Developing immersive VR experience for visualizing cross-cultural relationships in music

Kaustuv Kanti Ganguli* Oscar Gomez† Leonid Kuzmenko‡ Carlos Guedes§

Music and Sound Cultures research group
New York University Abu Dhabi, Abu Dhabi, UAE 129188

Abstract

With the advent in advanced computing methodologies and forward-thinking data storage needs over the last decade, there has been a major drive for reformatting archival collections to enable advanced computational analysis. In this paper, we present our digital compendium of music from the Arab Mashriq and Western Indian Ocean comprising two music collections drawn from Library materials and field recordings at the New York University Abu Dhabi. This is at once the product and object of our ongoing research at the intersection of cultural heritage preservation and computational analysis. Through computational-ethnomusicological research, we explore the cross-cultural similarities, interactions, and patterns from the music excerpts in order to understand their similarity space by employing audio analysis, machine learning, and visualization techniques. Besides the digital artifactual value, pedagogical/educational and scholarly outcomes, we focus on attracting user-friendly and community engagement into appreciating the music from this region. This is done by providing interactive visualizations of the musical features on a dashboard application and 3-D rendering of the mappings in a VR environment. The VR experience is not only immersive but also provides a scope for appreciation, learning, and dissemination of the music from the region.

Index Terms: H.5.2 [User Interfaces]: User Interfaces—Graphical user interfaces (GUI); H.5.m [Information Interfaces and Presentation]: Miscellaneous

1 Introduction

The development of the field of Music Information Retrieval (MIR) over the past 20 years and the continued development of content- and feature-extraction algorithms have brought to the fore the many meaningful dimensions by which music can be organized and described. Over this period, there were also dramatic changes in computational speed, storage, and streaming technologies that substantially changed the way humans interact with the technologies to access and listen to music. Developing novel intelligent interfaces for browsing and discovering music through its many dimensions has been an important element in MIR research, and a fertile field for future development [6].

Recent breakthroughs in display technologies and modern spatial tracking techniques have enabled the design of VR (Virtual Reality) systems that introduced a new age of immersive experiences. The role of sound in establishing a convincing sense of immersion in such experiences is not a new phenomenon. Thus there is an opportunity that exists for audio and music researchers to elaborate about VR audio, and define the role of sound as a component of the ultimate displays of the future. It is the interdisciplinary nature of the paradigm that feeds contemporary artistic research focused on the role of sound and music within virtual spaces. Advances in VR technology inherently provoke new research questions across disciplines that deal with, or benefit from, simulated experiences. For audio and music researchers, these questions range from tools and techniques to those that deal with experiential and aesthetic qualities of immersive experiences.

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. Hosted out of the Center for Digital Scholarship (CDS) [1] in the New York University Abu Dhabi (NYUAD) Library, the Music and Sound Cultures research group (MaSC) [2] focuses on the interdisciplinary study of music from the Arab Mashriq and adjacent regions by using computational analysis and understanding of these musics and other forms of sound culture from large digital collections, and to the development of methodologies for cross- cultural mapping and comparison of these materials. 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 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. Our two-pronged research partnership uses the preservation of audio, video and ephemera to establish a repository of materials dedicated for the study of music from the Arab Mashriq, East Africa Swahili coast and South India.

By engaging with these three research streams, namely Digital Humanities, Music Information Retrieval, and Virtual Reality, we make possible a broad exploration of the music of the Arab world and Swahili coast by building a large repository of music intended for research and consultation. 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. This has been facilitated by developments in deep learning, where stacking layers in neural networks with large number of data has yielded high classification accuracy. However, there has been limited progress in exploring computational analysis methods in music collections of non-Eurogenetic music [2, 3, 10–12]. Previous research was based on the use of off-the-shelf feature extractors, such as Mel-Frequency Cepstral Coefficients (MFCC) and rhythm based features. With this view in mind, we wish to explore the effectiveness of machine learning and representation learning for non-Western music excerpts. We primarily focus on two models: (i) MFCCs extracted from the audio excerpts and t-distributed stochastic neighbor embedding (t-SNE);

*e-mail: kaustuvkanti@nyu.edu
†e-mail:oscar.gomez@nyu.edu
‡e-mail:lk1640@nyu.edu
§e-mail:carlos.guedes@nyu.edu

[1] Center for Digital Scholarship

[2] http://masc.hosting.nyu.edu/

978-1-7281-6532-5/20/$31.00 ©2020 IEEE
DOI 10.1109/VRW50115.2020.0-190

and (ii) Deep stacked autoencoders learnt from the spectrogram and t-SNE. We employ the log-frequency spectrograms to learn the acoustic representations of the music excerpts.

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; and (ii) To develop computational analysis methods and interactive tools to study the cross-cultural differences and similarities between non-Eurogenetic music excerpts. The paper, an extension of our previous work [14], 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 [4] 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 [9]. 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 [15].

3 The 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 [13] 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. In addition to the information available on the album cover-art, culture-specific elements (e.g. terminology for a given concept in a repertoire), in consultation with domain experts, add value to the metadata. Serra also mentions the usefulness of sharing the corpora and associated metadata on open platforms, so researchers of diverse background can use them for subtasks. This facilitates discovery of similarity and differences across music cultures; or to address the bigger question of cross-cultural universality of music concepts.

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. The first two series of the Eisenberg Collection arrived as born-digital music files, having been digitized from cassette tapes. The third series in the Eisenberg Collection collection is being digitized from commercial audio cassettes released in the late 1990s and early 2000s.

3.2 Music Compendium from the Arab Mashriq

Using a collections 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. These CDs were purchased for the Library's collection to support the teaching and research at the university, based on their connection to the research parameters of music from the Gulf, and the Arab Mashriq, and reformatted into digital files in order to make them computationally accessible to the research group. Access to the compendium is restricted to only those in the research group.

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 [4], 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

In our study we explore the cross-cultural similarities, interactions and patterns of the music excerpts from the different regions and understand these similarities by employing visualization and dimensionality reduction techniques to the data. On the one hand, as a baseline model we extracted standard feature extractors, such as MFCCs to investigate how these features correlate with the music excerpts. MFCCs are spectral representations and can best used to describe the instrumentation and genre/style of the recordings and have been used in MIR extensively in the past.

On the other hand, there has been very little work done on representation learning on lower level features, such as the time domain waveform/spectrogram for these types of non-western corpus. We tested our baseline model against an unsupervised analysis of the raw representation of the spectrogram that is fed to a deep autoencoder and investigate if its able to learn more complicated relationships and patterns of attributes of the music structure. In our deep learning model we are using a series of hidden layers that encode and decode the spectrogram to 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

For both the Eisenberg collection and the Music Compendium of the Arab Mashriq excerpts we extracted 13 Mel coefficients. We used

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

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

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 5 seconds taken from the middle of each excerpt so that we have a representative sample of each excerpt.

4.2 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. In the model, the weights for all layers in the network are trained jointly through backpropagation.

The input dataset of N data points $\{x_{i=1:N}\}$ is passed into a feed-forward neural network of one or more hidden layers, where the hidden layer is a bottleneck layer with activations . The activations that we used for our model were the *relu* and were obtained as:

$$y_i = f(Wx_i + b) \tag{1}$$

The output of the autoencoder can be obtained from the activations of the autoencoder as:

$$z_i = W' y_i + b' \tag{2}$$

Often the model is trained in a stacked fashion where there are multiple hidden layers with weight W^k for the k^{th} hidden layer where the activation for the different layers can be obtained by:

$$y_i{}^k = f(W^{k-1} y_i{}^{k-1} + b^{k-1}) \tag{3}$$

4.3 Model Architecture

The architecture of the autoencoder model in our study consisted of 2 encoding layers and 3 decoding layers (see Figure 1). 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 high-dimensional bottleneck encoding representation of the neural network into an 2-D embedding 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 [5] and multidimensional scaling methods [7] 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.

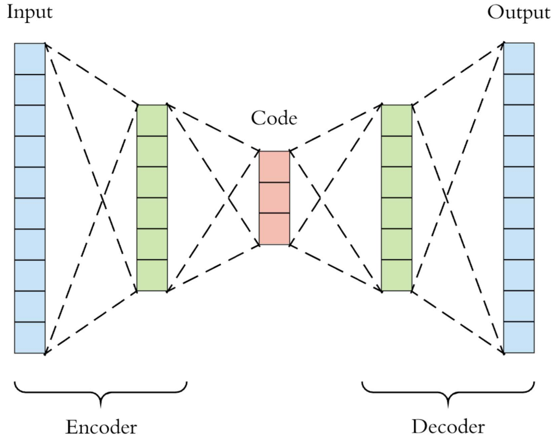


Figure 1: Autoencoder architecture

In our analysis, we used the so called t-SNE for visualizing the resulting similarities of the feature representations [8]. 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.

Given a set of high-dimensional dataset $\{x_1, x_2, \cdots, x_N\}$, t-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects x_i and x_j, as follows:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{4}$$

The similarity of datapoint x_j to the datapoint x_i is conditional probability $p_{j|i}$ defined as:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k\neq i} exp((-||x_i - x_k||^2/2\sigma^2))} \tag{5}$$

where σ^2 denotes the variance of the Gaussian Kernel centered on the datapoint x_i. t-SNE aims to learn a d-dimensional map $\{y_1, y_2, \cdots, y_N\}$, that represent the similarities $p_{i|j}$ as closely as possible. To this end, it measures the similarities of q_{ij} between the points y_i and y_j in the map.

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq m} (1 + ||y_k - y_m||^2)^{-1}} \tag{6}$$

The locations of the points y_i in the map are determined by minimizing the KL divergence, given by:

$$KL(P||Q) = \sum_{i\neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{7}$$

5 Results and Discussion

In order to better visualize the 2-D and 3-D embeddings produced by t-SNE, the data points were separated into two clusters corresponding to the collections to which they belong. This can be seen in Figures 4a and 4c, where the Eisenberg Collection is colored in green and the Compendium from the Arab Mashriq is colored in red. While we start from two broad music cultures and assume the same to be the expected number of clusters, the intra-repertoire diversity may cause genuine separation of the feature space. Locally observed sub-clusters give rise to a question about possibly performing a more in depth segmentation of the collection. For this purpose, and in order to offer a way to visualize clusters formed in the original, high-dimensional space, k-means was performed and the resulting clusters colored in the final 2-D and 3-D visualizations. The choice of k for the computation of clusters was made by computing the inertia (sum of squared distances of samples to their closest cluster center) and choosing a k such that improvements diminish after it - the so called "elbow method", resulting in an optimal $k = 6$ clusters.

Figure 4 shows the different visualization results obtained for the two different computational approaches followed. For the MFCC with t-SNE embedding, there is a clear separation of the two corpus with the Eisenberg collection being clustered at the right. Another prominent cluster is opposite to this one on the left, which consists mostly of music from Dariush Talai. On the other hand, the autoencoders and t-SNE embedding shows more interesting clusters between the two corpus, as seen in Figure 2.

The first cluster circled in purple color in the graph includes Persian instrumental music of Tar and Sitar with artists such as Dariush Talai. There is also another cluster circled in blue color which includes traditional vocal music from Syria, Lebanon, and Palestine of the first 3 decades of 1900's with artists such as Ahmad al-Sheikh, Ahmad al-Mir, and Antoine al-Shawwa.

There is a third cluster circled in black that includes modern electronic and pop Arab music including artists such as Abdul Majeed Abdulla, Ahlam, Mesaed Al-Belushi, and Abdullah Al Rowaished. Finally, there is another cluster circled in green color that includes modern contemporary and classical Arab excerpts including artists such as Sabreen. Folk music artists from both the Music Compendium of the Arab Mashriq and Eisenberg collection are clustered together at the center of the map. This might be the case due to the similar instrumentation of these excerpts.

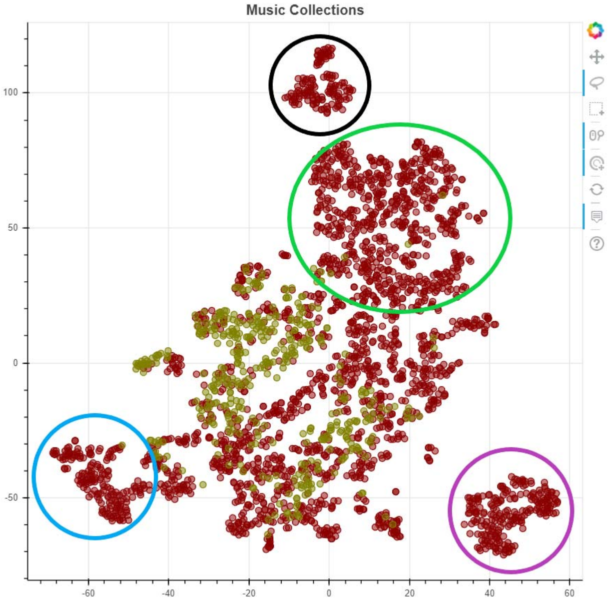


Figure 2: Bottleneck layer of autoencoder model using t-SNE.

5.1 Dashboard Application

To get meaningful insights about the structural similarities of the different corpus, a dashboard browsing application was developed. In the dashboard application, statistics and frequency distributions of the artists for the different areas of the mapping are presented. This is particular helpful to get insights regarding the similarities of the artists within the clusters and different areas. We can also get information regarding which artists are the prevailing representations for these clusters and use these tools as a recommendation engine to find similarities between artists. Users can 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.

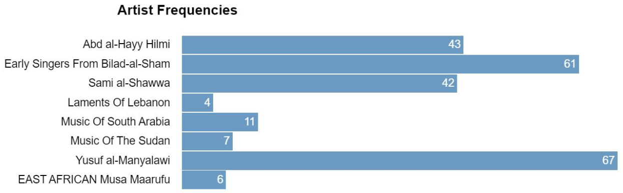
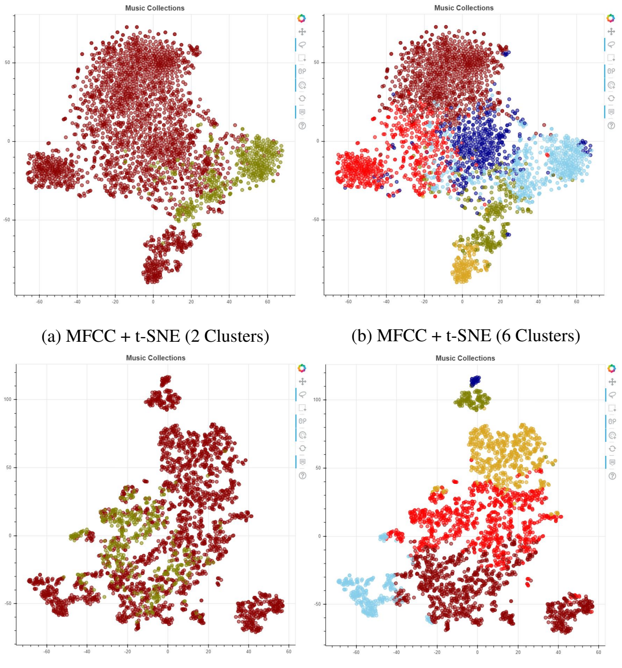


Figure 3: Summary statistics in the dashboard for the blue selection above



(a) MFCC + t-SNE (2 Clusters)

(b) MFCC + t-SNE (6 Clusters)

(c) STFT + Autoencoder (2 Clusters)

(d) STFT + Autoencoder (6 Clusters)

Figure 4: Visualizations for learned representations

5.2 VR Implementation

Virtual reality is a fast growing and powerful medium that allows the user to engage with the content on the next level of immersion by enabling a six depth-of-freedom motion inside the virtual environment. The scatter plot representation of the corpus was easily transformed from 2-D space into virtual reality, as the mapping produced by t-SNE for 3 dimensions kept the overall structure as in 2-D, while adding a extra degree for freedom of movement.

Unreal Engine 4 was used as a development platform. First, the audio clips from the corpus were assigned with unique IDs. Next, instead of colors, different materials were used to visually differentiate the clusters. With a help of UE4 Data Table asset that read from the excel sheet, so called "Audio Spheres" were spawned each time user runs an application. Each sphere had an audio clip and material embedded in it. The user could point at any nearby sphere and pull the trigger of the controller to activate the corresponding audio clip. Multiple clips could be triggered at once to appreciate the difference between them. To add more immersion to the experience, Steam Audio was used as a spatialization method. Steam Audio does an HRTF-based binaural rendering of audio. It means that the sound is altered based on the user head's position towards the sound source. It not only adds to the realism of the experience but helps the user to better orient themselves in the virtual environment as well.

The user could either move freely in the virtual environment and explore audio clips in whatever order wished, or, with a press of a button, teleported to the beginning of one of the predefined paths. Consequent presses of the button move the user along the path until it is completed. The user is teleported back to the starting position then. The experience was also designed to let the user engage a view from a distance - the user is teleported to a position, from which it is possible to see the whole corpus and appreciate the visual differences between the clusters.

The virtual reality headset of choice was Oculus Quest. It is standalone and wireless, which means that it could be easily transported. Oculus Quest also has good positional loudspeakers, which do not block user's ears and provide with a good recognition of an audio source in virtual environment when using binaural rendering feature of Steam Audio. It has two "touch" controllers that act as user hands. In the future the experience could be shaped so that the user would be able to use their own hands instead as Oculus already delivered hand tracking support. If required, Oculus Quest could be connected to a PC or a laptop running UE4 to utilize full processing power of modern computers. Figure 5 and Figure 6 show a screenshot of the VR rendering of the 3-D t-SNE space made on Oculus Quest.

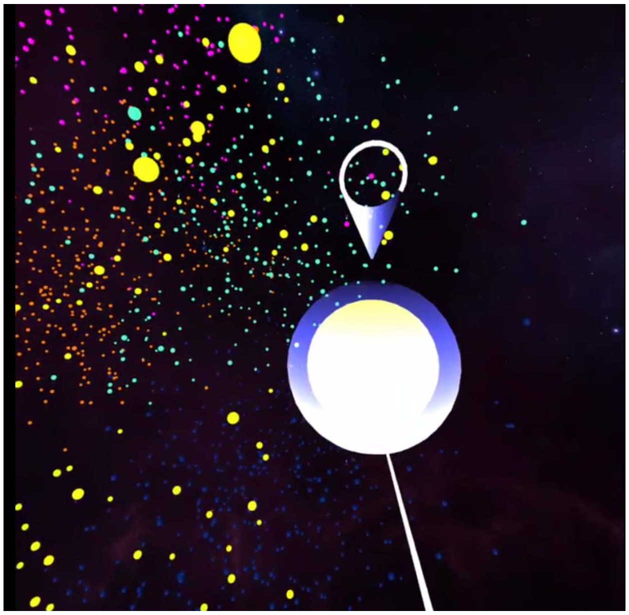


Figure 5: Selecting a song to play in the VR rendering of the 3-D map.

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 the different regions and try to understand these similarities by employing computational audio analysis, machine learning, and visualization. Finally, we want to use this analysis to build intelligent applications so that users can interact, explore and get meaningful insights about the structural similarities of these data. In our computational analysis we used a baseline representation by extracting off-the-shelf MFCC features to model the spectral characteristics of the music excerpts in conjunction with t-SNE to create a 2-D embedding of these features onto a lower-dimensional similarity space. Next, we compare this representation with a more sophisticated approach to feature engineering of acoustic features by extracting the logSTFT of the music excerpts and train a deep autoencoder neural network to learn the relationships and structure of the excerpts by compressing the raw representation of the STFT of the acoustic signal into a compact low dimensional vector the bottleneck of the network. Finally, we use the compressed layer representation of the neural network as a new feature representation of each excerpt and map it with t-SNE to a 2-D embedding to compare the clustering with the baseline method. Overall the separation of clusters using the autoencoder model is more interesting compared with the baseline method.

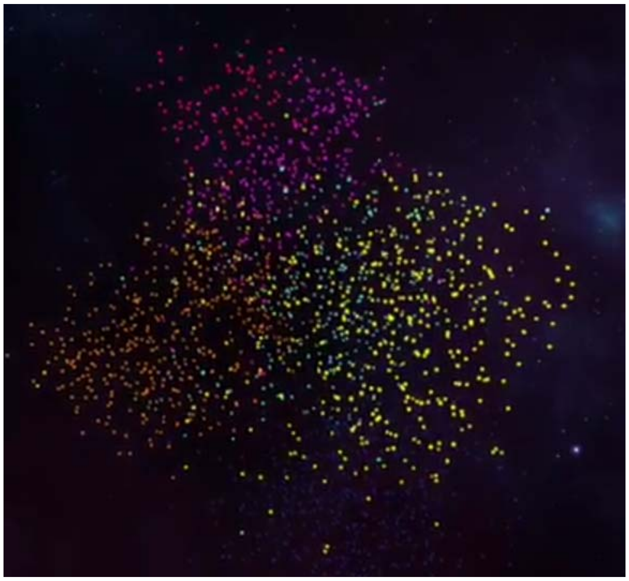


Figure 6: Screenshot of the VR rendering of the 3-D map as seen from a position away from the clusters.

The model can separate the data into a number of clusters. There is one cluster that includes traditional instrumental string music and also two others with 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.

Acknowledgments

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] P. Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49, 2012.

[2] K. K. Ganguli. How do we 'See' & 'Say' a raga: A Perspective Canvas. *Samakalika Sangeetham*, 4(2):112–119, Oct. 2013.

[3] K. K. Ganguli and P. Rao. Discrimination of melodic patterns in Indian classical music. In *Proc. of National Conference on Communications (NCC)*, Feb. 2015.

[4] S. J. Ipert. Preservation and digitization activities in arabic- and farsi-speaking countries. *Preservation, Digital Technology & Culture*, 45(2):63–75, 2016.

[5] I. Joliffe. Principle component analysis. 2nd, 2002.

[6] P. Knees, M. Schedl, and M. Goto. Intelligent user interfaces for music discovery: The past 20 years and what's to come. In *Proceedings of the 20th International Society for Music Information Retrieval (ISMIR) Conference, Delft, The Netherlands*, 2019.

[7] J. B. Kruskal and M. Wish. Multidimensional scaling. number 07–011 in sage university paper series on quantitative applications in the social sciences, 1978.

[8] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[9] E. Muhanna. *The Digital Humanities and Islamic & Middle East Studies*. Walter de Gruyter GmbH & Co KG, 2016.

[10] P. Rao, J. C. Ross, and K. K. Ganguli. Distinguishing raga-specific intonation of phrases with audio analysis. *Ninaad*, 26-27(1):59–68, Dec. 2013.

[11] X. Serra. 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), 2011.

[12] X. Serra. Exploiting domain knowledge in music information research. In *Stockholm Music Acoustics Conference 2013 and Sound and Music Computing Conference*. Logos Verlag Berlin, 2013.

[13] X. Serra. 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, 2014.

[14] K. Trochidis, B. Russell, A. Eisenberg, K. K. Ganguli, O. Gomez, C. Plachouras, C. Guedes, and V. Danielson. Mapping the sounds of the swahili coast and the arab mashriq: Music research at the intersection of computational analysis and cultural heritage preservation. In *6th International Conference on Digital Libraries for Musicology (DLfM), Delft, The Netherlands*, 2019.

[15] M. Urberg. Pasts and futures of digital humanities in musicology: Moving towards a "bigger tent". *Music Reference Services Quarterly*, 20(3-4):134–150, 2017.