Voice Responsive Virtual Reality

Adinda van 't Klooster
School of Digital Arts
Manchester Metropolitan University
UK
A.vant.Klooster@mmu.ac.uk

Nick Collins
Music Department
Durham University
UK
nick.collins@durham.ac.uk

ABSTRACT

Our audio-reactive Virtual Reality (VR) interface encourages a performer to explore extended vocal techniques within an alternative visual aesthetic experience. We are interested in whether this can support creative vocal expression, and whether the sound-responsive animated hand-drawn aesthetics in the VR scene can help immersion and wellbeing. As an interface for musical expression, there is no extraneous sound output, but rather, an attempt to encourage investigation of a user's own voice through a visual feedback system. We describe the technicalities of building the system as an Oculus Quest app, and evaluation through various existing theories pertinent to VR musical interaction.

Author Keywords

NIME, Machine Listening, Virtual Reality, Voice-controlled graphics

CCS Concepts

• Applied computing → Arts and humanities → Sound and music computing; **Performing arts; Media arts**;

1. INTRODUCTION

It is well established that singing can benefit physical and mental health [3, 17]. In the present project, a virtual reality (VR) interface is presented where graphics are reactive to analysis of audio input, with the aim of encouraging a playful vocal interaction and potentially increased wellbeing. The interface was built for the Oculus Quest 2 and 3 using Unity and a realtime audio analysis library.

Although current VR has nothing near the mass population reach of mobile apps, devices such as the Oculus Quest have made access to VR straight forward, with high frame rate and inside-out tracking now standard. Artists have exploited VR tech since the late 1980s [19, 16, 18] and more recent projects utilize high frame rates, high graphics resolution, spatial audio and fast realtime reactions.

VR has been used in different contexts to improve wellbeing and health, from combating phobias or permanent pain [14] to the measurement of emotions following a VR-based museum visit [5]. The potential impact that the arts can have on emotional wellbeing also been recognised [1, 12], and is a growing area of practice for artists and cultural organisations. The link between singing and wellbeing is partially due to physiological factors [8] and vocal exploration helps build a sense of identity [4, 11]. Although the community aspect of singing is often highlighted, we feel there is a strong benefit to the solitary experience of being immersed in VR when using the voice, promoting less hesitation about being observed by others [7]. This project encourages people to free their voice, going beyond singing and humming to individual vocal expression with graphics to encourage vocal exploration.

NIME'24, September 4–6, 2024, Utrecht, The Netherlands

The interface is introduced in section 2, and evaluated with existing NIME design frameworks in section 3; section 4 presents discussion including plans for future work.

2. The Interface

2.1 Overview

The user is located on a terrain in a large virtual space (see Figure 1). The graphical style is based upon real world 2D drawings by the first author, digitized at high resolution and used for surrounding skybox backgrounds and textured objects. The user can move around the space using the joystick on the right Oculus controller combined with head movement; direction of the gaze determines the direction of the movement.

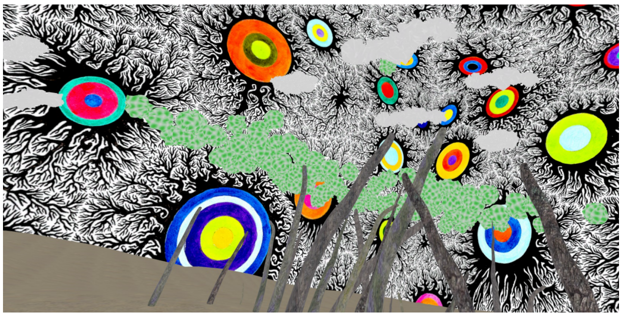

Figure 1: Screengrab from in-headset experience, © Adinda van 't Klooster 2023

A performer uses their voice to control a set of graphical objects generated as they navigate the scene; which audio features control which objects is dynamically allocated. Table 1 lists the main graphical components. Aside from being moved by audio features, the shapes are programmed to follow the user's gaze.

Graphical element	Description	Mapping
Trees	Recursively constructed trees from a set of trunk and branch models. Branches appear gradually over the depth of the tree, from trunk to smallest branch, continuously scaling up to their full size.	Tree branches rotate based on audio feature values (branch at depth N uses feature N%4); player proximity determines the amount of the tree visible.
Clouds	Bulbous cloud models with associated animations played at varying speeds. Vertices for the cloud can be perturbed along the normal direction at each	Audio features mapped to step animation, and to drive mesh vertex deformation. Clouds are spawned in reaction to audio onsets if enough

	point to swell or squeeze the object. Clouds travel continuously, adjusting to stay perpendicular to the gaze of the user.	time has passed since a previous spawn.
Flock	A flock of blob objects swarm in front of the player in the virtual world, acting to move away whenever approached.	Two new flock members are spawned on reaction to a percussive onset. The flock's movement is driven by brightness of sound, pitch and volume.
Terrain	The player is restricted to stay above a large terrain, which also grounds the trees and the flock	No audio feature reactivity

Table 1 Graphical elements and mappping

2.2 Artistic process

This practice-as-research incorporates the overlapping areas of gaming, fine art, coding, music and audiovisual improvisation. Aiming for a truly immersive experience, the initial exploration focused on bringing 2D drawings by the first author into the VR space, devising mappings from sound input to visual output. Three ink drawings were made specifically for the interface (Figures 2 to 4). They stand on their own as fine art drawings, and due to their size and pattern are already visually immersive as 2D artworks.

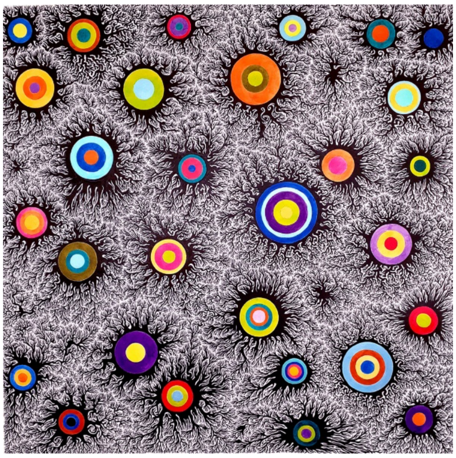

Figure 2: *The Roaming Eye*, Ink on Paper, 73.5 x 73.5 cm, © Adinda van 't Klooster, 2022

The drawings were made in a square format to allow a seamless transfer into the digital landscape (textures need to be square when using the six-sided cube as a base for the Skybox). Figure 3 didn't work as a 6-sided cube, so we instead used the panorama Skybox, requiring further digital manipulation in Photoshop.

The drawings are each based on almost fractal patterns, reflecting on forms found in nature. *Ectopic* took nine months to create and was made after an ectopic pregnancy; The pattern is one of leaves that could be part of a flower but are falling apart. The negative spaces of the leaves are created by small branching trees. The time-consuming nature of the drawing fitted with wanting to express a sense of time passing and the ageing process without directly referring to it. The other two drawings are more abstract, though all were envisaged from the start as never ending landscapes.

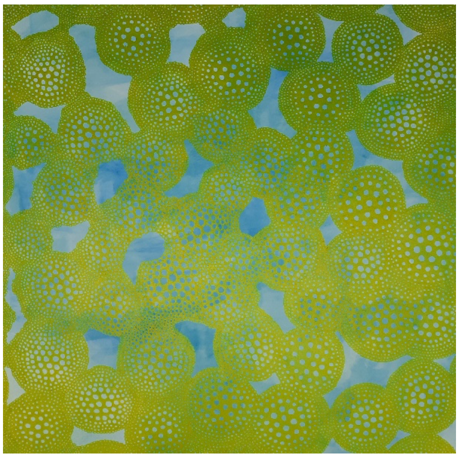

Figure 3: *Cell Sky*, Ink on Paper, 73.5 x 73.5 cm, © Adinda van 't Klooster, 2022

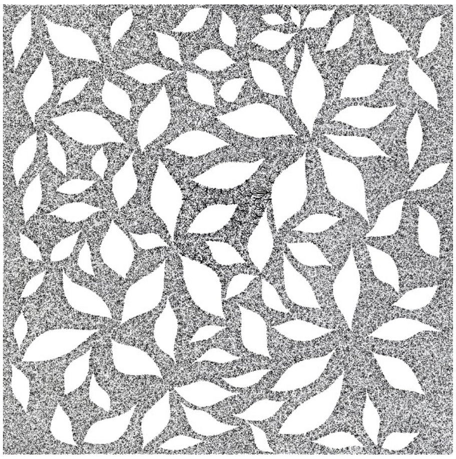

Figure 4: *Ectopic*, ink on paper, 100 cm x 100 cm, © Adinda van 't Klooster 2020

The hand drawn aesthetic is effectively maintained in the virtual landscape and a strong sense of immersion is created through the Skybox feature: it allows the drawings to be explored close-up by the interested user, and they are simultaneously trapped inside of it. The only way to exit the drawing is to press the A button which replaces the drawing with another Skybox.

To avoid the drawings visually overpowering the whole scene the terrain needed to be of a reasonably large scale. The trees are generated by the user when s/he navigates through the landscape using the controllers. The stylized trees have realistic textures on them from barks and mosses the first author found in their daily walks. The combination of abstract drawings with semi-realistic trees and strangely shaped clouds leads to an overall surreal landscape that is clearly manmade, even if inspiration is drawn from nature (Figure 5).

2.3 Audio Feature Analysis

An open-source machine listening plug-in for Unity[1][9] is used to extract audio features continuously from the Quest microphone. Four primary features were selected (from ten candidates) which had been the most effective in testing, namely the spectral centroid, the Root Mean Square power, sensory dissonance, and a constant Q pitch detector. In order to cope with different users and maintain parity of feature values, real-time adaptive max-min normalization adjusts the range of each feature independently given the observed values so far. This keeps all features within a [0,1] interval, and though not as robust

[1] https://github.com/sicklincoln/MusicalMachineListeningUnityPlugin

to outliers as quartile normalisation, is straight forward to calculate just by keeping variables for the largest and smallest values recorded so far for each feature. In addition, a live onset detector creates audio triggers to introduce graphical objects.

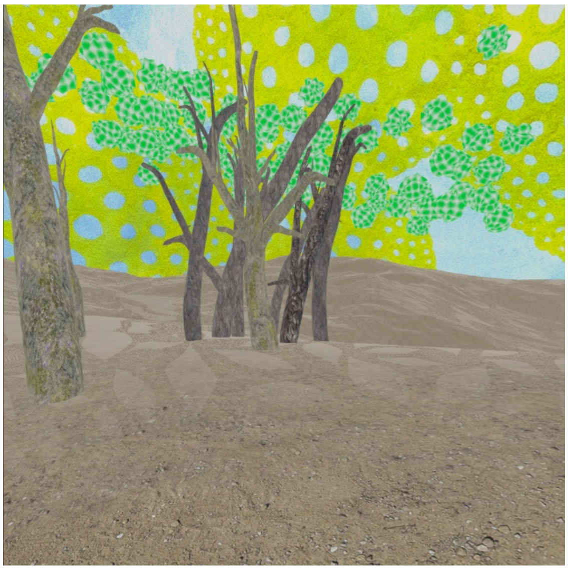

Figure 5: Screengrab from in-headset experience, © Adinda van 't Klooster 2023

2.4 Re-Mapping

The scene can be refreshed by pressing the A button on the right-hand controller, bringing in different graphics for the background skybox, different textures and coloration for objects, and a different mapping between audio features and graphical controls.

An intermediate layer in the code assigns audio features to four slots, allowing the polling by objects of the currently assigned Nth feature without foreknowledge of which audio feature this corresponds to. The adaptive normalization ensures that the ranges of the audio features are congruent (to a common [0,1] interval) for this to work. Although certain critical mappings were fixed to particular features, this indirection allowed other graphical elements to take on fresh behaviors each time the scene loaded.

2.5 Building in Unity

Developing a voice interactive animated app for the Quest is as much a question of graphics programming as audio. The most awkward issue in getting the app accepted by the Oculus Store for the App Lab was maintaining a consistently high frame rate (the device expects an average 90 fps). The generative graphics, particularly the recursive trees, were expensive and restrictions had to be placed on the branch factors and depth permitted, and the numbers of trees that could grow at any one time around the player (a maximum of 95 was established empirically for Quest 2; the Quest 3 allowed around double this). The actual models for trunks and branches had to utilize Level Of Detail (LOD) Groups to reduce triangle count whenever further away from the camera. One main straighter trunk model and three more curved branch models were used, fitted with piecewise linear segments to set the 3D locations where scaled smaller child models would connect to a parent without obvious joins.

Both authors of this paper preferred development within a previous project on the tethered HTC Vive Pro, which was more comfortable to wear (especially for those with glasses) and already allowed for higher resolution graphics. However, the convenience of being able to show the work requiring only a headset made the Quest a strong contestant when trying to bring the work to a wider audience.

3. Evaluation

We used a number of tools published in the NIME literature to evaluate the project design, applying VR-specific criteria from Serafin et al [15], and two-dimension spaces [2,10]. These frameworks help to triangulate the work. In terms of O'Modhrain [13], the Performer's Perspective is the most helpful viewpoint. The present interface is intended to be accessible and stimulating, but is not adaptive to high virtuosity performance, and there is no expectation of intensive practice. It is more of an installation experience or explorative art game rather than a virtuosic instrument, accessible to all who can safely wear a VR headset [20].

3.1 VR Design Principles

Serafin et al. [15] posit nine design principles for Virtual Reality Musical Instruments (VRMIs), that can be further applied as evaluative dimensions. Table 2 treats each of the nine across the rows.

Design Principle	Analysis
DP 1: Feedback and Mapping	Kinaesthetic sense of voice. Haptic touch of controllers but no vibrational/contact feedback in the world. No audio output but for own voice, but synced to visual behaviours
DP 2: Latency	Not noticeable. Designed to have a baseline 90fps graphics frame rate, as demanded by the Quest app design requirements
DP 3: Cybersickness	Objects can move fast, but exploration of the scene by looking and moving wasn't sickening. At times when being immersed in designing the interface for a whole day we did experience headaches, but not cybersickness. When sticking to the normal amount of interaction time with the interface that a user would have (between 5 and 15 minutes) no headaches were experienced.
DP 4: Do Not Copy/Make Use of Existing Skills	Abstract graphics manipulated via voice; however, use of tree and cloud imagery shows some connection to reality. Existing experience of using the voice is leveraged, but exploration of new vocal sounds to get new virtual reactions is encouraged
DP 5: Interaction	More 'magical', flying to traverse the world and provoking energy in remote objects through vocalising
DP 6: Ergonomic Design	Oculus Hand Controllers, and Quest 2 or 3 headset
DP 7: Sense of Presence	Immersed in graphics, may feel adrift without body representation
DP 8: Body Representation	No virtually represented hands, body or nose
DP 9: Social Experience	Single user

Table 2 VR Design Principles

3.2 Phenomenological and Epistemic Dimension Space

Birnbaum et al. [2] provide a useful phenomenological dimension space to characterize musical instruments; Magnusson [10] extends this to become an epistemological space concerning the embedded musical knowledge in an interface. In terms of the former, the musical control is timbral, there are four primary audio features mapped to visual motion giving four degrees of freedom, feedback is real-time to the user via visual response, there is a single interactor, a stationary boundary is used for performance so that the distribution in space is not wide (though the virtual space gives the illusion of a much larger landscape of action). The role of sound is expressive, and there are no prerequisites in particular musical expertise, only a willingness to use the voice. The epistemic dimensions are treated in Table 3.

Dimension	Analysis
Expressive Constraints	Tries to encourage varied vocal behaviour, but not strongly constrained into stylistically varied scenes.
Autonomy	No music generation autonomy, but influence over visuals through vocal exploration.
Music Theory	No particular music theory except for the spectromorphological properties tracked via the chosen priority features. Lends itself best to free timbral improv.
Explorability	Roaming the landscape generates trees. When refreshing the scene, mappings and the scale of trees shift and the backgrounds are interchangeable, to provide variation for the user. There are certain constants between the scenes, such as the clouds, the terrain and the blobs being controlled by onsets.
Required Foreknowledge	No specific theoretical foreknowledge required except basic use of the voice
Improvisation	Graphics respond to real-time vocalizing/singing, improvisation is the primary mode
Generality	Relatively limited graphical range, doesn't distinguish divergent vocal styles at a higher level, just mid- and lower level audio features
Creative-Simulation	Based solidly on the voice, but with novel audiovisual outcomes

Table 3 Interface analysis in terms of Magnusson's epistemic dimension space

Though we intend a positive effect on wellbeing from using this interface, an evaluation to substantiate this is forthcoming. We are planning to collect audience feedback at NIME 2024 and scheduled exhibitions of the interface in Manchester in 2025 such as at the Modal gallery in MMU and the Manchester Science Museum.

4. CONCLUSIONS

VR provides a fertile option for bringing 2D artworks into 3D environments in much more interesting ways than the more common VR art galleries that were especially popular in lockdown. We hope to interest the general public and live vocal performers interested in having a responsive visual backdrop to their performance. The app and video can be found at: https://www.meta.com/en-gb/experiences/5775768202525310/.

Future work would make interaction and graphical output more varied. We'd also like to make a follow-on interface more intelligently equipped to respond to the spoken word by using speech recognition and sentiment analysis.

5. ETHICAL STANDARDS

Intending face to face evaluations of the interface this year, university ethics clearance will be sought to run all user tests.

6. REFERENCES

[1] Ander, E., Thomson, L., and Chatterjee, H. (2010), Culture's Place in Wellbeing: Measuring Museums Wellbeing Interventions. In *Landscape, Well-Being and Environment*. Edited By R. Coles and Z. Millman (eds.), Abingdon, Routledge, 2013, pp. 161-180

[2] Birnbaum, D., Fiebrink, R., Malloch, J., and Wanderley, M., Towards a Dimension Space for Musical Artifacts. In *NIME 2005 Proceedings*.

[3] Clift, S., Singing, Wellbeing, and Health. In *Music, Health and Wellbeing*, R. MacDonald (ed.), Oxford University Press, Oxford, (2012), 113–124

[4] Elgar, A., Culturally Speaking: The Rhetoric of Voice and Identity in a Mediated Culture, The Ohio State University Press, 2019.

[5] Gatto, C., Calabrese, L. et al. Wellbeing Assessment of a Museum Experience in Virtual Reality through UCL Measurement Tool Kit and Heart Rate Measurement: a Pilot Study. 2022 *IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering*

[6] Grau, O., *Virtual Art: From Illusion to Immersion*. MIT Press, Cambridge, MA, 2003

[7] Harma, V. Interaction and Performativity in Digital Art Exhibitions. *Nordisk Museology,* 1 (2011), 98-105

[8] Kang J., Scholp A., Jiang J.J. A Review of the Physiological Effects and Mechanisms of Singing. *Journal of Voice*, 32, 4 (2018), 390-395

[9] van't Klooster A., Collins N., Virtual Reality and Audiovisual Experience in the AudioVirtualizer, *EAI Endorsed Transactions on Creative Technologies*, 8, 27 (2021),e2

[10] Magnusson, T. An Epistemic Dimension Space for Musical Devices. In *Proceedings of the 2010 Conference on New Interfaces for Musical Expression*, 43–46.

[11] Monks, S. Adolescent singers and perceptions of vocal identity. *British Journal of Music Education* 20, 3, (2003), 243-256.

[12] Montana, J.I., et al., The Benefits of Emotion Regulation Interventions in Virtual Reality for the Improvement of Wellbeing in Adults and Older Adults: A Systematic Review. *Journal of Clinical Medicine* 9, 2 (2020), 500.

[13] O'Modhrain, S., A Framework for the Evaluation of Digital Musical Instruments. *Computer Music Journal* 35,1, 28–42.

[14] Schroeder, D. et al Creating Widely Accessible Spatial Interfaces: Mobile VR for Managing Persistent Pain. *IEEE Computer Graphics and Applications*, 33, 2 (2013), 82-89

[15] Serafin, S., Erkut, C., Kojs, J., Nilsson, N.C and Nordahl, R., Virtual Reality Musical Instruments: State of the Art, Design Principles, and Future Directions. *Computer Music Journal* 40, 3 (2016), 22–40.

[16] Tatar, K. Prpa, M. and Pasquier, P., Respire: Virtual Reality Art with Musical Agent Guided by Respiratory Interaction. *Leonardo Music Journal* 29 (2019), 19–24.

[17] Trivedi G, Sharma K, Saboo B, et al., Humming (Simple Bhramari Pranayama) as a Stress Buster: A Holter-Based Study to Analyze Heart Rate Variability (HRV) Parameters During Bhramari, Physical Activity, Emotional Stress, and Sleep. *Cureus* 15, 4 (2023)

[18] Weinel, J., Cyberdream VR: Visualizing Rave Music and Vaporwave in Virtual Reality, In: *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound;* September 2019; Nottingham, 277-281.

[19] Wilson, S., *Information Arts: Intersections of Science, Art and Technology*. MIT Press, Cambridge, MA, 2002.

[20] U.S. PIRG Education fund, VR risks for kids and teens (2023). Available at : <https://pirg.org/edfund/resources/vr-risks-for-kids>, Accessed 29 Jan 2024