SOUND OF(F): Contextual storytelling using machine learning representations of sound and music

Zeynep Erol[1], Zhiyuan Zhang[2], Eray Uzgunay[1], and RAY LC[2]

[1] Hong Kong University of Science and Technology, Hong Kong
zeynep.erol@metu.edu.tr, eouzgunay@connect.ust.hk
[2] City University of Hong Kong, Hong Kong
zzhang452-c@my.cityu.edu.hk, LC@raylc.org

Abstract. In dreams, one's life experiences are jumbled together, so that characters can represent multiple people in your life and sounds can run together without sequential order. To show one's memories in a dream in a more contextual way, we represent environments and sounds using machine learning approaches that take into account the totality of a complex dataset. The immersive environment uses machine learning to computationally cluster sounds in thematic scenes to allow audiences to grasp the dimensions of the complexity in a dream-like scenario. We applied the t-SNE algorithm to collections of music and voice sequences to explore the way interactions in immersive space can be used to convert temporal sound data into spatial interactions. We designed both 2D and 3D interactions, as well as headspace vs. controller interactions in two case studies, one on segmenting a single work of music and one on a collection of sound fragments, applying it to a Virtual Reality (VR) artwork about replaying memories in a dream. We found that audiences can enrich their experience of the story without necessarily gaining an understanding of the artwork through the machine-learning generated soundscapes. This provides a method for experiencing the temporal sound sequences in an environment spatially using nonlinear exploration in VR.

Keywords: Spatial Audio; Virtual Reality Art; Machine Learning; t-SNE; Sound Visualization; Nonlinear Listening.

1 INTRODUCTION

Our dreams are full of unexplored data, from memories that we seem to have forgotten about to sounds that we hear incompletely but feel completely at home with. Dreams are expressions of our collective memories, much like the way machine learning represents large data sets using a memory-network-based model. To explore the way we represent music and sound in our dreams, we applied machine learning to spatially cluster both whole works of music and sound recordings, converting temporal representations to spatial interactions, These interactions allow audiences to grasp a long piece of music without having to listen to it sequentially, but rather explore fragments of the piece put in related locations by the similarity of their sound features. We first designed these spatial interactions of sound in two related case

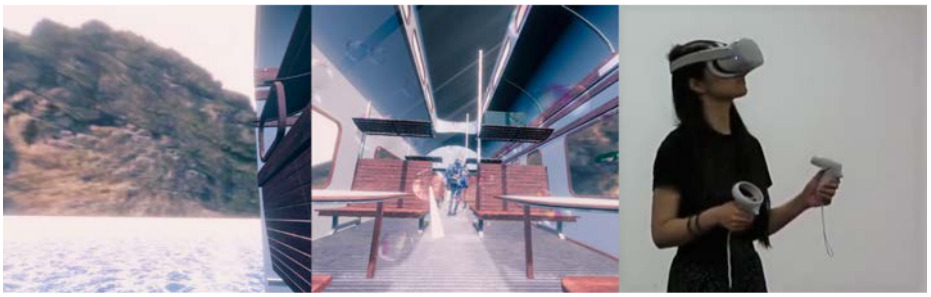

Fig. 1. (**Left**) The natural landscape outside of the train generated from a 360 photo dataset by machine learning algorithm StyleGANS2. (**Middle**) The interior of the train with translucent bubbles being pointed at by the controller as the source of sound clustered by t-SNE. (**Right**) An audience member experiencing Sound Of(f) using the VR headset and controllers.

studies, then uses these prototype findings to create a Virtual Reality (VR) work that narrates a particular looping memory in dreams. Evaluation with exhibition audiences revealed that the sound interactions enriched the experience of the story without affecting the understanding of particular scenes, showing that machine learning-based spatial interactions of sound and music can foster a new way to perceive diverse and potentially incongruent sources of data as one may find in a dream-like state.

Current approaches to understanding music and sound include song-content-based visualization using graphs [19] and reduced dimensionality visualization using musical features [9], part of a general attempt to understand music by visualizing the computation of features. In addition, visualizations of music databases allow intuitive exploration of music catalogs. Music and sounds, however, are part of an environment that evokes nuances that are 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 [3]. To better grasp the experience of long works of music and large sets of sound recordings, we used the t-SNE algorithm [16] to cluster related sounds for interaction in 3D space in a VR environment.

In the testing phase, we prototyped both segmentation of a *single* work of music to convert a temporal experience to a spatial one, as well as clustering of *different* sound recordings in a single environment to grasp the scene soundscape. We designed spatial interactions like 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. In the creative phase, these design principles allowed us to create a VR artwork exhibited at a local gallery that attempts to blend different clustered sound and music fragments with particular scenes that convey the mood and nuance of "saying goodbye," "hope," "longing," "misunderstanding," and "silence," conveying the way rich sources of information blend together in a dream.

2 BACKGROUND

Previous attempts at understanding complex audio data must deal with a large amount of information under consideration, and have included metrics that make the retrieval process more efficient [5]. These approaches rely on efficient classification schemes that resonate with human perception [14] but require a user-centered design perspective to implement. Machine learning has been applied to high dimensional audio classification using features of the sound [23], but these computational approaches do not always produce the phenomenological separations in human sound classification[8]. Similarly, environmental sounds have also been classified using convolutional neural networks [22]. 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 [25].

One way to overcome the divide between the 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 workflows [6], visualization relationship amongst scientific paper corpora [10], musical catalog visualization [7], cultural analyses of musical patterns [9], and previewing audio samples using dimension reduction techniques [4]. 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.

Previous artworks like Blortasia have explored the effect of soundscapes on the unreal state of a virtual environment [17] but uses abstract shapes and colors to represent the abstract world instead of a reality-based transfiguration in dreams.

3 TECHNOLOGY VALIDATION

To test our machine learning technology, we first collected 117 sounds of subway street musicians in New York City to form a sound collection that can be grouped according to musical features by machine learning. We then use t-SNE to populate these sounds in a 2D sphere around the audience. In addition, we also used a single 16:45-long performance of Gershwin's *Rhapsody in Blue* to use for the application of segmenting of a single musical work. For this work, we put sounds in 3D space using t-SNE also clustered by similarity. This allows us to test both the clustering of different sound samples to understand an environment, and the segmentation of a single musical work to transfer temporal experience to spatial experience.

3.1 Audio Processing

For segmenting the single piano piece *Rhapsody in Blue*, we used an onset detection algorithm [2] embedded in a MIR python package (librosa [18]). This algorithm divides the piece into chunks by detecting their onsets, i.e. beginning of the transient parts. For the collection of sound recordings in the New York Subway stations, we do

not break them into chunks because we directly use the recordings for clustering, but instead, we take segments of 10 seconds from each sample for subsequent analysis.

Next, we generate the feature vectors of the sounds to capture the parameters of the recordings and segments. One way to capture these features is to obtain the Mel-frequency cepstrum coefficients and their time derivatives of the sounds, which are used in speech and music information retrieval (MIR) and processing [13,15,20]. We get these coefficients, 13 for each recording, and their first and second-time derivatives, called first and second delta features, using librosa [18]. 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**

Using the 39-dimensional vectors for each recording, we transform them into 2D or 3D spaces for human visualization. We use the scikit-learn's tool [21] dimension reduction method t-SNE [16] for this purpose. t-SNE is a machine learning technique that assigns every point located in a higher dimension onto a location in the two or three-dimensional space by considering pairwise similarities between these points. t-SNE essentially clusters the audio data while maintaining the local structure of the data by transforming the similarities between vectors into the models of joint probability distributions and minimizing the Kullback–Leibler (KL) divergence [12] (a measure of how much a probability distribution is divergent from another) between them. Compared with linear dimensionality reduction methods like PCA, which focus on separating the low-dimensional representation, t-SNE puts similar neighbors close together in low-dimensions, making it easier to cluster sounds to different genres.

In detail, one problem with clustering algorithms involves the "crowding problem", in which the points that are far away in the high-dimensional space are all packed together in the low-dimensional space because there is less space to work with in low-dimensions [16]. A direct consequence of the crowding problem is that the separated clusters in the high-dimensional space are not clearly divided in the low-dimensional space. t-SNE solves the crowding problem by replacing the joint Gaussian model distribution being minimized with a Student-T distribution. Other non-linear methods like Isomap[1] and LLE [24] are more suitable for unfolding a single continuous low-dimensional manifold, but are not as desirable as t-SNE in this case due to the multiple manifold structure of audio data.

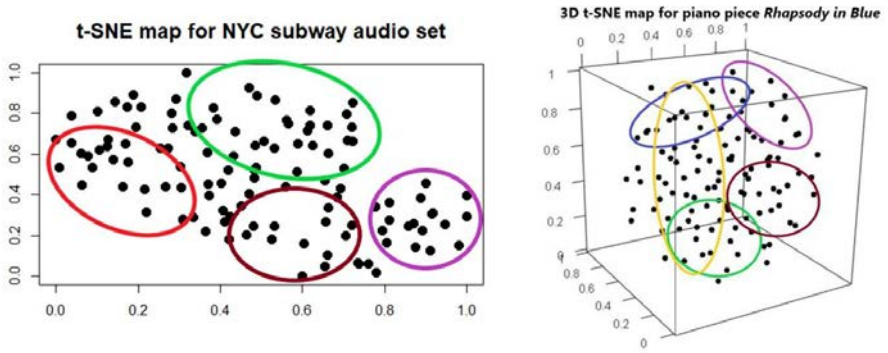

Fig. 2. (**Left**) The 2D point cloud for New York Subway street recordings after applying t-SNE. The red ellipse indicates a cluster of percussive sounds, the green ellipse includes vocals, the burgundy ellipse is string sounds, and the purple ellipse includes brass sounds. (**Right**) The 3D point 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 blue ellipsoid includes brisk sounds in the piano segments.

After applying t-SNE to the extracted MFCC and its time derivatives of the recordings, we obtained 2D and 3D point clouds of the sound data in a dispersed form. Some of the points are extremely close to each other while others are far away. In the 2D output case, they can be transformed into a regular raster cloud without losing their neighborhoods. We used Raster Fairy [11] to assign this diverse point cloud into a circular 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 putting each point in the skybox environment. This would provide a regularized set of points for use in 3D for an evenly distributed encoding.

3.3 Virtual Reality Configuration

Using the coordinates obtained in the t-SNE and the Raster Fairy reduction process (in the 2D case), we place sound sources onto 3D locations in the Unity (2019.4.9f1 URP) development environment. Audio sources are either placed on a sphere (2D) 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.

To see how a spatial view of audio collections and segments can facilitate user experience, we prototyped two VR designs as case studies for the final artistic output. In one case, we let users experience environmental sounds of the NY subway system by putting sounds clustered by t-SNE around a sphere. In the other case, we put fragments segmented from a single work of music and grouped by t-SNE in 3D space.

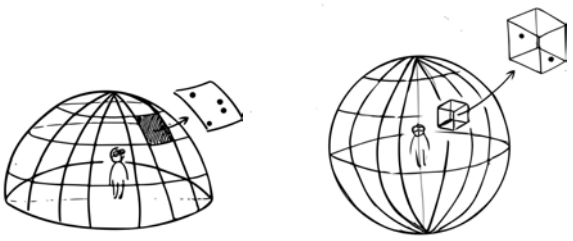

Fig. 3. Different embeddings of audio sources in 2D (left) and 3D (right) spatial projections. Points represent the sounds in the collection. Half-sphere surface 2D mapping is used for the NYC subway music case study (**Left**). The number of sources per unit area is 0.75 given a radius of 5 units. Full-sphere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

3.4 **Testing Study: Sounds of Street Performers in the New York City Subway**

Environmental sounds are a strong determinant of the way we experience space. To test an immersive platform for audiences to experience the sonic environment of a soundscape, 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 photos) of the station where the street performer is recorded. 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 similarities in the features of the sound using the original t-SNE encoding.

Fig. 4. (**Left**) VR design for Case Study 1: the spheres located on the upper surface of the outer sphere (2D) 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 images of this station. The audio source spheres are equidistance away at the radius of the 360 photos of the subway station where the sound originated. (**Right**) VR design for Case Study 2: the floating spheres represent the 8.5 seconds segments of the piano performance of *Rhapsody in Blue* as audio sources, with the selected spheres highlighted with light blue in 3D. The sizes of the spheres indicate how far they are away in 3D space.

3.5 Testing 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. In particular the 3D sense of space allowed users to play better with sounds structurally and provide an immersive experience outside the screen-like sphere encoding in the previous test study. 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.

4 DESIGN OF SOUND OF(F)

We learned from our previous test studies that placing sounds in 3D using a controller-based interaction best facilitates the music exploration process. Hence we decided to use the 3D t-SNE encoding for our subsequent sound placement.

Sound Of(f) is a narrative VR experience that uses machine learning clustered soundscapes and 360 video landscapes to build an environment around different themes in a dream-like exploration on a metaphorical train. The inspiration for the work comes from a dream where someone close to the dreamer exited the train at her exit without ever saying goodbye. The dreamer then wakes up to find that he has no recollection of the exact person involved but rather a continuous fusing of one person's face with another person's personality, the product of a dream that, like machine learning models, fuse together personalities in creating characters and landscape. The character is intended to have within a single person the characteristics of multiple people in our lives, just as the landscape transitions between many places we know.

The experience takes around five minutes and loops repeatedly. The audience begins inside a train where she can move around using the joystick in the controller but cannot exit the train entirely. The train is going nowhere and everywhere. Noises in the form of music and sounds selected for each theme and clustered by t-SNE are found in the environment. They may be sounds that we don't want to hear waking us up, or fake news that we don't want to pay attention to, or misinformation designed to trigger our behaviors, but each set of sound revolves around a theme. The audience can listen to the spatial audio presented as translucent spheres in the scene grouped by t-SNE so that similar-feature audio clips are close to each other in space. The sounds are played back (while the sphere colors change) when you hover the mouse over individual clips. Audiences can feel like they are inside sound sequences, with the ability to explore them spatially rather than passively listen to them temporally.

The sounds for each thematic scene consist of sets of both music and sound recording fragments. The themes, presented in order, are: *Goodbye*, *Hope*, *Longing*, *Misunderstanding*, and finally *Silence*, each with its own associated character animation of a different way for the character to leave the train. For the *Goodbye* theme, we used Hiroshi Sato (佐藤博)'s *Say Goodbye* (セイグッバイ) and the goodbye final scene from the movie *Casablanca* where the main characters go their separate ways at when Ilsa boards the plane. Both are nostalgic and have different approaches to say goodbye: one is hopeful and the other one is dramatic and sad. For the *Hope* theme, we used a sound recording of a song sung by local Hong Kong people and Martin Luther King's "I have a dream" speech, which is inviting the dreamer to hope for future. For the *Longing* scene, we used the song *Apo Mesa Pethamenos* and the dialog from the car scene in the movie *Before Sunset*, both of which are about longing for someone, focusing after a breakup. For the *Misunderstanding* theme, we used Animals' song *Don't let me be Misunderstood* and the movie *The Switch* that puts the audience inside a fight. The songwriter shouts out about how everybody doesn't understand him and in the movie we hear one couple's dialog consistently fail to understand each other. In this scene, we also change the perspective so that the audience is looking at the character and sound bubbles from above, narrating the misunderstanding idea. For the *Silence* theme, we leave it to similar sounding

meditative sounds. This is the literal “sound off” for all noises as well as a goodbye to our character, who now stands outside the train for the first time while the train has stopped moving and the landscape outside has stopped moving. After our intimate character has repeatedly stepped off the train without saying goodbye, we ourselves finally say goodbye ourselves, in order to turn off the sound.

Fig. 5. Installation of the VR artwork. (Upper Left) Poster for the exhibition. (Upper Right) Location of the poster and headset relative to the wall. (Lower Left) Positioning of the headset on top of a plinth with cable entering the box, and two controllers on hinge brackets. (Lower Right) An audience member interacting with the work in the layout designed for the show.

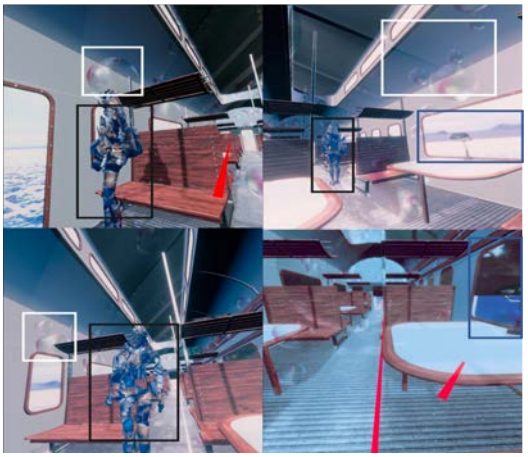

Fig. 6. Interior view of the train. The black rectangle contains the character. The dark blue rectangle contains the landscape, which is the 3D video generated by the machine learning model. The rectangle circle contains the bubbles in the air, which can be triggered to play t-SNE sounds.

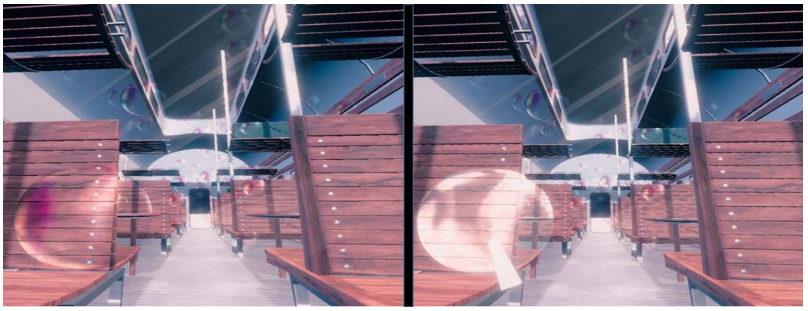

Fig. 7. Interaction with sound by pointing: audience can hear and explore the spatial audio (**Left**) grouped by machine learning using the red laser pointing at the bubble (**Right**).

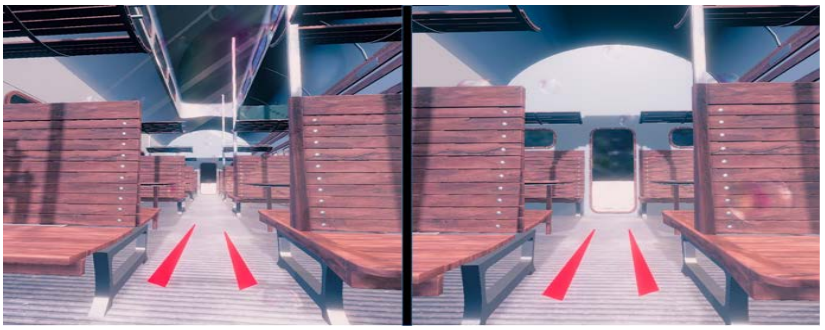

Fig. 8. Interaction with the environment by joystick: audience can walk and explore the inside of the train. (**Left**) Before walking movement. (**Right**) After walking movement.

Fig. 9. An intimate character contains the characteristics of multiple people in our lives. During the journey, the character is changing repeatedly and walking off the train. In the last scene our intimate character is seen outside the train and walking off in a new direction without saying goodbye. (**Left**) Close-up of the character to see the shader working. (**Right**) Standing just under the character in an early scene.

In terms of the context, the train is running on a moving ocean. The interior decoration of the train is nostalgic and given bloom effects to emphasize the dreamy scene. As the train moves, the view outside of the windows will transit between many places seamlessly. The landscape skybox is a looping 360 video generated by state space traversal through a StyleGAN2 machine learning model trained using 478 total 360 photos taken by the authors at local landscape locations. As the audience walk in the train, turn around, look out of the windows, and explore the spatial sounds grouped by t-SNE using the controller, they find the ever-changing character on the train that acts on her own to leave the train. Everytime she leaves, the scene transitions to the next segment. The previous character that stepped off the train without saying goodbye is replaced by the same character at a different position. The audience is also relocated to a different location of the train, while the sound bubbles are updated. To see the VR experience, see this link: https://youtu.be/yMyR5DKjGA0

5 EVALUATION

5.1 Methodology

We surveyed 26 people at the opening of the exhibition immediately after they experienced all five scenes of the art work (14 female). The sample included 18-25 year-olds (13), 26-35 year-olds (5), 36-50 year-olds (5), and over 50 year-olds (3). The experimenters alerted the visitors that the controller can only be used for navigation and pointing and warned about possible dizziness. Visitors were allowed to immediately stop playing when experiencing strong dizziness. If the visitor completed the game, they were asked to participate in the survey immediately after removing the headset. The 12-question survey was given on a tablet, and took approximately five minutes, including 6 likert scale ranking questions (1-7) and 6 open-ended questions.

5.2 Quantitative findings

Likert scale questions revealed that participants experienced the environment as a dream-like experience rather than a realistic one, and that the clustering of the sounds appear to be well-organized. Interestingly, the sound organization appear to contribute to the experience of the story but not to the understanding of what the scenes mean." This may be due to the difficulty of inferring the theme for each scene from the audio information alone, or to the lack of other identifying information that reveals the theme (goodbye, misunderstanding, etc.) We were also encouraged by the ability of the sounds to capture the mood of the scene, suggesting the effectiveness of spatial interactions to form a milieu of environments rather than sequential sounds like playing music from beginning to end.

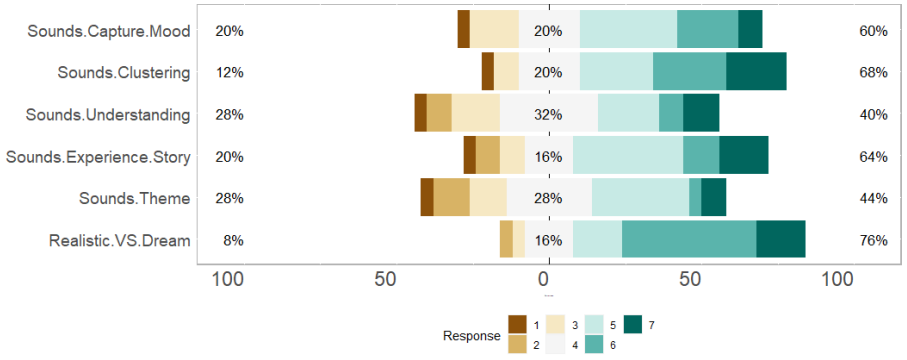

Fig. 10. Quantitative audience evaluation following playthrough of the entire set of 5 scenes. Sounds Capture Mood rated 1-7 for "How well do the sounds of each scene capture the mood of the particular scene?" Sounds Clustering rated 1-7 for "How well are the sounds in each scene clustered into related close-by fragments?" Sounds Understanding rated 1-7 for "How strongly do the sound fragments facilitate your understanding of what is happening in the scene?" Sounds Experience rated 1-7 for "How well do the audio fragments contribute to your experience of the story in VR?" Story Theme rated 1-7 for " Based on playing through each scene, how much have you grasped the theme of the audio in each scene?" Realistic VS. Dream rated 1-7 for "How much does the environment evoke a realistic vs an abstract, dream-like state?"

5.3 Qualitative findings

Most audiences found their experience to be dream-like, as intended by the intervention. Almost half of the participants (11 of 25) chose "a dream-scape" as the answer to "Which do you think is the topic of this VR experience?" For example, one participant described the story as "*dreamy meta good-bye to self*," while another wrote that "*you are on a train in a dreamy situation and there is a perturbed shape of a woman who moves around; everything feels half a dream, half real*." The dream-like experience may come from the evolving landscape, the moving and stationary train, the eerie soundscape and postprocessing, the change in perspective in one scene, etc, but one unexpected source of the dreaminess is the ever-present character. One audience member noticed that when they get close to the character, she does not interact with them, which gives an eerie feeling of an unreal dream space. Interestingly, participants tended to be attached to the intimate figure. They had a connection with the character and made comments like "*I have liked that person*," going so far as to attempt to follow her out of the train in some scenes. One person said "*the user (me) wants to speak to the character really badly but she leaves the train and cannot be reached!*"

The interactions of different participants reflected their experience and personality. For example, some visitors moved their hands often and heard sound fragments only briefly, while others were more deliberate, hearing the entire sound one bubble at a time patiently before moving on. Some stayed in one place for the duration of the experience while others tried to go outside first and found themselves being stuck inside. Younger visitors appear to adapt themselves easily, having the most fun and interaction throughout the exhibit session. Older visitors take time to get adjusted, and tend to tire and get dizzy quickly. A few did not finish all five scenes, and were left with partial knowledge based on their incomplete experience, holding different opinions about the scenes they did see. However, some older participants found the scene relaxing as they slowly went through the sounds, especially the silence scene.

While participants understood a difference in the sounds and music used in each scene, they often didn't perceive the theme being portrayed. They variously described the sounds heard as "*political scene*," to "*noisy restaurants*," and "*orchestral music in old movies*." One of the most common descriptions, however, involves comparing the experience to a radio. One audience member described the experience as "*this is like searching for a channel on the radio.. clustering the sounds, and trying to find the correct one.*" The idea of the radio strongly reflects the idea of spatial navigation of sound, in that people can turn a dial spatially and explore nearby channels to hear fragments of sonic experience instead of temporally listening to the entire piece. The way participants gravitated towards this type of interaction may reflect the need to turn temporal sonic events into spatial movement events for a global view of the soundscape, instead of listening to an entire sonic experience beginning to end.

6 CONCLUSIONS

In this work, we created an immersive environment for storytelling using spatial interactions for canonically temporal audiom, shaped by machine learning clustering technique and 360 degree panoramic video generated by the machine learning. First, we prototyped a musical soundscape of the subways in New York using a spherical embedding of the sound collection and a reticle-based pointing system. This interaction puts the audio sources in a sphere around 360 photos to provide context to the machine learning representation. Next, we prototyped a contrasting case where a single musical work is broken down into segments that 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 musical work in VR.

Using the findings in these prototypes, we then created an artwork that applied the t-SNE strategy of clustering sounds for spatial interaction into a narrative context, exploring a way of interacting with audio data spatially. We have used the interactable 3D space and combined the two case studies' approach: using single audio and multiple sounds, since both approaches have their own outcomes which help telling a story in the VR environment. However we chose to use controller-based operation for its more precise control over sound selection. We further supported the work with GAN-generated 360 video landscapes. The sounds are key elements for storytelling, with a set of five different themes inside the dream-like setting with a unique design of the character that represents the multiplicity of intimate people in our dreams.

Audience evaluation further showed how the experience of the story can be enhanced by the spatial sound interactions while the understanding of the scenes may not be affected. It also showed how spatial interactions of sound may already be present in a simplified form in the case of the radio, and points out the general complementarity of spatial and temporal interactions for sound. By using machine learning to pre-categorize our audio data, we envision a future where single glances and fast spatial exploration in 3D are utilized to convey the essence of entire musical works. It thus allows us to experience the story and its thematic elements as sonic spaces of different sound recordings or fragments of a long piece of music using an augmented form of intuitive understanding in space, in short, the "sound of" an environment.

Supplemental Materials

To see the interactions in VR during gameplay, see: https://youtu.be/yMyR5DKjGA0

References

1. M. Balasubramanian. 2002. The Isomap Algorithm and Topological Stability. *Science* 295, 5552: 7a–77.
2. Sebastian Böck, Florian Krebs, and Markus Schedl. *Evaluating the Online Capabilities of Onset Detection Methods*. .
3. Georgina Born. 2013. *Music, Sound and Space: Transformations of Public and Private Experience*. Cambridge University Press.
4. CJ Carr and Zack Zukowski. 2019. Curating Generative Raw Audio Music with D.O.M.E. *Los Angeles*: 4.
5. Michael Casey, Christophe Rhodes, and Malcolm Slaney. 2008. Analysis of Minimum Distances in High-Dimensional Musical Spaces. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 5: 1015–1028.
6. M. Cavallo, M. Dholakia, M. Havlena, K. Ocheltree, and M. Podlaseck. 2019. Dataspace: A Reconfigurable Hybrid Reality Environment for Collaborative Information Analysis. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 145–153.
7. A. Flexer. 2015. Improving Visualization of High-Dimensional Music Similarity Spaces. *ISMIR*.
8. Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, et al. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.
9. Oscar Gomez, Kaustuv Kanti Ganguli, Leonid Kuzmenko, and Carlos Guedes. 2020. Exploring Music Collections: An Interactive, Dimensionality Reduction Approach to Visualizing Songbanks. *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, Association for Computing Machinery, 138–139.
10. 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. *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Association for Computing Machinery, 1–6.
11. Mario Klingemann. 2016. *Raster Fairy*. .
12. S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1: 79–86.
13. Franz de Leon and Kirk Martinez. Enhancing Timbre Model Using MFCC and Its Time

Derivatives for Music Similarity Estimation. 5.
14. Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. 2001. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters* 22, 5: 533–544.
15. Beth Logan. 2000. Mel Frequency Cepstral Coefficients for Music Modeling. *Proc. 1st Int. Symposium Music Information Retrieval*.
16. Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86: 2579–2605.
17. Kevin Mack. 2017. Blortasia: a virtual reality art experience. *ACM SIGGRAPH 2017 VR Village*, Association for Computing Machinery, 1–2.
18. Brian McFee, Colin Raffel, Dawen Liang, et al. 2015. librosa: Audio and Music Signal Analysis in Python. 18–24.
19. Chris Muelder, Thomas Provan, and Kwan-Liu Ma. 2010. Content Based Graph Visualization of Audio Data for Music Library Navigation. *2010 IEEE International Symposium on Multimedia*, 129–136.
20. Meinard Müller. 2007. *Information Retrieval for Music and Motion*. Springer-Verlag, Berlin Heidelberg.
21. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*: 6.
22. Karol J. Piczak. 2015. Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.
23. Feng Rong. 2016. Audio Classification Method Based on Machine Learning. *2016 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, 81–84.
24. S. T. Roweis. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290, 5500: 2323–2326.
25. Yi Yu, Samuel Beuret, Donghuo Zeng, and Keizo Oyama. 2018. Deep Learning of Human Perception in Audio Event Classification. *2018 IEEE International Symposium on Multimedia (ISM)*, 188–189.