IMPROVING TEMPO-SENSITIVE AND TEMPO-ROBUST DESCRIPTORS FOR RHYTHMIC SIMILARITY

Andre Holzapfel, Arthur Flexer and Gerhard Widmer
Austrian Research Institute for Artificial Intelligence (OFAI)
aholza@inescporto.pt, arthur.flexer@ofai.at, gerhard.widmer@jku.at

ABSTRACT

For the description of rhythmic content of music signals usually features are preferred that are invariant in presence of tempo changes. In this paper it is shown that the importance of tempo depends on the musical context. For popular music, a tempo-sensitive feature is improved on multiple datasets using analysis of variance, and it is shown that also a tempo-robust description profits from the integration into the resulting processing framework. Important insights are given into optimal parameters for rhythm description, and limitations of current approaches are indicated.

1. INTRODUCTION

Determining the similarity between two pieces of music is one of the core problems in Music Information Retrieval (MIR). Methods to estimate such similarity usually consider the timbre of music, *i.e.* the instantaneous sound characteristics that are contained in a sample. Similarity measures based on timbre can be improved by adding the aspect of rhythmic similarity [1]. However, while the meaning of timbre similarity is somehow intuitive, rhythmic similarity is a more abstract concept. In Cooper and Meyer [2], rhythm is defined as the way one or more unaccented beats are grouped in relation to an accented one. Furthermore, meter is defined as the measurement of the number of pulses between more or less regularly occurring accents. Even though rhythm can be perceived without the existence of a meter, in this paper we will restrict to music signals that have a meter. As soon as we impose this restriction, each piece of music is characterized by a frequency of pulsation (*i.e.* a pulse-tempo), that determines how fast the accents in the metrical structure are performed. Thus, in order to achieve high similarity values for similar pieces that are performed at different tempi, one approach is to make descriptions of rhythmic content independent of this pulse-tempo. Such descriptors were *e.g.* proposed by Peeters [3] and Jensen *et al.* [4]. These descriptors are based on periodicity representations: Given a music signal, the periodicities caused by its regularly occurring accents are estimated. Then it is tried to make either these representations or the applied similarity measure between them robust to tempo changes. An important question that will be addressed in this paper is whether such invariance is desirable in every context, or if there are cases in which a certain sensitivity of the descriptors to changing pulse-tempo is of advantage for the rhythmic similarity measurements.

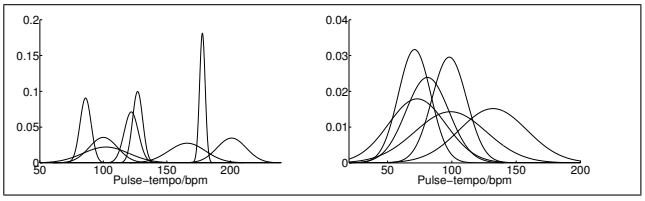

Figure 1. Tempi of the Ballroom (left) and the Turkish Art music (right) datasets modeled by Gaussian distributions.

In order to get an understanding of the significance of this question, let us have a look at two music collections: First, a collection of eight western Ballroom dances that is widely used in experiments in the MIR research community (*e.g.* in [3]), and, second, a collection of Turkish Art music divided into six metric classes that has been compiled by Holzapfel and Stylianou [5]. The pulse-tempo of all pieces is known for these collections. In Figure 1, the pulse-tempo in beats per minute (bpm) of all pieces in each contained class was modeled by a Gaussian distribution. For the Ballroom collection it is obvious that tempo can serve as a valuable information in order to differentiate between samples of different dances, a fact that was observed by Dixon *et al.* [6] for this dataset. The tempo distributions of the Turkish Art music collection, however, reveal opposite conditions for a good similarity measure. On this collection, it appears to be a good choice not to consider tempo information, because distributions have large overlaps and standard deviations.

Thus, depending on the type of music samples we want to compare we would either choose to discard tempo information, or to use it for improving our similarity measure. However, in most cases an annotated ground truth of the pulse-tempo is not given, and it must be estimated from the audio signal. Including estimated instead of annotated tempo information will lead to a decreased performance of the similarity measure, as shown recently by Peeters [3]. This is due to the fact that the tempo estimation is subject to halving- and doubling errors, and its accuracy depends strongly on the signal characteristics [7]. For that reason, it would be desirable to have two types of descriptors at

hand. In the first case, when we want to discard tempo information, a descriptor that completely ignores tempo information would be preferred, as *e.g.* for Turkish and Arabic art music. In the second case, we would prefer a descriptor which remains invariant for a small range of tempo changes, and which automatically varies in presence of larger tempo changes. To give an example, for Hip Hop music one would like to have descriptors that do not vary when the same beat is used in another track with a difference of only 5 beats per minute, but the contained shuffled grooves would appear altered and of different character when changed by 20 bpm.

For that reason, it was chosen to contrast two different techniques for rhythmic similarity estimations. The first was presented by Holzapfel and Stylianou [8] and is based on the Scale Transform Magnitudes (STM). This method was shown to be invariant to tempo changes. The second method was introduced by Pohle *et al.* [1], and applies descriptors that are referred to as Onset Patterns (OP). Large changes in tempo lead to a shift in these descriptors, but small changes in tempo leave this representation almost unchanged as shown in an example in Section 3. For both descriptors, no estimation of the pulse-tempo from the signal is necessary.

In this paper, with the availability of multiple datasets, it was feasible to conduct a series of analyses of variance (ANOVA) [9] in order to find improved parameters for rhythm descriptors. Improvements are related to optimal multi-band processing schemes, length of applied analysis windows, and the resolution which is necessary to obtain a good similarity descriptor. Our experimental setup can serve as an example of how to obtain optimal system parameters when several data sources are given. Until now, such parameters are usually found in a trial and error procedure, and not in a rigorous statistical setting as in our contribution.

Optimal parameters will be obtained by performing ANOVA on the OP computation, but it will be shown that the STM based rhythm descriptors profit from the obtained system improvement in the same way. This confirms that the found processing framework is generalizable and can be applied to other descriptors as well. We will then contrast the performance of the OP and STM descriptors on various datasets in order to verify the correctness of our hypothesis about the context-dependent meaning of pulse-tempo for rhythmic similarity.

The following Sections of this paper are structured as follows: In Section 2 experimental methods are detailed. Datasets are described, it is detailed how conclusions about the accuracy of descriptors are obtained, and STM and OP descriptors will be outlined, with emphasis on the method to improve the OP framework. Then, in Section 3, the different degree of robustness to tempo changes of OP and STM descriptors will be clarified in some examples. The results of the analyses of variance and comparisons between STM and OP features are given in Section 4, and Section 5 concludes the paper.

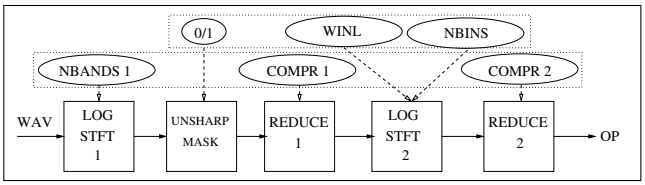

Figure 2. OP computation and system parameters.

2. EXPERIMENTAL SETUP

2.1 Rhythm Content Description

In Figure 2 the computation of Onset Patterns (OP) is summarized. The computations are symbolized by the bold rectangular boxes, and the dotted rectangular boxes show the parameters that will be evaluated in separate ANOVAs. As indicated by the dotted boxes, parameters are grouped into two sets. The first set, `NBANDS1`, `COMPR1` and `COMPR2` are related to the multi-band processing. The second set, `WINL`, `NBINS` and the usage of unsharp masking are related to the OP computation. A simultaneous analysis of those factors and all their interactions would be too challenging. We believe that factors in the two parameter sets are sufficiently independent to be improved separately. The rather coarse grid of factor levels (see Section 4) is also due to considerations of tractability. In the following paragraph, the computation of OP will be described and the meaning of the mentioned parameters will be explained.

Input to the first computation in Figure 2 is a monophonic piece of music sampled at 22050 Hz. The input is transformed into the frequency domain using a STFT with a 46.4 ms length Hanning window with half overlap. The magnitude of the transform is then processed by a filterbank in order to obtain coefficients on a logarithmic axis. The number of bands on this axis is denoted as `NBANDS1` in Figure 2, and was set to 85 by Pohle *et al.* [1]. In each of these bands, a masking can be computed in order to accentuate instrument onsets by emphasizing transient regions in the signal. This masking applies a moving average filter with a length of 0.25 s to each band and then half-wave rectifies the output. In this paper we will retain the notation of unsharp masking for this process which was used in [1]. Then the logarithm of the signal is computed and the `NBANDS1` bands can be reduced by the factor `COMPR1`. In [1], 85 bands were reduced to 38, which results in a compression of $\texttt{COMPR1} = 85/38 \approx 2.24$. Then, a second STFT is computed on each band in order to obtain a description of the periodicities contained in this band. Such a description will be referred to as periodicity spectrum. The periodicity spectral magnitudes are mapped onto a logarithmic axis by applying a filter bank. In this computation, it was decided to evaluate the optimal analysis window length of the STFT (`WINL`) and the number of bins per octave that are obtained from the filter bank, the original values were 6s and 5 bins per octave [1]. The periodicities are described in five octaves from 30 to $960bpm$. It should be pointed out that no zero padding was used in the STFT's, and a Hanning window of `WINL` length in seconds with a shift of half a second was applied to obtain the periodicity spectra. In the final stage of the OP computation,

it was tried to reduce the number of bands again, in order to obtain more compact descriptors. This results in a two stage compression scheme, starting from NBANDS1 bands. The rhythm of a whole sample is described by the mean of the OP obtained from the various segments of this sample.

A method that is robust to tempo variance in a very wide range is the description based on Scale Transform Magnitudes (STM) as proposed by Holzapfel and Stylianou in [8]. The computation of these descriptors was left exactly as explained therein, and its basic computation steps are depicted in Figure 3. The first step is a computation of a spectral flux based Onset Strength Signal (OSS). Within moving windows of eight seconds length, autocorrelation coefficients are computed and then transformed into the scale domain by applying a discrete Scale Transform. For one sample, the mean of Scale Transform Magnitudes (STM) obtained from all the analysis windows are the STM descriptors of the rhythmic content of a sample. For the exact computation parameters please refer to [8]. However, it should be pointed out that the final descriptors do not contain separate information from various bands as for the OP. In order to improve the existent descriptors, the two pa-

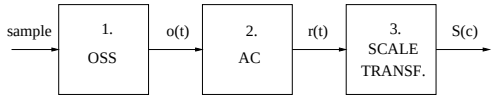

Figure 3. Computational steps of STM rhythm descriptors.

rameter groups in the OP computation depicted in Figure Figure 2 are evaluated in a two stage analysis of variance. In the first stage, the optimal parameters for the multi-band parameter set (NBANDS1, COMPR1, COMPR2) are evaluated in an ANOVA with the observations being the 1-nearest-neighbor classification accuracies on three datasets using Onset Patterns. For this, the parameters from the second set were set to the values applied in [1] (usage of unsharp masking, WINL = 6s, NBINS = 5). After deciding on the values for the parameters for the first set, these values are fixed and optimal values for the second set are found using a second ANOVA again with the observations being the accuracies on the same three datasets using Onset Patterns. Then, the improved processing framework for OP will be applied to STM in order to prove the validity of the obtained parameters also for these descriptors.

2.2 Datasets and evaluation

In order to improve the system depicted in Figure 2, three data sets will be used. The first, DBall, is the widely used Ballroom dataset, consisting of 8 classes with 698 ballroom dance excerpts of 30s length. The second dataset, DLat, was presented by Silla *et al.* [10], and contains 3226 files of Latin dance music in 10 classes. Finally, a third dataset, DPop, was compiled that consists of 347 excerpts from popular music samples organized into 15 different classes that are related to rhythmic concepts (*e.g. Break Beat* and *Jive*). In order to investigate the different demands on the context of traditional music, two more datasets will be used to compute similarity measurements. The first, DCrete, was used by Holzapfel and Stylianou [8] and contains 180 short excerpts of six different dances commonly encountered in the island of Crete in Greece. The second traditional dataset, DTurk, contains 288 audio samples synthesized from melodies of Turkish art music. The pulse-tempo distributions of this dataset are shown in Figure 1, and further details on the dataset and the synthesis method are given in [8].
As the application for the proposed features is music similarity, the features will be evaluated in a 1-Nearest-Neighbor classification in a *leave-one-out* scheme. The distance between features will be Euclidean distance in all cases. The obtained classification accuracies will be used to find an optimal computation setup, and will serve as a way to contrast the performance of different features when applied to music of different style.

3. TEMPO ROBUSTNESS

In order to show the influence of tempo changes on the OP and STM descriptors, a simple experimental setting was chosen. From each class of DBall, DLat and DPop one song was chosen and its tempo was manipulated without changing pitch using the audacity audio editor. The tempo of each song was changed by +/-20%, +/-10%. This results in five tempo variants for each song, including the original tempo. Following this procedure, 22 songs in five tempo variants were obtained. Note that for classes which appear in several datasets (*e.g. Tango*), only one sample was used. The accuracy of correctly identifying a song in a 1-NN classification was determined. This means that it was determined how often the nearest neighbor is indeed a tempo changed version. Additionally, the average ratio of the distances between a song and all different songs and distances between a song and its tempo variations was computed. For example, if this ratio equals 2, the distance of one song to a different song is on average two times larger than the distance of a song to its tempo variants. Hence, larger numbers of this ratio indicate a better robustness to the tempo changes. The applied features are the OP and STM with the original parameters as presented in [1] and [8], respectively. The results shown in Table 1 clearly show the supe-

Table 1. Song identification in presence of tempo changes

	OP	STM
ACCURACY	70.0	83.6
RATIO	2.26	3.03

rior tempo robustness of the STM features, both in terms of ratio and in terms of accuracies. However, we should have a closer look at the effect of small tempo changes on the OP features. This effect is visualized in Figure 4, where the low coefficients of a periodicity spectral magnitude of a Cha-cha-cha sample is shown as a bold line. This piece has

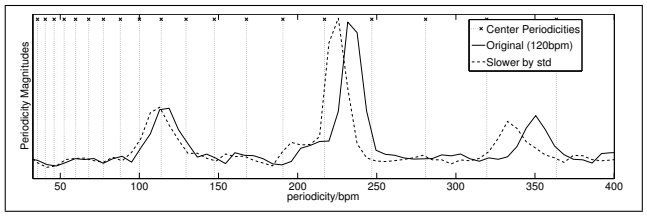

Figure 4. Excerpt of periodicity spectral magnitude of a Cha-cha-cha sample. Small changes in tempo lead to minimal change in OP due to the log-filterbank (center frequencies shown as dotted vertical lines).

an annotated tempo of 120 bpm, and the standard tempo deviation of this class in DBall is 5.6bpm. The dashed periodicity magnitude has been derived from the same piece, when its tempo has been changed by this standard deviation using audacity. The dotted vertical lines denote the positions of the log-filterbank center frequencies that map the periodicity spectral magnitudes to a logarithmic axis (LOG-STFT2 in Figure 2). It is obvious that this change in tempo leads to a minimal change in the resulting descriptors due to the coarse frequency resolution. This confirms that the OP descriptors are robust to tempo changes within certain limits that are determined by the NBINS parameter in Figure 2.

4. RESULTS

As explained in Section 2, two ANOVAs were performed as indicated by the dotted boxes in Figure 2. The results will be analyzed starting with the multi-band processing scheme.

4.1 Multi-band Processing ANOVA

We performed a four-way analysis of variance (ANOVA) with the following factors: "Number of bands" (NBANDS1, 4 levels: 16, 32, 64, 128), "Compression 1" (COMPR1, 3 levels: 1, 2, 4), "Compression 2" (COMPR2, 3 levels: 1, 2, 4), "Data set" (DS, 3 levels: DBall, DLat, DPop). We also looked into possible two-factor interactions. The dependent variable is the accuracy resulting from 1-Nearest-Neighbor classification. As can be seen in Table 2, all main effects as well as two-factor interactions are significant at the .05 error level (see last column, Prob>F smaller than 0.05).

Table 2. Result table of multi-band ANOVA

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
NBANDS1	0.0369	3	0.0123	76.82	8.469e-22
COMPR1	0.0104	2	0.005	32.42	1.296e-10
COMPR2	0.0037	2	0.0018	11.6	4.624e-05
DS	0.5807	2	0.2903	1810.14	1.079e-59
NBANDS1*COMPR1	0.0179	6	0.0029	18.64	1.099e-12
NBANDS1*COMPR2	0.0052	6	0.0008	5.48	1.125e-04
NBANDS1*DS	0.0405	6	0.0067	42.17	4.803e-21
COMPR1*COMPR2	0.0031	4	0.0007	4.86	0.0017
COMPR1*DS	0.0115	4	0.0028	18.03	3.916e-10
COMPR2*DS	0.0081	4	0.0020	12.68	9.152e-08
Error	0.0109	68	0.0001		
Total	0.7294	107			

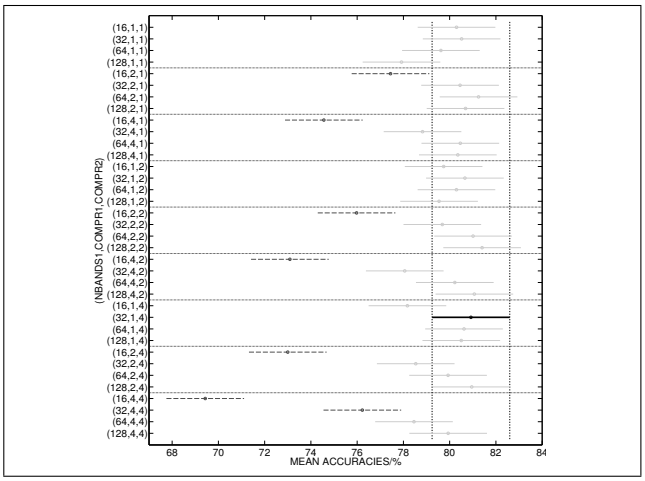

Figure 5. Influence of the number of bands and the compression factors. Chosen parameter shown as a bold line, the significantly different means are depicted by dashed lines. Additional horizontal lines improve legibility.

Therefore there are significant effects on the classification accuracy caused by the number of frequency bands, the first and second compression rates as well as the data set. The fact that the type of data set used has an influence on the accuracies achieved is clear since the three data sets have different levels of difficulty (mean accuracies for the datasets are 87.1% (DBall), 80.1% (DLat) and 69.3% (DPop)). Since all two-factor interactions are also significant, we have to investigate all factors together to find out which combinations of factors are optimal in terms of achieved accuracy. It is important to point out that the two-factor interactions with the datasets (DS) in this as well as in the second ANOVA only influence the degree but not the direction a factor has on the dependent variable. Therefore we can analyze aggregate results across all three datasets. In Figure 5 we plotted mean accuracies and 95% confidence intervals for all combinations of factors "Number of bands" (NBANDS1), "Compression 1" (COMPR1) and "Compression 2" (COMPR2). The mean accuracies are based on the results from all three data sets. A considerable number of combinations of factors is able to achieve similar levels of mean accuracy of around 80% and more. We concentrate on one combination that achieves good accuracy and compact representation at the same time: NBANDS = 32, COMPR1 = 1, COMPR2 = 4, *i.e.* this combination uses a log-filterbank with 32 filters after the first STFT, and reduces the resulting number of bands to eight in the second reduction in Figure 2. The periodicities in each of these eight bands are described using 25 coefficients (5 NBINS $\times$ 5 octaves). This specific combination is shown using a bold line in Figure 5. Based on the results from the ANOVA, we compare the mean accuracy for this one combination with all other combinations with a series of t-tests (level of significance $\alpha = .05$). Tukey's HSD adjustment was used to account for the effect of multiple comparisons. All combinations significantly different from the chosen combination are shown as dashed lines. These combinations start

at a low number of bands (NBANDS1=16), and then further reduce this representation, except one case in which starting with 32 bands and reducing them to 2 bands leads to significant decrease. Compared to the chosen scheme, no other higher dimensional combination can significantly improve the results. This shows that a number of bands much smaller than the number of semitone bands (85) is sufficient. This number can be further compressed to obtain a more compact descriptor; A lower bound for the number of bands to start with is at about 32, and a lower bound of bands to keep at the end is 4.

At this point it should be pointed out that instead of the second reduction in Figure 2, also usage of a two dimensional DCT was considered, which resulted in the Onset Coefficients proposed in [1]. The first DCT reduces the number of bands, while the second DCT reduces dimensionality of the periodicity content description in every band. However, it was found that by using DCT the dimensionality of periodicity content description cannot be further reduced, and application of a DCT to the dimension of the bands leads to no performance gain compared to our simple reduction based on linear combination of neighboring bands. Moreover, when applying a DCT results are no longer nicely interpretable as log-periodicity spectra, and for those reasons it appears to be preferable to refrain from using Onset Coefficients.

4.2 Processing parameter ANOVA

We performed a four-way analysis of variance (ANOVA) with the following factors: "Unsharp mask" (`MASK`, 2 levels: 0, 1), "Window length" (`WINL`, 4 levels: 6, 8, 10, 12), "Number of bins" (`NBINS`, 4 levels: 3, 4, 5, 6), "Data set" (`DS`, 3 levels: DBall, DLat, DPop). We also looked into possible two-factor interactions. The dependent variable is again the accuracy resulting from 1-Nearest-Neighbor classification. As can be seen in Table 3, all main effects as well as the two-factor interactions "MASK*DS" and "NBINS*DS" are significant at the .05 error level. There

Table 3. Result table of processing parameter ANOVA

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
MASK	8224.3	1	8224.28	5495	2.2382e-58
WINL	171.5	3	57.15	38.19	1.1558e-13
NBINS	239.7	3	79.89	53.38	1.4701e-16
DS	6772.5	2	3386.23	2262.48	5.0402e-55
MASK*WINL	10.9	3	3.63	2.43	0.0747
MAKS*NBINS	10.6	3	3.53	2.36	0.0811
MASK*DS	671.2	2	335.61	224.23	9.7103e-28
WINL*NBINS	18.5	9	2.05	1.37	0.2229
WINL*DS	29.5	6	4.92	3.29	0.0075
NBINS*DS	50.9	6	8.49	5.67	0.0001
Error	85.3	57	1.5		
Total	16284.8	95			

is a strong positive effect of using the "Unsharp mask" on the accuracy in all tested combinations of parameters. To be precise, mean accuracies improved by 22.4% for DBall, 10.9% for DLat, and 22.1% for DPop. In Figure 6 we plotted mean accuracies and 95% confidence intervals for all levels of factor "Window length" (`WINL`). The mean accuracies are based on the results from all three data sets. The

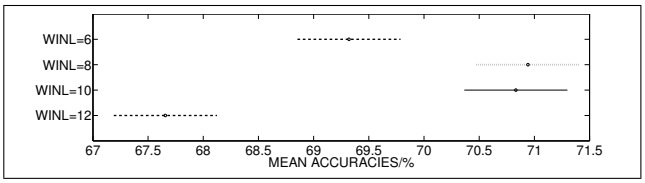

Figure 6. Influence of WINL in the LOG-STFT 2. Chosen parameter shown as a dotted line, the significantly different means are depicted by dashed lines.

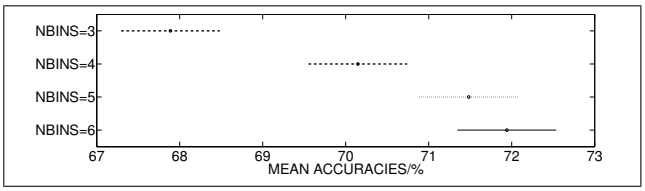

Figure 7. Influence of NBINS in the LOG-STFT 2. Chosen parameter shown as a dotted line, the significantly different means are depicted by dashed lines.

result for using a window length of `WINL` = 8 is shown as a dotted line Figure 6. Using `WINL` = 8 is significantly better than using `WINL` = 6 or 12 and equally good as using `WINL` = 10 (based on t-tests, $\alpha = .05$, Tukey's HSD adjustment). In Figure 7 we plotted mean accuracies and 95% confidence intervals for all levels of factor "Number of bins" (`NBINS`). The mean accuracies are based on the results from all three data sets. The result for using `NBINS` = 5 number of bins is shown as a dotted line Figure 7. Using `NBINS` = 5 is significantly better than using `NBINS` = 3 or 4 and equally good as using `NBINS` = 6 (based on t-tests, $\alpha = .05$, Tukey's HSD adjustment).

The most important conclusion from the processing parameter ANOVA concerns the effect of the window length. By using STFT, we are bound to the stationarity constraint. In this paper, the finding is that increasing window lengths to values of more than 8s leads to problems. This phenomenon was described in [8] as well, but on partly different datasets and using different descriptors. However, the common aspect is that the processed signal had to be stationary within their analysis window as well. This leads to the conclusion that rhythmic aspects of music performances tend to be non-stationary beyond this temporal limit.

4.3 Summary and comparison

In order to quantify the performance gain that is achieved when using the optimal parameters, we will contrast the accuracies of the improved features (OP_{opt}) with the original setting from [1], denoted as OP_{org}. The exact parameters of the improved and the original setups are listed in Table 4.

In Table 5.(a), accuracies on all five datasets using the OP features are depicted, while Table 5.(b) shows the accuracies for the STM_{org} features computed as described in Section 2 together with the STM_{opt}, which were obtained by integrating the STM computation into the optimized multi-band processing scheme: Instead of the logarithmic

Table 4. Comparison of system parameters

Parameter	Improved	Original
NBANDS1	32	85
COMPR1	1	2.2
COMPR2	4	1
MASK	1	1
WINL	8	6
NBINS	5	5

filterbank applied in the LOG-STFT 2 step in Figure 2, we input the linear axis periodicity spectral magnitudes into a Discrete Scale Transform and keep the magnitude. All the other processing steps are the same as for OP_{opt} (except of NBINS, which is specific to the OP computation). Bold numbers in Tables 5.(a) and 5.(b) indicate significant changes, at a .05 error level, by applying the improved multiband processing to either STM or OP. Underlined numbers indicate significant differences between the different features, thus comparing either OP_{opt} with STM_{opt} or OP_{org} with STM_{org}.

Table 5. Classification Accuracies

	(a)			(b)	
Dataset	OP_{opt}	OP_{org}		STM_{opt}	STM_{org}
DBall	**88.4**	86.1		84.1	85.1
DLat	81.0	81.8		**79.7**	65.2
DPop	**74.4**	68.6		**70.3**	60.5
DCrete	**70.4**	64.0		**68.2**	61.5
DTurk	46.2	45.2		56.3	58.3

It can be seen that both for OP and STM descriptors, introducing the improved multiband processing leads to significant improvement in three cases (bold numbers in Table 5). Only for DTurk it shows no effect, which is related to the fact that this dataset contains melodies synthesized from MIDI, and is characterized by less spectral diversity than the other datasets. Observing the underlined accuracies in Table 5, it can be seen that only for DTurk there is a significant advantage of the STM over the OP features, whereas OP appear to represent a more accurate similarity measure on the three popular music datasets. The superior performance on the popular music datasets related to the small variance in pulse-tempo in the classes. The small set of Cretan dances has similar standard deviations as the DBall (for exact values refer to [8]), but larger overlaps between distributions. The accuracies on this set are not significantly different for OP and STM descriptors (70.4% and 68.2%, respectively). However, in a musical context where we have to face huge variance of tempo for one and the same rhythmic class, such as in Turkish and Arabic art music, the tempo robustness of STM lead to a significant improvement over OP (56.3% compared to 46.2%).

5. CONCLUSIONS

In this paper, a crucial problem for rhythmic similarity estimation in music was addressed: Depending on the tempo variances inherent in classes of a musical style it is either of advantage to encode larger tempo changes in the descriptors, or to use descriptors that are robust for even large tempo changes. The former case was addressed by descriptors based on Onset Patterns, while for the latter Scale Transform based descriptors were shown to be more adequate. An advantage of both descriptor types is that no tempo estimation has to be performed on the audio signal, which is an error-prone step in almost all styles of music. Another important contribution of this paper is the improvement of system parameters using an analysis of variance (ANOVA). It is shown that the obtained parameters lead to improvements even for different approaches (STM). The conclusions drawn from the ANOVA are related to the numbers of bands to be used for rhythm description, and the limitation of a STFT analysis window length to 8 seconds. This limitation also limits the possible resolution of the OP descriptors, because for a higher resolution (NBINS) longer windows would be necessary. Thus, the stationarity requirement for the STFT limits the possible parameter space of the OP description. A possible approach to explore the effects of going beyond this border is the usage of transforms that can deal with non-stationary signals.

Acknowledgements

This research was supported by the Austrian Research Fund (FWF), project no. Z159 (Wittgenstein Award), and the Vienna Science and Technology Fund (WWTF), project MA09-024 (Audiominer).

6. REFERENCES

[1] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *Proc. of ISMIR*, 2009.

[2] G. Cooper and L. Meyer, *The Rhythmic Structure of Music*. University of Chicago Press, 1960.

[3] G. Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 19, no. 5, pp. 1242–1252, 2011.

[4] J. Jensen, M. Christensen, and S. Jensen, "A tempo-insensitive representation of rhythmic patterns," in *Eusipco*, Glasgow, Scotland, 2009.

[5] A. Holzapfel and Y. Stylianou, "Rhythmic similarity in traditional turkish music," in *Proc. of ISMIR*, 2009.

[6] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proc. of ISMIR*, 2004.

[7] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[8] A. Holzapfel and Y. Stylianou, "Scale transform in rhythmic similarity of music," *IEEE Trans. Speech and Audio Processing*, vol. 19, no. 1, pp. 176–185, 2010.

[9] R.A.Bailey, *Design of Comparative Experiments*. Cambridge, UK: Cambridge University Press, 2008.

[10] C.N. Silla Jr., A.L. Koerich, and C.A.A. Kaestner, "The Latin Music Database," in *Proc. of ISMIR*, 2008.