# SOUND OF(F):

Exploring contextual representations of sound and music data sets using machine learning

Zeynep Erol
HKUST

Eray Ozgunay
HKUST

RAY LC
City University of Hong Kong

## ABSTRACT

Complex datasets like sound collections and musical performances are difficult to experience intuitively. Machine Learning provides a way to computationally cluster large audio collections, but context-dependent forms of interaction are required to allow audiences to grasp the dimensions of the complexity. We applied the t-SNE algorithm to collections of subway street music in New York, as well as to a live performance of Gershwin's Rhapsody in Blue, to explore the way interactions in immersive space can be used to explore complex and large audio collections. We found that 2D and 3D interactions, as well as headspace vs. controller interactions can differentially affect the experience of different sound spaces by prototyping these interactions in VR using data processed through t-SNE. This provides a method for experiencing the sounds of a city or environment via intuitive navigation, and nonlinear exploration of a work of music using joystick manipulations in VR.

## CCS Concepts

**•Applied computing→Sound and music computing; •Human-centered computing→Human computer interaction (HCI);**

## Keywords

spatial audio; virtual reality; machine learning; t-SNE; music collection visualization; nonlinear listening.

## 1. INTRODUCTION

As data becomes more ubiquitous in our lives, it becomes necessary to experience the full dimensionality of datasets as opposed to merely understanding them. Perhaps most evident are the examples of music collections and environmental sounds. The former illustrates the problem of grasping a whole genre of music or works by particular artists as a whole without having to listen to each work song by song. The latter concerns particularly our current inability to travel to far-off exotic vacation spots as well as experience the sounds and environments of nostalgia and the quarantined home.

Current approaches to understanding music and sound collections include song-content-based visualization using graphs [17] and reduced dimensionality visualization using musical features [8], part of a general attempt to understand music by visualizing the computation of features. In addition, Visualizations of music databases are a popular form of interface allowing intuitive exploration of music catalogs. Music and sounds, however, are part of an environment that evokes nuances of presence, which is divorced from visualization based only on computed dimensions. To evoke the feeling of one's home city, for example, we need to consider the spatial relationships between the sounds collected as well as the interactions with these sounds in space [2]. To better grasp the experience of music and sound collections, we used the t-SNE algorithm [15] to cluster related sounds for interaction in 3D space in a Virtual Reality environment. We designed spatial interactions that include both gestural and pointing methods for experiencing the sounds in space, providing an immersive method to experience sounds in the context of the environment they originate from.

## 2. BACKGROUND

Previous attempts at understanding high dimensional audio data must deal with the shear amount of information under consideration, and have included metrics that make the retrieval process more efficient [4]. These approaches rely on efficient classification schemes that resonate with human perception [13], but requires a user-centered design perspective to implement. Machine learning has been applied to high dimensional audio classification using features of the sound [21], but these computational approaches do not always produce the phenomenological separations in human sound classification [7]. Similarly, environmental sounds have also been classified using convolutional neural networks [20]. Recent approaches have included using human biometrics data like EEG to automatically and computationally classify the experience of the sound itself rather than its physical properties [22].

One way to overcome the divide between classification of the sound's features and classification of its experience involves using immersive techniques to allow human interaction with the sound's computational classification. The immersive experience of data has been applied to domains such as data analysis work flows [5], visualization relationship amongst scientific paper corpuses [9], musical catalog visualization [6], cultural analyses of musical patterns [8], and previewing audio samples using dimension reduction techniques [3]. While these works have shown the promise of using immersive techniques like VR to help users experience complex audio data, they have yet to focus on the diverse set of gestural and spatial interactions that are possible.

## 3. METHODS

### 3.1 Audio Processing

We collected sounds of subway street musicians in New York City for application of environmental sounds collection, and one long performance of Gershwin's *Rhapsody in Blue* to use for the application of segmenting of a single musical work.

For the piano piece *Rhapsody in Blue,* we broke down the piece into segments with the technique called onset detection algorithm [1] embedded in a MIR python package (librosa [16]). This algorithm divides the piece into chunks by detecting their onsets, i.e. beginning of the transient parts. For the recordings that are recorded in the New York Subway stations, we don't break them into chunks because we use the entire recordings, but take

segments of 10 seconds from each sample for subsequent analysis.

Then, we generate the feature vectors of the sounds to capture the features of the recordings. One way is to capture these features is to obtain the Mel-frequency cepstrum coefficients and their time derivatives of the sounds, which are widely used in speech and music information retrieval (MIR) and processing [12,14,18]. We get these coefficients, 13 for each recording, and their first and second-time derivatives, called first and second delta features, using librosa [16]. Then we concatenate them to get the feature vectors of each recording; in total, we have a vector with a length of 39 for each recording.

## 3.2 Dimensionality Reduction

Now that we have 39-dimensional vectors for each recording, we need to transform them into 2D and 3D spaces for human visualization. We use the scikit-learn's tool [19] dimension reduction method t-SNE [15] for this purpose. T-SNE is a machine learning technique that is developed particularly for data visualization by assigning every point located in a higher dimension onto a location in the two or three-dimensional space and considering pairwise similarities between these points. T-SNE essentially clusters the audio data. It tries to maintain the local structure of the data by transforming the similarities between vectors into the models of joint probability distributions and then trying to minimize the Kullback–Leibler (KL) divergence [11], which is a measure of how much a probability distribution is divergent from another, among them. In this way, we obtain 2D and 3D point clouds of the MFCC of recordings carrying the intrinsic similarities and dissimilarities between them.

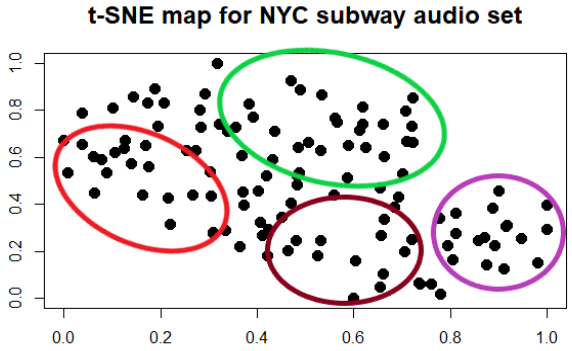


Figure 1: The 2D point cloud for New York Subway street recordings after applying t-SNE. The red ellipse indicates a cluster of percussive sounds, the green ellipse includes vocals, the burgundy ellipse are string sounds, and the purple ellipse includes brass sounds.

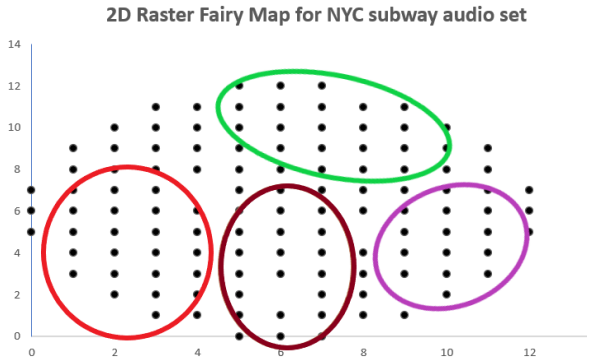


Figure 2: 2D Raster Fairy cloud map for New York Subway, for an evenly distributed encoding. Clusters are the same as in Figure 1.

After applying t-SNE to the extracted MFCC and its time derivatives of the recordings, we obtained 2D and 3D point clouds in a dispersed form; some of the points are extremely close to each other where some of them are far. Therefore, they can be transformed into a regular raster cloud without losing their neighborhoods. We used Raster Fairy [10] to assign this diverse point cloud into a circle point cloud while preserving the similarities and dissimilarities between points. Essentially, the Raster Fairy encoding gives us an alternative 2D embedding that can be transformed to a 3D embedding in VR by raising each point in the latter environment. This would provide a regularized set of points for use in 3D for an evenly distributed encoding.

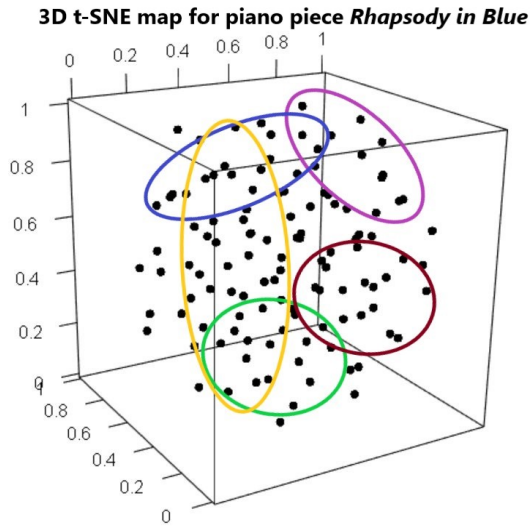


Figure 3: The 3D point cloud for a piano performance of *Rhapsody in Blue* after applying t-SNE to 8.5-second segments. Yellow ellipsoid includes fast sounds, green ellipsoid includes mellow sounds, burgundy ellipsoid includes monotonic sounds, purple ellipsoid includes rich sounds and the blue ellipsoid includes brisk sounds in the piano segments.

## 3.3 Virtual Reality Configuration

Using the coordinates obtained in the t-SNE and Raster Fairy reduction process, we place sound sources onto 3D locations in the Unity (2019.4.9f1 URP) development environment. Audio sources are either placed on a sphere corresponding to a 360 photo mapping in the New York subway case study or in 3D locations corresponding to a 3D t-SNE encoding in the piano performance case study. An application was built to add environmental context as detailed in the design section, then built for Oculus Quest 2 VR headset for subsequent prototyping. Additional augmented content was added as 360 photos or navigable assets in Unity.

# 4. DESIGN

To show how a spatial view of audio collections and segments can facilitate user experience, we prototyped two VR designs as case studies for enveloping sound experience. The first case study is dedicated to experiencing environmental sound and the other one is for understanding a song's internal structure. For each case, the building envelope, selecting tool, and encoding of the data are determined according to its purpose.

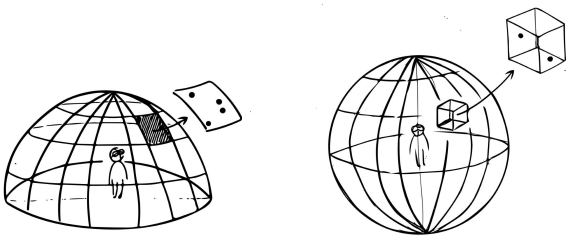

Figure 4: Different embeddings of audio sources in 2D (left) and 3D (right) spatial projections. Points represent the sounds in the collection. Half-sphere surface 2D mapping is used for the NYC subway music case study (left). The number of sources per unit area is 0.75 given a radius of 5 units. Full-sphere 3D mapping is used for the piano performance case study (right). The number of sources per unit volume is 0.22 when the radius is 5 units.

## 4.1 Case Study: Sounds of Street Performers in the New York City Subway System

Environmental sounds are a strong determinant of the way we experience space. To provide an immersive platform for audiences to experience the sonic environment of a cityscape, we recorded 117 clips of street music in the different NYC subway stations and applied the t-SNE / Raster Fairy strategies previously described.

We found that representation of the location-specific properties of the sound requires a spatial distribution of the t-SNE returned samples in a 2D sphere around the user. In such a format, using the reticle of the looking direction of the user provides optimal ability to discern and select different fragments of sound. The user selects the desired spheres that represent the recordings by directly looking at them. We calculate using ray tracing which of the audio source spheres are selected/highlighted. After that, the user finds herself in the virtual panorama (360 photo) of the station where the performer is recorded.

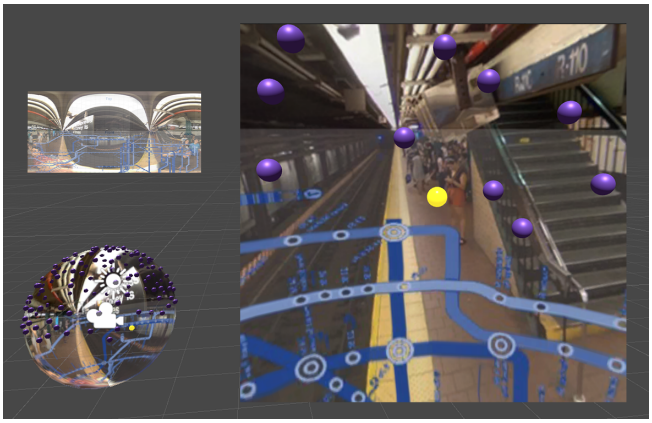

Figure 5: VR design for Case Study 1 (developer view on the left, user view on the right): the spheres are located on the upper surface of the outer sphere are representing the recordings of the street performers, and the selected spheres turn yellow. A transparent NYC Subway map is shown in 3D and the station where the performer is recorded is highlighted on the Subway map, along with the background of the 360 image of this station. The audio source spheres are equidistance away at the radius of the 360 photo of the subway station where the sound originated.

Moreover, on the bottom half of the sphere, an NYC Subway map in the equirectangular form that surrounds the user appears. The station that the recording took place is highlighted on this NYC Subway map conveying the impression of traveling between different subway stations.

In the case of the Raster Fairy constructed 2D grid, we found that the interaction was not as telling of the subjective distances between the sounds, since Raster Fairy imposes equal distances between the sources. In VR therefore it would be better to perform clustering that optimizes the ability to tell between similarity in the features of the sound using the original t-SNE encoding.

## 4.2 Case Study: The Structure of a Piano Performance

Longer works of music require sustained attention over the course of the performance on the part of the audience, whose mental state may vary at the beginning and end of the piece. To allow audiences to physically play with, and listen to each part of a complex work of music in an interactive, self-directed, nonlinear manner, we divided a 16:45 long work of professional performance of Gershwin's *Rhapsody in Blue* into 117 segments (8.5-second fragments) and produced a 3D embedding of the data using t-SNE. Using a controller it is possible to let the user interact with the sonic environment in a creative way. We used a 3D encoding here after we found that this allows for the most natural gestures for exploring the structure of a single piece interactively since in this case, the piece does not rely on contextual information, but rather on the ability to finely navigate the particular structure of the piece.

Moreover, in this case, we found that using the joystick controllers to identify and play the fragments best allowed for fine-tuned nonlinear navigation over the different fragments of work. To enhance the sense of place, the following design issues are taken into consideration: 1. two controllers can be used together and selected spheres are highlighted, 2. multiple sounds can be heard at the same time by using triggers in the controllers simultaneously, 3. navigating in 3D space leads to different views of the sources and different ways to use the controllers.

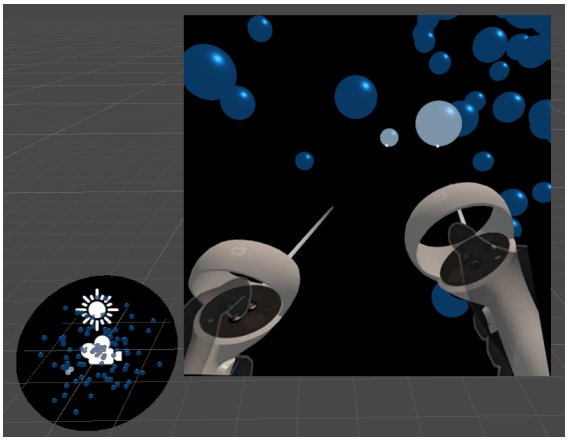

Figure 6: VR design for Case Study 2 (developer view on the left, user view on the right): the floating spheres represent the 8.5 seconds segments of the piano performance of *Rhapsody in Blue* as audio sources, with the selected spheres are highlighted with light blue. The sizes of the spheres indicate how far they are away in 3D space.

## 5. CONCLUSIONS

In this work, we prototyped audio spatial interaction with large data sets using machine learning representations. Firstly, we provided a musical soundscape of the subways in New York by transforming the perceived intrinsic features to the interactable and visualizable form. This interaction puts the audio sources in a sphere around 360 photos of the subway station to provide context to the machine learning representation. We then prototyped a contrasting case where a single musical work is broken down into segments which are then interactable in 3D space. Here the separate expressive parts of the music are selected and played using controllers to better allow nonlinear exploration of the single performance in VR.

The prototypes we explored prompted a new way of interacting with complex data. By using machine learning to pre-categorize our audio data, we envision a future where single glances and fast spatial exploration in 3D is utilized to convey the essence of entire musical and sound collections. It thus allows us to experience nuanced regimes like the sonic space of a place or the structure of a long piece of music using an augmented form of intuitive understanding in space, in short the "sound of" an environment or formal structure.

## 6. REFERENCES


[1] Sebastian Böck, Florian Krebs, and Markus Schedl. *Evaluating the Online Capabilities of Onset Detection Methods*.

[2] Georgina Born. 2013. *Music, Sound and Space: Transformations of Public and Private Experience*. Cambridge University Press.

[3] CJ Carr and Zack Zukowski. 2019. Curating Generative Raw Audio Music with D.O.M.E. *Los Angel.* (2019), 4.

[4] Michael Casey, Christophe Rhodes, and Malcolm Slaney. 2008. Analysis of Minimum Distances in High-Dimensional Musical Spaces. *IEEE Trans. Audio Speech Lang. Process.* 16, 5 (July 2008), 1015–1028. DOI:https://doi.org/10.1109/TASL.2008.925883

[5] M. Cavallo, M. Dholakia, M. Havlena, K. Ocheltree, and M. Podlaseck. 2019. Dataspace: A Reconfigurable Hybrid Reality Environment for Collaborative Information Analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 145–153. DOI:https://doi.org/10.1109/VR.2019.8797733

[6] A. Flexer. 2015. Improving Visualization of High-Dimensional Music Similarity Spaces. *ISMIR* (2015). Retrieved May 14, 2021 from /paper/Improving-Visualization-of-High-Dimensional-Music-Flexer/6861648ea009eec227b2d53c0da03ad8e3e9c183

[7] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. DOI:https://doi.org/10.1109/ICASSP.2017.7952261

[8] Oscar Gomez, Kaustuv Kanti Ganguli, Leonid Kuzmenko, and Carlos Guedes. 2020. Exploring Music Collections: An Interactive, Dimensionality Reduction Approach to Visualizing Songbanks. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (IUI '20), Association for Computing Machinery, New York, NY, USA, 138–139. DOI:https://doi.org/10.1145/3379336.3381461

[9] Stanislav Klimenko, Michael Charnine, Oleg Zolotarev, Nadezhda Merkureva, and Aida Khakimova. 2018. Semantic approach to visualization of research front of scientific papers using web-based 3D graphic. In *Proceedings of the 23rd International ACM Conference on 3D Web Technology* (Web3D '18), Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/3208806.3208825

[10] Mario Klingemann. 2016. *Raster Fairy*. Retrieved from www.underdestruction.com/2016/02/25/raster-fairy-2016

[11] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Ann. Math. Stat.* 22, 1 (March 1951), 79–86. DOI:https://doi.org/10.1214/aoms/1177729694

[12] Franz de Leon and Kirk Martinez. Enhancing Timbre Model Using MFCC and Its Time Derivatives for Music Similarity Estimation. 5.

[13] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. 2001. Classification of general audio data for content-based retrieval. *Pattern Recognit. Lett.* 22, 5 (April 2001), 533–544. DOI:https://doi.org/10.1016/S0167-8655(00)00119-7

[14] Beth Logan. 2000. Mel Frequency Cepstral Coefficients for Music Modeling. *Proc 1st Int Symp. Music Inf. Retr.* (November 2000).

[15] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 86 (2008), 2579–2605.

[16] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. Austin, Texas, 18–24. DOI:https://doi.org/10.25080/Majora-7b98e3ed-003

[17] Chris Muelder, Thomas Provan, and Kwan-Liu Ma. 2010. Content Based Graph Visualization of Audio Data for Music Library Navigation. In *2010 IEEE International Symposium on Multimedia*, 129–136. DOI:https://doi.org/10.1109/ISM.2010.27

[18] Meinard Müller. 2007. *Information Retrieval for Music and Motion*. Springer-Verlag, Berlin Heidelberg. DOI:https://doi.org/10.1007/978-3-540-74048-3

[19] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON*, 6.

[20] Karol J. Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. DOI:https://doi.org/10.1109/MLSP.2015.7324337

[21] Feng Rong. 2016. Audio Classification Method Based on Machine Learning. In *2016 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, 81–84. DOI:https://doi.org/10.1109/ICITBS.2016.98

[22] Yi Yu, Samuel Beuret, Donghuo Zeng, and Keizo Oyama. 2018. Deep Learning of Human Perception in Audio Event Classification. In *2018 IEEE International Symposium on Multimedia (ISM)*, 188–189. DOI:https://doi.org/10.1109/ISM.2018.00-11