Design of an Interactive Virtual Reality System for Ensemble Singing

GAVIN KEARNEY, HELENA DAFFERN, LEWIS THRESH, HAROOM OMODUDU, CALUM ARMSTRONG AND JUDE BRERETON

Audiolab, Department of Electronics, University of York, UK
e-mail: gavin.kearney@york.ac.uk

September 23rd 2016.

Abstract

This paper outlines a prototype of an immersive virtual reality reconstruction of a choral singing scenario, where a player can sing along with the rest of a virtual quartet using an immersive audio-visual system. The virtual simulation is intended to be a conduit for data collection for auditory perception in virtual reality (VR) reproduction as well as performance measures and effects on singing health and well-being. Live performance capture of a quartet is conducted at St. Olave's Church in York. Acoustic measurements and 360° video are used to create an interactive virtual environment of the real performance space. This paper outlines the technical recording and production workflow of the prototype VR system.

1 Introduction

Singing (particularly singing in a choir) is often reported as being beneficial to both physical and psychological health, although empirical research in this area is in its infancy, with an acknowledged need for robust interdisciplinary studies [1, 2]. Whilst initial studies have indicated psychological benefits, including a significant clinical improvement on mental wellbeing for people suffering enduring mental health issues, there are few quantitative studies into the physical benefits of singing in a choir [3].

The aim of this paper is to create a prototype virtual reality system that allows the user to replace a singer within a vocal quartet performance in a given space. Wearing an Oculus Rift [4] headset and headworn DPA microphone and either within a speaker array or wearing headphones, the user should experience the venue of the original quartet performance from the viewpoint of one of the singers, including visual and audio effects, and be able to hear themselves in that space as they sing live with the recorded performance. This system is intended for use in testing the wellbeing and health benefits of singing in a group and improving accessibility to the benefits of singing in a group via remote technological solutions, with future potential for adapting the system to optimise the teaching and learning of singing.

The paper is organised as follows: An overview of the VR system is first presented. Data capture of the ensemble performances and the subsequent workflow to create an interactive VR version is then discussed. Finally limitations of the current implementation and future directions of the work are presented.

2 System Overview

The virtual reality system is comprised of two parts:

- **Audio**: One machine running Max/MSP is used to playback 2-D 3rd order Ambisonic [5] audio files (can also be used for 3D and up to 4th order) as well as convolve the real-time audio input from the DPA microphone with 3rd order Ambisonic impulse responses (IRs). These are then output to a horizontal array of 8 loudspeakers. By standing in the centre of the array and wearing a DPA microphone, the user can hear themselves and other members of the quartet as though they were in the church as part of the original recordings.

- **Visuals**: Wearing an Oculus Rift, the user is able to freely look around and visually experience ensemble singing from the perspective

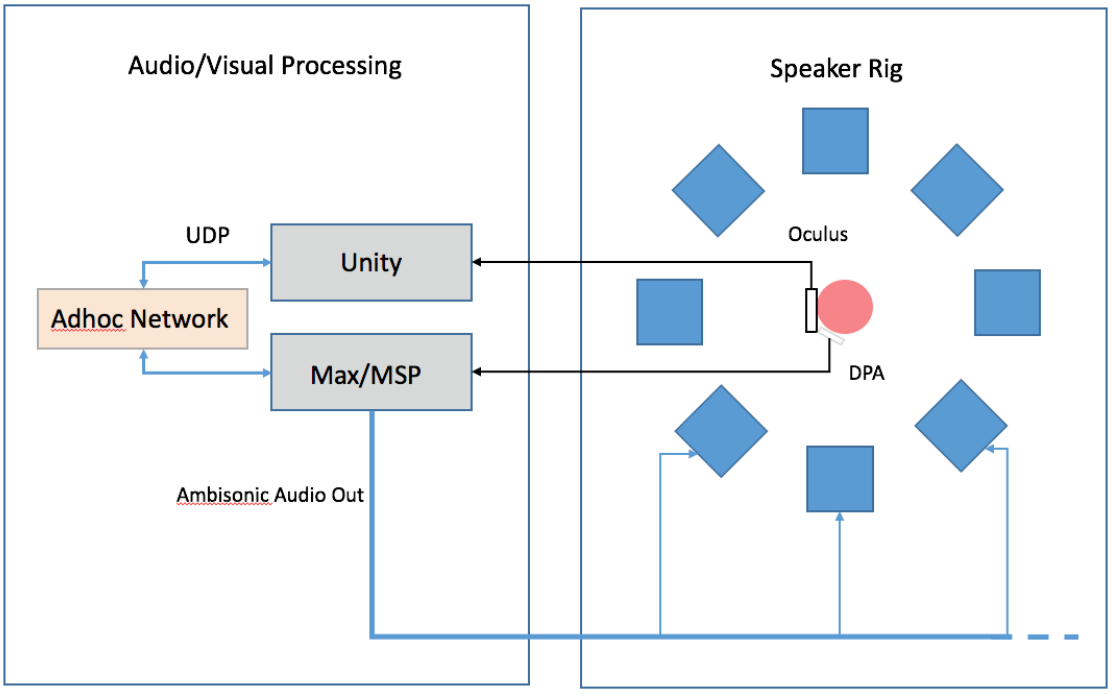

Figure 1: Illustration of the virtual reality system set up.

of the singer they have replaced in the quartet. 360° video footage was recorded in each of the performance positions using a GoPro [6] rig consisting of six, synchronized, cameras placed in a cubic array, shown in Figure 2.

Each element is processed on separate machines which communicate via UDP over an ad-hoc network. A block diagram of the system is shown in Figure 1.

3 Data Capture

3.1 Venue and repertoire

St Olaves, a 15th century church set in York's city centre, was chosen as the venue for the VR experience. It is known for its pleasing acoustic for live classical music making it ideal for an a cappella vocal quartet performance.

Two pieces were chosen: *'Amazing Grace'* [7] sung in unison, to allow non-singers to engage with the system through a well known song that doesn't require musical expertise; and *'If Ye Love Me'* [8] by Thomas Tallis, a relatively simple four-part piece that is commonly well-known by experienced ensemble singers. The singers were asked in advance to memorise the music and were given rehearsal time prior to the recording.

3.2 Session Objectives

There were 3 sets of data captured at St. Olave's:

1. **Physiological Data**: An initial recording session captured the singers performing in the venue for use in future work on health and well-being in singing. A number of sensors were used to measure various physiological parameters, including heart rate, skin conductance through a Shimmer device [9] and vocal fold activity through laryngography [10], as they performed as a quartet. This experimental data will be reported elsewhere with future work to compare parameters when singing in the real and virtual environments.

2. **Singing Audio**: To make the virtual reality system the material was repeated four times. In each performance three members of the quartet performed with the fourth singer replaced by the Eigenmike and GoPro camera rig shown in Figure 2. The Eigenmike is a 32 capsule microphone array which allows for capture of up to 4th Order Ambisonic recordings [11].

3. **Sweep Capture**: Sinusoidal sweeps were captured in each of the singing positions in order to produce spatial impulse responses. These allow

the user in the virtual environment to experience the correct acoustic when in position as part of the quartet .

Ethical approval was obtained prior to the start of this project from the Physical Sciences Ethics Committee at the University of York.

3.3 Performance Capture Procedure

In order to enable the user to interact in the virtual environment, each singer was replaced with the Eigenmike and 360° GoPro rig one at a time, while the other three singers performed both the songs as if the fourth singer was still present (interacting with the camera as though it were the fourth singer). For each GoPro/Eigenmike position, the singers were asked to sing two songs and two chords:

1. Tallis - If Ye Love Me
2. Amazing Grace (in unison)
3. Blended Chord (Triad with note missing)
4. Blended Full Chord (Triad)

The GoPro rig was positioned in the same place and at the same height as each of the singers eyes as shown in Table 1. The Eigenmike was placed just behind the GoPro rig and slightly higher. This was to capture the soundfield as close to the ear positions as possible while trying to minimise capture of the microphone in the video. The setup can be seen in Figure 2.

Singer	GoPro Height (eyes) (m)	Eigenmike Height (ears) (m)
Soprano	1.62	1.55
Alto	1.65	1.57
Tenor	1.69	1.62
Bass	1.58	1.58

Table 1: GoPro and Eigenmike recording positions

Figure 2: Eigenmike and GoPro rig set up for replacing each of the singers.

Figure 3: Sweep recording setup with rotated Eigenmike

3.4 Acoustic Measurement Procedure

Acoustic impulse responses were captured using the inverse swept sinusoidal technique [12]. For each of the singing positions, four 30s long sinusoidal sweeps were played through a Genelec 8030 loudspeaker. The initial position of the speaker was facing in the same direction as the singer. For each measurement, the loudspeaker was then rotated at 90° increments, in order to capture directional impulse responses that better reflected the directivity of the singing voice and that could be interpolated later at runtime depending on the user's orientation [13]. The Eigenmike was used to capture the 3D soundfield from which a higher order Ambisonic spatial impulse response was produced. As these IRs were used to recreate how a singer should sound in the given position, the microphone and loudspeaker were placed as close to each other as possible to try and replicate the position of the mouth and ears. To do this, the Eigenmike was rotated so that the front of the mic was facing the roof of the church, and the top of the mic was facing the 0° position, illustrated in Figure 3.

4 Post Processing

4.1 Audio

All of the recordings taken with the Eigenmike went through a similar post-production procedure, however the sine sweeps were subjected to several additional steps:

Impulse Response Procedure

1. Encoded into 3rd order Ambisonic files using MH Acoustic EigenUnits plug-in [11] in Reaper [14].
   - Channel Order: ACN
   - Normalisation: SN3D
2. Deconvolved with the inverse sine sweep in Matlab
3. 40ms trimmed off the start of the IR (see 5.4)
4. Divided into two separate files: 1) Early reflections 2) Diffuse field (See section 5.4)

Singing files procedure

1. Encoded into 3rd order Ambisonic files using mh acoustic EignUnits plug-in.
   - Channel Order: ACN
   - Normalisation: SN3D
2. Divided into appropriate song lengths

These singing files were then imported into another Reaper project including the GoPro video footage in order to synchronise and edit the video files with the audio clips.

4.2 Video

There were four aspects to the video editing process:

1. The video footage was stiched together to create a 360° panorama.
2. Masks were created to hide the microphones from the video.
3. The videos were rendered to ultra HD.
4. Footage was cut into song sections and aligned with an Eigenmike audio track.

Table 2 contains a list of software used for video processing.

Software	Description
Kolor Autopano Giga [15]	Image editing.
Kolor Autopano Video [15]	Video stitching and rendering.
Reaper 5	Audio and Video synchronization

Table 2: Video processing software utilised.

The Autopano video software includes built-in support for GoPro cameras. A GoPro preset was used to facilitate the stitching process. Other useful actions such as *blend* and *color correction* were performed to enhance the visual quality of the panorama.

Autopano Video also allows the user to edit a newly created panorama through integration of Autopano Giga (panoramic image editor). The latter program allows the user to manually select the masking and control points to stitch the videos together.

At the time of filming, additional footage was taken post performance with the microphones removed. This was in order to create masking material for the video editing process. Snapshots of the video stills were created in VLC and imported into Giga as layers. The masking tool was then applied allowing the user to define how much of each layer was to be retained or removed from the panorama.

The projects were rendered at 4K UHDTV (1920 x 1080) resolution and the frame rate was set to the original value (48 fps). Video specifications are shown in Table 3.

Key	Description
Container format	MP4
Resolution	1920x1080.
Frame rate	48 fps
Codec	H.264/MPEG-4 AVC

Table 3: Video specifications of VR system.

Each panoramic video file was split into four parts, corresponding to the four songs performed by the singers. The native audio of each video was overdubbed by an Eigenmike audio track. In order to synchronize the audio for each song with its associated video, both tracks were imported into Reaper. The digital audio workstation has support for both mp4 video and 32-channel audio. With the tracks visible as waveforms it was simple to align the video with audio and choose appropriate cutting regions ensuring both the audio and video for each song started/finished at the same time. The audio clips were played through Max/MSP whilst the video clips (rendered through reaper) were played using Unity.

5 VR System Implementation

The following sections describe how the appropriate software was implemented to produce the VR system.

5.1 Unity Project

360° video was achieved in Unity by applying the video as a mesh texture and then mapping the texture to the inside of a sphere. In the game world, the main camera was placed in the centre of the sphere, as shown in Figure 4. A store asset - *AV pro windows media* [16] was used to playback HD video and facilitate the sphere mapping procedures. *K-lite codec pack*[17] was installed to support the playback of MP4 files within the application.

Unity has introduced built in headset support for Oculus Rift since the release of version 5.X. As such, enabling VR in the settings menu allows the headset to control the main camera. Hence placing the main camera in the centre of the video sphere creates an Oculus compatible 360° video player.

A script, *Export Handler* was created in Visual Studio to handle the Oculus rotation data and send it to Max/MSP via UDP. Conversely a second script, *Import Handler* was created to handle any incoming messages (such as media playback commands) from Max/MSP. Two Unity assets were installed to facilitate the scripting - *UDP messenger* [18] and *UnityOSC* [19].

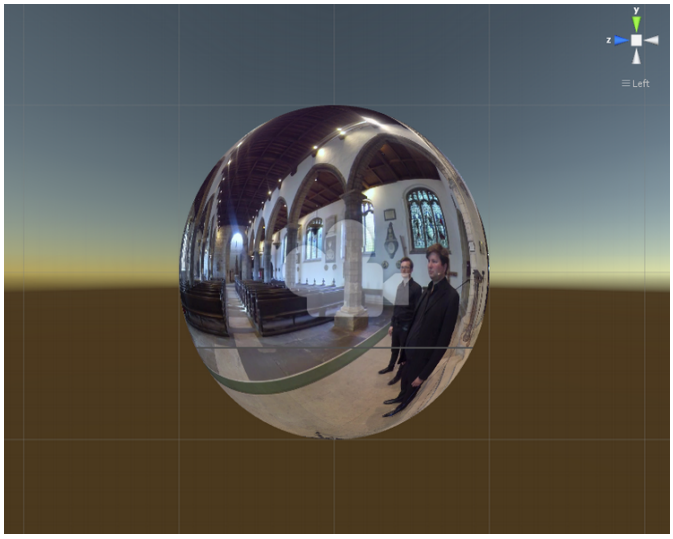

Figure 4: Unity - Mapping the 360° video.

5.2 Max Patch User Interface

Max/MSP is used to process real-time audio input and playback appropriate singing files whilst also communicating with Unity to ensure the correct videos are loaded and played in sync with the audio.

The user interface shown in Figure 5 can be used to load the appropriate audio files and video files simultaneously, as well as operations such as play, pause and stop. Alternatively, a menu screen can be used to load the appropriate files when wearing the Oculus Rift. However, only the user interface in Max/MSP can be used to select options such as which speaker layout to use, which audio input to use and such settings.

The patch takes two inputs, **Audio** (taken from the audio interface input) and **Data from Unity**. Two types of data are sent to MAX from Unity. The first of these are commands that determine which audio file to load into the system and when to play or stop the specified file. This information is handled in the second section in Figure 6 (highlighted in blue). The second type of data is azimuth rotation data taken from the Oculus headset, handled in the first section in Figure 6 (highlighted in red).

5.3 Convolution/Order selection

The red section of the UI shown in Figure 5 shows three drop-down boxes. These can be used to select 1^{st}, 2^{nd} or 3^{rd} order Ambisonic rendering, allowing a lower CPU load if required. Binaural rendering can also be selected when using headphones and uses the virtual loudspeaker approach [20]. The other two

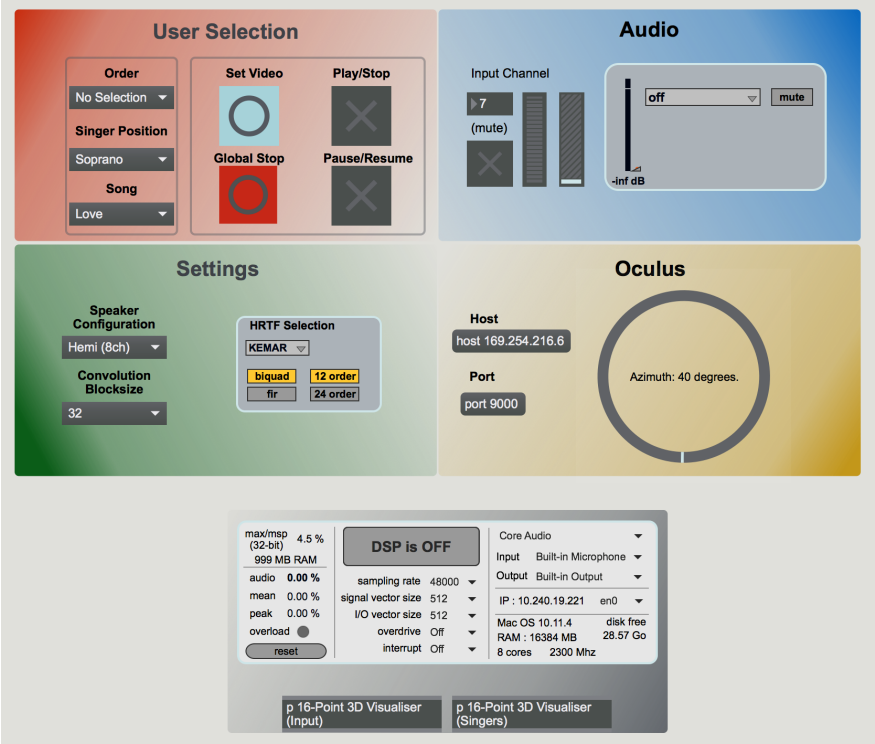

Figure 5: Max patch user interface

drop-down menus allow the user to select which position they would like to sing in and which song.

The head-tracking data from the Oculus is used to pan the audio input between four multi-channel convolution objects, each of which is using one of the directional IRs taken in the church, effectively simulating the acoustics of singing in any direction. The IRs used for simulating this however are only the first 200ms of the IR file. A fifth convolution object is used to convolve the audio input with the diffuse part of the IR. This means the system only has to use four 200ms long IR files and a single 3s long file, as opposed to four 3s long files (See section 5.4).

The output of the convolution objects are then rotated to counter the endfire rotation of the Eigenmike used in the measurements.

It is possible to route the output from Max to different loudspeaker configurations. For the development of the prototype, a horizontal rig consisting of 8 loudspeakers was set up. The patch provides the option to use a 16 channel loudspeaker rig and can easily be changed to accommodate larger sized rigs in either 2D or 3D. The option of a a 3D rig can be chosen from the **Settings** menu, which is the green section of the UI shown in figure 5.

As previously mentioned, the blue section in Figure 6 indicates where the data from Unity is handled. In essence, every message that is sent to Max via UDP is also sent back to Unity, meaning either of them can be used to control the system.

5.4 System Latency

An undesirable consequence of real-time audio processing inside of MAX/MSP is inherent latency introduced by block processing. In order to ensure the correct perceived reverberation time, there is a trade off between the processing overhead and the amount of truncation at the start of the spatial impulse responses. Signal Vector and I/O Vector sizes of 1024 samples at 44.1kHz sampling rate will have a combined effect of producing a round-trip latency of 46ms. Full length IRs can be used at these block sizes, but either the latency is unacceptable or too many of the early reflections are truncated to compensate.

To overcome this issue, we employ two strategies: First, we reduce the processing overhead by separating the early reflections and diffuse portions of the impulse responses. The early reflections were taken from each spatial impulse response over a 200ms window after the direct sound. A single diffuse-only spatial impulse response was utilised after 200ms. In this way we are only interpolating across the early reflection rendering during head-movement.

Secondly, we consider the large initial time delay gap between the direct sound and the first reflection already exhibited in the impulse responses due

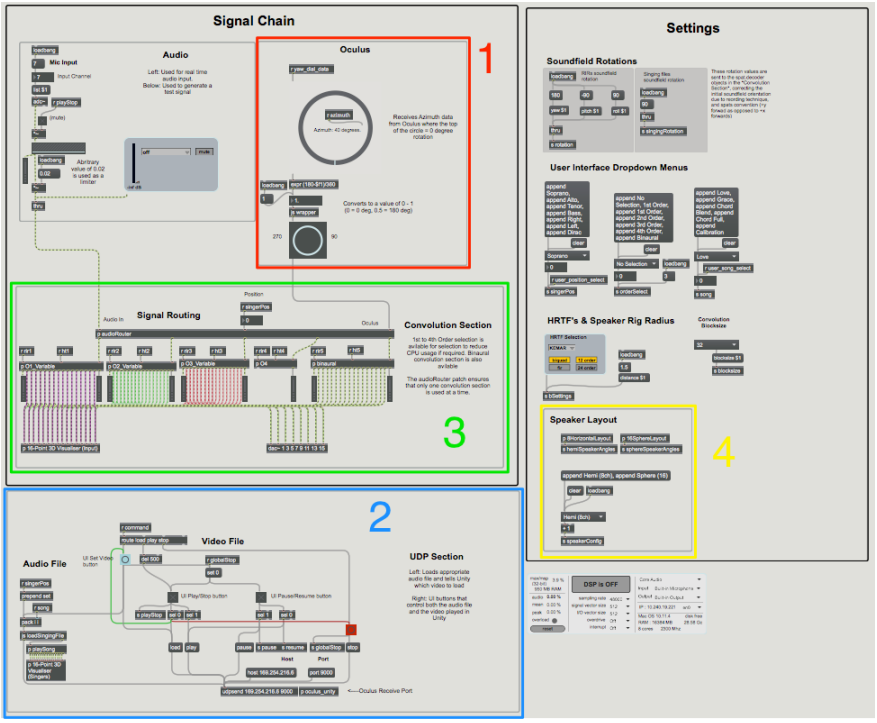

Figure 6: Overview of the max patch when not in presentation mode

to the size of the performance space. Although immediate floor reflections from the loudspeaker to the Eigenmike were apparent in the impulse responses, we consider such reflections, which represent those coming from a singers voice to their own ears to be insignificant due to the shadowing effect of the torso and directivity of the human voice. The significant first arriving reflections come from the ceiling and side walls which have a total propagation length greater than 13m from any direction in the performance space. This means that up to 40ms could readily be truncated from the impulse responses without loss of significant spatial information at the singer position. (Note that this only corresponds to the user's own position. Important spatial information is conveyed from the recordings of the other singers to the user due to floor reflections occurring within 10ms after the direct sound). By employing these two optimisation strategies across the early reflections of the spatial impulse responses, Signal Vector and I/O Vector sizes of 512 were utilised with smooth operation.

6 User testing

The singers who took part in the recordings for the virtual environment were invited to test the system set up in the music studio in the AudioLab at the University of York. They also took part in a further experiment to measure the same parameters assessed during performance at the venue, but as they sang in the virtual environment. The results of this experiment, to assess whether the singers responded the same way to the virtual and real environments are currently being analysed. Feedback regarding the system as a virtual reality experience and tool for singing performance was obtained from the singers via a questionnaire to gather information regarding the comfort of the equipment and the level of believability of the system. The singers were given ten minutes to play with the virtual reality environment, trying out different positions and songs that they recorded, and were then asked to sing their part in the piece by Tallis four times.

The original singers gave unanimously positive feedback to the system as a whole. They found the experience immersive and positive, all of them commenting that they forgot that they were in a studio and felt like they were in the original performance space. In addition to finding it a fun experience the singers saw lots of potential in the system as a tool for rehearsal, practice and learning singing, particularly in terms of memorising music or improving sight-singing. The singers didn't comment on the lack of peripheral vision which was a predicted problem with the VR headset, and when asked about this limitation they said that they just adapted by moving their head more than usual. They did comment on the auditory rendering and felt that it was a con-

vincing representation of the original space.

7 Known issues and future work

There are several parts of the system that have proved problematic. One issue was the Eigenmike's noise floor. When using the system, there is some noticeable noise that would not be present in the real environment, thus detracting from the immersive experience. Other issues currently include the quality of the 360° videos used which also detract from the immersive experience. This is a result of the software being used to produce the videos and the display resolution of the VR headset.

As well as being relatively heavy, the Oculus Rift (DK 2) only provides a 100° field of view, thus cutting out peripheral vision. Other potential headsets that provide a wider field of view include the HTC Vive (110°) [21] and the StarVR (210°) [22].

Reading sheet music was another concern of the project, as it is not possible to view external visual material whilst wearing the Oculus. This could be remedied by using the hand sensors that come with VR kits such as the HTC Vive, where the user can hold two sensors in their hand for which virtual hands can be rendered as well as sheet music.

8 Conclusion

In this paper we have presented a framework for capture and reproduction of interactive virtual reality simulation of a singing ensemble. The system is capable of reproducing convincing representations of real performance scenarios using 360° video and higher order Ambisonic audio capture and reproduction. The workflow demonstrated utilised commercially available software and hardware to create the VR experience, alongside customised MAX/MSP patches and offline acoustic response processing. Considerations were given to real-time rendering of the system and its limitations discussed. Future work will employ this framework into assessing the health and wellbeing of singers in both real and VR environments.

9 Acknowledgements

This work is supported by the University of York Creativity Priming Fund. It is also supported by the EPSRC funded SADIE (Spatial Audio for Domestic Interactive Entertainment) project (EPSRC Ref: EP/M001210/1).

References

[1] Mary L Gick. Singing, health and well-being: A health psychologist's review. *Psychomusicology: Music, Mind and Brain*, 21(1-2):176, 2011.

[2] Rosie Stacy, Katie Brittain, and Sandra Kerr. Singing for health: an exploration of the issues. *Health Education*, 102(4):156–162, 2002.

[3] Stephen Clift and Ian Morrison. Group singing fosters mental health and wellbeing: findings from the east kent "singing for health" network project. *Mental Health and Social Inclusion*, 15(2):88–97, 2011.

[4] Oculus. *Oculus Rift, DK2*, 2014 (accessed September 1, 2016).

[5] Michael A Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973.

[6] GoPro. *GoPro Omni Rig*, 2015 (accessed September 1, 2016).

[7] John Newton. *Amazing Grace*, 1779.

[8] Thomas Tallis. *If Ye Love Me*, c. 1547-1553.

[9] Shimmer. *Shimmer Sensor Technologies*, 2008 (accessed September 1, 2016).

[10] Evelyn RM Abberton, David M Howard, and Adrian J Fourcin. Laryngographic assessment of normal voice: a tutorial. *Clinical Linguistics & Phonetics*, 3(3):281–296, 1989.

[11] MH Acoustics. *Eigenmike*, 2002 (accessed September 1, 2016).

[12] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th Convention of the Audio Engineering Society*, Paris, France, 2000. Paper 5093.

[13] Gavin Kearney. *Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction*. PhD thesis, Trinity College, Dublin, Ireland, 2010.

[14] Cockos Inc. Reaper Digital Audio Workstation v 5.24, 2016 (accessed September 1, 2016).

[15] Kolor. *Autopano*, 2016 (accessed September 1, 2016).

[16] Renderheads. *AV Pro Windows Media: Unity Plugin for Windows Desktop Video Playback*, 2016 (accessed September 1, 2016).

[17] Codec Guide. *K-Lite Codec Pack v12.3.5*, 2015 (accessed September 1, 2016).

[18] Ping Pong Technologies. UDP Messenger v1.0, 2015 (accessed September 1, 2016).

[19] Jorge Garcia. UnityOSC v1.2, 2016 (accessed September 1, 2016).

[20] Gavin Kearney, Claire Masterson, Stephen Adams, and Frank Boland. Towards efficient binaural room impulse response synthesis. In *EAA Symposium on Auralization*, Espoo, Finland, June 2009.

[21] HTC and Valve Corporation. *HTC Vive*, 2015 (accessed September 1, 2016).

[22] Starbreeze Studios. *Star VR*, 2015 (accessed September 1, 2016).