tempo annotations of 85.7% presented in [11] on this data set, the accuracy presented here using d_{scale} is only slightly worse while the best accuracy of 86.9% is better. The improvement of $d_{cos}(P)$ in comparison to [6] must be assigned to the changed sample rate of the onset strength signal

which in general improved results throughout the experiments. The good performance of the cosine measure can be assigned to the small range of intra-class tempo variation [2]. Summing up, the high classification accuracy using d_{scale} for D1 confirm the validity of the proposed scale transform based distance measure.

On D2, Table 1 shows a considerable advantage for the proposed scale distance measure d_{scale}: on this data set it outperforms the cosine measures by $40.7\%/73.9\%$. This clear improvement can be assigned to the robustness to tempo changes of the distance measure. The scale distance measure also improves compared to the previous presented d_{DPW} [6]. Another advantage is related to computational load: While in DPW there is the need to compute a warping path using dynamic programming, the most time consuming operation in the scale distance measure is the FFT.

In Table 2 a confusion matrix of the experiment resulting in the best score of the scale distance d_{scale} (77.4%) on D1 is shown. The row denotes the correct class and the column the assignment by the classifier. The only two classes that are confused often are the dances Maleviziotis and the dance fast Pentozalis. This has been observed to be the most difficult distinction in our listening tests as well. Both dances have a similar range of tempo and are characterized by a similar style of dancing. The other dances are classified more easily; The dances Syrtos and Siganos have the best classification results. These dances have isolated tempo distributions and can also be easily recognized by their scale representations.

Table 2. Confusion matrix for D2

	Kal.	Sig.	Mal.	Pen.	Sous.	Syrt.
Kal.	232	0	0	14	32	22
Sig.	0	289	0	0	11	0
Mal.	0	0	206	83	1	10
Pen.	4	10	79	171	16	20
Sous.	11	12	19	12	239	7
Syrt.	11	0	10	10	6	263

As mentioned in Section 3, the number of used scale coefficients has been set to 40. To confirm the correctness of this choice, the number of coefficients has been varied on both data sets. The accuracies depicted in Table 3 show that there is little change when varying the number of coefficients in the range between 20 and 100. As it is expected, using a few coefficients leads to a strong decrease in accuracy. While the decrease towards higher number of coefficients appears less strong, this decrease can be assigned to the spurious peaks in the higher scale bands of the STM's.

Table 3. Accuracies for different number of scale coefficients

N_{coeff}	10	20	40	70	100	200	400
D1	70.1	82.8	85.1	84.0	84.0	83.6	83.6
D2	65.0	77.5	77.4	76.2	73.3	71.8	70.0

7. CONCLUSION

A novel method for rhythmic similarity in presence of large tempo variations has been introduced. It is based on scale transforms, which makes it robust to tempo changes while it is not time consuming in terms of computational load. By using a data set that is not separable by using tempo, meter, or other kind of information (*e.g.* instrumentation) but the periodicities related to the style of dance, the validity of the method for similarity of rhythm is demonstrated. Further research on this subject are the application to music of other regions, and the usage of the scale respresentations in a more sophisticated classification framework.

8. REFERENCES

[1] Irene Loutzaki, "Audio report: Greek folk dance music," *Yearbook for traditional music*, vol. 26, pp. 168–179, 1994.

[2] Jonathan Foote, Matthew D. Cooper, and Unjung Nam, "Audio retrieval by rhythmic similarity," in *Proc. of ISMIR 2002 - 3rd International Conference on Music Information Retrieval*, 2002.

[3] Geoffroy Peeters, "Rhythm classification using spectral rhythm patterns," in *Proc. of ISMIR 2005 - 6th International Conference on Music Information Retrieval*, 2005, pp. 644–647.

[4] Jouni Paulus and A.P. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. of ISMIR 2002 - 3rd International Conference on Music Information Retrieval*, 2002.

[5] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 14, no. 1, 2006.

[6] Andre Holzapfel and Yannis Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *ICASSP 2008*, 2008.

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

[8] Iasonas Antonopoulos, Angelos Pikrakis, Sergios Theodoridis, Olmo Cornelis, Dirk Moelants, and Marc Leman, "Music retrieval by rhythmic similarity applied on greek and african traditional music," in *Proc. of ISMIR 2007 - 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[9] L. Cohen, "The scale representation," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3275–3292, 1993.

[10] Antonio De Sena and Davide Rocchesso, "A fast mellin and scale transform," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 75–75, 2007.

[11] Simon Dixon, Fabien Gouyon, and Gerhard Widmer, "Towards characterisation of music via rhythmic patterns," in *Proc. of ISMIR 2004 - 5th International Conference on Music Information Retrieval*, 2004.

[12] Andre Holzapfel and Yannis Stylianou, "Beat tracking using group delay based onset detection," in *ISMIR 2008*, 2008.

[13] Fabien Gouyon and Simon Dixon, "Dance music classification: A tempo based approach," in *Proc. of ISMIR 2004 - 5th International Conference on Music Information Retrieval*, 2004.

[14] Haris Sarris, Tassos Kolydas, and Panagiotis Tzevelekos, "A framework of structure analysis for instrumental folk music," in *Proc. of CIM08, 4th Conference on Interdisciplinary Musicology*, Thessaloniki, Greece, 2008.

[15] George Tzanetakis, Ajay Kapur, Andrew Schloss, and Matthew Wright, "Computational ethnomusicology," *Journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 1–24, 2007.