Mapping the Sounds of the Swahili coast and the Arab Mashriq: Music research at the intersection of computational analysis and cultural heritage preservation

Konstantinos Trochidis
konstantinos.trochidis@gmail.com
New York University Abu Dhabi
Abu Dhabi, UAE

Beth Russell
beth.russell@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

Andrew Eisenberg
andrew.eisenberg@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

Oscar Gomez
oag229@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

Kaustuv Kanti Ganguli
kaustuvkanti@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

Carlos Guedes
carlos.guedes@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

Virginia Danielson
ginny.danielson@gmail.com
New York University Abu Dhabi
Abu Dhabi, UAE

Christos Plachouras
cplachouras@nyu.edu
New York University Abu Dhabi
Abu Dhabi, UAE

ABSTRACT

This paper discusses an overview of an ongoing research that combines the preservation of musical heritage with ethnomusicological research driven by computational analysis. The two collections of non-Eurogenetic music under study are a curated collection of East African Swahili coast music and commercial recordings of Arab music. We explore the cross-cultural similarities, interactions and patterns of the music excerpts from the different regions and understand the similarities by employing computational audio analysis, machine learning and visualization techniques. We used a baseline model of representation by extracting Mel-Frequency Cepstral Coefficients (MFCC) features to model the spectral characteristics of the music excerpts in conjunction with t-distributed stochastic neighbor embedding (t-SNE) to create 2-D mappings of these features into a lower dimensional space of similarity. We compare this representation with more sophisticated approaches of acoustic feature representation. Principal Component Analysis (PCA) was used on the mel-scaled spectrograms of the music excerpts, and t-SNE was used to map the Principal Components to a 2-D space. The logarithmic short-time Fourier transform (STFT) of the music excerpts were extracted and a deep autoencoder neural network was trained to learn the relationships and structure of the excerpts by compressing the raw representation of the STFT. The results from the analysis show that PCA and the autoencoder model can reveal more interesting cluster representation than MFCCs by generating more complex clusters between the different styles of the corpus.

KEYWORDS

Computational analysis, Preservation of cultural heritage, Pattern recognition, Ethnomusicology.

ACM Reference Format:
Konstantinos Trochidis, Beth Russell, Andrew Eisenberg, Oscar Gomez, Kaustuv Kanti Ganguli, Carlos Guedes, Virginia Danielson, and Christos Plachouras. 2019. Mapping the Sounds of the Swahili coast and the Arab Mashriq: Music research at the intersection of computational analysis and cultural heritage preservation. In *Proceedings of 6th International Conference on Digital Libraries for Musicology (DLfM 2019).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital humanities is progressively coming to the forefront as a field of scholarly inquiry in the non-Western world. Similarly, non-Western scholars are increasingly looking to academic libraries as partners in this field, as they seek to use their resources in innovative ways to produce non-traditional, and often digitally-focused scholarship. It also focuses on the qualitative study of the ways in which music and sound cultures, and their attendant epistemologies (including those of the sonic digital humanities themselves), have been shaped by digital technologies. MaSC [1] has a particular focus on the Arab world and regions, including East Africa's Swahili coast, that have a long history of contact with the Arab world.

The role of the CDS's [2] involvement in this project is driven by established library practices in metadata description, archival preservation treatments, intellectual property awareness and research data management. This guides the development of both a

6th International Conference on Digital Libraries for Musicology (DLfM 2019), November 9, 2019, Delft, The Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Music and Sound Cultures research group
[2] Center for Digital Scholarship in New York University Abu Dhabi (NYUAD) Library

digital compendium of music from the Arab Mashriq (broadly defined here to include the Gulf), the East Africa Swahili coast and South India as well as a preservation archive of Arab music from the same regions.

Using the computational analysis of the audio we can develop methods to explore and interact with those data by employing techniques from Music Information Retrieval (MIR), machine learning and data visualization. The field of machine learning and representation learning has witnessed a remarkable progress over recent years, particularly in the automated machine perception of images, text, and music. However, there has been limited progress in exploring computational analysis methods in music collections of non-Eurogenetic music [7, 8].

The aim of this work is twofold:

(i) To facilitate the study of both musicological and cultural heritage of this rich musical history by building a large repository;

(ii) To develop computational analysis methods and interactive tools to study the cross-cultural differences and similarities between non-Eurogenetic music excerpts.

The paper is organized as follows. Section 2 presents background information about preservation of Arab cultural heritage; Section 3 provides information on the music collections used in this study; Section 4 presents details on the audio feature representation analysis and representation learning; Section 5 presents the results of the proposed approach; and Section 6 discusses the conclusion and future work.

2 PRESERVATION OF ARAB CULTURAL HERITAGE

The preservation of Arab cultural heritage materials is not a recent development. There are numerous libraries and museums committed to archiving this valuable historical material. Ipert [2] outlines several of these programs at institutions in Qatar, the United Arab Emirates, Kuwait, Oman, Egypt, Morocco, and several others. Noted digitization projects that capture and preserve these materials and publish them online include the Arabic Collections Online book digitization project at the NYUAD, the Afghanistan Digital Library, the Aga Khan Documentation Center at the Massachusetts Institute of Technology, the proprietary Early Arabic Printed Books from the British Library catalog offered by the Gale database provider, al-Maktaba I-Shamila and many others. A recent review of digital humanities projects in Islamic Studies is covered in [6]. The vast majority of these archives focus on visual materials.

While the application of digital humanities methods to cultural heritage collections continues to be a growing field of interest, it is a relatively new field within Arabic-speaking countries and even more so within the realm of musicology research in the Arab world. The vast majority of this work focuses on Western-language visual- and text-based collections, however Urberg reminds us that musicologists were among the earliest adopters of what we now call digital humanities methods, citing Fujinaga and Weiss's computational and digital archiving projects involving music and sound, as well as the development of one of the first musical databases in the 1970s, the Hymn Tune Index out of the University of Illinois [10].

3 CORPORA

The corpora for analysis consist of two collections: the Eisenberg Collection and the Music Compendium from the Arab Mashriq. We provide a short description of each corpus. Serra [9] stressed on the fundamental differences between a research corpora and a test corpus in terms of the ability of the former to capture the essence of a particular music culture. The author advocated the relevance of five criteria that were taken care of during compilation of the CompMusic [3] collection, namely purpose, coverage, completeness, quality, and reusability. The types of data collection that a research corpora demands are audio recordings and editorial metadata.

3.1 Eisenberg Collection

Our first archival collection for the MaSC research group is the Eisenberg Collection of East African Commercial Sound Recordings. This collection contains 500 sound files and associated metadata of commercial recordings produced for East African Swahili coast audiences between the late 1920s and the first decade of the twenty-first century. Most of the sounds in the collection fall within the realm of Swahili-language urban popular music from the Swahili coast's major urban centers (Mombasa, Dar es, Salaam, and Zanzibar. There are also examples of rural music traditions, colonial-era, martial music, recited Swahili poetry, and Swahili comedy sketches.

3.2 Music Compendium from the Arab Mashriq: Collections as Data

Using a collections [4] as data approach, the second corpus of this work consist of a digital compendium of 2827 recordings collected from the Library's collection of Arab audio on compact disc. The ethnic group and region of the digital compendium comes from Jordan, Kurdistan, Turkey, Lebanon, Morocco, Egypt, UAE, Bahrain, Yemen, Afghanistan, Beirut, Azerbaijan. Using a metadata-driven software [5], they are then described using a controlled vocabulary of metadata terms defined by the group to aid in their analysis. The metadata tags include descriptive elements such as melodic mode, cultural/ethnic group, instrumentation, component bass, rhythm or melody, meter, tempo, pitch collection, all of which are used in the models generated as part of the computational aspects of the work.

4 COMPUTATIONAL ANALYSIS OF AUDIO

We extracted standard feature extractors, such as MFCCs to investigate how these features correlate with the music excerpts. MFCCs are spectral representations and can be best used to describe the instrumentation and genre/style of the recordings and have been used in MIR extensively in the past. We tested our baseline model against the extracted principal components from the mel-scale spectrogram and against an unsupervised analysis of the raw representation of the spectrogram that is fed to a deep autoencoder to investigate if these methods are able to learn more complicated relationships and patterns of attributes of the music structure. For the PCA, we are utilizing the first 20 components that we compute from the mel-scale spectrogram, while in our deep learning model we are using a series of hidden layers that encode and decode the spectrogram to

[3] https://compmusic.upf.edu/

[4] http://dlib.nyu.edu/findingaids/html/nyuad/ad_mc_035/index.html

[5] https://www.markvapps.com/metadatics

learn a compressed representation of the important features of the spectrogram. We are using the bottleneck of this autoencoder layer as our final feature representation and fingerprint for the music excerpts.

4.1 Feature Representation of Audio

We extracted 13 Mel coefficients with a frame window of 20 ms and an overlap of 10 ms to compute the MFCC descriptors. We also derived the log magnitude/frequency short-time Fourier transform (logSTFT) which is a raw representation of spectral information for the recordings. To compute this feature, we first resample the audio to 22050 Hz and peak-normalize it. We then compute the linear-frequency STFT on 1024-sample frames with a ~10 ms (221 samples) hop size. The magnitudes of the linearly spaced frequency bins are then grouped into log-spaced bins using triangular frequency-domain filters – 8 octaves of 8 bins per octave, starting at 40 Hz (i.e. 64 bins). We then log-scale these features. The feature extraction for both MFCCs and spectrograms were calculated for a duration of 15 seconds taken from the middle of each excerpt so that we have a representative sample of each excerpt.

4.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure which is often utilized to reduce the dimension of a dataset and increase its interpretability. Using an orthogonal transformation it converts a set of possibly correlated variables to a set of linearly uncorrelated variables called principal components. The first principal component captures the direction of largest variance in the dataset, and subsequent components capture the highest variance possible under the constraint that it is orthogonal to the preceding components.

4.3 Deep Autoencoders

The autoencoder [1] is a neural network which is trained to learn a lower-dimensional representation of the input data. In a deep autoencoder, the network is trained to reconstruct the input using an encoder and decoder architecture that have a series of shallow layers. The architecture of the autoencoder model in our study consisted of 2 encoding layers and 3 decoding layers. The encoding layers had dimensions of 2000, 100, and 50 nodes respectively and the decoding layers had the same dimensions in reverse. The model was trained using backpropagation and we used the binary cross entropy loss function to optimize and find the optimal weights for the network. For both the encoding and decoding hidden layers of the model we used the *relu* activation function. The last hidden layer of the encoder called the bottleneck was used as a final feature representation of each music excerpt. This was a 50-D representation of the learned important attributes and features of each music excerpt.

4.4 Visualization and Mapping of the Learned Expressions

The next step was to convert the feature vector of 20 principal components as well as the high-dimensional bottleneck encoding representation of the neural network into 2-D embeddings to visualize interesting clusters of the music excerpts using the feature representations learned from the model. Traditional dimensionality reduction techniques such as Principal Components Analysis [3] and multidimensional scaling methods [4] are linear techniques that focus on keeping the low-dimensional representations of dissimilar data points far apart. For high-dimensional data that lie on or near a low-dimensional, non-linear manifold it is usually more important to keep the low-dimensional representations of very similar data points close together, which is typically not possible with a linear mapping.

In our analysis, we used the so called t-SNE for visualizing the resulting similarities of the feature representations [5]. Compared to methods discussed previously, t-SNE is capable of capturing much of the local structure of the high-dimensional data, while also revealing global structure such as the presence of clusters at several scales. In a second step, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the *Kullback-Leibler* (KL) divergence between the two distributions with respect to the locations of the points in the map.

5 RESULTS AND DISCUSSION

To get meaningful insights about the structural similarities of the different corpus, a dashboard browsing application was developed. Users could interact and explore the corpus by navigating through a 2-D similarity space. The user could select and hover over different points in the space and listen to the corresponding music excerpts. Points closer to each other reveal timbral and instrumentation similarities between different music excerpts.

Figure 1 presents the three different approaches used. The leftmost graph presents the 2-D embedding of the MFFC features of the music excerpts using t-SNE, the middle graph represents the mapping of the first 20 principal components of the mel-scaled spectrogram using t-SNE, and the rightmost graph represents the mapping of the bottleneck of the autoencoder method for learning a representation of the mel-spectrogram of the excerpts, again using t-SNE. For each graph, KMeans, a method for vector quantization used for assigning data points to the cluster with the nearest mean, is used in order to make the graph more interpretable and provide insight into the potential groupings of music excerpts with similarity in some musical aspect.

While the mapping of MFCCs separates 2 clusters which contain music pieces with the same instrumentation, it struggles to separate pieces within the larger cluster. Both the mapping of the principal components and that of the autoencoder method provide more complex clusters that contain more insightful separation of instrumentation and audio intensity.

Figure 2 presents the intensity of the audio files, which is represented by the shade of blue for each point in the graph; lighter shades indicate smaller intensity, while darker shades indicate higher intensity. Here we see clear directions of intensity variance across the music pieces in all graphs.

6 CONCLUSION AND FUTURE WORK

This work presented our approach towards preservation of Arab cultural heritage and the computational analysis of two collections of non-Eurogenetic music. In our study we explore the cross-cultural similarities, interactions, and patterns of the music excerpts from

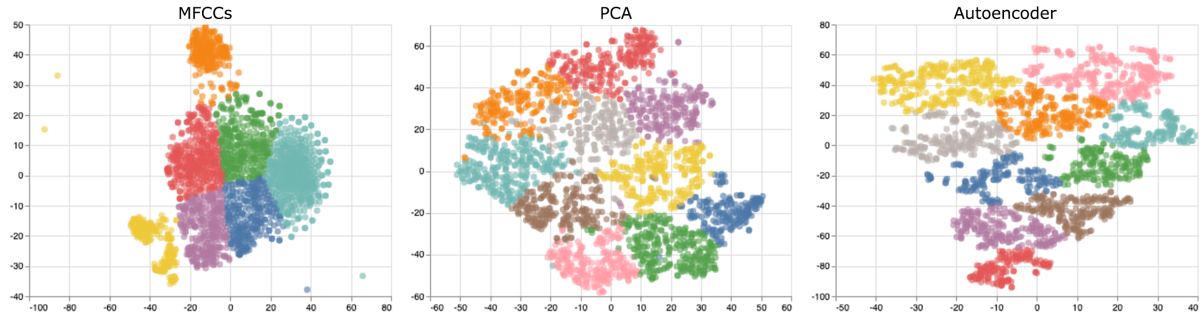

Figure 1: Statistics of artists for a given region in the space embedding.

Figure 2: Intensity mapping of the clustered data points on the t-SNE space.

the different regions and try to understand these similarities by employing computational audio analysis, machine learning, and visualization. Overall from the feature set, the separation of clusters using Principal Component Analysis and the autoencoder model is more interesting compared to the baseline method.

The model can separate the data into a number of clusters. While certain clusters include traditional instrumental string music, others include traditional vocal, electronic, and pop Arab music. Folk music excerpts with similar instrumentation from both two archives are clustered together in the mapping. The drawback of this approach is that it can only serve as a high-level exploratory data analysis tool, since there are still not enough metadata regarding the style, genre, and structure of the archives.

One of the main challenges of future work is to find appropriate metadata descriptions of genre and structural categorization of these music traditions using domain knowledge expertise. Future work will also entail a systematic annotation of this content in collaboration with experts of these genres regarding the collection of metadata about performance style, prevailing rhythmic cycles, melodic modes, instrumentation, ethnic and social groups, and structural segmentation. This will allow us to build and evaluate supervised training models with labelled data for MIR tasks such as genre classification, instrument recognition, rhythmic and melodic analysis to name a few.

7 ACKNOWLEDGEMENT

This research is part of project “Computationally engaged approaches to rhythm and musical heritage: Generation, analysis, and performance practice," funded through a grant from the Research Enhancement Fund at the New York University Abu Dhabi.

REFERENCES

[1] Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*. 37–49.
[2] S. J. Ipert. 2016. Preservation and Digitization Activities in Arabic- and Farsi-Speaking Countries. *Preservation, Digital Technology & Culture* 45, 2 (2016), 63–75.
[3] IT Joliffe. 2002. Principle Component Analysis. 2nd.
[4] Joseph B Kruskal and Myron Wish. 1978. Multidimensional scaling. Number 07–011 in Sage University Paper series on quantitative applications in the social sciences.
[5] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[6] Elias Muhanna. 2016. *The Digital Humanities and Islamic & Middle East Studies.* Walter de Gruyter GmbH & Co KG.
[7] Xavier Serra. 2011. A multicultural approach in music information research. In *Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24-28; Miami, Florida (USA). Miami: University of Miami; 2011.* International Society for Music Information Retrieval (ISMIR).
[8] Xavier Serra. 2013. Exploiting domain knowledge in music information research. (2013).
[9] Xavier Serra. 2014. Creating research corpora for the computational study of music: the case of the Compmusic project. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.* Audio Engineering Society.
[10] Michelle Urberg. 2017. Pasts and Futures of Digital Humanities in Musicology: Moving Towards a âĂIJBigger TentâĂİ. *Music Reference Services Quarterly* 20, 3-4 (2017), 134–150.