The Virtual Singing Studio (VSS) is a loudspeaker-based room acoustics simulation for real-time musical performance under development at the AudioLab, University of York. An objective evaluation of the VSS found that relative differences between T30 values of the input (real) and output (virtual) Room Impulse Responses fall within the double tolerance subjective limen (10%). In addition a consideration of the most suitable microphone type and placement for the simulation is carried out by recording a singer with a number of different microphones. Whilst a microphone placed overhead is most realistic for the singer since unseen, it has been shown to cause a pitch shifting artefact due to the presence of high frequency energy.

1 Introduction

Recent advances in the research and development of Virtual Acoustic Environments (e.g. 1, 2) allow the user to interact with the virtual space, usually by controlling their position within the virtual environment via joystick or other movement controls. More recently virtual acoustic environments have been developed which allow the user to physically move about the virtual space and hear the resulting changes in their own sound (hand claps, speech, singing) according to the geometry and room acoustics of the acoustically rendered environment.

This paper reports on the on-going design and implementation of the Virtual Singing Studio - a real-time interactive loudspeaker based room acoustics simulation for musical performance. Whereas others (e.g. 3, 4, 1) have used synthetic room impulse responses, the Virtual Singing Studio (VSS) is based on real-time convolution with Ambisonic B-format room impulse responses which have been measured in an existing performance space.

An objective evaluation of the VSS has been made by comparing room acoustic parameters of the input RIR (real performance space) and the output RIR (simulated space). A consideration of the most suitable microphone type, and best positioning is carried out by comparing recordings of a singer made in an anechoic room with a number of different microphones.

Figure 1: Interior of real performance space
2 The Virtual Singing Studio

2.1 The Real Performance Space

The performance space simulated in this work is the National Centre for Early Music (York, UK), [Figure 1] which has an internal volume of 3600$m^3$ and is equipped with a passive variable system which gives performers the opportunity, via manipulation of wall panels and ceiling drapes, to adjust the overall absorption and reverberation time.

Three of the five possible acoustic configurations were chosen for the study. Mean T30 values at 1kHz, measured across 26 receiver positions for the three acoustic configurations considered are: “Speech” 1.32s, “Music Recital” 1.83s and “Large Choral” 2.22sec. Figure 2 shows mean T30 values for the mid-way “Music Recital” setting in the space. As is common in a mid-sized church building T30 values rise with frequency up to a peak in the 500Hz octave band, and decrease with frequency in the octave bands above 500 Hz. Further details of the acoustic properties of the performance space are given in (5).

![Figure 2: Mean T30 values for real performance space](image)

2.2 Design and Implementation of the VSS

Input RIR measurement

The B-format room impulse responses were measured in the real performance space at a sampling rate of 96 kHz based on the Exponential Swept Sine (ESS) Method as described in (6). A Genelec 8040 loudspeaker emitted a 15 second long log sine sweep which was captured by a Soundfield SPS422B Microphone. The recorded sine wave sweeps were then deconvolved in MATLAB, and room acoustic parameters were computed via Aurora plug-ins (7). Performer RIRs were measured specifically in an attempt to replicate the performer’s topology with the Soundfield microphone positioned directly above the source loudspeaker.

Playback in the VSS

The Virtual Singing Studio comprises a Digital Audio Workstation which outputs to a 3-D loudspeaker array consisting of 16 Genelec 8040 loudspeakers in an acoustically-damped listening room measuring 4.5m x 5m x 2.5 m.

When used for musical performance the input channel (singer’s microphone) is convolved over four channels (Ambisonic B-format) with an edited version of the measured Performer RIR in Reaper (8) using the convolution reverb audio processor “Reverberate” plugin by Liquidsonics (9). The convolved channels are decoded for loudspeaker array (octagonal plus cube) using the VST plug-in "B-dec High resolution First Order Ambisonic B-format decoder" (10) and output to the loudspeakers which are arranged in an 8-loudspeaker horizontal ring and two 4-loudspeaker rings (ceiling A-D and floor E-H) – see Figure 3. The performer stands at the central point of the loudspeaker array, at a distance of 1.95m from the loudspeakers.
3 Objective Evaluation of VSS

An initial objective evaluation of the Virtual Singing Studio was carried out in order to assess any errors which were introduced due to the signal processing involved, playback methods and spatial properties of the listening room in which the VSS is located. For evaluation the input RIRs were convolved and output to the loudspeakers. In turn the output RIR was measured at the central point of the array and room acoustic parameters measured and compared.

T30 and EDT values were computed for seven octave bands from 125Hz to 8kHz for the input and output RIRs. Since T30 is measured in seconds the percentage errors are calculated between the output and input RIR (virtual vs real). An initial assessment of the perceptual relevance of the errors found is based on a method used by Favrot (11,12) which compares potential errors introduced by the room acoustics simulation system to single and double tolerance subjective limen i.e. 1 and 2 times the widely quoted “just noticeable differences” respectively (13,14).

3.1 T30 errors

Relative differences (%) in T30 values between the input (real) and output (virtual) RIRs for the Performer RIR are presented in Figure 4 (dashed horizontal lines represent the double tolerance subjective limen). All T30 errors lie within the double tolerance subjective limen (10% - dashed line) but the greatest errors occur at 250Hz and 500Hz octave bands in the “Speech” setting. The “Large Choral” and “Music Recital” settings are most closely matched to the
input RIR in terms of T30 with error values lying within the single tolerance subjective limen (5%) at all but 500Hz and 2000Hz octave bands.

Further results of this objective evaluation are reported elsewhere (15) but it is noted that errors are greatest at 250 and 500Hz octave bands, which is most probably due to a ceiling reflection and room modes in the listening room where the VSS is situated. A second objective analysis is planned once further acoustic treatment has been carried out in the listening room.

4 Microphone choice for the VSS

At present, for performance in the VSS, the singer’s sound is captured by a head-mounted omni-directional microphone, which allows a sufficiently high relative level of direct sound in order to avoid acoustic feedback loops arising from the reverberant field output at the loudspeakers. However, capturing the singer’s voice at such close quarters means that the voice can sound too “present” in the mix, mouth noises are exaggerated and the singer experiences what Laird refers to as the ‘PA effect’ (16), i.e. the singer has the impression that their voice is amplified through a PA system before being played into the virtual space.

As an alternative to the head-mounted microphone, a cardioid microphone attached to the ceiling over the head of the singer has been trialled for use in the VSS by a small number of singers. Although the overhead microphone affords a high degree of realism, due mostly to being unseen by the performer and therefore not contributing visually to the ‘PA effect’, some singers reported informally that they perceived the reverberated sound field was played back at a higher level in pitch. It was thought that this perceptual effect might arise from spectral balance of the reverberant field produced in the VSS and coloration in the auralization chain due to microphone type and placement.

4.1 Method

In order to compare a number of different microphone placements and types a recording of a female singer was made in an anechoic chamber simultaneously on 3 different microphones:

- head-mounted DPA 4066 (as used in the VSS) positioned 5 cm from singer’s mouth
- DPA 4066 attached to the peak of a baseball cap worn by the singer (a possible alternative placement) at 10 cm from singer’s mouth
- AudioTechnica PRO45 Cardiod Microphone positioned at 20cm directly above the singer’s head
- AKG C414 XLS at a distance of 1m from the singer’s mouth

The frequency response plots of the overhead (AudioTechnica) and head-mounted (DPA 4066) microphones are shown in Figure 5 – data from (17)

Figure 5: Frequency responses of AudioTechnica Pro45 Cardiod (upper) and DPA 4066 (lower) microphones.
The singer was asked to perform an excerpt from a prepared piece (song1 song2), and 3 English language spoken word tongue-twisters on a sustained pitch (e.g. “Peter Piper picked a peck of pickled pepper”). Recordings were made using a TASCAM DR-680 recorder at a sampling frequency of 96kHz. Wideband spectrograms are plotted of the recorded singing for each microphone, using the voicebox MATLAB toolbox (18) for a frequency range up to 10,000 Hz using a Hamming window with bandwidth of 200 Hz. Long-term average spectra are calculated by computing the average power spectra across short time frames, to provide a ‘typical’ spectral envelope of the sound source recorded (19) and allows longer term spectral characteristics of the recordings to be more easily compared. LTAS were computed using a MATLAB function based on Monson’s (19) analyzed over 2048 data points resulting in a frequency resolution of 46.87Hz (96000Hz/2048).

5 Results

Comparisons are made here between recordings captured at the head-mounted microphone and the overhead microphone. It is striking in Figure 7 that although there seem to be ‘gaps’ in the spectrum of the head-mounted microphone recording around 5kHz and 8kHz, energy is still present in these frequency regions during the impulsive release of the plosive “t” – seen as vertical lines in the spectrogram above for example at around 1.5secs and just before 6 secs.

Figure 7: Spectrograms of sung phrase recorded at head-mounted (left) and overhead (right) microphones.
In addition it can also be seen in the spectrograms of the phrase “Peter Piper”, recorded by the overhead microphone show generally more energy in the spectrum above 5 kHz than the head-mounted microphone recording. These differences at higher frequencies can be more readily distinguished when long-term average spectra are plotted for the same recorded phase at different microphones (Figure 8) whereas at lower frequencies the microphone responses are very similar.

![Figure 8: LTAS of sung phrase recorded at overhead, head-mounted and 1m distance microphones.](image)

If the difference in response between the overhead and head-mounted microphone is plotted (Figure 9) the higher energy in frequencies above 5000Hz is more easily seen.

![Figure 9: Difference between the overhead and head-mounted microphones for the phrase “Peter Piper”](image)

The overhead microphone placement also appears to boost frequencies below 100 Hz due to the proximity effect caused by the cardioid microphone. On the other hand there are prominent dips in the spectrum around 1500Hz and 3000Hz which are most probably due to the shadowing effect of the singer’s head.
6 Discussion

The two different microphones already trialled in the VSS (Head-mounted DPA4066 and Overhead cardioid microphones) have elicited different informal feedback from singers who have used the system.

Although the overhead cardioid allows a realistic feel to the simulation because it is out of sight, some singers have perceived the simulated reverberant field to be ‘sharper’ in pitch than expected. This could be due to colouration introduced by this microphone in the region above 5kHz, which shifts the spectral locus, - the balance between higher frequencies and the region around the fundamental frequency - of the reverberated sound. A shift in spectral locus has been shown to effect listener’s perception of the pitch of a tone (20) even if the fundamental frequency remains unchanged.

The head-mounted DPA 4066 microphone is generally well liked for use in the VSS by singers, but one concern is that this microphone gives undue emphasis to mouth noises and plosive sounds (“p”, “t”, “k” etc), as can be seen in spectrograms in Figure 7.

7 Conclusion

The initial objective evaluation of the VSS gives overall good results in terms of matching input and output RIRs and T30 values are generally closely matched across the seven octave bands considered. Although for some octave bands the errors lie outside of Just Noticeable Differences quoted, it must be noted that the VSS is designed to be used for live musical performance, and as yet JNDs values and the perceptual aspects of room acoustic parameters which are relevant to the musician during performance have not yet been fully investigated.

8 Future Work

Only the monaural parameters T30 and EDT have been evaluated thus far, but since the goal of the VSS is to give the singer an impression of being in a spatially realistic virtual performance space, future work will evaluate the VSS simulation in terms of parameters which may be more relevant to the performer such as measures of Support. (21,22) Additional acoustic treatment is being carried out in the listening room in which the VSS is located in order to ameliorate the effects of the ceiling reflection and room modes below 500Hz.

It has been noted that the acoustic treatment in the listening room means that the ambient noise level of the VSS is very low in comparison to the real performance space. The contrast between the reverberant space being simulated and the dry acoustic of the room produces an unnatural effect. For example, Kearney (23) measured long-term averaged A-weighted levels in Trinity College Chapel, Dublin and found that the noise floor inside the church was 44 dBA. A further improvement to the VSS will be to replicate the ambient noise of the real performance space within the simulation which will also help to mask the ‘exaggerated’ mouth noises from the close head-mounted DPA microphone used in the system.

9 Acknowledgements

Many thanks go to my supervisors for their continuing support and encouragement, to Aglaia Foteinou for assistance with impulse response recordings, to Andrew Chadwick for invaluable technical support, assistance and for building the listening room, and to those singers who have already given informal feedback on the VSS.
References


   http://www.liquidsonics.com/software_reverberate.htm

    http://www.dmalham.freeserve.co.uk/vst_ambisonics.html


