SOUNDSCAPE AURALISATION AND VISUALISATION: A CROSS-MODAL APPROACH TO SOUNDSCAPE EVALUATION

Francis Stevens
Audio Lab
Department of Electronic Engineering
University of York
York, United Kingdom
frank.stevens@york.ac.uk

Damian T Murphy
Audio Lab
Department of Electronic Engineering
University of York
York, United Kingdom
damian.murphy@york.ac.uk

Stephen L Smith
Intelligent Systems Group
Department of Electronic Engineering
University of York
York, United Kingdom
stephen.smith@york.ac.uk

ABSTRACT

Soundscape research is concerned with the study and understanding of our relationship with our surrounding acoustic environments and the sonic elements that they are comprised of. Whilst much of this research has focussed on sound alone, any practical application of soundscape methodologies should consider the interaction between aural and visual environmental features: an interaction known as cross-modal perception. This presents an avenue for soundscape research exploring how an environment's visual features can affect an individual's experience of the soundscape of that same environment. This paper presents the results of two listening tests[1]: one a preliminary test making use of static stereo UHJ renderings of first-order-ambisonic (FOA) soundscape recordings and static panoramic images; the other using YouTube as a platform to present dynamic binaural renderings of the same FOA recordings alongside full motion spherical video. The stimuli for these tests were recorded at several locations around the north of England including rural, urban, and suburban environments exhibiting soundscapes comprised of many natural, human, and mechanical sounds. The purpose of these tests was to investigate how the presence of visual stimuli can alter soundscape perception and categorisation. This was done by presenting test subjects with each soundscape alone and then with visual accompaniment, and then comparing collected subjective evaluation data. Results indicate that the presence of certain visual features can alter the emotional state evoked by exposure to a soundscape, for example, where the presence of 'green infrastructure' (parks, trees, and foliage) results in a less agitating experience of a soundscape containing high levels of environmental noise. This research represents an important initial step toward the integration of virtual reality technologies into soundscape research, and the use of suitable tools to perform subjective evaluation of audiovisual stimuli. Future research will consider how these methodologies can be implemented in real-world applications.

1. INTRODUCTION

To provide a context for the methods used in the two listening test presented in this paper, this section includes a summary of the various research areas informing this study. This includes soundscape theory and evaluation, cross-modal perception, and green infrastructure.

1.1. Soundscape Theory

In his seminal text 'The Soundscape: Our Sonic Environment and the tuning of the World', R. Murray Schafer defines a soundscape as [1]:

> *'The sonic environment. Technically, any portion of the sonic environment regarded as a field for study. The term may refer to actual environments, or to abstract constructions such as musical compositions and tape montages, particularly when considered as an environment.'*

Soundscape analysis looks at the holistic experience of all sound in a given location, and aims to explore an individual's perception of, and interaction with, that environment [2]. In this way, soundscape analysis describes both the physical and perceptual properties of an environment [3]. This explains soundscape research's position as a convergence of multiple disciplines, including acoustic ecology, musicology, sociology, psychology, architecture, and acoustics [4,5].

1.2. Cross-modal Perception

Cross-modal perception is where the stimulation of one sensing modality (for example vision) can influence the experience of another (e.g. hearing). A famous example of this phenomenon is the McGurk effect [6] where a change in the appearance of mouth movement can alter the phoneme heard in recorded speech.

In a soundscape context, cross-modal perception has been considered as a way of understanding how the visual setting of an environment can change the perception of that environment's soundscape. For example, Lercher and Schulte-Fortkamp showed living on a 'pretty' street could reduce noise annoyance [7] and Viollon et al. found that exposure to still images of natural environments incorporating natural features reduced the perceived 'noisiness' of a soundscape [8]. Research into this area is of great importance to human health and well-being, in terms of reduced stress due to lower levels of noise annoyance and other health effects (for example, a patient's recovery following an operation has been shown to be faster if the patient has access to a window with a pleasant view [9]).

1.3. Green Infrastructure

Broadly speaking, when considering noisy soundscapes, the kind of visual features that may be present to improve one's experience of noise can be collected under the term Green Infrastructure. A definition of Green Infrastructure is given in [10]:

[1] This work is part of an EPSRC supported doctoral training studentship: reference number 1509136.

> *'It can be considered to comprise of all natural, semi-natural and artificial networks of multifunctional ecological systems within, around and between urban areas, at all spatial scales.'*

Whilst the acoustic impact (noise level reduction, acoustic absorption to reduce reverberation times etc.) of green infrastructure may be minimal, the impact on perception of sound may be much more pronounced [11]. An underlying motivation for this research is to investigate to what extent the presence of green infrastructure and other natural, pleasant, visual features can reduce the negative effects of acoustic noise in a soundscape. This aligns with the Biophilia thesis, originating from the field of environmental psychology, which posits that human beings have an innate appreciation for, and affinity with, natural environmental features: particularly water and vegetation [12].

The motivation for the work presented here is to make use of visualisation and soundscape methodologies to understand how the presence of certain visual features can change the emotional response evoked by a soundscape. This includes a preliminary test making use of still panoramic images and ambisonic UHJ renderings of soundscape stimuli, and a main test making use of panoramic videos and dynamic binaural rendering of FOA soundscape recordings.

2. METHODS

This section will consider the research methods and approaches applied to this study, including the soundscape evaluation methodologies used, and the data collection process.

2.1. Subjective Evaluation

2.1.1. The Self-Assessment Manikin

A previous study [13] made a direct comparison between semantic differential (SD) pairs and the Self-Assessment Manikin (SAM) as methods for measuring a test participant's experience of a soundscape.

The use of SD pairs is a method originally developed by Osgood to indirectly measure a person's interpretation of the meaning of certain words [14]. The method involves the use of a set of bipolar descriptor scales (for example 'calming-annoying' or 'pleasant-unpleasant') allowing the user to rate a given stimulus. SD pairs are a well established aspect of listening test methodology in soundscape research [15–17]. Whilst useful in certain scenarios, they can be time-consuming and unintuitive [13]. An alternative subjective assessment tool to use is the SAM.

The SAM is a method for measuring emotional responses developed by Bradley and Lang in 1994 [18]. It was developed from factor analysis of a set of SD pairs rating both aural [19] and visual stimuli [20] (using, respectively, the International Affective Digital Sounds database, or IADS, and the International Affective Picture System, or IAPS). The three factors developed for rating emotional response to a given stimuli are:

- **Valence:** How positive or negative the emotion is, ranging from unpleasant feelings to pleasant feelings of happiness.
- **Arousal:** How excited or apathetic the emotion is, ranging from sleepiness or boredom to frantic excitement.
- **Dominance:** The extent to which the emotion makes the subject feel they are in control of the situation, ranging from not at all in control to totally in control.

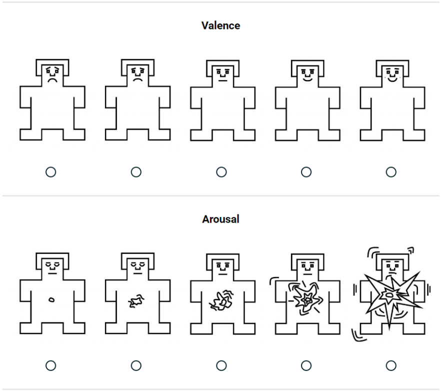


Figure 1: The Self-Assessment Manikin (SAM) as used in this study, after [18].

These results were then used by Bradley and Lang to create the SAM itself as a set of pictorial representations of the three identified factors. The version of the SAM used in this experiment (as shown in Fig. 1) contained only the Valence and Arousal dimensions following results from a previous study [13].

2.1.2. Soundscape Categorisation

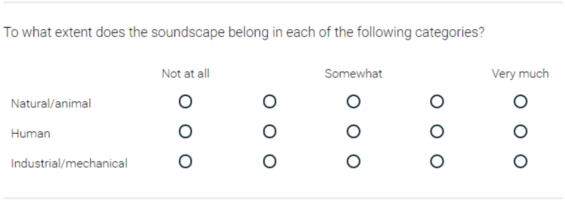
The soundscape recordings used in this test were selected in order to cover as wide a range of sound sources as possible. In order to determine what such a set of soundscape recordings would contain, a review of soundscape research indicated that in a significant quantity of the literature [21–24] three main groups of sounds are identified:

- **Natural:** These include animal sounds (such as bird song), and other environmental sounds such as wind, rustling leaves, and flowing water.
- **Human:** Any sounds that are representative of human presence/activity that do not also represent mechanical activity. Such sounds include footsteps, speech, coughing, and laughter.
- **Mechanical:** Sounds such as traffic noise, industrial and construction sounds, and aeroplane noise.

Following results from a previous test [25] it was decided to include ratings scales for the test participants to evaluate the soundscape in terms of the three above categories. Fig. 2 shows the category ratings question as presented to the test participants. The purpose of including this question, in both the preliminary and main listening tests, was to see how the presence of visual features can alter the perceived category of an environment, and how this relates to evoked emotional state.

2.2. Data Collection

The data used in this study were collected from various locations around the North of the United Kingdom, including: Dalby forest, a natural environment; Pickering, a suburban/rural environment; and Leeds city centre, a highly developed urban environment. All of the soundscape recordings were made in FOA using a Soundfield STM 450 microphone [26]. Concurrent A-weighted noise level measurement were taken to allow for calibration of later auralisation.



To what extent does the soundscape belong in each of the following categories?

	Not at all		Somewhat		Very much
Natural/animal	○	○	○	○	○
Human	○	○	○	○	○
Industrial/mechanical	○	○	○	○	○

Figure 2: The category ratings question as presented to test participants.

Table 1 gives details of the sound sources present in each of the 16 clips used in the listening test. These clips were 30 seconds long and extracted from the 10 minutes of soundscape recording made at each location. These clips have been used in previous stages of this research [13, 27].

The visual data was collected at each recording location using six GoPro cameras mounted as the faces of a cube in a Freedom360 rig [28]. At each location a still image was taken immediately before recording began, and then full motion video recordings were made alongside the FOA sound recordings.

3. PRELIMINARY LISTENING TEST

This section covers the content creation and test procedure for the preliminary listening test, as well as its results. This includes the conversion of the FOA soundscape recordings to stereo UHJ format, and the stitching of the still GoPro photographs to create panoramic images of the recording location.

3.1. Stereo UHJ Conversion

In order to present the recorded soundscape material over headphones without head-tracking, the FOA signals had to be converted to a suitable two-channel format. It was decided to make use of Ambisonic UHJ stereo format, where the $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}$ channels of an FOA recording are used to translate the horizontal plane of the soundfield into two-channels [29]. The resultant signal can the be shared online and reproduced over headphones, allowing the FOA recordings to be used with the spatial content of the $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}$ channels preserved in reproduction. The use of this format has been established as ecologically valid in a prior stage of this research [30], where it was shown that emotional states evoked by exposure to the stereo UHJ format soundscape recordings were significantly similar to those evoked by full FOA renderings in a 16-loudspeaker listening rig.

The following equations are used to convert from the $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}$ channels of the FOA signal to two stereo channels:

$$S = 0.9397\mathbf{W} + 0.1856\mathbf{X} \quad (1)$$

$$D = j(-0.342\mathbf{W} + 0.5099\mathbf{X}) + 0.6555\mathbf{Y} \quad (2)$$

$$L = 0.5(S + D) \quad (3)$$

$$R = 0.5(S - D) \quad (4)$$

where j is a $+90^\circ$ phase shift and L and R are the left and right channels respectively of the resultant stereo UHJ signal [31]. Note that the Cartesian reference for FOA signals is given by ISO standard 2631 [32], and the $\mathbf{Z}$ channel of the FOA recording is not used.

3.2. Preliminary Test Procedure

The listening test was presented using Qualtrics [33] to administer the questions to the test participants, and using MATLAB to play the stereo UHJ audio and present the panoramic images using FSPViewer [34] (a freely downloadable viewer for spherical panoramic images). Presenting the images in this way allowed participants to click-and-drag the panoramic image to 'look' around the environment (which they were encouraged to do). These images were created using Kolor Autopano [35] to stitch together the still images from the GoPro cameras into single equirectangular spherical panoramic images. An example image can be accessed online [36].

All 16 soundscape clips were presented to the test participants in both the aural and audiovisual stages. These were presented in a random order each time and were preceded by two orienting stimuli. The audio-only test was completed by 31 test participants, and the audiovisual test was completed by 11 participants. Of the 31 audio-only test participants, 20 were male, and 16 were aged under 26. No demographic data were collected for the audiovisual test, as analysis of previous results did not indicate any significant effect on test results due to demographic factors. The next section includes an evaluation and discussion of the test results.

3.3. Preliminary Test Results

A Shapiro-Wilk's test was applied to all of the rating scales for each test stimuli as a test for normality [37]. Only a handful were identified as normally distributed. As such, in order to make comparisons between the results for the different stimuli, the Mann-Whitney test was used [38]. This test is suitable for comparing the values of two variables that are not normally distributed [39]. It is also suitable for comparing variables with small, arbitrary, sample sizes, including where the sample sizes of the two variables are different.

The purpose of applying the Mann-Whitney test was to indicate where the test results were significantly different for each of the five rating scales (Valence, Arousal, Natural, Human, and Mechanical) when comparing the results for the audiovisual stimuli with the audio alone. Fig. 3 shows the Mann-Whitney test results for the preliminary listening test data, indicating these significant differences. The next section will discuss theses results.

3.3.1. Significant Differences

The three clips showing a significant difference in arousal values are 6A, 6B, and 7B. For all three of these clips the arousal rating value was significantly larger when the clip was presented with the visual stimuli. Both of these recording locations were in Leeds city centre: one next to a main road (location 7); one on a pedestrianised street (location 6). This increase in arousal is therefore possibly due to the presences of cars and people in the images of the scenes that are not so pronounced in the soundscape recordings.

The 6 clips showing a significant difference in valence values are 1A-2A, 3A-3B, and 8A. As with the arousal results, for all of these clips the presence of visual stimulus results in an increase in valence. For clips 1A and 1B this is unsurprising: the soundscape clips contain some birdsong and insect noise, but despite their hi-fidelity (where the sound sources present are clearly defined with little background noise [1]) there is little information given to indicate the features of the recording location. As such it is to be anticipated the presence of the visual features with the soundscape results in an increased valence rating.

Location	Site	Clip A Sound Sources	Clip B Sound Sources
Dalby Forest (Rural/Natural)	1. Low Dalby Path	Birdsong, Owl Hoots, Wind	Birdsong and honking, Insects, Aeroplane flyby
	2. Staindale Lake	Birdsong, Wind, Insects, Single car	Insects, Birdsong, Water
North York Moors (Rural/Suburban)	3. Hole of Horcum	Birdsong, Traffic, Bleating	Birdsong, Traffic, Conversation
	4. Fox & Rabbit Inn	Traffic, Car door closing, Car starting	Traffic, Footsteps, Car starting
	5. Smiddy Hill, Pickering	Traffic, Car door starting, Conversation	Birdsong, Distant traffic
Leeds City Centre (Urban)	6. Albion Street	Busking, Footsteps, Conversation, Distant traffic	Workmen, Footsteps, Conversation, Distant traffic
	7. Park Row	Traffic, Buses, Wind, Busking	Busking, Footsteps, Conversation, Distant traffic
	8. Park Square	Birdsong, Traffic, Conversation, Shouting	Workmen, Traffic, Conversation, Birdsong

Table 1: Details of the sound sources present in the two 30 second long clips (labelled A and B) recorded at each of the eight locations.

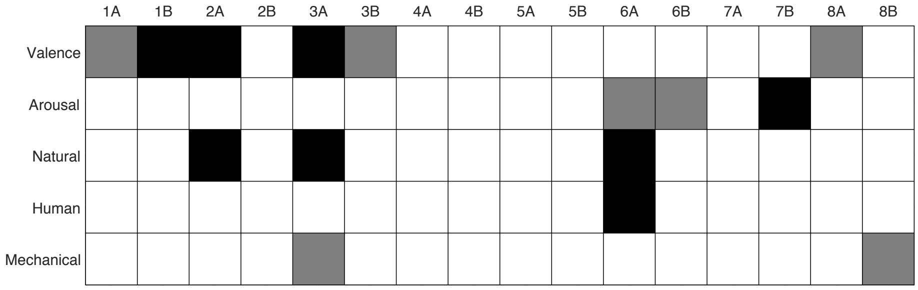


Figure 3: Mann-Whitney test results for the preliminary listening test, comparing results for each of the five rating scales for each of the 16 test stimuli when presented as the soundscape alone and with accompanying still panoramic images. Dark squares indicate a significant difference at 95% confidence ($p < 0.05$), and Light marked squares at 90% confidence ($p < 0.1$). White squares indicate no significant difference at either confidence level.

For clip 2A a similar effect can be observed, due to the presence of single car driving past. These results suggest that the visual setting (greenery and trees, peaceful lake, big sky) results in a significantly increased valence rating.

The significant increases in valence value for the audiovisual presentation of clips 3A and 3B also show the same effect: the aural information in these clips contains some natural sounds and traffic noise that indicate little about of the surrounding countryside of the North York Moors national park.

Likewise the soundscape of clip 8A contains some birdsong alongside quiet traffic noise (and some sounds of human activity), but the visuals recorded at that location show an inner city park with foliage, flowers, and some trees. This green infrastructure is clear when viewing the scene, but not evident in any explicit way in the audio-only presentation, and is likely responsible for evoking an alternative emotional state where reported valence levels (i.e. how pleasant the scene is) are higher.

The significant differences in the natural rating scale support this argument in part: clips 2A and 3A show a significant increase in the natural rating with the presence of visual stimuli, which includes a forest and countryside respectively. Clip 6A (recorded on a pedestrianised shopping centre street) also shows a significant increase in the natural rating with the presence of visual information. This environment contains some very minor elements of green infrastructure in the form of a couple of trees in some small pots. Whilst this cannot directly be correlated with a change in the valence rating for the environment, it does indicate how even a very slight presence of green infrastructure can change an individual's experience and perception of a location. This location also sees a significant decrease in the human category rating for the audiovisual presentation of the clip relative to the soundscape alone. This is possibly due to the difference between reality and expectation of the visual setting: the dominant sound sources in this clip are human sounds (including very loud conversation, footsteps, and some shouting) with only some distant traffic noise. However the visual setting is dominated by concrete in the form a pavement, shop-fronts and some larger inner city buildings reducing the impact of the human activity.

The two soundscapes showing a significant difference in the mechanical category rating are 3A and 8B, both of which saw a decrease in mechanical rating with the introduction of visual stimuli. In a way these two clips can be considered as the corollary of one another: clip 3A shows a natural environment 'interrupted' by the presence of a busy road; and clip 8B shows a green-infrastructure (a park) in the context of a large city. As such both of these soundscape clips indicate little about the features of the visual settings, resulting in a decreased mechanical rating for the audiovisual presentation.

3.3.2. Perceptual Noise Impact Rating

In order to further investigate the effect of certain visual features on the emotional state evoked by a soundscape, the valence and arousal rating scales can be combined to form a single measure of the emotional state evoked by a noisy soundscape. This new measure is called the Perceptual Noise Impact Rating (PNIR) and was introduced as part of this body of research in [40]. It is formulated by:

$$\mathrm{PNIR} = 1 - 0.5(1 - \mathrm{A} + \mathrm{V}) \qquad (5)$$

where A and V represent the Arousal and Valence scores respectively (where the scores are normalised between 0 and 1).

Fig. 4 shows a summary of PNIR results from the preliminary listening test. Indicated in this plot are the mean PNIR values across all participants for each of the 16 stimuli for both the audio-only and audiovisual listening conditions. These results show a trend in the data towards three groups of PNIR values:

1. Clips 1A-2B: These soundscapes were recorded at two locations in Dalby forest, and are comprised of many natural

sounds (birdsong, insects, wind) and visual features (trees, a lake, open sky).

2. Clips 4A-7B: These soundscapes were recorded in highly developed environments, including various locations in the centre of the city of Leeds, and next to a road in the town of Pickering. The most commonly identified sound sources in these clips were traffic noise, other mechanical noise, and human sounds (footsteps and conversation).
3. Clips 3A-3B and 8A-8B: These soundscapes were recorded in environments that can be considered as being on the interface between the recording locations of the two above categories. Location 3 was next to a country road overlooking a wide expanse of countryside, and location 8 was in a park in Leeds city centre. Both of these environments contained a mixture of mechanical and natural sounds (i.e. relatively quiet traffic noise and birdsong) and visual features (i.e. flowers, trees and other greenery alongside the roads and buildings).

These three emotional groups were used alongside the Mann-Whitney test results to identify which of the soundscape clips to use in the main listening test.

Clips 1B and 2A were chosen to represent group 1: clip 1B was recorded in Dalby forest and contains natural sounds and visual elements; clip 2A was recorded at a nearby lake and again presents many natural sounds and visual elements, as well as a single car drive by.

Clips 6A and 7B were chosen to represent group 2: clip 6A was recorded on a pedestrianised street lined with shops; clip 7B was recorded next to a busy road in Leeds city centre. Both of these clips contain mainly human and mechanical sounds, with little in the way of natural sounds or visual elements.

Clips 3A and 8A were chosen to represent group 3: clip 3A was recorded next to a road in the North York Moors national park; clip 8A was recorded in a small park in the centre of Leeds. As stated above, these locations both represent something of an interface between natural and developed habitats and contain both human and natural sounds and visual elements, including the presence of green infrastructure.

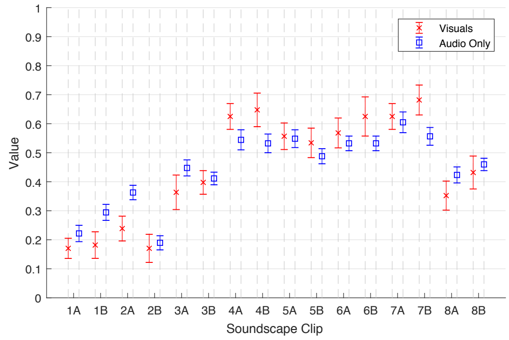


Figure 4: A summary of PNIR ratings from the preliminary listening test results.

4. MAIN LISTENING TEST

This section covers the creation of VR content and the test procedure methodologies used in the main listening test.

4.1. Virtual Reality Content Creation

Fig. 5 depicts a flow diagram for the creation of full motion spherical audiovisual content ready for playback on YouTube, either via a VR headset or on a standard computer monitor. Firstly Kolor Autopano is used to stitch together the six feeds of GoPro footage into a single equirectangular panoramic video [35]. FFMPEG [41], a free software project designed for handling multimedia data, is then used to add the FOA audio (with its channels in ACN, rather than Furse-Malham, order) to the panoramic footage [42]. In order for this file to then be uploadable to YouTube [43] the Spatial Media Metadata Injector [44] is used to indicate that the file contains a panoramic video. For the 'audio-only' stimuli a still image of equirectangular perspective lines was used as the visual component, in order to give the test participants some sense of orientation [45]. The resultant content can be viewed in the following two YouTube playlists: the audio-only playlist [46]; and the full audiovisual playlist [47].

4.2. Main Test Procedure

For the main listening test there were 20 participants, split into two groups of 10. Each group was exposed to the six chosen soundscape recordings: one group experienced the audio-only soundscapes first, and then experienced them with accompanying video footage; the other group of participants experienced the stimuli with the order reversed. Within each listening condition the presentation order was randomised. As with the audiovisual stage of the preliminary listening test no demographic data were collected here. In each viewing condition participants were encourage to pan and 'look around' the environment, with YouTube updating the binaural rendering of the FOA audio according to the visual perspective.

The soundscapes were presented as YouTube content embedded in Qualtrics. The presentation order within each set of stimuli was randomised. As with the preliminary test, each stimulus was rated in terms of valence and arousal, and in terms of the three established soundscape categories. Test participants were also asked to list the sound sources and visual elements in the scene.

4.3. Main Test Results

This section presents an evaluation and analysis of the results of the main listening test. As with the preliminary listening test, a Shapiro-Wilks test for normality was used. Similarly only a very small number of variables were shown to demonstrate a non-normal distribution. The main listening test results were therefore suitable to be compared using the Mann-Whitney U-test.

Initially the results for all test participants are all compared with no consideration of the order in which the two sets of stimuli were presented. Further analysis is then presented in order to investigate how the order in which test participants were exposed to the aural and audiovisual stimuli has affected their experience of the soundscape.

4.3.1. Overall Comparison

Fig. 6a shows the results from Mann-Whitney U-test applied to the main listening test results, comparing the results for the audio-only soundscape presentations with the audiovisual ones.

As this figure indicates, there are relatively few significant differences in any of the rating scales when comparing the two

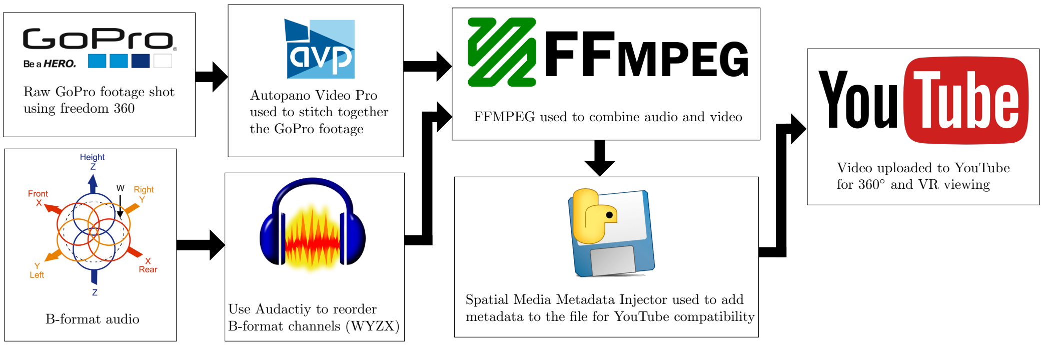

GoPro
Be a HERO.
Raw GoPro footage shot using freedom 360
avp
Autopano Video Pro used to stitch together the GoPro footage
FFMPEG
FFMPEG used to combine audio and video
YouTube
Video uploaded to YouTube for 360° and VR viewing
Height
Z
Front
X
W
Right
Y
Y
Left
X
Rear
Z
B-format audio
Use Audactiy to reorder B-format channels (WYZX)
Spatial Media Metadata Injector used to add metadata to the file for YouTube compatibility

Figure 5: A flow diagram showing the method used in this study for VR content creation.

listening conditions. The clip that shows the most significant differences are for clip 7B, which was recorded next to a busy road in Leeds city centre. Compared to the audio only presentation of this soundscape clip, the ratings for the audiovisual presentation show significantly increased valence and human ratings, and a significantly reduced PNIR rating.

There are two aspects of the visual setting of this clip that have likely contributed to these differences: firstly, it is hard from listening to the soundscape alone to get a sense of how close to the road the listener is, as the traffic sounds are very loud, whilst the visual setting makes it clear that recording position is safely away from the road; secondly, the square that this recording was made at is lined with some trees which were clearly identified by test participants as a major visual feature of the scene.

The only other significant difference shown in Fig. 6a is for clip 3A, where the presence of visuals alongside the soundscape results in a significantly higher natural rating (as expected from the preliminary test results).

4.3.2. Order Dependence

Having now considered all of the results for both listening conditions for both groups of test participants, a breakdown of results by presentation order will now be considered.

Fig. 6b shows the results of applying the Mann-Whitney U-test to just the first listening condition experienced by each group: i.e. the audio-only results for the group that experienced those clips first compared with the audiovisual results from the other group.

Firstly it is interesting to note that the significant differences shown in this figure are not the same as those shown in Fig. 6a. These results show that for clip 1B, recorded at Dalby forest, the version of the clip presented with the accompanying visuals received a significantly greater valence rating, and a significantly lower mechanical rating. As with the preliminary test results, the change in valence rating is most likely due to the pleasantness of the trees and open sky in the visual setting. The mechanical rating is also lower with the presence of visuals for this clip. The soundscape contains some ambiguous noise that may be distant traffic, wind, or aircraft flying overhead. When presented with visual features this ambiguity is resolved and the natural visual elements take precedence.

A significant difference in mechanical rating can also be seen for clip 7B; this is most likely due to the human elements (people walking past) and minor elements of green infrastructure (some trees lining the square) that reduce the impact of the mechanical noise on the audiovisual experience of the soundscape.

Also shown in Fig. 6b are two significant differences in the ratings for clip 6A: the audiovisual presentation of this clip received significantly lower valence and human ratings than the audio-only version. This is most likely due to, again, elements of the visual environment that are not manifest in the soundscape itself: in this case the inner city shopping district buildings. In the audio-only presentation the dominant features are conversation and footsteps, whilst in the visual presentation the large buildings are the dominant feature. The presence of these buildings and paved streets also possibly gives some orientation for the background noise in the clip, grounding its otherwise ambiguous nature and indicating to participants that there is some distant traffic noise present.

Fig. 6c shows the Mann-Whitney U-test results comparing the two listening conditions for the group who experienced the audio-only soundscapes first, followed by audiovisual presentation. For clip 3A, recorded next to the Hole of Horcum in the North York Moors national park, there is a significant increase in the natural rating for the audiovisual presentation of the clip relative to the audio-only version due to the rolling countryside (something not obviously present in the soundscape itself).

The category ratings for all other soundscapes show no significant differences between listening conditions, but for clips 7B and 8A there are some differences in the emotion ratings. For clip 7B this means a significantly higher valence rating, and a significantly lower PNIR, once again showing how the presence of a relatively small amount of green infrastructure can improve the experience of a location.

Also of note in Fig. 6c is that for clip 8A, recorded at an inner city park in Leeds, there is indicate a significant decrease in the PNIR for the clip presented with visuals relative to the audio alone. This is interesting as neither the valence nor arousal ratings on their own show significant differences, but when these ratings are combined a significant difference can be demonstrated.

4.4. Discussion

When taken together the above results can be summarised as three main findings. Firstly, many of the significant differences in emotional or categorical ratings for the different soundscape clips are (perhaps unsurprisingly) due to the visual features that are not manifest in the soundscape clips. This makes clear the need for a cross-modal approach to soundscape evaluation as any real-life soundscape evaluation procedure will have to consider the visual context of that soundscape.

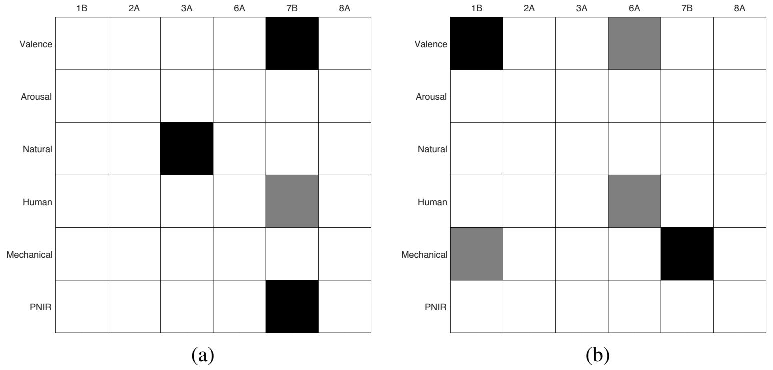


Figure 6: Mann-Whitney test results indicating significant differences between the two listening conditions. Plot (a) compares all of the results from both groups for each clip (b) compares the results for the first listening condition experienced by each group, and (c) compares only the results from the participants that experienced the soundscapes as audio-only first and then audiovisually. Dark marked squares indicate a difference at 95% confidence ($p < 0.05$), and light marked squares indicate a difference at 90% confidence ($p < 0.1$).

Secondly, for many of the differences in perception of the soundscape clips, the presence of elements of green infrastructure can be identified. This lends credence to the idea that green infrastructure, whilst not necessarily resulting in a significant change to an environment's acoustic properties, can improve the experience of that location.

Thirdly, the SAM, which has been examined thoroughly throughout this research in terms of its usefulness for soundscape evaluation, has been shown to be very useful in examining differences between the emotional states evoked by different soundscape. The PNIR, a combination of the valence and arousal dimensions of the SAM into a single perceptual rating, has also been shown to be useful in this study for discerning significant differences between emotional states evoked by soundscapes.

5. CONCLUSION

This paper has presented the results of two listening tests, each making use of soundscape recordings and images of the recording locations to investigate how a cross-modal approach to soundscape evaluation can be use to measure the impact of green infrastructure. The SAM and category ratings were used to conduct this evaluation: first in a preliminary test making use of stereo-UHJ renderings of the soundscape clips and still images; and then in a main listening test presenting the soundscapes in dynamically rendered binaural audio accompanied by full motion panoramic video footage.

Whilst the results presented in this paper show some significant differences in emotion and category rating between the audio only and audiovisual clip presentation, further work should be conducted comparing ratings for audiovisual soundscape presentation where the visual setting is altered, for example through the addition of trees or other aspects of green infrastructure. Such research would build on the results presented here, which validate the methodology in terms of the rating scales used, and the VR content creation and presentation methods.

6. REFERENCES

[1] R. Schafer, *The Soundscape: Our Sonic Environment and the Tuning of the World*. Inner Traditions/Bear, 1993. [Online]. Available: http://books.google.co.uk/books?id=ltBrAwAAQBAJ

[2] S. Payne, W. Davies, and M. Adams, "Research into the practical policy applications of soundscapes concepts and techniques in urban areas. DEFRA report NANR200, june 2009," 2009.

[3] E. Thompson, *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933*. MIT Press, 2004. [Online]. Available: http://books.google.co.uk/books?id=7jvtvGbatv4C

[4] G. Keizer, *The Unwanted Sound of Everything We Want: A Book About Noise*. PublicAffairs, 2010. [Online]. Available: https://books.google.co.uk/books?id=yZ44DgAAQBAJ

[5] B. Truax, *Acoustic Communication*. Greenwood Publishing Group, 1984.

[6] J. Macdonald and H. McGurk, "Visual influences on speech perception processes," *Perception & Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978. [Online]. Available: http://dx.doi.org/10.3758/BF03206096

[7] P. Lercher and B. Schulte-Fortkamp, "The relevance of soundscape research to the assessment of noise annoyance at the community level," in *Proceedings of the Eighth International Congress on Noise as a Public Health Problem*, 2003, pp. 225–231.

[8] S. Viollon, L. C., and C. Drake, "Influence of visual setting on sound ratings in an urban environment," *Applied Acoustics*, vol. 63, no. 5, pp. 493 – 511, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/\S0003682X01000536

[9] R. Ulrich, "View through a window may influence recovery," *Science*, vol. 224, no. 4647, pp. 224–225, 1984.

[10] K. Tzoulas, K. Korpela, S. Venn, V. Yli-Pelkonen, A. Kaźmierczak, J. Niemela, and P. James, "Promoting ecosystem and human health in urban areas using green infrastructure: A literature review," *Landscape and urban planning*, vol. 81, no. 3, pp. 167–178, 2007.

[11] D. T. Murphy, A. Southern, and F. Stevens, "Sounding our smart cities: Soundscape design, auralisation and evaluation for our urban environment," in *Sound + Environment 2017*, Hull, UK, 2017.

[12] L. Steg, A. van den Berg, and J. de Groot, *Environmental Psychology: An Introduction*, ser. BPS textbooks in psychology. Wiley, 2012. [Online]. Available: http://books.google.co.uk/books?id=RFHmw57kiNwC

[13] F. Stevens, D. T. Murphy, and S. L. Smith, "Emotion and soundscape preference rating: using semantic differential pairs and the self-assessment manikin," in *Sound and Music Computing conference, Hamburg, 2016*, Hamburg, Germany, 2016.

[14] C. Osgood, "The nature and measurement of meaning." *Psychological bulletin*, vol. 49, no. 3, p. 197, 1952.

[15] J. Kang and M. Zhang, "Semantic differential analysis of the soundscape in urban open public spaces," *Building and environment*, vol. 45, no. 1, pp. 150–157, 2010.

[16] W. Davies, N. Bruce, and J. Murphy, "Soundscape reproduction and synthesis," *Acta Acustica United with Acustica*, vol. 100, no. 2, pp. 285–292, 2014.

[17] S. Viollon and C. Lavandier, "Multidimensional assessment of the acoustic quality of urban environments," in *Conf. proceedings "Internoise", Nice, France, 27-30 Aug*, vol. 4, 2000, pp. 2279–2284.

[18] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[19] M. Bradley and P. J. Lang, *The International affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings*. NIMH Center for the Study of Emotion and Attention, 1999.

[20] M. Bradley, B. Cuthbert, and P. Lang, "Picture media and emotion: Effects of a sustained affective context," *Psychophysiology*, vol. 33, no. 6, pp. 662–670, 1996.

[21] A. Léobon, "La qualification des ambiances sonores urbaines," *Natures-Sciences-Sociétés*, vol. 3, no. 1, pp. 26–41, 1995.

[22] W. Yang and J. Kang, "Acoustic comfort and psychological adaptation as a guide for soundscape design in urban open public spaces," in *Proceedings of the 17th International Congress on Acoustics (ICA)*, 2001.

[23] L. Anderson, B. Mulligan, L. Goodman, and H. Regen, "Effects of sounds on preferences for outdoor settings," *Environment and Bevior*, vol. 15, no. 5, pp. 539–566, 1983.

[24] G. Watts and R. Pheasant, "Tranquillity in the scottish highlands and dartmoor national park–the importance of soundscapes and emotional factors," *Applied Acoustics*, vol. 89, pp. 297–305, 2015.

[25] F. Stevens, D. T. Murphy, and S. L. Smith, "Soundscape categorisation and the self-assessment manikin," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK, 2017.

[26] E. Benjamin and T. Chen, "The native b-format microphone," in *Audio Engineering Society Convention 119*, 10 2005.

[27] F. Stevens, D. T. Murphy, and S. L. Smith, "Soundscape auralisation and perception for environmental sound modelling," in *Sound + Environment 2017*, Hull, UK, 2017.

[28] "Freedom 360 mount," 2015. [Online]. Available: http://freedom360.us/shop/freedom360/

[29] R. Elen, "Ambisonics: The surround alternative," in *Proceedings of the 3rd Annual Surround Conference and Technology Showcase*, 2001, pp. 1–4.

[30] F. Stevens, D. T. Murphy, and S. L. Smith, "Ecological validity of stereo uhj soundscape reproduction," in *In Proceedings of the 142nd Audio Engineering Society (AES) Convention*, Berlin, Germany, 2017.

[31] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[32] ISO, *Mechanical Vibration and Shock: Evaluation of Human Exposure to Whole-body Vibration. Part 1, General Requirements: International Standard ISO 2631-1: 1997 (E).* ISO, 1997.

[33] J. Snow and M. Mann, "Qualtrics survey software: handbook for research professionals," 2013.

[34] "FSPViewer," 2017. [Online]. Available: http://www.fsoft.it/FSPViewer/

[35] "Kolor autopano," 2015. [Online]. Available: http://www.kolor.com/autopano-video/#start

[36] F. Stevens, "Dalby forest panoramic image," 2017. [Online]. Available: http://www.dermandar.com/p/bIrnfi

[37] S. Shapiro and M. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, pp. 591–611, 1965.

[38] H. Mann and D. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[39] S. Harriet, "Application of auralisation and soundscape methodologies to environmental noise," Ph.D. dissertation, University of York, 2013.

[40] A. Southern, F. Stevens, and D. T. Murphy, "Sounding out smart cities: Auralization and soundscape monitoring for environmental sound design," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3880–3880, 2017.

[41] "FFMPEG." [Online]. Available: https://www.ffmpeg.org/

[42] B. Wiggins, "Youtube, ambisonics and vr," 2016. [Online]. Available: https://www.brucewiggins.co.uk/?p=666

[43] "Google support: Upload 360-degree videos." [Online]. Available: https://support.google.com/youtube/answer/6178631\?hl=en-GB

[44] "Spatial media metadata injector," 2016. [Online]. Available: https://github.com/google/spatial-media/releases

[45] D. Swart, "Equirectangular perspective lines," 2016. [Online]. Available: https://www.flickr.com/photos/dmswart/26363697850

[46] F. Stevens, "Final test audio stimuli YouTube playlist," 2017. [Online]. Available: https://www.youtube.com/playlist?list=PL-3kCuZ4n30QM5zUhzqfn9vkiwZPl2QzD

[47] ——, "Final test visual stimuli YouTube playlist," 2017. [Online]. Available: https://www.youtube.com/playlist?list=PL-3kCuZ4n30TIn40XSXdz5Y-sPr5brNet