ON THE EVALUATION OF RHYTHMIC AND MELODIC DESCRIPTORS FOR MUSIC SIMILARITY

Maria Panteli, Simon Dixon
Centre for Digital Music, Queen Mary University of London, United Kingdom
{m.panteli, s.e.dixon}@qmul.ac.uk

ABSTRACT

In exploratory studies of large music collections where often no ground truth is available, it is essential to evaluate the suitability of the underlying methods prior to drawing any conclusions. In this study we focus on the evaluation of audio features that can be used for rhythmic and melodic content description and similarity estimation. We select a set of state-of-the-art rhythmic and melodic descriptors and assess their invariance with respect to transformations of timbre, recording quality, tempo and pitch. We create a dataset of synthesised audio and investigate which features are invariant to the aforementioned transformations and whether invariance is affected by characteristics of the music style and the monophonic or polyphonic character of the audio recording. From the descriptors tested, the scale transform performed best for rhythm classification and retrieval and pitch bihistogram performed best for melody. The proposed evaluation strategy can inform decisions in the feature design process leading to significant improvement in the reliability of the features.

1. INTRODUCTION

With the significant number of music information retrieval techniques and large audio collections now available it is possible to explore general trends in musical style evolution [11, 16]. Such exploratory studies often have no ground truth to compare to and therefore any conclusions are subject to the validity of the underlying tools. In music content-based systems this often translates to the ability of the audio descriptors to correctly and sufficiently represent the music-specific characteristics.

In this study we focus on the evaluation of audio features that can be used for rhythmic and melodic content description and similarity estimation. We are particularly interested in audio features that can be used to describe world music recordings. We propose an evaluation framework which aims to simulate the challenges faced in the analysis of recorded world music collections, such as processing noisy recordings or audio samples exhibiting a variety of music style characteristics. In particular, we define transformations with respect to timbre, recording quality, tempo and key and assess the invariance of a set of state-of-the-art rhythmic and melodic descriptors.

A number of studies have dealt with the evaluation of audio features and specifically of rhythmic and melodic descriptors. Robustness is usually addressed in the design process where certain decisions ensure tempo or key invariance of the features. For example, rhythmic descriptors have been designed to achieve partial [5,6] or complete [7] tempo invariance, and melodic descriptors exist which are tempo and/or key invariant [1,22,24]. Robustness to audio quality has been also addressed for MFCC and chroma features (describing timbre and harmony respectively) [21]. The relevance of the features is not guaranteed even if a classification task seems successful. For example, unbalanced datasets can lead to high accuracies in genre classification tasks [18], or high rhythm classification accuracies can be achieved with (only) tempo information [4, 6] indicating that other audio features used had limited relevant contribution to the task.

To be perceptually valid, and useful in real-world collections, the representations need to be invariant to subtle changes in tempo, key (or reference pitch), recording quality and timbre. Additionally, to be usable in cross-cultural studies, the features need to be agnostic to properties of particular music cultures. For instance, pitch representations should not depend on the 12-tone equal temperament tuning, and rhythm representations should not depend on specific Western metric structures such as $\frac{4}{4}$ metre.

We examine a selection of audio features for rhythm and melody to assess their suitability for scientific studies of world music corpora subject to the above constraints. To achieve this we test classification and retrieval performance of multiple rhythm and melody features on a controlled dataset, which allows us to systematically vary timbre, tempo, pitch and audio quality. The main contributions of the paper are the controlled dataset, which we make freely available, and the proposed evaluation strategy to assess robustness and facilitate the feature design and selection process.

2. FEATURES

We present details of three descriptors from each category (rhythm and melody), chosen from the literature based on their performance on related classification and retrieval tasks. In the paragraphs below we provide a short sum-

Attribution: Maria Panteli, Simon Dixon. “On the evaluation of rhythmic and melodic descriptors for music similarity”, 17th International Society for Music Information Retrieval Conference, 2016.

mary of these features and discuss further considerations of their design. Our implementations of the features follow the specifications published in the corresponding research papers but are not necessarily exact replicas.

2.1 Rhythm

We start our investigation with state-of-the-art rhythmic descriptors that have been used in similarity tasks including genre and rhythm classification [5, 7, 12]. The rhythmic descriptors we use here share the general processing pipeline of two consecutive frequency analyses [15]. First, a spectrogram representation is calculated, usually with frequencies on the mel scale. The fluctuations in its "rows", i.e. the frequency bands, are then analysed for their rhythmic frequency content over larger windows. This basic process has multiple variations, which we explain below.

For comparison purposes we fix the sampling rate at 44100 Hz for all features. Likewise, the spectrogram frame size is 40 ms with a hop size of 5 ms. All frequency bins are mapped to the mel scale. The rhythmic periodicities are calculated on 8-second windows with a hop size of 0.5 seconds. In the second step we compute the periodicities within each mel band then average across all bands, and finally summarise a recording by taking the mean across all frames.

Onset Patterns (OP). The defining characteristic of Onset Patterns is that the mel frequency magnitude spectrogram is post-processed by computing the first-order difference in each frequency band and then subtracting the mean and half-wave rectifying the result. The resulting onset function is then frequency-analysed using the discrete Fourier transform [5, 6, 14]. We omit the post-processing step of transforming the resulting linear fluctuation frequencies to $\log_2$-spaced frequencies. The second frame decomposition results in an $F \times P_O$ matrix with $F = 40$ mel bands and $P_O = 200$ periodicities linearly spaced up to 20 Hz.

Fluctuation Patterns (FP). Fluctuation patterns differ from onset patterns by using a log-magnitude mel spectrogram, and by the additional application of psychoacoustic models (e.g. loudness and fluctuation resonance models) to weight perceptually relevant periodicities [12]. We use the MIRToolbox [8] implementation of fluctuation patterns with the parameters specified at the beginning of Section 2.1. Here, we obtain an $F \times P_F$ matrix with $F = 40$ mel bands and $P_F = 1025$ periodicities of up to 10 Hz.

Scale Transform (ST). The scale transform [7], is a special case of the Mellin transform, a scale-invariant transformation of the signal. Here, the scale invariance property is exploited to provide tempo invariance. When first introduced, the scale transform was applied to the autocorrelation of onset strength envelopes spanning the mel scale [7]. Onset strength envelopes here differ from the onset function implemented in OP by the steps of post-processing the spectrogram. In our implementation we apply the scale transform to the onset patterns defined above.

2.2 Melody

Melodic descriptors selected for this study are based on intervals of adjacent pitches or 2-dimensional periodicities of the chromagram. We use a chromagram representation derived from an NMF-based approximate transcription.

For comparison purposes we fix the following parameters in the design of the features: sampling rate at 44100 Hz, variable-Q transform with 3 ms hop size and pitch resolution at 60 bins per octave (to account for microtonality), secondary frame decomposition (where appropriate) using an 8-second window and 0.5-second hop size, and finally averaging the outcome across all frames in time.

Pitch Bihistogram (PB). The pitch bihistogram [22] describes how often pairs of pitch classes occur within a window d of time. It can be represented as an n-by-n matrix P where n is the number of pitch classes and element p_{ij} denotes the count of co-occurrences of pitch classes i and j. In our implementation, the pitch content is wrapped to a single octave to form a chromagram with 60 discrete bins and the window length is set to $d = 0.5$ seconds. The feature values are normalised to the range $[0, 1]$. To approximate key invariance the bihistogram is circularly shifted to $p_{i-\hat{\imath},j-\hat{\imath}}$ where $p_{\hat{\imath}\hat{\jmath}}$ denotes the bin of maximum magnitude. This does not strictly represent tonal structure but rather relative prominence of the pitch bigrams.

2D Fourier Transform Magnitudes (FTM). The magnitudes of the 2-dimensional Fourier transform of the chromagram describe periodicities in both frequency and time axes. This feature renders the chromagram key-invariant, but still carries pitch content information, and has accordingly been used in cover song recognition [1, 9]. In our implementation, chromagrams are computed with 60 bins per octave and no beat-synchronisation. The FTM is applied with the frame decomposition parameters stated above. We select only the first 50 frequency bins which correspond to periodicities up to 16 Hz.

Intervalgram (IG). The intervalgram [24] is a representation of chroma vectors averaged over different windows in time and cross-correlated with a local reference chroma vector. In the implementation we use,we reduce this to one window size $d = 0.5$, and cross-correlation is computed on every pair of chroma vectors from successive windows.

In this study we place the emphasis on the evaluation framework and provide a baseline performance of (only) a small set of features. The study could be extended to include more audio descriptors and performance accuracies could be compared in order to choose the best descriptor for a given application.

3. DATA

For our experiments we compiled a dataset of synthesised audio, which allowed us to control transformations under which ideal rhythmic and melodic descriptors should be invariant. In the sections below we present the dataset of selected rhythms and melodies and detailed description of their transformations.

Melody		Rhythm	
Description	No.	Description	No.
Dutch Folk (M)	5	Afro-American (M)	5
Classical (M)	5	North-Indian (M)	5
Byzantine (M)	5	African (M)	5
Pop (M)	5	Classical (M)	5
Classical (P)	5	EDM (P)	5
Pop (P)	5	Latin-Brazilian (P)	5

Table 1: The dataset of rhythms and melodies transformed for feature robustness evaluation. (M) is monophonic and (P) polyphonic as described in Section 3.1.

3.1 Material

We compiled 30 melodies and 30 rhythms extracted from a variety of musical styles with both monophonic and polyphonic structure (Table 1). In particular, we collect MIDI monophonic melodies of classical music used in the MIREX 2013: Discovery of Repeated Themes and Sections task[1], MIDI monophonic melodies of Dutch folk music from the Meertens Tune Collections [23], fundamental frequency (F0) estimates of monophonic pop melodies from the MedleyDB dataset [2], fundamental frequency (F0) estimates of monophonic Byzantine religious music [13], MIDI polyphonic melodies of classical music from the MIREX 2013 dataset, and fundamental frequency (F0) estimates of polyphonic pop music from the MedleyDB dataset. These styles exhibit differences in the melodic pitch range, for example, classical pieces span multiple octaves whereas Dutch folk and Byzantine melodies are usually limited to a single octave range. Pitch from fundamental frequency estimates allows us to also take into account vibrato and microtonal intervals. This is essential for microtonal tuning systems such as Byzantine religious music, and for melodies with ornamentation such as recordings of the singing voice in the Dutch folk and pop music collections.

For rhythm we collect rhythmic sequences common in Western classical music traditions [17], African music traditions [20], North-Indian and Afro-American traditions [19], Electronic Dance Music (EDM) [3], and Latin-Brazilian traditions.[2] These rhythms span different metres such as $\frac{11}{8}$ in North-Indian, $\frac{12}{8}$ in African, $\frac{4}{4}$ in EDM, and $\frac{6}{8}$ in Latin-Brazilian styles. The rhythms for Western, African, North-Indian, Afro-American traditions are constructed from single rhythmic patterns whereas EDM and Latin-Brazilian rhythms are constructed with multiple patterns overlapping in time. We refer to the use of a single pattern as 'monophonic' and of multiple patterns as 'polyphonic' for consistency with the melodic dataset.

3.2 Transformations

Intuitively, melodies and rhythms retain their character even if the music is transposed to a different tonality, played at a (slightly) different tempo or under different recording conditions. These are variations that we expect to find in real-world corpora, and to which audio features should be reasonably invariant. Indeed, the cover song identification literature suggests that invariance of features in terms of key transpositions and tempo shifts is desirable [1, 22, 24]; for rhythm description, the existing literature mainly focuses on tempo invariance and robustness against recording quality [5, 6]. We add to this list the requirement of invariance to slight changes in timbre for both melody and rhythm description[3]. Overall, we test our features for robustness in tempo, pitch, timbre and recording quality by systematically varying these parameters to produce multiple versions of each melody and rhythm (Table 2). We apply only one transformation at a time while keeping the other factors constant. The 'default' version of a rhythm or melody is computed using one of the 25 timbres available, fixing the tempo at 120 bpm, and, for melody, keeping the original key as expressed in the MIDI or F0 values. The dataset is made available online[4].

Timbre (Timb): For a given sequence of MIDI notes or fundamental frequency estimates we synthesise audio using sine waves with time-varying parameters. The synthesised timbres vary from harmonic to inharmonic sounds and from low to high frequency range. For a given set of rhythm sequences we synthesise audio using samples of different (mainly percussive) instruments[5]. Beyond the typical drum set sounds (kick, snare, hi-hat), we include percussive instruments of different music traditions such as the Indian mridangam, the Arabic daf, the Turkish darbuka, and the Brazilian pandeiro. Overall, we use 25 different timbres for each melody and rhythm in the dataset.

Recording Quality (RecQ): Large music archives usually contain material recorded under a variety of recording conditions, and are preserved to different degrees of fidelity. We use the Audio Degradation Toolbox [10] to create 25 audio degradations that we expect to be representative of what is found in such archives. Amongst the degradations we consider are effects of prominent reverb (live recordings), overlaid random noise (old equipment), added random sounds including speech, birds, cars (field recording), strong compression (MP3), wow sampling, high or low pass filtering (vinyl or low quality microphone).

Global tempo shifts (GTemp): We define 'small' variations the tempo changes of up to 20% of the original tempo (in this case centred at 120 bpm), which we assume will leave the character of melodies and rhythms intact. In particular, we use 25 tempo shifts distributed in the range $[-20, 20]$ (excluding 0) percent slower or faster than the original speed.

[1] http://www.tomcollinsresearch.net/mirex-pattern-discovery-task.html

[2] http://www.formedia.ca/rhythms/5drumset.html

[3] The timbre transformations we consider are not expected to vastly alter the perception of a rhythm or melody.

[4] https://code.soundsoftware.ac.uk/projects/rhythm-melody-feature-evaluation

[5] http://www.freesound.org

Transformations	Values
Timbre	25 distinct timbres (similar frequency range and instrument)
Rec. Quality	25 degradations including reverb, compression, wow, speech, noise
Global Tempo	25 values in $[-20, 20]$ percent deviation from original tempo
Key Transp.	25 values in $[-10, 10]$ semitones deviation from original key
Local Tempo	25 values in $[-20, 20]$ percent deviation from original tempo

Table 2: Transformations for assessing feature invariance.

Key transpositions/Local tempo shifts (KeyT/LTemp): For melodic descriptor robustness we consider transposing the audio with respect to 25 key transpositions in the range $[-10, 10]$ (excluding 0) semitones from the original key. These shifts include microtonal intervals e.g. a transposition of 1.5 semitones up as one expects to find in world music singing examples. For rhythmic descriptor robustness we consider instead small step changes of the tempo. We introduce a local tempo change for a duration of 2 (out of 8) seconds centred around the middle of the recording. This is common in, for example, performances of amateur musicians where they might unintentionally speed up or slow down the music. Similar to global tempo transformation we use 25 shifts in the range $[-20, 20]$ percent slower or faster than the original speed.

While the above transformations do not define an exhaustive list of effects and variations found in world music corpora they provide a starting point for assessing feature robustness. The dataset can be expanded in future work to include more transformations and parameter values. For this study we restrict to the abovementioned 4 transformations with 25 values each (Table 2). For our dataset of 30 rhythms and 30 melodies this results in a total of 3000 transformed rhythms and 3000 transformed melodies.

4. EVALUATION STRATEGY

With the proposed evaluation strategy we would like to assess feature robustness with respect to the transformations and transformation values presented above in Section 3.2. Additionally we would like to check whether the performance of the features relates to particularities of the music style for the styles presented in Section 3.1. Lastly, since our dataset consists of monophonic and polyphonic melodies and rhythms, we would also like to check whether the features are influenced by the monophonic or polyphonic character of the audio signal.

Robustness evaluation is performed on the dataset of 3000 transformed rhythms and 3000 transformed melodies (Section 3.1). Considering the variety of MIR tasks and corresponding MIR models, we choose to assess feature performance accuracy in both classification and retrieval experiments as explained below. In our experiments we include a variety of classifiers and distance metrics to cover a wide range of audio feature similarity methods.

We first verify the power of the features to classify different melodies and rhythms. To do so we employ four classifiers: K-Nearest Neighbors (KNN) with 1 neighbor and Euclidean distance metric, Support Vector Machine (SVM) with a linear kernel, Linear Discriminant Analysis (LDA) with 20 components, and Gaussian Naive Bayes. We use 5-fold cross-validation for all classification experiments. In each case the prediction target is one of the 30 rhythm or melody 'families'. For each of the 3000 transformed rhythms or melodies we output the classification accuracy as a binary value, 1 if the rhythm or melody was classified correctly and 0 otherwise.

As reassuring as good classification performance is, it does not imply that a melody or rhythm and its transformations cluster closely in the original feature space. Accordingly, we choose to use a similarity-based retrieval paradigm that more directly reflects the feature representations. For each of the 30 rhythms or melodies we choose one of the 25 timbres as the default version of the rhythm or melody, which we use as the query. We rank the 2999 candidates based on their distance to the query and assess the recall rate of its 99 transformations. Each transformed rhythm or melody is assigned a score of 1 if it was retrieved in the top $K = 99$ results of its corresponding query and 0 otherwise. We compare four distance metrics, namely Euclidean, cosine, correlation and Mahalanobis.

For an overview of the performance of the features we compute the mean accuracy across all recordings for each classification or retrieval experiment and each feature. To better understand why a descriptor is successfull or not in the corresponding classification or retrieval task we further analyse the performance accuracies with respect to the different transformations, transformation values, music style and monophonic versus polyphonic character. To achieve this we group recordings by, for example, transformation, and compute the mean accuracy for each transformation. We discuss results in the section below.

5. RESULTS

The mean performance accuracy of each feature and each classification or retrieval experiment is shown in Table 3. Overall, the features with the highest mean classification and retrieval accuracies are the scale transform (ST) for rhythm and the pitch bihistogram (PB) for melody.

5.1 Transformation

We consider four transformations for rhythm and four for melody. We compute the mean accuracy per transformation by averaging accuracies of recordings from the same transformation. Results for rhythm are shown in Table 4 and for melody in Table 5. Due to space limitations we present results for only the best, on average, classifier (KNN) and similarity metric (Mahalanobis) as obtained in Table 3. We observe that onset patterns and fluctuation patterns show, on average, lower accuracies for transformations based on global tempo deviations. This is expected as the aforementioned descriptors are not tempo invariant.

	Rhythm			Melody		
Metric	ST	OP	FP	PB	IG	FTM
Classification						
KNN	**0.86**	0.71	0.68	**0.88**	0.83	0.86
LDA	**0.82**	0.66	0.59	**0.83**	0.82	0.82
NB	**0.80**	0.62	0.58	**0.84**	0.76	0.81
SVM	**0.87**	0.66	0.59	0.86	0.86	**0.87**
Retrieval						
Euclidean	**0.65**	0.47	0.42	**0.80**	0.56	0.67
Cosine	**0.66**	0.47	0.42	**0.80**	0.55	0.68
Correlation	**0.66**	0.47	0.42	**0.80**	0.54	0.67
Mahalanobis	**0.61**	0.48	0.40	**0.81**	0.60	0.72

Table 3: Mean accuracy of the rhythmic and melodic descriptors for the classification and retrieval experiments.

Metric	Feature	Timb	GTemp	RecQ	LTemp
Classification					
KNN	ST	**0.98**	**0.90**	**0.93**	0.62
KNN	OP	0.97	0.20	0.92	**0.75**
KNN	FP	0.91	0.18	0.92	0.71
Retrieval					
Mahalan.	ST	**0.95**	**0.36**	**0.91**	**0.25**
Mahalan.	OP	0.94	0.00	0.88	0.13
Mahalan.	FP	0.62	0.01	0.87	0.09

Table 4: Mean accuracies of the rhythmic descriptors under four transformations (Section 3.1).

In the rhythm classification task, the performance of the scale transform is highest for global tempo deviations but it is lowest for local tempo deviations. We believe this is due to the scale transform assumption of a constant periodicity over the 8-second frame, an assumption that is violated when local tempo deviations are introduced. We also note that fluctuation patterns show lower performance accuracies for transformations of the timbre compared to the onset patterns and scale transform descriptors.

5.2 Transformation Value

We also investigate whether specific transformation values affect the performance of the rhythmic and melodic descriptors. To analyse this we compute mean classification accuracies averaged across recordings of the same transformation value (there are 25 values for each of 4 transformations so 100 mean accuracies in total). Due to space limitations we omit the table of results and report only a summary of our observations.

Onset patterns and fluctuation patterns exhibit low classification accuracies for almost all global tempo deviations whereas scale transform only shows a slight performance degradation on global tempo deviations of around $\pm 20\%$. For local tempo deviations, scale transform performs poorly at large local deviations (magnitude $> 15\%$) whereas onset patterns and fluctuation patterns show higher accuracies for these particular parameters. All descriptors seem to be robust to degradations of the recording quality with the exception of a wow effect that causes all rhythmic descriptors to perform poorly. Onset patterns and fluctuation patterns perform poorly also in the degradation of a radio-broadcast compression.

Metric	Feature	Timb	GTemp	RecQ	KeyT
Classification					
KNN	PB	0.97	**0.99**	**0.78**	0.76
KNN	IG	0.95	**0.99**	0.62	0.77
KNN	FTM	**0.98**	0.96	0.71	**0.79**
Retrieval					
Mahalan.	PB	**0.94**	**0.98**	**0.78**	0.53
Mahalan.	IG	0.70	0.91	0.33	0.46
Mahalan.	FTM	0.87	0.88	0.57	**0.57**

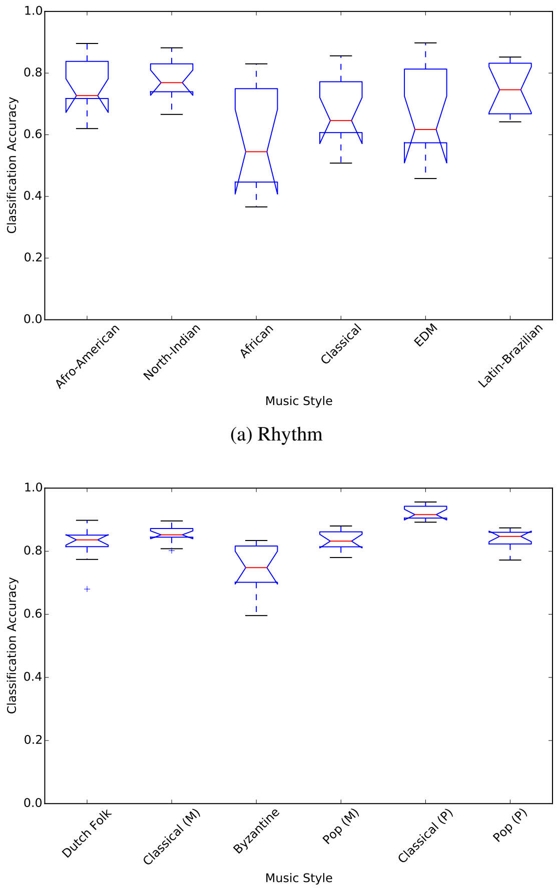
Table 5: Mean accuracies of the melodic descriptors under four transformations (Section 3.1).

For melody classification, all features perform poorly on key transpositions of more than 6 semitones up amd a wow effect degradation. Pitch bihistogram also performs poorly in transpositions between $2.5 - 5$ semitones down. Intervalgram and Fourier transform magnitudes perform badly also for reverb effect degradations and noisy recordings with overlaid wind, applause, or speech sound effects.

5.3 Music Style

Our dataset consists of rhythms and melodies from different music styles and we would like to test whether the robustness of the features is affected by the style. To achieve this we average classification accuracies across recordings of the same style. We have 6 styles for rhythm with 500 recordings in each style and likewise for melody. This gives us 6 mean accuracies for each feature and each classification experiment. We summarise results in a boxplot as shown in Figure 1. We also perform two sets of multiple paired t-tests with Bonferroni correction, one for rhythmic and one for melodic descriptors, to test whether mean classification accuracies per style are significantly different.

Using the paired t-tests with multiple comparison correction we observe that the majority of pairs of styles are significantly different at the Bonferroni significance level $alpha = 0.003$ for both the rhythmic and melodic descriptors. In particular the accuracies for classification and retrieval of African rhythms are significantly different from all other styles. Western classical rhythms are significantly different from all other styles except the EDM rhythms, and North-Indian rhythms are significantly different from all other styles except the EDM and Latin-Brazilian rhythms. For melody, the accuracies for the Byzantine and polyphonic pop styles are significantly different from all other styles. The descriptors that perform particularly badly with respect to these styles are the fluctuation patterns for rhythm and the intervalgram for melody. We use our current results as an indication of which styles might possibly affect the performance of the features but leave the analysis of the intra-style similarity for future work.



(a) Rhythm

(b) Melody

Figure 1: Box plot of classification accuracies of the rhythmic (top) and melodic (bottom) descriptors for each style.

5.4 Monophonic versus Polyphonic

Our dataset consists of monophonic and polyphonic melodies and rhythms and we would like to test whether the performance of the features is affected by the monophonic or polyphonic character. Similar to the preceding analysis, we average performance accuracies across all monophonic recordings and across all polyphonic recordings. We perform two paired t-tests, one for rhythmic and one for melodic descriptors, to test whether mean classification accuracies of monophonic recordings are drawn from a distribution with the same mean as the polyphonic recordings distribution. At the $\alpha = 0.05$ significance level the null hypothesis is not rejected for rhythm, $p = 0.25$, but is rejected for melody, $p < 0.001$. The melodic descriptors achieve on average higher classification accuracies for polyphonic ($M = 0.88$, $SD = 0.02$) than monophonic recordings ($M = 0.82$, $SD = 0.04$).

6. DISCUSSION

We have analysed the performance accuracy of the features under different transformations, transformation values, music styles, and monophonic versus polyphonic structure. Scale transform achieved the highest accuracy for rhythm classification and retrieval, and pitch bihistogram for melody. The scale transform is less invariant to transformations of the local tempo, and the pitch bihistogram to transformations of the key. We observed that the descriptors are not invariant to music style characteristics and that the performance of melodic descriptors depends on the pitch content being monophonic or polyphonic.

We have performed this evaluation on a dataset of synthesised audio. While this is ideal for adjusting degradation parameters and performing controlled experiments like the ones presented in this study, it may not be representative of the analysis of real-world music recordings. The latter involve many challenges, one of which is the mix of different instruments which results in a more complex audio signal. In this case rhythmic or melodic elements may get lost in the polyphonic mixture and further preprocessing of the spectrum is needed to be able to detect and isolate the relevant information.

Our results are based on the analysis of success rates on classification or retrieval tasks. This enabled us to have an overview of the performances of different audio features across several factors: transformation, transformation value, style, monophonic or polyphonic structure. A more detailed analysis could involve a fixed effects model where the contribution of each factor to the performance accuracy of each feature is tested individually.

In this evaluation we used a wide range of standard classifiers and distance metrics with default settings. We have not tried to optimise parameters nor use more advanced models since we wanted the evaluation to be as independent of the application as possible. However, depending on the application different models could be trained to be more robust to certain transformations than others and higher performance accuracies could be achieved.

7. CONCLUSION

We have investigated the invariance of audio features for rhythmic and melodic content description of diverse music styles. A dataset of synthesised audio was designed to test invariance against a broad range of transformations in timbre, recording quality, tempo and pitch. Considering the criteria and analyses in this study the most robust rhythmic descriptor is the scale transform and melodic descriptor the pitch bihistogram. Results indicated that the descriptors are not completely invariant to characteristics of the music style and lower accuracies were particularly obtained for African and EDM rhythms and Byzantine melodies. The performance of the melodic features was slightly better for polyphonic than monophonic content. The proposed evaluation framework can inform decisions in the feature design process leading to significant improvement in the reliability of the features.

8. ACKNOWLEDGEMENTS

MP is supported by a Queen Mary Principal's research studentship and the EPSRC-funded Platform Grant: Digital Music (EP/K009559/1).

9. REFERENCES

[1] T. Bertin-Mahieux and D. P. W. Ellis. Large-scale cover song recognition using the 2D Fourier transform magnitude. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 241–246, 2012.

[2] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 155–160, 2014.

[3] M. J. Butler. *Unlocking the Groove*. Indiana University Press, Bloomington and Indianapolis, 2006.

[4] S. Dixon, F. Gouyon, and G. Widmer. Towards Characterisation of Music via Rhythmic Patterns. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 509–516, 2004.

[5] T. M. Esparza, J. P. Bello, and E. J. Humphrey. From Genre Classification to Rhythm Similarity: Computational and Musicological Insights. *Journal of New Music Research*, 44(1):39–57, 2014.

[6] A. Holzapfel, A. Flexer, and G. Widmer. Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity. In *Proceedings of the 8th Sound and Music Computing Conference*, pages 247–252, 2011.

[7] A. Holzapfel and Y. Stylianou. Scale Transform in Rhythmic Similarity of Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):176–185, 2011.

[8] O. Lartillot and P. Toiviainen. A Matlab Toolbox for Musical Feature Extraction From Audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[9] M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Transactions on Multimedia*, 10(8):1617–1625, 2008.

[10] M. Mauch and S. Ewert. The Audio Degradation Toolbox and Its Application to Robustness Evaluation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 83–88, 2013.

[11] M. Mauch, R. M. MacCallum, M. Levy, and A. M. Leroi. The evolution of popular music: USA 19602010. *Royal Society Open Science*, 2(5):150081, 2015.

[12] E. Pampalk, A. Flexer, and G. Widmer. Improvements of Audio-Based Music Similarity and Genre Classification. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 634–637, 2005.

[13] M. Panteli and H. Purwins. A Quantitative Comparison of Chrysanthine Theory and Performance Practice of Scale Tuning, Steps, and Prominence of the Octoechos in Byzantine Chant. *Journal of New Music Research*, 42(3):205–221, 2013.

[14] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 525–530, 2009.

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

[16] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos. Measuring the Evolution of Contemporary Western Popular Music. *Scientific Reports*, 2, 2012.

[17] S. Stober, D. J. Cameron, and J. A. Grahn. Classifying EEG recordings of rhythm perception. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 649–654, 2014.

[18] B. L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

[19] E. Thul and G. T. Toussaint. A Comparative Phylogenetic-Tree Analysis of African Timelines and North Indian Talas. In *Bridges Leeuwarden: Mathematics, Music, Art, Architecture, Culture*, pages 187–194, 2008.

[20] G. Toussaint. Classification and phylogenetic analysis of African ternary rhythm timelines. In *Meeting Alhambra, ISAMA-BRIDGES Conference*, pages 25–36, 2003.

[21] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra. What is the effect of audio quality on the robustness of MFCCs and chroma features. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 573–578, 2014.

[22] J. van Balen, D. Bountouridis, F. Wiering, and R. Veltkamp. Cognition-inspired Descriptors for Scalable Cover Song Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 379–384, 2014.

[23] P. Van Kranenburg, M. de Bruin, L. P. Grijp, and F. Wiering. *The Meertens Tune Collections*. Meertens Institute, Amsterdam, no. 1 edition, 2014.

[24] T. C. Walters, D. A. Ross, and R. F. Lyon. The Intervalgram: An Audio Feature for Large-scale Melody Recognition. In *9th International Conference on Computer Music Modeling and Retrieval*, pages 19–22, 2012.