LEARNING A FEATURE SPACE FOR SIMILARITY IN WORLD MUSIC

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

ABSTRACT

In this study we investigate computational methods for assessing music similarity in world music. We use state-of-the-art audio features to describe musical content in world music recordings. Our music collection is a subset of the Smithsonian Folkways Recordings with audio examples from 31 countries from around the world. Using supervised and unsupervised dimensionality reduction techniques we learn feature representations for music similarity. We evaluate how well music styles separate in this learned space with a classification experiment. We obtained moderate performance classifying the recordings by country. Analysis of misclassifications revealed cases of geographical or cultural proximity. We further evaluate the learned space by detecting outliers, i.e. identifying recordings that stand out in the collection. We use a data mining technique based on Mahalanobis distances to detect outliers and perform a listening experiment in the 'odd one out' style to evaluate our findings. We are able to detect, amongst others, recordings of non-musical content as outliers as well as music with distinct timbral and harmonic content. The listening experiment reveals moderate agreement between subjects' ratings and our outlier estimation.

1. INTRODUCTION

The analysis, systematic annotation and comparison of world music styles has been of interest to many research studies in the fields of ethnomusicology [5, 14, 20] and Music Information Retrieval (MIR) [7, 12, 28]. The former studies rely on manually annotating musical attributes of world music recordings and investigating similarity via several clustering techniques. The latter studies rely on automatically extracting features to describe musical content of recordings and investigating music style similarity via classification methods. We focus on research studies that provide a systematic way of annotating music; a method that often disregards specific characteristics of a music culture but makes an across-culture comparison feasible. We are interested in the latter and follow a computational approach to describe musical content of world music recordings and investigate similarity across music cultures.

This study falls under the general scope of music corpus analysis. While several studies have focused on popular (mainly Eurogenetic) music corpus analysis, for example, the use of modes in American popular music [21], pitch, loudness and timbre in contemporary Western popular music [23], harmonic and timbral aspects in USA popular music [16], only a few studies have considered world or folk music genres, for example, the use of scales in African music [17]. Research projects have focused on the development of MIR tools for world music analysis [1], but no study, to the best of our knowledge, has applied such computational methods to investigate similarity in a world music corpus.

While the notion of world music is ambiguous, often mixing folk, popular, and classical musics from around the world and from different eras [4], it has been used to study stylistic similarity between various music cultures. We focus on a collection of folk recordings from countries from around the world, and use these to investigate music style similarity. Here we adopt the notion of music style by [19], 'style can be recognized by characteristic uses of form, texture, harmony, melody, and rhythm'. Similarly, we describe music recordings by features that capture aspects of timbral, rhythmic, melodic, and harmonic content [2].

The goal of this work is to infer similarity in collections of world music recordings. From low-level audio descriptors we are interested to learn high-level representations that project data to a music similarity space. We compare three feature learning methods and assess music similarity with a classification experiment and outlier detection. The former evaluates recordings that are expected to cluster together according to some ground truth label and helps us understand better the notion of 'similarity'. The latter evaluates examples that are different from the rest of the corpus and is useful to understand 'dissimilarity'. Outlier detection in large music collections can also be applied to filter out irrelevant audio or discover music with unique characteristics. We use an outlier detection method based on Mahalanobis distances, a common technique for detecting outliers in multivariate data [1]. To evaluate our findings we perform a listening test in the 'odd one out' framework where subjects are asked to listen to three audio excerpts and select the one that is most different [27].

Amongst the main contributions of this paper is a set

[1] Digital Music Lab (http://dml.city.ac.uk), CompMusic (http://compmusic.upf.edu/node/1), Telemata (https://parisson.github.io/Telemeta/)

[2] The use of form is ignored in this study as our music collection is restricted to 30-second audio excerpts.

of low-level features to represent musical content in world music recordings and a method to assess music style similarity. Our results reveal similarity in music cultures with geographical or cultural proximity and identify recordings with possibly unique musical content. These findings can be used in subsequent musicological analyses to track influence and cultural exchange in world music. We performed a listening test with the purpose of collecting similarity ratings to evaluate our outlier detection method. In a similar way, ratings can be collected for larger collections and used as a reference for ground truth similarity. The data and code for extracting audio features, detecting outliers and running classification experiments as described in this study are made publicly available [3].

The paper is structured as follows. First a detailed description of the low-level features used in this study is presented in Section 2. Details of the size, type, and spatio-temporal spread of our world music collection are presented in Section 3. Section 4 presents the feature learning methods with specifications of the models and Section 5 describes the two evaluation methods, namely, classification and outlier detection. In Section 5.2 we provide details of the listening test designed to assess the outlier detection accuracy. Results are presented in Section 6 and finally a discussion and concluding remarks are summarised in Section 7 and 8 respectively.

2. FEATURES

Over the years several toolboxes have been developed for music content description and have been applied for tasks of automatic classification and retrieval [13, 18, 25]. For content description of world music styles, mainly timbral, rhythmic and tonal features have been used such as roughness, spectral centroid, pitch histograms, equal-tempered deviation, tempo and inter-onset interval distributions [7, 12, 28]. We are interested in world music analysis and add to this list the requirement of melodic descriptors.

We focus on state-of-the-art descriptors (and adaptations of them) that aim at capturing relevant rhythmic, melodic, harmonic, and timbral content. In particular, we extract onset patterns with the scale transform [10] for rhythm, pitch bihistograms [26] for melody, average chromagrams [3] for harmony, and Mel frequency cepstrum coefficients [2] for timbre content description. We choose these descriptors because they define low-level representations of the musical content, i.e. less abstract representations but ones that are more likely to be robust with respect to the diversity of the music styles we consider. In addition, these features have achieved state-of-the-art performances in relevant classification or retrieval tasks, for example, onset patterns with scale transform perform best in classifying Western and non-Western rhythms [9, 15] and pitch bihistograms have been used successfully in cover song (pitch content-based) recognition [26]. The low-level descriptors are later used to learn high-level representations using various feature learning methods (Section 4).

The audio features used in this study are computed with the following specifications. For all features we fix the sampling rate at 44100 Hz and compute the (first) frame decomposition using a window size of 40 ms and hop size of 5 ms. We use a second frame decomposition to summarise descriptors over 8-second windows with 0.5-second hop size. This is particularly useful for rhythmic and melodic descriptors since rhythm and melody are perceived over longer time frames. For consistency, the timbral and harmonic descriptors considered in this study are summarised by their mean and standard deviation over this second frame decomposition.

Rhythm and Timbre. For rhythm and timbre features we compute a Mel spectrogram with 40 Mel bands up to 8000 Hz using Librosa [4]. To describe rhythmic content we extract onset strength envelopes for each Mel band and compute rhythmic periodicities using a second Fourier transform with window size of 8 seconds and hop size of 0.5 seconds. We then apply the Mellin transform to achieve tempo invariance [9] and output rhythmic periodicities up to 960 bpm. The output is averaged across low and high frequency Mel bands with cutoff at 1758 Hz. Timbral aspects are characterised by 20 Mel Frequency Cepstrum Coefficients (MFCCs) and 20 first-order delta coefficients [2]. We take the mean and standard deviation of these coefficients over 8-second windows with 0.5-second hop size.

Harmony and Melody. To describe melodic and harmonic content we compute chromagrams using variable-Q transforms [22] with 5 ms hop size and 20-cent pitch resolution to allow for microtonality. Chromagrams are aligned to the pitch class of maximum magnitude for key invariance. Harmonic content is described by the mean and standard deviation of chroma vectors using 8-second windows with 0.5-second hop size. Melodic aspects are captured via pitch bihistograms which denote counts of transitions of pitch classes [26]. We use a window $d = 0.5$ seconds to look for pitch class transitions in the chromagram. The resulting pitch bihistogram matrix is decomposed using non-negative matrix factorization [24] and we keep 2 basis vectors with their corresponding activations to represent melodic content. Pitch bihistograms are computed again over 8-second windows with 0.5-second hop size.

3. DATASET

Our dataset is a subset of the Smithsonian Folkways Recordings, a collection of documents of "people's music", spoken word, instruction, and sounds from around the world [5]. We use the publicly available 30-second audio previews and from available metadata we choose the country of the recording as a proxy for music style. We choose a minimum number of $N = 50$ recordings for each country to capture adequate variability of its style-specific characteristics. For evaluation purposes we further require the

[3] https://code.soundsoftware.ac.uk/projects/feature-space-world-music

[4] https://bmcfee.github.io/librosa/

[5] http://www.folkways.si.edu

dataset to have the same number of recordings per country. By manually sub-setting the data we observe that an optimal number of recordings is obtained for $N = 70$, resulting in a total of 2170 recordings, 70 recordings chosen at random from each of 31 countries from North America, Europe, Asia, Africa and Australia. According to the metadata these recordings belong to the genre 'world' and have been recorded between 1949 and 2009.

4. FEATURE LEARNING

For the low-level descriptors presented in Section 2 and the music dataset in Section 3, we aim to learn feature representations that best characterise music style similarity. Feature learning is also appropriate for reducing dimensionality, an essential step for the amount of data we currently analyse. In our analysis we approximate style by the country label of a recording and use this for supervised training and cross-validating our methods. We learn feature representations from the 8-second frame-based descriptors.

The audio features described in Section 2 are standardised using z-scores and aggregated to a single feature vector for each 8-second frame of a recording. A recording consists of multiple 8-second frame feature vectors, each annotated with the country label of the recording. Feature representations are learned using Principal Component Analysis (PCA), Non-Negative Matrix Factorisation (NMF) and Linear Discriminant Analysis (LDA) methods [24]. PCA and NMF are unsupervised methods and try to extract components that account for the most variance in the data. LDA is a supervised method and tries to identify attributes that account for the most variance between classes (in this case country labels).

We split the 2170 recordings of our collection into training (60%), validation (20%), and testing (20%) sets. We train and test our models on the frame-based descriptors; this results in a dataset of 57282, 19104, and 19104 frames for training, validation, and testing, respectively. Frames used for training do not belong to the same recordings as frames used for testing or validation and vice versa as this would bias results. We use the training set to train the PCA, NMF, and LDA models and the validation set to optimise the number of components. We investigate performance accuracy of the models when the number of components ranges between 5 and the maximum number of classes. We use the testing set to evaluate the learned space by classification and outlier detection tasks as explained below.

5. EVALUATION

5.1 Objective Evaluation

To evaluate whether we have learned a meaningful feature space we perform two experiments. One experiment aims at assessing similarity between recordings from the same country (which we expect to have related styles) via a classification task, i.e. validating recordings that lie close to each other in the learned feature space. The second experiment aims at assessing dissimilarity between recordings by detecting 'outliers', i.e. recordings that lie far apart in the learned feature space.

Classification. For the classification experiment we use three classifiers: K-Nearest Neighbors (KNN) with $K = 3$ and Euclidean distance metric, Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM) with a Radial Basis Function kernel. We report results on the accuracy of the predicted frame labels and the predicted recording labels. To predict the label of the recording we consider the vote of its frame labels and select the most popular label.

Outlier Detection. The second experiment uses a method based on squared Mahalanobis distances to detect outliers in multivariate data [1,8]. We use the best performing feature learning method, as indicated by the classification experiment, to transform all frame-based features of our dataset. For each recording we calculare the average of its transformed feature vectors and use this to compute its Mahalanobis distance from the set of all recordings. Using Mahalanobis, an n-dimensional feature vector is expressed as the distance to the mean of the distribution in standard deviation units. Data points that lie beyond a threshold, here set to the 99.5% quantile of the chi-square distribution with n degrees of freedom [6], are considered outliers.

5.2 Subjective Evaluation

To evaluate the detected outliers we perform a listening experiment in the 'odd one out' fashion [27]. A listener is asked to evaluate triads of audio excerpts by selecting the one that is most different from the other two, in terms of its musical characteristics. For the purpose of evaluating outliers, a triad consists of one outlier excerpt and two inliers as estimated by their Mahalanobis distance from the set of all recordings.

To distinguish outliers from inliers (the most typical examples) and other excerpts which are neither outliers nor inliers, we set two thresholds for the Mahalanobis distance. Distances above the upper threshold identify outliers, and distances below the lower threshold identify inliers. The thresholds are selected such that the majority of excerpts are neither outliers nor inliers. We randomly select 60 outliers and for each of these outliers we randomly select 10 inliers, in order to construct 300 triads (5 triads for each of 60 outliers), which we split into 10 sets of 30 triads. Each participant rates one randomly selected set of 30 triads.

The triads of outlier-inlier examples are presented in random order to the participant and we additionally include 2 control triads to assess the reliability of the participant. A control triad consists of two audio excerpts (the inliers) extracted from the first and second half, respectively, of the same recording and exhibiting very similar musical attributes, and one excerpt (the outlier) from a different recording exhibiting very different musical attributes. At the end of the experiment we include a questionnaire for demographic purposes.

We report results on the level of agreement between the computational outliers and the audio excerpts selected as the odd ones by the participants of the experiment. We

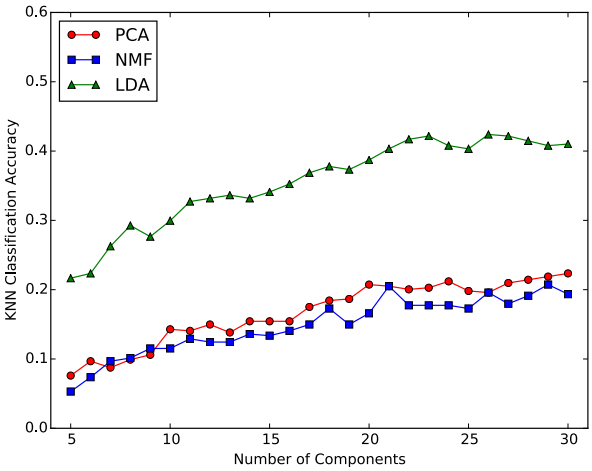

Figure 1. Classification accuracy for different numbers of components for PCA, NMF, and LDA methods (random baseline is 0.03 for 31 classes).

focus on two metrics; first, we measure the average accuracy between detected and rated outlier across all 300 triads used in the experiment, and second, we measure the average accuracy for each outlier, i.e. for each of 60 outliers we compute the average accuracy of its corresponding rated triads. Further analysis such as how the music culture and music education of the participant influences the similarity ratings is left for future work.

6. RESULTS

In this section we present results from the feature learning methods, their evaluation and the listening test as described in Sections 4 and 5.

6.1 Number of Components

First we present a comparison of classification performance when the number of components for PCA, NMF and LDA methods ranges between 5 and 30. For each number of components we train a PCA, NMF and LDA transformer and report classification accuracies on the validation set. The accuracies correspond to predictions of the label as estimated by a vote count of its predicted frame labels. We use the KNN classifier with $K = 3$ neighbors and Euclidean distance metric. Results are shown in Figure 1. We observe that the best feature learning method is LDA and achieves its best performance when the number of components is 26. PCA and NMF achieve optimal results when the number of components is 30 and 29 respectively. We fix the number of components to 30 as this gave good average classification accuracies for all methods.

6.2 Classification

Using 30 components we compute classification accuracies for the PCA, NMF and LDA transformed testing set. We also compute classification accuracies for the non-transformed testing set. In Table 1 we report accuracies

Classifier	Transform. Method	Frame Accuracy	Recording Accuracy
KNN	–	0.175	0.281
	PCA	0.177	0.279
	NMF	0.139	0.214
	LDA	0.258	**0.406**
LDA	–	**0.300**	0.401
	PCA	0.230	0.283
	NMF	0.032	0.032
	LDA	**0.300**	0.401
SVM	–	0.038	0.035
	PCA	0.046	0.044
	NMF	0.152	0.177
	LDA	0.277	0.350

Table 1. Classification accuracies for the predicted frame labels and the predicted recording labels based on a vote count (– denotes no transformation).

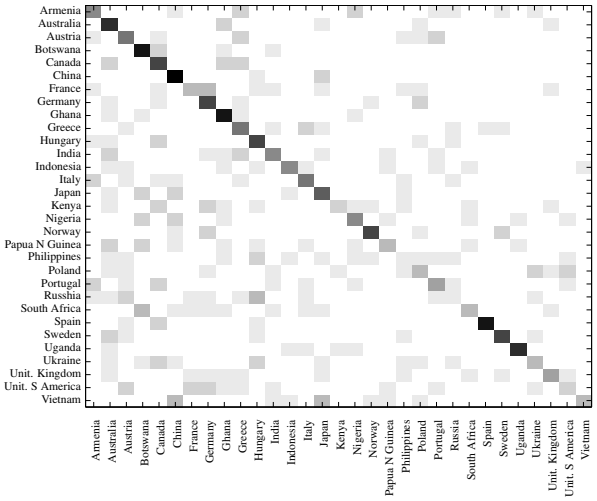

Figure 2. Confusion matrix for the best performing classifier, KNN with LDA transform (Table 1).

for the predicted frame labels and the predicted recording labels as estimated from a vote count (Section 4). The KNN classifier with the LDA transform method achieved the highest accuracy, 0.406, for the predicted recording labels. For the predicted frame labels the LDA classifier and transform was best with an accuracy of 0.300. In subsequent analysis we use the LDA transform as it was shown to achieve optimal results for our data.

For the highest classification accuracy achieved with the KNN classifier and the LDA transformation method (Table 1), we compute the confusion matrix shown in Figure 2. From this we note that China is the most accurate class and Russia and Philippines the least accurate classes. Analysing the misclassifications we observe the following: Vietnam is often confused with China and Japan, United States of America is often confused with Austria, France and Germany, Russia is confused with Hungary, and South Africa is confused with Botswana. These cases are characterised by a certain degree of geographical or cultural proximity which could explain the observed confusion.

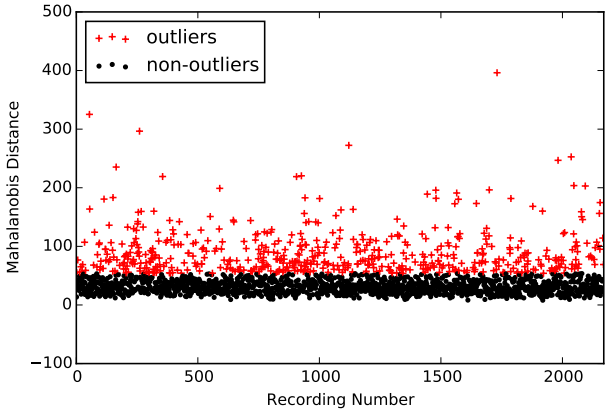

Figure 3. Mahalanobis distances and outliers at the 99.5% quantile of chi-square distribution.

6.3 Outlier Detection

The second experiment to evaluate the learned space aims at detecting outliers in the dataset. In this experiment we are not interested in how close music recordings of the same country are to each other, but we are rather interested in recordings that are very different from the rest. We use the LDA method as found optimal in the classification experiment (Section 6.2) to transform all frame-based feature vectors in our collection. Each recording is characterised by the average of its transformed frame-based descriptors.

From our collection of 2170 recordings (70 recordings for each of 31 countries), 557 recordings (around 26%) are detected as outliers at the chi-square 99.5% quantile threshold. In Figure 3 we plot the Mahalanobis distances for all samples in our dataset and indicate the ones that have been identified as outliers. The three recordings with maximum distances, i.e. standing out the most from the corpus, are identified as follows (in order of high to low Mahalanobis distance): 1) A recording of the dav dav instrument from the culture group 'Khmu' from Vietnam [6], 2) a rather non-musical example of bells from Greece [7], 3) an example of the angklung instrument from Indonesia [8]. These recordings can be characterised by distinct timbral and harmonic aspects or, in the case of the second example, by a distinct combination of all style attributes considered.

We plot the number of detected outliers per country on a world map (Figure 4) to get an overview of the spatial distribution of outliers in our music collection. We observe that Germany was the only country without any outliers (0 outliers out of 70 recordings) and Uganda was the country with the most outliers (39 outliers out of 70 recordings). Other countries with high number of outliers were Nigeria (34 outliers out of 70 recordings), Indonesia and Botswana (each with 31 outliers out of 70 recordings). We note that Botswana and Spain had achieved a relatively high classification accuracy in the previous evaluation (Section 6.2) and were also detected with a relatively high number of outliers (31 and 26 outliers, respectively). This could indicate that recordings from these two countries are consistent in their music characteristics but also stand out in comparison with other recordings of our world music collection.

[6] http://s.si.edu/1RuJfuu
[7] http://s.si.edu/21DgzP7
[8] http://s.si.edu/22yz0qP

6.4 Listening Test

The listening test described in Section 5.2 aimed at evaluating the outlier detection method. A total of 23 subjects participated in the experiment. There were 15 male and 8 female participants and the majority (83%) aged between 26 and 35 years old. A small number of participants (5) reported they are very familiar with world music genres and a similar number (6) reported they are quite familiar. The remaining participants reported they are not so familiar (10 of 23) and not at all familiar (2) with world music genres.

Following the specifications described in Section 5.2, participant's reliability was assessed with two control triads and results showed that all participants rated both these triads correctly. From the data collected, each of the 300 triads (5 triads for each of 60 detected outliers) was rated a minimum of 1 and maximum of 5 times. Each of the 60 outliers was rated a minimum of 9 and maximum of 14 times with an average of 11.5.

We received a total of 690 ratings (23 participants rating 30 triads each). For each rating we assign an accuracy value of 1 if the odd sample selected by the participant matches the 'outlier' detected by our algorithm versus the two 'inliers' of the triad, and an accuracy of 0 otherwise. The average accuracy from 690 ratings was 0.53. A second measure aimed to evaluate the accuracy per outlier. For this, the 690 ratings were grouped per outlier, and an average accuracy was estimated for each outlier. Results showed that each outlier achieved an average accuracy of 0.54 with standard deviation of 0.25. One particular outlier was never rated as the odd one by the participants (average accuracy of 0 from a total of 14 ratings). Conversely, four outliers were always in agreement with the subjects' ratings (average accuracy of 1 for about 10 ratings for each outlier). Overall, there was agreement well above the random baseline of 33% between the automatic outlier detection and the odd one selections made by the participants.

7. DISCUSSION

Several steps in the overall methodology could be implemented differently and audio excerpts and features could be expanded and improved. Here we discuss a few critical remarks and point directions for future improvement.

Numerous audio features exist in the literature suitable to describe musical content in sound recordings depending on the application. Instead of starting with a large set of features and narrowing it down to the ones that give best performance, we chose to start with a small set of features selected upon their state-of-the-art performance and relevance and expand the set gradually in future work. This way we can have more control of what the contribution is from each feature and each music dimension, timbre, rhythm, melody or harmony, as considered in this

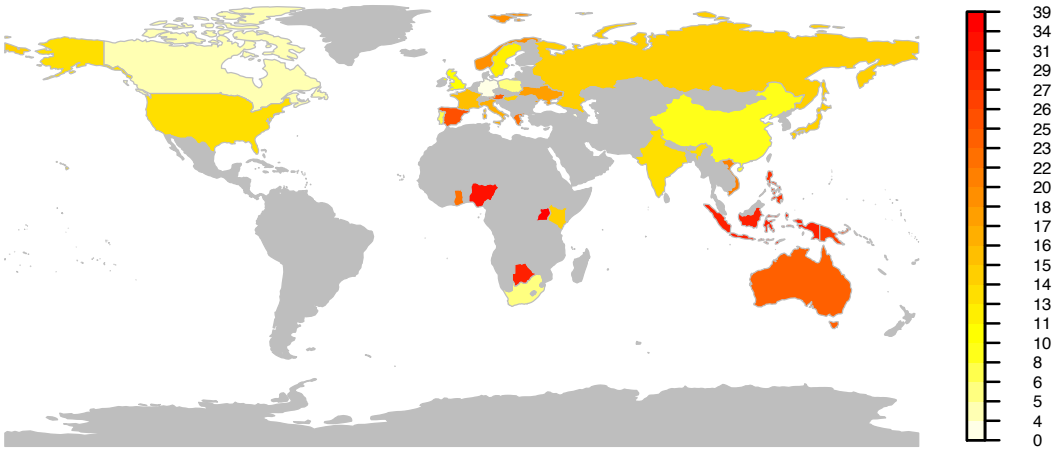

Figure 4. Number of outliers for each of the 31 countries in our world music collection (grey areas denote missing data).

study. The choice of features and implementation parameters could be improved, for example, in this study we have assumed descriptor summaries over 8-second windows but the optimal window size could be investigated further.

We used feature learning methods to learn higher-level representations from our low-level descriptors. We have only tested three methods, namely PCA, NMF, LDA, and did not exhaustively optimise parameters. Depending on the data and application, more advanced methods could be employed to learn meaningful feature representations [11]. Similarly, the classification and outlier detection methods could be tuned to give better accuracies.

The bigger aim of this work is to investigate similarity in a large collection of world music recordings. Here we have used a small dataset to assess similarity as estimated by classification and outlier detection tasks. It is difficult to gather representative samples of 'all' music of the world but at least a larger and better geographically (and temporally) spread dataset than the one used in this study could be considered. In addition, more metadata can be incorporated to define ground truth similarity of music recordings; in this study we have used country labels but other attributes more suitable to describe the music style or cultural proximity can be considered. An unweighted combination of features was used to assess music similarity. Performance accuracies can be improved by exploring feature weights. What is more, analysing each feature separately can reveal which music attributes characterise most each country or which countries share aspects of rhythm, timbre, melody or harmony.

Whether a music example is selected as the odd one out depends vastly on what it is compared with. Our outlier detection algorithm compares a single recording to all other recordings in the collection (1 versus 2169 samples) but a human listener could not do this with similar efficiency. Likewise, we could only evaluate a limited set of 60 outliers from the total of 557 outliers detected due to time limitations of our subjects. We evaluated comparisons from sets of three recordings and we used computational methods to create 'easy' triads, i.e. select three recordings from which one is as different as possible compared to the other two. However in some cases, as also reported by some of the participants, the three recordings were very different from each other which made it difficult to select the odd one out. In future work this could be improved by restricting the genre of the triad, i.e. selecting three audio examples from the same music style or culture. In addition the selection criteria could be made more specific; in our experiment we let participants decide on 'general' music similarity but in some cases it is beneficial to focus on, for example, only rhythm or only melody.

8. CONCLUSION

In this study we analysed a world music corpus by extracting audio descriptors and assessing music similarity. We used feature learning techniques to transform low-level feature representations. We evaluated the learned space in a classification manner to check how well recordings of the same country cluster together. In addition, we used the learned space to detect outliers and identify recordings that are different from the rest of the corpus. A listening test was conducted to evaluate our findings and moderate agreement was found between computational and human judgement of odd samples in the collection.

We believe there is a lot for MIR research to learn from and to contribute to the analysis of world music recordings, dealing with challenges of the signal processing tools, data mining techniques, and ground truth annotation procedures for large data collections. This line of research makes a large scale comparison of recorded music possible, a significant contribution for ethnomusicology, and one we believe will help us understand better the music cultures of the world.

9. ACKNOWLEDGEMENTS

EB is supported by a RAEng Research Fellowship (RF/128). MP is supported by a Queen Mary Principal's research studentship and the EPSRC-funded Platform Grant: Digital Music (EP/K009559/1).

10. REFERENCES

[1] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *International Conference on Management of Data (ACM SIGMOD)*, pages 37–46, 2001.

[2] J. J. Aucouturier, F. Pachet, and M. Sandler. "The way it sounds": Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005.

[3] M. A. Bartsch and G.H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[4] P. V. Bohlman. *World Music: A Very Short Introduction*. Oxford University Press, 2002.

[5] S. Brown and J. Jordania. Universals in the world's musics. *Psychology of Music*, 41(2):229–248, 2011.

[6] P. Filzmoser. A Multivariate Outlier Detection Method. In *International Conference on Computer Data Analysis and Modeling*, pages 18–22, 2004.

[7] E. Gómez, M. Haro, and P. Herrera. Music and geography: Content description of musical audio from different parts of theworld. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 753–758, 2009.

[8] V. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

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

[10] 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.

[11] E. J. Humphrey, A. P. Glennon, and J. P. Bello. Nonlinear semantic embedding for organizing large instrument sample libraries. In *Proceedings of the International Conference on Machine Learning and Applications*, volume 2, pages 142–147, 2011.

[12] A. Kruspe, H. Lukashevich, J. Abeßer, H. Großmann, and C. Dittmar. Automatic Classification of Musical Pieces Into Global Cultural Areas. In *AES 42nd International Conference*, pages 1–10, 2011.

[13] 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.

[14] A. Lomax. *Folk song style and culture*. American Association for the Advancement of Science, 1968.

[15] U. Marchand and G. Peeters. The modulation scale spectrum and its application to rhythm-content description. In *International Conference on Difital Audio Effects*, pages 167–172, 2014.

[16] M. Mauch, R. M. MacCallum, M. Levy, and A. M. Leroi. The evolution of popular music: USA 1960-2010. *Royal Society Open Science*, 2(5):150081, 2015.

[17] D. Moelants, O. Cornelis, and M. Leman. Exploring African Tone Scales. In *Proccedings of the International Society for Music Information Retrieval Conference*, pages 489–494, 2009.

[18] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *Technical Report. IRCAM*, 2004.

[19] S. Sadie, J. Tyrrell, and M. Levy. *The New Grove Dictionary of Music and Musicians*. Oxford University Press, 2001.

[20] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 112(29):8987–8992, 2015.

[21] E. G. Schellenberg and C. von Scheve. Emotional cues in American popular music: Five decades of the Top 40. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):196–203, 2012.

[22] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler. A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution. In *AES 53rd Conference on Semantic Audio*, pages 1–8, 2014.

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

[24] L. Sun, S. Ji, and J. Ye. *Multi-Label Dimensionality Reduction*. CRC Press, Taylor & Francis Group, 2013.

[25] G. Tzanetakis and P. Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(3):169–175, 2000.

[26] 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.

[27] D. Wolff and T. Weyde. Adapting Metrics for Music Similarity Using Comparative Ratings. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 73–78, 2011.

[28] F. Zhou, Q. Claire, and R. D. King. Predicting the Geographical Origin of Music. In *IEEE International Conference on Data Mining*, pages 1115–1120, 2014.