Classification of soundscapes of urban public open spaces

Abstract: It is increasingly acknowledged by landscape architects and urban planners that the soundscape contributes significantly to the perception of urban public open spaces. Describing and classifying this impact, however, remains a challenge. This article presents a hierarchical method for classification that distinguishes between *backgrounded* and foregrounded, *disruptive* and supportive, and finally *calming* and *stimulating* soundscapes. This four-class classification is applied to a growing collection of immersive audio-visual recordings of sound environments from around the world that could be explored using virtual reality playback. To validate the proposed methodology, an experiment involving 40 participants and 50 soundscape stimuli collected in urban public open spaces worldwide was conducted. The experiment showed that (1) the virtual reality headset reproduction based on affordable spatial audio with 360-degree video recordings was perceived as ecologically valid in terms of realism and immersion; (2) the proposed classification method results in well-separated classes; (3) membership to these classes could be explained by physical parameters, both regarding sound and vision. Moreover, models based on a limited number of acoustical indicators were constructed that could correctly classify a soundscape in each of the four proposed categories, with an accuracy exceeding 88% on an independent dataset.

Keywords: soundscape, classification, urban space

1. Introduction

Soundscape, as defined by the International Organization for Standardization (ISO), is an "acoustic environment as perceived or experienced and/or understood by a person or people, in context" (ISO, 2014). The urban soundscape contributes to the perceived quality of the urban environment and the identity of a city. Ambient sounds may evoke thoughts and emotions, may influence our mood or steer our behavior. Cities are comprised of many types of public outdoor spaces, each with their distinctive soundscape. Inspired by the potential positive effects a suitable acoustic environment may have on well-being of citizens and the attractiveness of the city, the challenge of designing the acoustic environment of urban public outdoor spaces has attracted attention since decades (Southworth, 1969; Schafer, 1994).

During the past decades, research on the urban sound environment and soundscape has grown, driven by increased population density and abundance of mechanical sounds in mega-cities across the world. Sound in outdoor environments has traditionally been considered in negative terms as both intrusive and undesirable (Jennings and Cain, 2013). However, sound may provide positive effects as well, such as enhancing a person's mood, triggering a pleasant memory of a prior experience, or encouraging a person to relax and recover (Payne, 2013). Where classical noise control exclusively focusses on reducing levels of unwanted sounds, soundscape design requires new tools. Hence the advent of realistic and affordable immersive audio-visual reproduction systems (head-mounted displays), backed by increasingly efficient and realistic acoustic simulation and auralization models (Vorländer, 2008) has been identified as a key enabling technology. Immersive virtual reality could also become a valuable tool for interactive participatory evaluation of the soundscape in urban planning and design projects (Puyana-Romero et al., 2017; Echevarria Sanchez et al., 2017), as virtual reality reproduction systems are rapidly becoming affordable and widely available.

Design is often inspired by good examples. As context is an important part of the soundscape and the visual setting is a string cue for context, examples of acoustic environments should be embedded in accurate 360-degree visualization. To date, however, no unique protocol or standards exist for immersive audio-visual recording and playback of urban environments with soundscape in mind (Hong et al., 2017). In addition to providing examples, high-quality immersive recordings of existing spaces are highly valuable to serve as an ecologically valid baseline for studying the perceptual outcome of noise control and soundscape measures. Hence, such recordings are now being collected in cities across the globe. To unlock such collections, a suitable classification is needed and best examples of each class need to be identified.

One could consider a purely acoustical categorization (Rychtáriková and Vermeir, 2013). However, according to the soundscape definition (ISO, 2014), soundscape evaluation should not be restricted to acoustical determinations only (Zannin et al., 2003), as the social context (Maris et al., 2007), visual context (Sun et al., 2018a) and individual differences need to be included (Dubois et al., 2006).

When asked to describe the urban acoustic environment, persons tend to name audible sounds and their sources and may relate the quality of the environment to the meaning given to these sounds (Dubois et al., 2006). In view of the importance of audible sounds, classification schemes based on urban sound source sorting have been proposed (Léobon, 1995; Brown et al., 2011). Such classifications can easily be applied to collections of audio-visual recordings through listening experiments conducted by sound specialists, yet one should remain aware that attention plays an important role in the perception of the acoustic environment in a real context (Oldoni et al., 2013). Classification based on audible sources does not capture the influence of the composition as a whole on persons and therefore should be complemented by more holistic indicators.

Holistic descriptors that have been proposed previously and that could be used for classification include: pleasantness, music-likeness, restorativeness, appropriateness. (Aletta et al., 2016; Botteldooren et al., 2006). A lot of research has focused on the soundscape descriptors inspired by emotion-denoting adjectives (Brown, 2012; Aletta et al., 2016). The well-known circumplex model of affect (Russell, 1980) identifies eight affective concepts that can be mapped to a two-dimensional plane. Previous research (Berglund and Nilsson, 2006; Axelsson et al., 2010) translated core affect to the physical environment that causes it and showed that outdoor soundscape quality may be represented by two main orthogonal components: pleasantness and eventfulness. In such a 2D model specific directions are labelled: exciting (45˚), chaotic (135˚), monotonous (225˚) and calm (315˚).

Although very popular, this assessment and classification framework has also been subject to some critique. Regarding the core affect model itself, research has identified a main problem with the two-dimensional approach offered by Russell: a variety of overlapping emotional concepts can be placed in the same quadrant of the model (e.g., Ekkekakis, 2008). Based on the 2D core affect model, Latinjak (2012) proposed a three-dimensional model, where a third dimension, namely “time perspective”, was added next to arousal and valence. In addition, the classification of soundscape in the pleasantness – eventfulness plane assumes that the environmental sound is attentively listened to. It assumes that perceiving the sonic environment is a main purpose of an individual visiting a place, which is not often the case. Unawareness of the surroundings (inattentional blindness (Simons and Chabris, 1999) and inattentional deafness (Macdonald and Lavie, 2011)) occurs especially during moments with reduced attention towards the environment. The sonic environment is thus often backgrounded.

Besides the soundscape descriptors and the 2D core affect model, a triangular qualitative urban sound environment mapping technique was recently proposed (Kamenický, 2018). This research used activities, mechanisms and presence to build an objective soundscape map based on composition of sound events. A significant correlation between qualitative cognitive-semantic variables clustering and quantitative acoustic and psychoacoustic parameters agglomerative clustering was proposed.

In an urban environment, the soundscape, the landscape, etc., and its users form an ecological entity. It might therefore be more suitable if the soundscape classification of existing urban sites could be treated within such a holistic context. With the aforementioned discussion in mind, we propose a coarse hierarchical classification that could be used for labelling audiovisual collections or as a first mapping of the city. The proposed classification, shown in Figure 1, was first suggested in De Coensel et al. (2017). In a first stage, soundscapes are classified according to whether they are backgrounded or contain foregrounded sound elements when perceived within context (Botteldooren et al., 2015) – where only visual context has been considered here. Foregrounded sound affects the overall perception of the environment. In a second stage, one could distinguish between sonic environments that are disruptive or supportive for the envisaged use. Disruptive sound environments could lead to annoyance. Finally, the sonic environment could be supportive for the overall experience of the living environment in many different ways. Here, the proposed classification follows the arousal dimension of core affect to distinguish between calming (reducing arousal) and stimulating (increasing arousal). We forward the hypothesis that the proposed classification system is strongly related to the sonic environment itself and less sensitive to differences between people than previous classification systems and therefore more appropriate for classifying the audio-visual representation of a place.

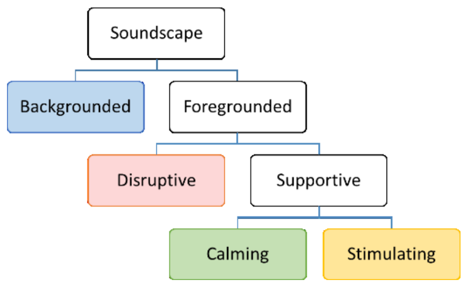

Figure 1 – Proposed hierarchical classification of urban soundscapes.

It is worth noticing that the proposed classification is not crisp; one could potentially mathematically formalize this classification using fuzzy set memberships.

In this article, the proposed classification will for the first time be made operational through a questionnaire that is administered to a panel of volunteers that is experiencing the immersive playback at the laboratory of a collection of audio-visual recordings at an urban site (Section 2.2.3). This will allow exploring the rationality of the proposed soundscape classification, the underlying affiliation between categories and its comparison with the 2D core affect model (Section 3.3). Classification of a collection achieved by questioning persons about the soundscape as experienced in the virtual reality environment has some drawbacks: because of the variability between persons (Sun et al., 2018b), this requires an assessment panel of sufficient size, which results in a large effort and cost for classifying new recordings. Hence this paper also proposes models based on acoustical parameters (Section 3.5).

2. Methodology

2.1 Methods for objective measurements and recording protocol

The methodological approach for the site selection, audio-visual recordings and post-processing of the for the Virtual Reality application are reported in Appendix I.

2.2 Experiment: Soundscape classification

2.2.1 Materials and participants

In total, 50 one-minute recordings were selected from the complete recording in this experiment (e.g.: Figure 3). One minute is very short for assuring that participants are not focusing on the sound, but this time interval was chosen as a compromise that still gave a good impression but would not take too much time from the users of the collection. The Table IV in Appendix III gives the overview of their basic characteristics namely location, time, and $L_{\mathrm{Aeq,\,1\,min}}$ (A-weighted equivalent sound pressure levels during the one-minute period). The L_{Aeq} of each stimulus was calculated on the basis of the binaural signal, applying an independent-of-direction (ID) equalization, and taking the energetic average between both ears.

To allow for completely independent validation of prediction models, the whole experiment was repeated two times. First, 25 stimuli (Table IV in Appendix III – collection 1) were chosen for participant group 1 (20 participants, 6 female, Age_{mean}=28.9 yr, standard deviation 2.8 yr, range: 25-35 yr). Five cities (Montreal, Boston, Tianjin, Hongkong and Berlin) were included in the experiment, and each city contributed with 5 stimuli. The stimuli were presented city by city to the participants. The city order and the order of stimuli in each city were randomized.

Another 25 recordings (Table IV in Appendix III – collection 2) were presented to participant group 2 (20 participants, 5 female, Age_{mean}=30.2 yr, standard deviation 5.6 yr, range: 22-46 yr). The number of stimuli per city was different now. These 25 recordings were grouped into 5 groups of 5 stimuli each, avoiding e.g. that one group contained only parks. The group order and the order of stimuli in each group were again fully randomized. To avoid social biases, the participants were a well balance in terms of occupation, nationality and education level.

All participants had normal hearing status which was assessed via pure tone audiometry (PTA) carried out in a soundproof room using a regularly calibrated AC5Clinical Computer Audiometer. All participants had normal color vision which was tested by the “Ishihara test for color deficiency” (Ishihara, 1957). The participants performed the perception experiment individually, and were offered a gift voucher as compensation.

Figure 3 – Example: snapshot of stimuli R0001. (more stimuli could be found in Supplement 1).

2.2.2 Experiment setup

Participants joined this experiment inside a soundproof booth (Figure 4), where the process was monitored through a double-glassed window from outside. Stimuli were played back using a PC (placed outside the booth), equipped with the GoPro VR Player 3.0 software, which allowed to play back video with spatial audio. The 360-degree video was presented through an Oculus Rift head-mounted display. The audio was played back through Sennheiser HD 650 headphones, driven by a HEAD acoustics LabP2 calibrated headphone amplifier. The gain of the ambisonics audio has been adjusted such that their level is as close as possible to that of the corresponding binaural audio tracks.

During the experiment, participants remained seated (seat height: 0.50m), which allowed them to freely move their head and look around in all directions but physically remained at a fixed position. The sensor for Oculus Rift was placed on a tripod (height: 1.20m), keeping approximately the same height as the participant's head position. A microphone was mounted on the tripod and was driven by a laptop, which was used to monitor the experiment from outside. When participants needed to answer questions during the experiment, they could do it by (verbal) talking and the experimenter could mark it from outside the booth. By this procedure, a holistic immersed experience was maintained throughout the full experiment.

Figure 4 – Experiment setup (*Left*: participant inside the listening booth; *Right*: view from monitoring position).

2.2.3 Procedure

Soundscape classification according to Figure 1 was achieved via a questionnaire. The questionnaire was designed to follow the hierarchical nature of the classification and with brevity in mind (Figure 5). To assess foregrounding/backgrounding of the sound within the holistic experience participants were asked: (Q3) *How much did the sound draw your attention?* To frame this question, a more general question (Q1) *In general, how would you categorize the environment you just experienced?* was added. The options for answering this question already focus attention on the more pleasurable evaluation: "*calming/tranquil*" to "*lively/active*" but with a clear option "*neither*" in between. The question distinguishing disruptive from supportive environments relates to possible activities: (Q4) *Would the sound environment prevent you from doing the activities above?* A question that again required some framing by listing possible activities in Q2 (see Figure 5). The answers to Q2 are not used and hence the choice of possible activities is not critical.

Finally, Q5 evaluates the contribution of the sonic environment as being supportive to the perception of the overall environment. This question *defines* the labels *calming* and *stimulating* as sonic environments that contribute to the *calmness/tranquility* and the *liveliness/activeness* of the place respectively.

Participants experienced the one-minute stimuli first, followed by the 5 questions presented in the VR screen with a black background (Figure 5). Participants needed to answer all 5 questions verbally. Hence also the choice for a 5-point answer scale with answering categories equidistantly spaced is in agreement with Fields et al. (2001). Note that question 5 has two versions, only one (5a or 5b) is presented to the participants. This is based on the answer in question 1: participants answering "very calming/tranquil" or "calming/tranquil" received question 5a, while participants answering one of the other choices got question 5b. After answering the questions, the next stimuli were presented. Thus, participants did not have to take off the headset between experiencing each stimulus.

The experiment was divided in 5 sections, each section contained 5 stimuli (in collection 1, one city is one section, while in collection 2, one group is one section, see Section 2.2.1). Between each section, there is a small break where participants could take the headset off. During this break, participants needed to answer additional questions regarding to the 5 stimuli they just experienced. Participants got 5 photos of the opening scenes of the stimuli in the same order as the stimuli play order. Below each photo, participants first needed to put a score on a 11-point scale (from 0: “not at all” to 10: “extremely”) on the following questions: “*How well do you remember the sound environment that goes with this picture?*” (which shows whether an environment is memorable), and “*How would you rate the sound environment of this place in terms of "full of life and exciting"/"chaotic and restless"/"calm and tranquil"/"lifeless and boring"?*”(Axelsson, 2015a), respectively. After this break, the next 5 stimuli were presented to the participants with the same procedure until all 25 stimuli (i.e. 5 sections) were evaluated.

After the participants finished the 25 stimuli, two questions regarding the overall reproduction quality were asked, specifically on the realism and immersion, using an 11-point scale. The questions presented during the break and at the end of experiment were answered on paper, thus an 11-point scale could be seen as continues scale.

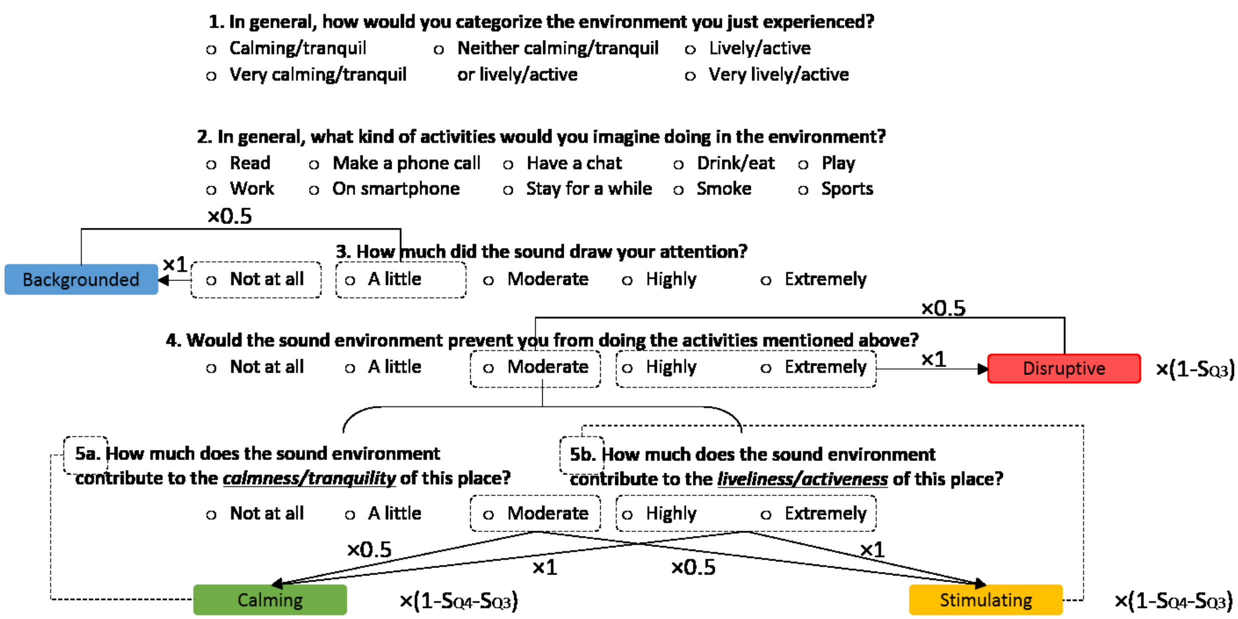

Figure 5 – List of questions asked to the participants in the classification experiment. Lines and multipliers denote the flow taken depending on the participants’ answers. Colored parts show possible outcomes of the classification.

2.2.4 **Data processing**

In this study, the fuzzy membership set of the four proposed classes *backgrounded, disruptive, calming,* and *stimulating* is based on the answers in question 3, 4, 5a and 5b, as marked in Figure 5, where $S_A(x)$ is the membership degree of soundscape *x* in the fuzzy set *A*. The fuzzy membership set, i.e. the correspondence between the answer on the question and the degree of belonging to each class, is given in Table 1.

Table 1 – The original fuzzy membership set for each class of soundscape.

Question	Answer	Fuzzy set

	Not at all	A little	Moderate	Highly	Extremely	
Question 3	1	0.5	0	0	0	$S_{backgrounded}(x)$
Question 4	0	0	0.5	1	1	$S_{disruptive}(x)$
Question 5a	0	0	0.5	1	1	$S_{calming}(x)$
Question 5b	0	0	0.5	1	1	$S_{stimulating}(x)$

To account for the hierarchical structure of the proposed classification scheme, exclusion rules should be implemented. For example, a soundscape cannot be disruptive if it is backgrounded or it cannot be supportive if it is disruptive. In mathematical form, this implies a transformation of the membership degree:

$$S'_{backgrounded} = S_{backgrounded}$$
$$S'_{disruptive} = S_{disruptive}(1 - S_{backgrounded})$$
$$S'_{calming} = S_{calming}(1 - S_{disruptive} - S_{backgrounded})$$
$$S'_{stimulating} = S_{stimulating}(1 - S_{disruptive} - S_{backgrounded})$$

where the AND and NOT operator were implemented as a probabilistic t-norm and fuzzy negation.

The membership data used in the analysis was performed after the above described mathematical transformation (i.e. all S'). The above procedure was applied to each soundscape-participant combination. For each soundscape, the average membership over all participants on the four classes was also calculated. Next to this, participants also evaluated each soundscape in terms of the 2D core affect model (“full of life and exciting”, “chaotic and restless”, “calm and tranquil” and “lifeless and boring”) on an 11-point scale during the small break in the experiment. Similarly, the average score using the 2D core affect model quadrant categories for each soundscape was also calculated.

2.2.5 **Psychoacoustical indicators and saliency**

A preliminary study (Appendix II) showed that either ambisonics or binaural recordings could be used for the reproduction. The gain of the ambisonics audio tracks has been adjusted such that their level is as close as possible to that of the corresponding binaural audio tracks. As the binaural tracks were recorded with a fully calibrated setup, the acoustical properties of the recordings are calculated on the basis of the one-minute binaural tracks using HEAD acoustics ArtemiS 8.3. The values for equivalent A-weighted sound pressure level (L_{Aeq}), percentile (L_{Axx}) and maximum sound levels (L_{AFmax}) were calculated as the energetic average of both left and right ears, whereas the values for loudness (N), sharpness (S) and corresponding percentile and maximum values were calculated as the arithmetic average between left and right ears.

Sounds that are noticed have a strong influence on the perception of soundscape (Kang et al., 2016, Terroir et al., 2013, De Coensel et al. 2009). Noticing of the sound is influenced by two interchanging processes: top-down and bottom-up attention. Top-down attention is voluntary: it assumes an active listening for the sounds occurring in the environment. On the other hand, bottom-up attention is involuntary and is influenced by the sonic environment alone. To investigate the bottom-up attention to sound, saliency as a concept is introduced. Saliency indicates how much the specific sound or a sound event stands out of its background. In consequence, the higher the saliency, the higher the probability of a sound being noticed. Although related to perception, it is possible to define the physical characteristics that contribute to saliency (Kaya and Elhilali, 2017). In this study, we used a computational model (Filipan et al., 2019) which calculates the saliency of the sound by simulating several aspects of the measured physiological response of the brain. This saliency model has two processing stages implemented: auditory periphery and brain processing. Auditory periphery simulates the initial transformation of the sound from the acoustic wave to the firing of neurons. The second stage of the model is related to the sensitivity of the human auditory cortex to spectrotemporal modulations (Santoro et al., 2017; Schönwiesner and Zatorre, 2009) that are frequently encountered in speech and biological vocalizations. This reaction is simulated by mapping the tonotopically spaced output of the periphery to both amplitude (AM) and frequency modulation (FM) space. The mapping is achieved by using resonator filters for the AM and summation of the differently delayed signals across frequency bands for the AM/FM combination space. These signals are then fed through the sensory activation stage, a part of the model that simulates defocusing of the attention (Xue et al., 2014, Krause et al. 2013) by inhibiting the excitatory input. To summarize the saliency of the sound in a single value indicator, all demodulated signals (spread over the frequency bands and AM/FM frequencies) are summed and saturated using a logarithm function. For the full overview of the saliency model used we refer to (Filipan et al., 2019).

One-minute indicators for the time-evolution of the overall saliency in this study calculated as: maximum (SL_max), average (SL_avg), median (SL_median) and 5, 10, 50, 90 and 95 percentile values (SL_xx).

2.2.6 Visual factors

The visual factors in each stimulus were also assessed, specifically the percentage of green pixels – a proxy for vegetation – and the number of people. The 50 stimuli were also labelled by the density of people appearing in the video using a qualitative 5-point scale, ranging from none to extremely dense. The proportion of person density grade in the dataset is 22%, 30%, 26%, 14%, 8% of the cases (from “none” to “extremely”), respectively.

The opening scene in each stimulus was used to calculate the green area percentage. The digital pictures consisted of 4096 × 1632 pixels and were saved in .png format. The “RGB greenness” parameter G_{RGB} (Crimmins and Crimmins, 2008; Richardson et al., 2007) is used and calculated as $G_{RGB} = (G\text{-}R) + (G\text{-}B)$, where G, R and B are the relative intensities of the green, red and blue channels in the RGB picture, respectively. A more robust assessment of green vegetation is the (broadband) normalized difference vegetation index (NDVI), however, requiring a measurement of near-infrared light. RGB greenness was shown to perform quite similar to NDVI in capturing the amount of vegetation as concluded by Richardson et al. (2007).

In a next step, an appropriate threshold was set. Note that all green is included when calculating G_{RGB}; so not only leaves from trees and bushes but also grass zones. Non-green vegetation is missed in this assessment. However, in this study, vegetation is predominantly green colored. Accidental non-vegetation green-colored objects were manually removed, typically accounting for only small zones in the photographs. Such a manual action was needed in less than 10% of the pictures. In Figure 6, examples are shown for a low, a moderate and a high vegetation percentage.

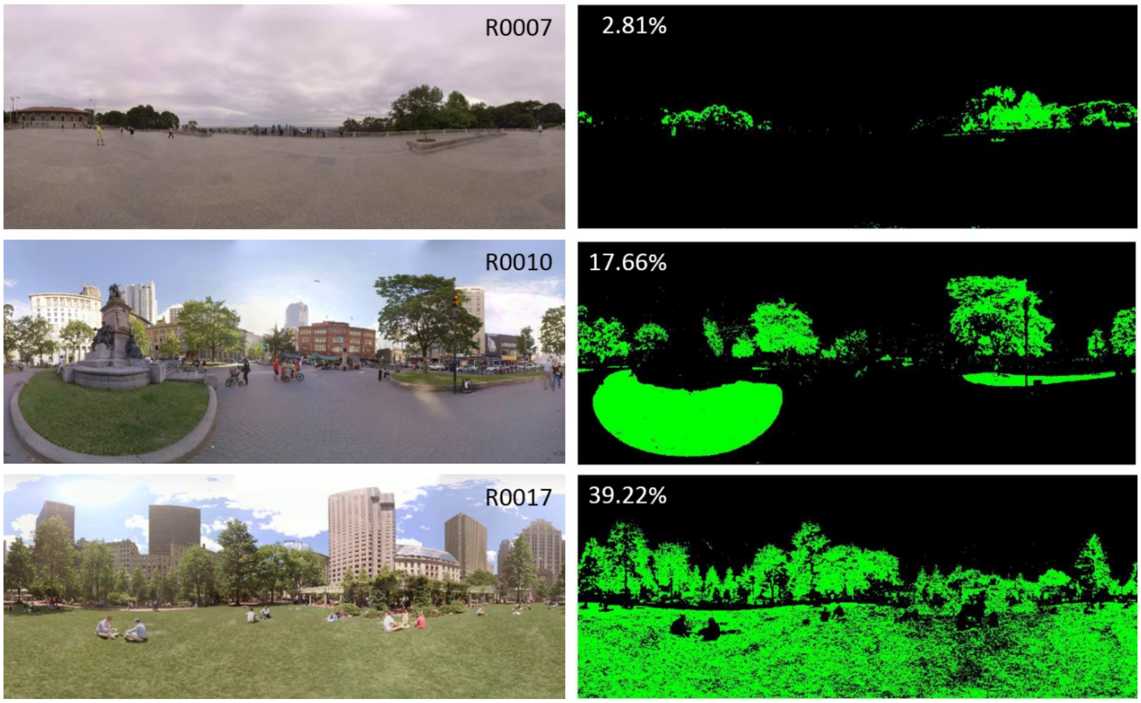

Figure 6 – Green coverage of opening scene in 360-degree videos. *Top to bottom*: low, moderate and high green percentage. (*Left*: original snapshots; *Right*: corresponding scene with pixels identified as green).

2.3 Statistical analysis

To observe relationships between the proposed soundscape categories, a principal component analysis (PCA) was performed. A PCA was also applied to the quadrant classifications in the 2D core affect model. A mixed factor generalized linear model (GLMM) fit was applied to check the relationship between memorization (question during the break, section 2.2.3) and fuzzy membership for each soundscape. Moreover, a GLMM was constructed for the four proposed categories to analyze the contribution of underlying physical parameters to the classification. The fittest model for each soundscape category was looked for, using the Akaike Information Criterion (AIC) as model quality indicator (models with smaller AIC values fit better). Finally, predicting models from collection 1 and 2 were built via linear regression, to predict the scores on four soundscape categories. A receiver operating characteristic (ROC) analysis was made to check the prediction quality. The statistical analysis in this study was conducted using the SPSS statistics software (version 25).

3. Results

3.1 Correlation between audiovisual perception and soundscape clustering

A crisp way to categorize the soundscapes is to compare the fuzzy membership to the proposed four classes. If the membership to one specific class is much larger than in the others, this soundscape is sorted in this class. Otherwise, this soundscape categorization remains unclear. Figure 8 shows the distribution of soundscapes that can be categorized into one of the four classes (i.e. 70.1% of cases), over the general audiovisual perception of the environment (answer to question 1). More specifically, *backgrounded* was found in 18% of the case, while *disruptive, calming, stimulating* was found in 18%, 14.5%, 19.6% of the cases, respectively.

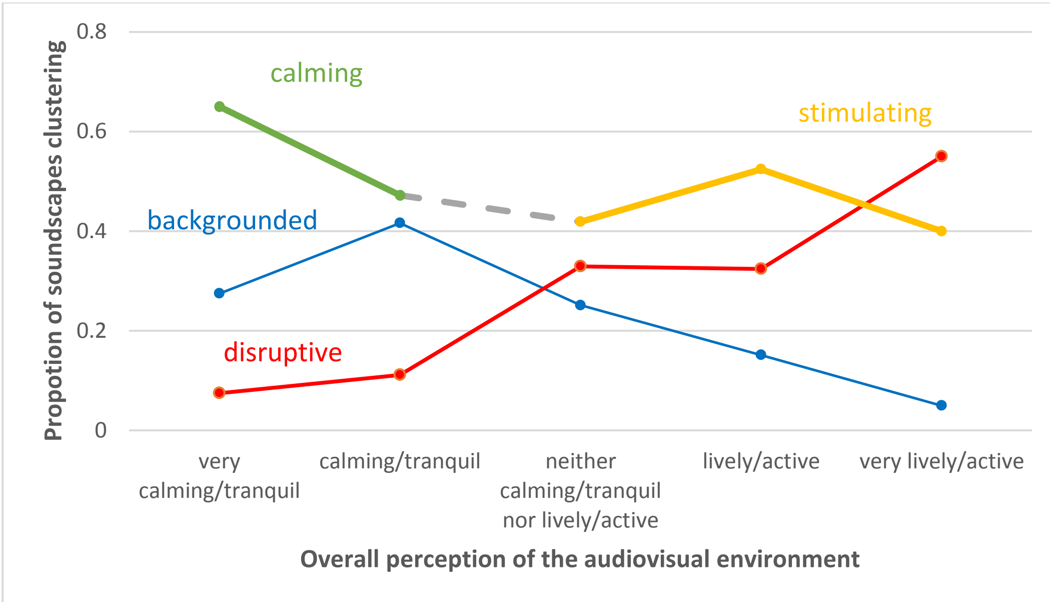

Figure 8 – Proportion of the fuzzy membership to soundscape classification category (Sections 2.2.3. and 2.2.4.) as a function of the overall perception of the audiovisual environment (Section 2.2.3.).

For the *backgrounded* category, the sound at the location does not lead to the awareness of the acoustical environment. The distribution shows that an overall "very lively/active" environment is very unlikely if the soundscape is *backgrounded* but then tends more towards a "calming/tranquil" environment. The *disruptive* category shifts the curve towards the "lively/active" side making a "very calming/tranquil" overall environment very unlikely. The supportive soundscape (*calming* and *stimulating*) pushes the curve towards the extremes in overall perception. A higher proportion of *calming* soundscapes appears in the overall perception cases of "very calming/tranquil". It is striking that for the option "very lively/active", the proportion of *disruptive* soundscapes is higher than the proportion of *stimulating* soundscapes, which might suggest that a relatively larger number of environments with a non-supportive soundscape were selected as stimuli.

3.2 Principal component analysis

In Figure 1, soundscapes are divided into *backgrounded* and foregrounded by attention causation. The foregrouded soundscapes consist of three categories, corresponding to the negative and positive effects. A principal component analysis (PCA) is applied to the average score on *disruptive*, *calming* and *stimulating* for 50 stimuli. Figure 9a shows the triangle of three foregrounded soundscape categories in the plane spanned by the two principal components. In particular, component 1 explains 71.1% of variance, while component 2 explains 22.1%.

The average score on the four proposed soundscape classifications forms a 4×50 size matrix, with values varying from 0 to 1. A threshold is set to the matrix for binary results to highlight the most pronounced 25% of the scores in the matrix. The threshold is set at 0.32, and 53 values out of 200 are greater than this threshold. It is found that 29 soundscapes clearly belong to one of the four proposed categories (*backgrounded*: 9, *disruptive*: 7, *calming*: 3, *stimulating*: 10), 12 soundscapes cover two categories and 9 soundscapes cannot be sorted into any of these categories. Figure 9a shows the distribution of 50 soundscapes in the PCA analysis, they are colored based on the binary results of the proposed classification.

As a comparison, the scores on four quadrant categories in the 2D core affect model (Axelsson et al., 2010) also forms a 4×50 size matrix. A threshold of 5.79 is set to the matrix to highlight the most pronounced 25% of the scores. 52 values out of 200 are greater than the threshold in the matrix. It is found that 28 soundscapes are determined by one of the four quadrant categories (chaotic: 6, exciting: 6, tranquil: 16, boring: 0), 12 soundscapes cover two categories and 10 soundscapes cannot be sorted into any of these categories. In Figure 9b, 50 soundscapes are colored based on the binary results in the 2D core affect model.

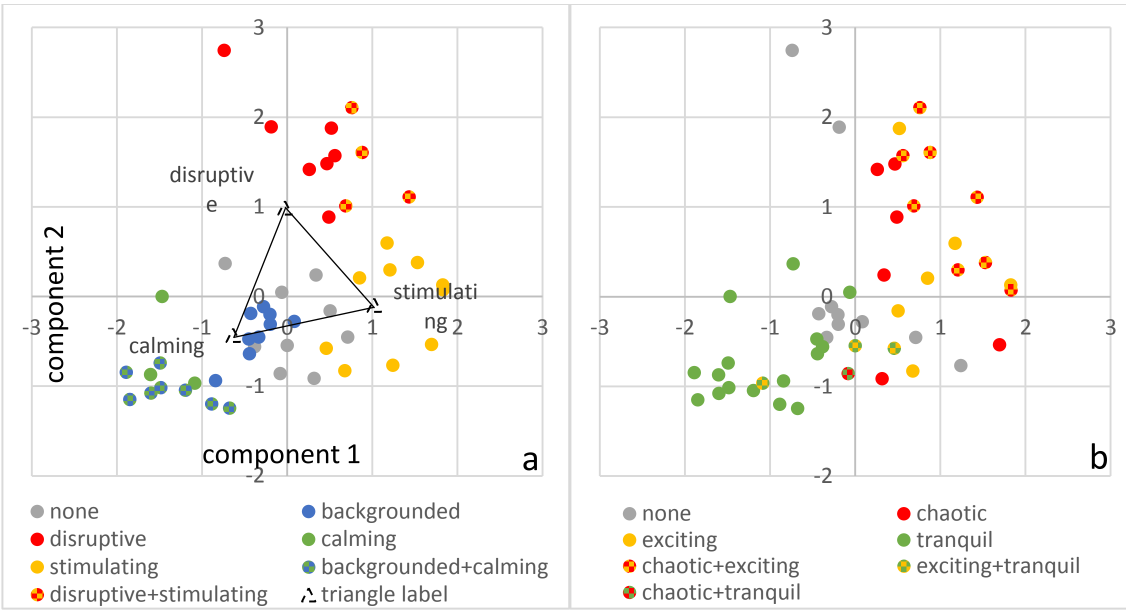

Figure 9 – Component plot based on fuzzy classification in a PCA rotated space: a) (triangle label) distribution of 50 soundscapes colored by the proposed classification; b) distribution of 50 soundscapes colored by the 2D core affect model classification (Axelsson et al., 2010).

Similarly, a PCA is also applied to the four quadrant categories in the 2D core affect model. In Figure 10a, component 1 explains 55.1% of variance, while component 2 explains 30.9%. Also, Figure 10 shows the distribution of 50 soundscapes in PCA analysis, colored by the 2D core affect model classification and the proposed classification, respectively.

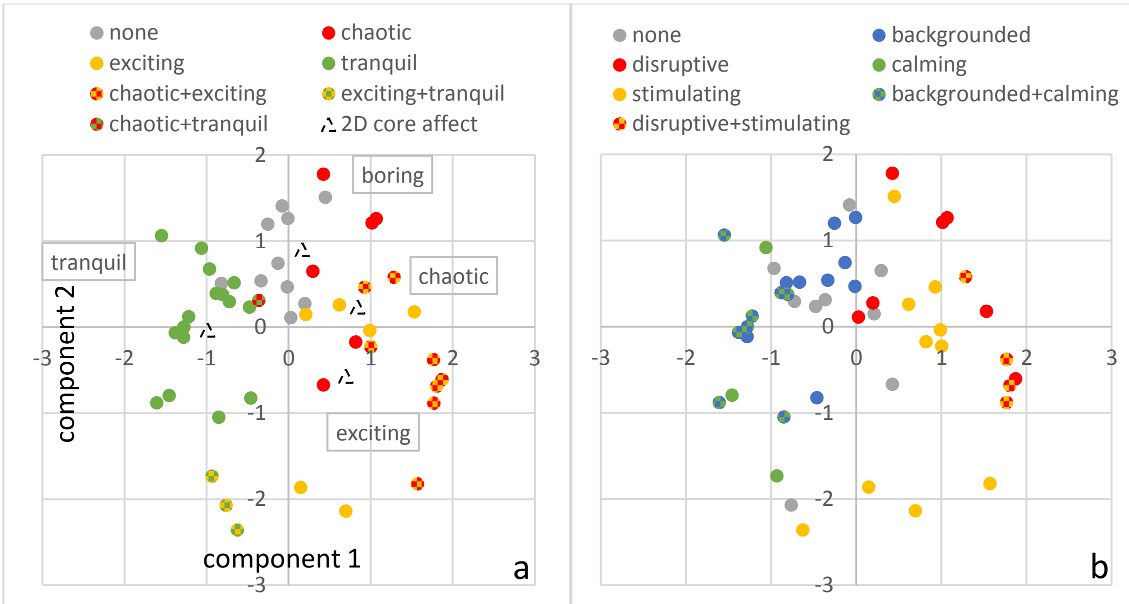

Figure 10 – Component plot based on answers to the core affect model question in a PCA rotated space: a) distribution of 50 soundscapes colored by the 2D core affect model classification (Axelsson et al., 2010); b) distribution of 50 soundscapes colored by the proposed classification.

3.3 Factor analysis

3.3.1 Relationships between soundscape class and memorization

During the small break in between experiencing 5 environments (see Section 2.2.3), a question about the memorization degree of the soundscape was asked, with the corresponding picture presented. There is a hypothesis that one tends to memorize foregrounded soundscapes better than backgrounded ones. To evaluate whether this memorization degree has a correlation with the scores on the proposed four soundscape categories, a mixed factor generalized linear model fit was applied, using participants as random factor. It is found that the memorization has significance in *backgrounded* ($F_{1,498}$=25.626; p<0.001) and *disruptive* ($F_{1,498}$=6.814; p<0.01), but not in *calming* ($F_{1,498}$=2.238; p>0.05) and *stimulating* ($F_{1,498}$=3.745; p>0.05). Naturally, the score of the *backgrounded* category has a negative correlation with memorization, while for the *disruptive* category, it is positively correlated.

3.3.2 Physical factors explaining soundscape classification

Taking into account all above-mentioned factors, a mixed factor generalized linear model fit was applied, with a stepwise method and using participant as random factor. Table 2 shows the fittest model results, with the Akaike Information Criterion (AIC) as a model quality indicator. The results suggest that the physical parameters that were tested fit the *backgrounded* category model best. All categories involve both acoustical factors and visual factors, except for the *disruptive* category. This might indicate that in a *disruptive* soundscape, the sound is dominating the perception.

Table 2 – Generalized linear mix model results of proposed soundscape categories.

glmm	AIC		F	df1	df2	coefficient	sig.
backgrounded	319.231	corrected model	48.081	5	994	0.458	0.000
		L_{A05}	55.591	1	994	-0.041	0.000
		N_{05}	30.428	1	994	0.023	0.000
		S_{max}	19.228	1	994	-0.068	0.000
		SL_median	10.011	1	994	-0.037	0.002
		Green pixels	6.827	1	994	-0.116	0.009
disruptive	511.113	corrected model	29.200	8	991	-1.432	0.000
		L_{A95}	45.799	1	991	-0.525	0.000
		L_{A90}	43.224	1	991	0.547	0.000
		SL_95	6.205	1	991	-0.035	0.013
		S_{50}	12.919	1	991	-0.480	0.000
		N_{05}	12.287	1	991	0.040	0.000
		N	5.469	1	991	-0.046	0.020
		S_{95}	6.886	1	991	0.302	0.009
		S_{05}	4.538	1	991	0.145	0.033
calming	591.150	corrected model	40.721	6	993	1.327	0.000
		L_{AFmax}	103.492	1	993	-0.020	0.000
		Person density	12.645	4	993	(=1)0.172 (=2)0.024 (=3)0.003 (=4)-0.057 (=5)0*	0.000
		S_{50}	22.805	1	993	0.106	0.000
stimulating	535.742	corrected model	40.829	5	994	0.755	0.000
		Person density	16.435	4	994	(=1)-0.196 (=2)-0.077 (=3)-0.064 (=4)0.091 (=5)0*	0.000
		SL_median	39.724	1	994	0.067	0.000

*:This coefficient is set to 0 because it is redundant.

3.4 Soundscape classification prediction

The previous section explored the factors that could modify the membership set of the proposed four categories. As stated before, an important challenge is to create models based on acoustical parameters that predict soundscape classification as accurately as possible within the context of the definition of soundscape. For this purpose, collection 1 and collection 2 (Table IV in Appendix III) that were conducted with two groups of totally different stimuli, and applied to two groups of different participants, will be treated as two independent data sets. Each soundscape gets an average membership score for each of the proposed soundscape classes. We will investigate whether a model based on physical parameters that is extracted from one of the classifications can predict this membership score for the other classification.

3.4.1 **Prediction models from collection 1**

A linear regression on 25 stimuli in collection 1 is applied, using a stepwise approach to access all possible acoustical parameters. Table 3 shows the remaining predictors, as well as the detailed model for each class membership.

Table 3 –Linear regression models for 25 stimuli in collection 1.

label	Soundscape category	R^2	SE	prediction equation – from collection 1	predictors	sig.
1-1	backgrounded	0.546	0.100	$y=-0.017x+1.393$	$x=L_{A05}$	0.000
1-2	disruptive	0.719	0.095	$y=0.029x_1-0.014x_2-0.922$	$x_1=L_{A05}$, $x_2=L_{A95}$	L_{A05}(0.000) L_{A95}(0.006)
1-3	calming	0.606	0.129	$y=-0.023x+1.936$	$x=L_{AFmax}$	L_{AFmax}(0.000)
1-4	stimulating	0.667	0.100	$y=0.105x+0.722$	x=SL_95	SL_95(0.001)

SE: Std. Error of the Estimate.

When applying the equations in Table 3, it is easy to get the predicted scores of proposed soundscape categories for 25 stimuli in collection 2. To compare this prediction with the experimental value in collection 2, a receiver operating characteristic (ROC) analysis is applied. Figure 11 shows the ROC curve of the prediction, referring the experimental binary results of collection 2 as criterion. The parameter in this ROC curve is the threshold for crisp classification. Table 4 further shows the detailed results of the model prediction quality.

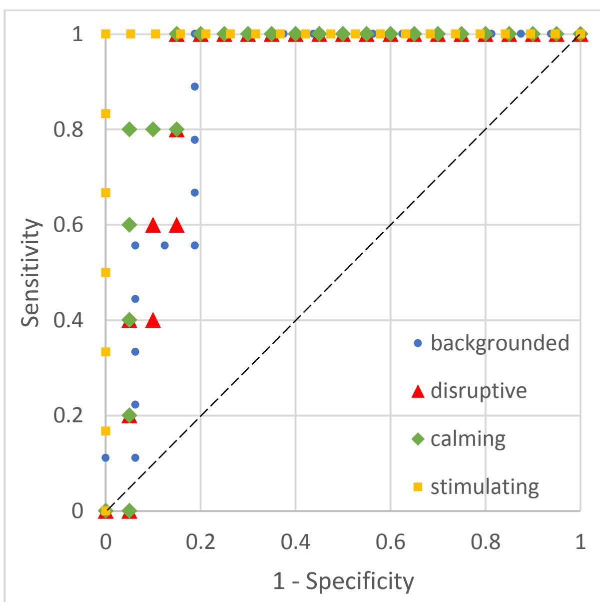

Figure 11 – Receiver operating characteristic (ROC) curve of prediction models for 25 stimuli in collection 1.

Table 4 – The ROC curve area analysis for prediction models from collection 1.

	Area Under the Curve				
	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
backgrounded	0.889	0.068	0.002	0.755	1.000
disruptive	0.900	0.063	0.007	0.777	1.000
calming	0.930	0.054	0.003	0.824	1.000
stimulating	1.000	0.000	0.000	1.000	1.000

a. Under the nonparametric assumption.
b. Null hypothesis: true area = 0.5.

As shown in Figure 11 and Table 4, the ROC curve shows the numeric results of the predictions. The Youden index (J) is often used as a criterion for selecting the optimum cut-off point (Schisterman et al., 2005). The Youden index is defined as shown in Eq. 1, and it ranges from -1 to 1. A higher value for J represents a lower proportion of totally misclassified results, i.e. a better prediction. Table 5 shows the maximum J value and its corresponding threshold.

$$J = sensitivity + specificity - 1 \qquad \text{(Eq. 1)}$$

Table 5 – Maximum Youden index for prediction models from collection 1.

label	soundscape category	Highest J	Recommended threshold	Accuracy
1-1	backgrounded	0.812	0.3101	0.88
1-2	disruptive	0.85	0.1592	0.88
1-3	calming	0.85	0.4659	0.88
1-4	stimulating	1	0.1916	1

3.4.2 **Prediction models from collection 2**

Vice versa, the same procedure applies to collection 2. Table 6 shows the results of linear regression (stepwise) applied to collection 2 and the model details for each category. The prediction for 25 stimuli in collection 1 is compared with the binary results of the experimental value in collection 1, using ROC analysis (Figure 12). Table 7 further shows the detailed results of the prediction quality. Similarly, Table 8 shows the maximum J value and the corresponding threshold for predictions from collection 2.

Table 6 – Linear regression models for 25 stimuli in collection 2.

label	Soundscape category	R^2	SE	prediction equation – from collection 2	predictors	sig.
2-1	backgrounded	0.603	0.113	y=-0.026x+1.894	$x=L_{A05}$	0.000
2-2	disruptive	0.360	0.148	y=0.020x-1.111	$x=L_{A05}$	0.002
2-3	calming	0.512	0.138	$y=-0.028x_1+1.161x_2+1.76$	$x_1=L_{AFmax}$, $x_2=S_{50}$	L_{AFmax}(0.000) S_{50}(0.027)
2-4	stimulating	0.663	0.090	y=0.023x-1.221	$x=L_{A10}$	L_{A10}(0.001)

SE: Std. Error of the Estimate

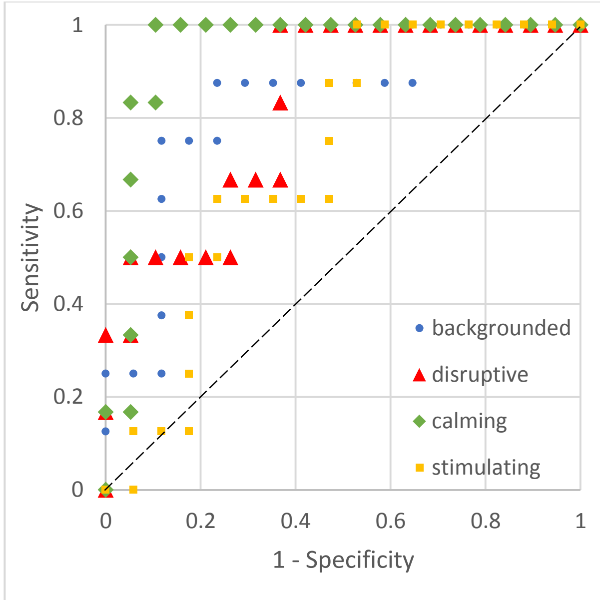

Figure 12 – Receiver operating characteristic (ROC) curve of prediction models for 25 stimuli in collection 2.

Table 7 – The ROC curve area analysis for prediction models from collection 2.

	Area Under the Curve				
	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
backgrounded	0.831	0.09	0.009	0.655	1.000
disruptive	0.825	0.089	0.019	0.65	0.999
calming	0.947	0.046	0.001	0.857	1.000
stimulating	0.713	0.103	0.091	0.511	0.915

a. Under the nonparametric assumption.
b. Null hypothesis: true area = 0.5.

Table 8 – Maximum Youden index for prediction models from collection 2.

label	Soundscape category	Highest J	Recommended threshold:	Accuracy
2-1	backgrounded	0.64	0.107	0.8
2-2	disruptive	0.632	0.2644	0.72
2-3	calming	0.895	0.1184	0.92
2-4	stimulating	0.471	0.3037	0.64

3.4.3 **Prediction quality comparison**

Taking the recommended threshold, the numeric result is transferred into a dichotomous result. As stated before, the experimental binary results are used as criterion. In the ROC analysis, the accuracy ($\frac{true\ positive+true\ negative}{total\ sample}$) is indicating the proportion of total correctly classified results. Tables 6 and 9 show the accuracy of each prediction taking the recommended threshold, respectively. They indicate that it is better to predict *backgrounded* soundscape with 1-1, and for *disruptive* and *stimulating* soundscape, 1-2 and 1-4 predicts better. Whereas for predicting a *calming* soundscape, 2-3 is clearly better. Another way to detect the quality of the predictions is considering the true positive to false positive rate (TPR to FPR). As shown in Figure 13, a smaller distance between prediction dots and point (0,1) indicates a higher prediction quality. The relative distance also indicates that for the proposed four categories, model 1-1, 1-2, 2-3 and 1-4 are optimized choices.

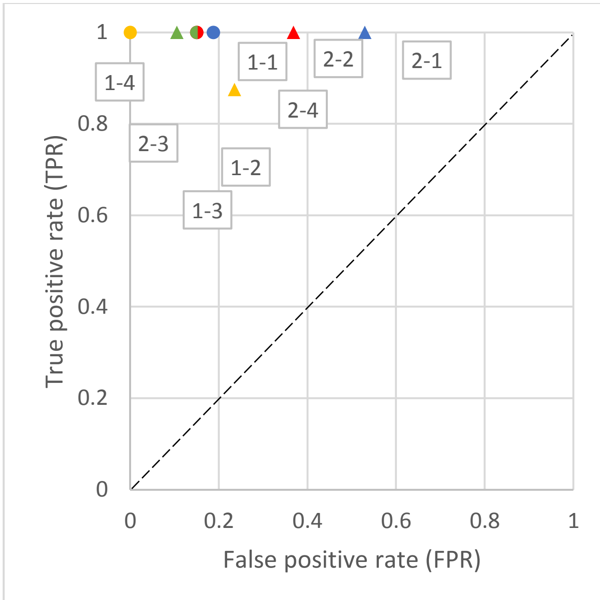

Figure 13– The Receiver operating characteristic (ROC) space with points of eight predictions at the thresholds recommended by the maximum Youden Index (Table 6 and Table 9).

3.4.4 **Models from all 50 stimuli**

Based on the above comparison, a better model is selected for each category (model 1-1, 1-2, 2-3, 1-4). Table 9 gives the models that are built on the dataset of all 50 stimuli, with the indicators obtained from the optimized models built on the subgroups that best extrapolated to an independent dataset. Within this study, we cannot test this model with other recordings as verification. However, it can serve as a guideline once the new recordings and new subjective assessment are done.

Table 9 – Linear regression models for all 50 stimuli used in the experiments.

label	Soundscape category	R^2	SE	prediction equation – from all 50 stimuli	predictors	sig.
0-1	backgrounded	0.521	0.112	$y=-0.018x+1.464$	$x=L_{A05}$	0.000
0-2	disruptive	0.488	0.128	$y=0.027x_1-0.015x_2-0.733$	$x_1=L_{A05}$, $x_2=L_{A95}$	$L_{A05}(0.000)$ $L_{A95}(0.006)$
0-3	calming	0.426	0.150	$y=-0.020x_1+0.079x_2+1.440$	$x_1=L_{AFmax}$, $x_2=S_{50}$	$L_{AFmax}(0.000)$ $S_{50}(0.098)$
0-4	stimulating	0.501	0.114	$y=0.078x+0.643$	x=SL_95	SL_95(0.000)

SE: Std. Error of the Estimate.

4. Discussion

4.1 Backgrounded soundscapes

A *backgrounded* soundscape is defined as one that does not contribute to the overall experience of the place. Thus, it is fair to assume that this class of soundscapes does not catch a lot of attention. If not heard, such a soundscape will neither leave an impression in memory which is supported by a significant negative correlation with memorization (Section 3.4.1). Earlier research (Axelsson, 2015b) categorized one outdoor space type as "my space", where crowds and mechanical sounds should be inaudible and sounds of nature and individuals should be only moderately audible. This supports the idea that *backgrounded* soundscapes are appropriate for “my space”. The distribution of this soundscape over general perception of environments shown in Figure 8, shows a trend towards an overall “calming/tranquil” perception of the environment. This reveals that a *backgrounded* soundscape is not often found in a lively and active environment. As the *backgrounded* soundscape does not attract attention, it covers a separate dimension and hence it was not included in the PCA (Section 3.3). In Figures 9&10, the stimuli labelled as *backgrounded* in the proposed classification scheme were labelled as “none” in 2D core affect model, i.e. not dominated by any category. This might be explained by the fact that a *backgrounded* soundscape could be allocated by all emotional components. It has been argued that a representative soundscape for the “lifeless and boring” label in the 2D core affect model seems rare (Axelsson, 2009; Bahalı and Tamer-Bayazıt, 2017), which is also the case in this study (Figure 10a). However, some *backgrounded* stimuli are located close to the “lifeless and boring” label in Figure 10b which might suggest that a “lifeless and boring” soundscape does not attract attention. Hence in an experiment that focusses attention on sound, either sonic environments that could lead to such a soundscape are not included or explicit foregrounding changes people’s perception. Note that this does not suggest that the *backgrounded* and “lifeless and boring” are completely overlapping since the two classifications are from different domains.

The generalized linearized model for individual soundscape classification with progressive inclusion of significant physical parameters shows that also visual factors contribute to the soundscape being *backgrounded*. Visible green reduced the chance for a soundscape to become labelled as *backgrounded*. This is consistent with previous work highlighting the importance of visual factors in the construct of annoyance at home – the place where *backgrounded* soundscapes may be most appropriate (Gidlöf-Gunnarsson and Öhrström, 2007; Van Renterghem and Botteldooren, 2016). While comparing the fittest model for each soundscape category (Table 3), it seems that physical parameters built the best model for *backgrounded* (with lowest AIC compared to other categories), thus it seems easier to predict on the basis of physics when the sound environment will not be noticed.

The stable model for predicting *backgrounded* soundscapes only retains L_{A5} as an acoustical indicator. To be *backgrounded*, sonic environments should simply not contain any loud sounds whatever their origin and duration. Focusing on the highest level using low percentile statistical indicators (or an equivalent level) is consistent with models for annoyance at home and the above observation that *backgrounded* soundscapes might be most appropriate for the environmental contribution to the private dwelling.

4.2 Disruptive soundscapes

Disruptive soundscapes are defined as sonic environments that prevent the users of the space from doing activities they would otherwise engage in. This conceptual soundscape relates very strongly to affordance and activity appropriateness as proposed in Nielbo et al. (2013) and Andringa and Van Den Bosch (2013). It is, to a certain extent, also aligned with the concept of “appropriateness”, which has been suggested as key determinant of soundscape evaluation (Axelsson, 2015a).

Among all three foregrounded categories, *disruptive* is the only one that significantly correlates to memorization (Section 3.4.1), suggesting that such a soundscape leaves a strong – albeit negative – impression. The distribution of *disruptive* soundscapes over categories of overall appreciation of the environment shows an increasing trend towards "lively/active" and neutral evaluation (Figure 8). A straightforward interpretation is that *disruptive* soundscapes prevent the overall environment to be "calming/tranquil", yet it could be compatible with an environment that is neither calming nor lively or even with a "lively/active" environment. Soundscapes in this category tend to be loud, accompanied by a high density of people (Supplement 2).

It seems that *disruptive* is close to "chaotic and restless" in the 2D core affect model from the description, as well as certain overlaps in binary results of stimuli (Figure 9&10). In the PCA (Figure 9a), *disruptive* determined soundscapes are concentrated in the upper part of the triangle, while two outliers are slightly deviated to the negative axes of component 1. When analyzing these two outliers (R0013 & R0029), a shared trait was found: both stimuli contain a (visually) peaceful park, there are nearly no human activities and the weather is nice. In R0029, a honk from a boat appears all of a sudden. In R0013, a sustained noise from a lawnmower (not visible) appears in the background. These unexpected occurrences trigger some participants to report a disturbance while others chose to ignore these two stimuli and focus on the calming aspects of the soundscape. These two stimuli were labelled as "none" in the PCA analysis based on the 2D core affect model (Figure 9b).

The generalized linear model combines many non-orthogonal factors to predict the *disruptive* category but does not contain visual factors in the fittest model (Table 3). The dominance of sound in such a case is in line with many studies dealing with the perception of "unpleasant" soundscapes (Guastavino, 2006; Davies et al., 2013). Moreover, *disruptive* leads to the best prediction model among the three foregrounded categories (Table 3, AIC), which supports the use of the disruptive-supportive subdivision as second stage division (Figure 1).

Finally, looking at the predictive models for average soundscape classification (see also Section 3.5), additional insight in this category of soundscape can be obtained. The predictive models contain L_{A5} and L_{A95} as acoustic descriptors, or looking in more detail at the signs and magnitude of the coefficients, L_{A5} and L_{A5}-L_{A95}, both with a positive trend. This indicates that in addition to the sound level – measured here as L_{A5} – that also appears in the classification of *backgrounded*, the temporal variability of the sound – measured here as L_{A5}-L_{A95} – is important for the soundscape to become disruptive. Previous work has suggested the importance of the latter difference or a similar indicator of fluctuation, sometimes referred to as *emergence*, for predicting the pleasantness of public place soundscapes (Nilsson et al., 2007; Liu and Kang, 2015), as well as for annoyance at home (Bockstael et al., 2011), but never found such strong effects.

4.3 Calming soundscapes

Supportive soundscapes are expected to contribute to the overall experience of a place. They should match expectations created by the context and purpose of the place. In a design phase the type of support expected could be put forward by the urban designer. In this study the type of support one may expect, *calming* or *stimulating*, is mainly evoked by visual information. Therefore, in the procedure (Figure 5), questions 5a and 5b were only asked based on the answer in question 1 (i.e. when the overall perception is "calming/tranquil", it is assumed the soundscape would support the "calming/tranquil" atmosphere). If a not very "calming/tranquil" soundscape appears in an overall "calming/tranquil" environment, the fuzzy scores will only give a lower score for *calming*, rather than categorizing the soundscape as *stimulating*. Thus, *calming* and *stimulating* are not opposites of each other. Because of this construction, the combined distribution of *calming* and *stimulating* soundscapes over overall perception (Figure 8) is not very informative, but at least shows a somewhat stronger importance of the soundscape in “very calming/tranquil” environments.

Stimuli identified as “calm and tranquil” in the 2D core affect model also appear in the *calming* region of the PCA based on the proposed classification (Figure 9) and vice versa (Figure 10). This is not surprising as the distinction between the *calming* and *stimulating* type of supportive environments is mainly in the arousal dimension of core affect. In addition, the pleasantness dimension seems to bare some resemblance with not being disruptive. It is also found that the *calming* category is close to *backgrounded*, as 8 stimuli out 12 were identified as belonging to these two categories (Figure 9a). One possible explanation, focusing on attention, is that as the stimuli in *calming* soundscapes lead to passive attention fading (Bradley, 2009). This shifts the perception towards *backgrounded*. This vacillates the soundscape perception along the attention causation, which makes it stringent to label a soundscape as *calming*. However, despite the crossover between *calming* and *backgrounded*, these two categories are still different. Firstly, *calming* soundscapes make the overall environment being perceived as “calm and tranquil” and “very calm and tranquil” (Figure 8). Secondly, the percentage of (visual) vegetation is not a significant factor for explaining *calming* soundscapes (Table 3 and Supplement 2). As for visual factors, a vegetation-dominated view is not a prerequisite for the soundscape to be classified as *calming* yet the visual presence of people plays a key role: too many people reduce the calmness of the soundscape. Sharpness (S_{50}) and the absence of strong peaks (L_{AFmax}) appear both in the explorative GLM and the predictive models. Sharpness is typically higher for natural sounds and lower for mechanical ones (Boes et al., 2018). A lot of research confirmed the positive effect of e.g. natural sounds (Payne, 2013, Van Renterghem, 2018) and the negative effect of mechanical sound (Bijsterveld, 2008).

4.4 Stimulating soundscapes

Finally, the *simulating* category is defined by the questionnaire as a soundscape that supports the liveliness and activeness of the environment. It is expected to arouse people, to encourage them to get involved. Music or music-like sound, for instance, could achieve such an effect (Botteldooren et al., 2006; Raimbault and Dubois, 2005), which was also found in some stimuli in this study (e.g., R0010, R0058, etc.). This type of soundscape helps the whole environment to be perceived as “lively/active” (Figure 8). However, compared to *disruptive*, a rather lower proportion of *stimulating* appears in an overall “very lively/active” perception. This might suggest that environments with such soundscapes attract people’s attention but is slightly more likely to cause activity interference. Given a closer look at the 4 stimuli that are crossing these two categories (Figure 9a), all of them contain a lot of people, so some people may judge this crowd disturbing for their envisaged activities. When putting *stimulating* soundscapes in the PCA plane of the 2D core affect model, they lay in between “chaotic and restless” and “full of life and exciting” (Figure 10a). As defined in the proposed classification, this category supports the liveliness and activeness of the environment. The GLM suggests that the presence of people is necessary (Table 3). It is consistent with previous research (van den Bosch et al., 2018; Aletta and Kang, 2018), which suggests that human sounds add to the eventfulness of a soundscape and the perceived audible safety. It is worth noting that only when the visual person density is high, this category seems to be favored while lower person densities tend to favor *calming* soundscapes.

Finally, both the explanatory GLM and the predictive models (See also Section 3.5) for *stimulating* soundscapes contain the continuous fraction of saliency. Saliency, as defined in the model based on amplitude and frequency modulations, focusses strongly on vocalisations. Hence it is also indicative of the presence of human sounds. Previous work showed that the second order time derivative of the level in the 500 Hz octave band – which is also an indicator for amplitude fluctuations – correlates well with the presence of human voices (Aumond et al., 2017).

4.5 The soundscape classification approach

This study proposed a holistic soundscape classification method as a labeling tool for audio-visual collections. This classification is not expected to be covering all details and further taxonomy could be used. The proposed classification is based on the contribution of the soundscape to the overall environmental perception.

This classification scheme recognizes that, in context, environmental sounds may remain backgrounded and that only sonic environments containing foregrounded elements may significantly contribute to the overall experience of the urban environment. Thus the *backgrounded* class is introduced as an orthogonal dimension. A good classification of the remaining foregrounded soundscapes: *disruptive*, *calming* and *stimulating* should be minimally overlapping and therefore form a triangle in the principle component space. This was proven to be indeed the case. Moreover, although the classes slightly overlap and soundscapes may have a finite fuzzy membership to multiple classes at the same time, a tendency for good separation is indeed visible (Figure 9a). Recent research (Kamenický, 2018) also uses a triangle (activities, mechanisms and presence) for classification, which suggests a spectrum evolution of soundscapes in between the extremes. The evolution between soundscape categories is also embodied by the stimuli crossing two categories. It suggests that the soundscape perception is fluid and could be modified by time, person and context (Maris et al., 2007; Sun et al., 2018b).

The proposed classification is compared to the popular classification in a 2D core affect plane. There are some obvious similarities between both classifications yet in the plane of the first two principle components classes, the latter seems less separated. This could be because another dimension is sampled and the core affect classification is richer, but as the variance explained by the first two components is even higher than for the proposed classification, this does not seem the case. This might suggest that in a given soundscape (with fixed physical parameters), detecting attention causation is easier than classifying emotion perception. It highlights the importance of involving attention causation in soundscape classification. None of the observed soundscapes is dominantly "boring" as observed above, which argues in favor of eliminating this dimension. It should be noted however that in this study, the data for the proposed classification were collected right after each stimulus, while the data of the 2D core affect model were collected afterwards (Section 2.2.3). This might introduce the deviation of acoustical memory in perception (Darwin and Baddeley, 1974). However, no significant correlation was found between memorization and any of the four categories in the 2D core affect model.

Understanding the soundscape needs to isolate it from the whole environment that contains more than the sonic environment, but it is also important to use the whole environment as a guideline to classify the soundscape. Visual context, specifically two items in this study (Supplement 2), were found significant in both whole environment perception and the crisp clustering, though the latter represents 70.1% of the variance (Section 3.2). This is not the case in some of proposed categories. For example, for *disruptive*, the visual factors do not influence significantly. On the other hand, the soundscape also modifies the overall perception (e.g., two outliers in *disruptive* category).

Although soundscape implies perception in context, a classification of sonic environments with soundscape in mind should benefit from capturing common understanding by society rather than personal preferences. Hence the proposed classification avoided the pleasantness dimension in affect which is expected to be more individual than the arousal dimension. If this attempt to remove individual differences from the classification was successful, it should be possible to construct predictive models solely based on physical parameters. This will be shown in the next Section.

4.6 Prediction models

The main goal of building prediction models is labelling new audio-visual recordings in the collection without the use of a panel. As the main application of the collection is to provide representative exemplars for each category, the prediction models do not need the refinement to resolve ambiguous situations and therefore could be based on a limited database of 50 samples. Another goal of building a model purely based on acoustical parameters could be to construct “soundscape maps”. Also for this application simple models are preferred. Of course other modelling options are available (Yu and Kang, 2015; Hong and Jeon, 2015), but this approach adds to the literature for explained reasons.

Thus, in this study, models predicting soundscape classification with a limited number of acoustical parameters were considered. The strongest possible model validation was assured by confirming model performance on the outcome of independent experiments. The linear models produce a membership degree for each of the four classes. Model comparison is done on sharp, binary classifications. The choice of threshold allows to balance between the risk of obtaining false positives and false negatives.

For model validation, the recommended threshold is based on the Youden Index which selects an optimal balance between sensitivity and specificity. This results in most crisp classification models combine the highest possible specificity with the highest possible sensitivity and appear in the upper left corner of Figure 13 (7 out of 8 dots). The recommended threshold for each model (Table 6&10), is lower than the value used to crisply classify the experimental results (0.32). This causes more than 25% data to be classified and therefore the model approach is less critical than the experimental approach. This may lead to false classification but it ensures that all possible example in each category are selected. Because it includes some soundscapes into one category unnecessarily, it might need additional panel tests to purify the selected soundscapes.

An alternative way to select the threshold is to push the outcome to maximal specificity (i.e. minimal FPR component). This method ensures that all automatically selected soundscapes are representative exemplars of a certain category, but it faces the fact that some soundscapes that could be a representative of a certain category, will be filtered out. As more audiovisual recordings are thus thrown out of the classification, this increases the work of site recording as a bigger collection is needed to start from. Thus, both methods for selecting the threshold have advantages and drawbacks. The choice depends on whether panel tests costs more than site recording or the other way around.

Besides the comparison between the models built on subgroups, Table 10 gives the models from the data of all 50 stimuli. Based on this study, they cannot be rigorously bilaterally verified. However, model parameter selection from the best models for the two subgroups are used without adding new parameters, which should reduce the risk of overfitting on the pooled data. Coefficients are nevertheless optimized for the pooled data. The models of Table 10 are therefore our suggestions for best available models.

4.7 Limitations

Although using audio-visual reproduction through virtual reality is a huge improvement over older methods to experience sonic environments in context, it still lacks other sensory context: odor, heat and humidity, etc. And, although the 360-degree visual scenery is a very strong cue for setting the context, it does not contain all information about a place, its use, its socio-cultural meaning, etc. The selection procedure for collecting the audio-visual recordings in each city was rather stringent and recordings from cities in different continents were included. Nevertheless, there might be some sampling bias: due to practical considerations, more recordings were made in less crowded environments like parks than in crowded places like shopping streets.

Additional indicators and alternative machine learning techniques could have been used while constructing prediction models. E.g. regarding visual factors, only two items were assessed, although many other aspects were shown to have an impact on soundscape perception (such as sound source visibility, number of vehicles, etc.). The database is open and will be extended in the future, allowing to test more hypotheses.

5. Conclusions

This study proposes a hierarchical soundscape classification methodology that is grounded in attention causation and reflects the contribution of the soundscape to the overall perception of the environment. The methodology is made operational through a brief questionnaire. The proposed hierarchical classification scheme offers an alternative to the 2D core affect model, and is based on how well the soundscape is noticed, how it interferes with possible activities at the site, and includes the overall appreciation of the environment. It (1) accounts for the existence of *backgrounded* soundscapes that do not catch attention; (2) forms a clear triangular construct between *disruptive*, *calming* and *stimulating*, which offers a clear separation of soundscape categories; (3) explores the multiple factors that might modify the four categories, both in terms of acoustics and vision. Finally, a set of models based on acoustical parameters is built to predict the partial membership to the proposed soundscape categories, which might be used to classify soundscapes without involving participants. It has a high proportion of correctly classified soundscapes, validated by verification on a completely independent dataset (other participants and other soundscapes). By using the proposed soundscape classification methodology, it is at least possible to identify the most pronounced examples in each category.

The methodology is developed with the classification of a repository of audiovisual recordings from around the world in mind, yet it could be applied in other application domains. It is tested on an ecologically valid, realistic and immersive soundscape reproduction system to be applied in a laboratory. This holistic method includes soundscape collection, on-site recordings and final playback.

Within the framework of the funded project, more soundscape recordings will gradually be added into the database. It is hoped that, together, this ecologically valid reproduction system and the models that automatically classify soundscapes as the recordings enter the database will allow building a growing international collection. This will offer urban planners the most interesting exemplars worldwide for each type of soundscape, inspiring and guiding future urban sound planning and design.

References

Aletta F, Kang J, Axelsson Ö. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape and Urban Planning, 149, 65-74.

Aletta F, Kang J. (2018). Towards an Urban Vibrancy Model: A Soundscape Approach. International Journal of Environmental Research and Public Health, 15(8), 1712.

Andringa TC, Van Den Bosch KA. (2013). Core effect and soundscape assessment: Fore-and background soundscape design for quality of life. In INTER-NOISE and NOISE-CON congress and conference proceedings (Vol. 247, No. 6, pp. 2273-2282). Institute of Noise Control Engineering.

Aumond P, Can A, De Coensel B, Botteldooren D, Ribeiro C, Lavandier C. (2017). Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. Acta Acustica united with Acustica, 103(3), 430-443.

Axelsson Ö. (2009). May information load be the key dimension underlying soundscape perception?. In Inter-Noise 2009. Institute of Noise Control Engineering of the USA.

Axelsson Ö. (2015a). How to measure soundscape quality. In Proceedings of the Euronoise 2015 conference, Maastricht, The Netherlands. pp. 1477-1481.

Axelsson Ö. (2015b). Towards guidelines for soundscape design. In AESOP Prague Annual Congress 2015: Definite Space–Fuzzy Responsibility, Prague, Czech Republic. pp. 802-808.

Axelsson Ö, Nilsson ME, Berglund B. (2010). A principal components model of soundscape perception. The Journal of the Acoustical Society of America, 128(5), 2836-2846.

Bahalı S, Tamer-Bayazıt N. (2017). Soundscape research on the Gezi Park–Tunel Square route. Applied Acoustics, 116, 260-270.

Berglund B, Nilsson ME. (2006). On a tool for measuring soundscape quality in urban residential areas. Acta Acustica united with Acustica, 92(6), 938-944.

Bijsterveld K. (2008). Mechanical sound: Technology, culture, and public problems of noise in the twentieth century. MIT press.

Bockstael A, De Coensel B, Lercher P, Botteldooren D. (2011). Influence of temporal structure of the sonic environment on annoyance. In 10th International Congress on Noise as a Public Health Problem (ICBEN-2011) (Vol. 33, pp. 945-952). Institute of Acoustics.

Boes M, Filipan K, De Coensel B, Botteldooren D. (2018). Machine Listening for Park Soundscape Quality Assessment. Acta Acustica united with Acustica, 104(1), 121-130.

Botteldooren D, Andringa T, Aspuru I, Brown AL, Dubois D, Guastavino C, Kang J, Lavandier C, Nilsson M, Preis A, Schulte-Fortkamp B. (2015). From sonic environment to soundscape. Soundscape and the Built Environment, 36, 17-42.

Botteldooren D, De Coensel B, De Muer T. (2006). The temporal structure of urban soundscapes. Journal of sound and vibration, 292(1-2), 105-123.

Bradley MM. (2009). Natural selective attention: Orienting and emotion. Psychophysiology, 46(1), 1-11.

Brown AL. (2012). A review of progress in soundscapes and an approach to soundscape planning. International Journal of Acoustics and Vibration, 17(2),73-81.

Brown AL, Kang J, Gjestland T. (2011). Towards standardization in soundscape preference assessment. Applied Acoustics, 72(6), 387-392.

Cain R, Jennings P, Poxon J. (2013). The development and application of the emotional dimensions of a soundscape. Applied Acoustics, 74, 232-239.

Crimmins MA, Crimmins TM. (2008). Monitoring plant phenology using digital repeat photography. Environmental Management, 41, 949-958.

Darwin CJ, Baddeley AD. (1974). Acoustic memory and the perception of speech. Cognitive Psychology, 6(1), 41-60.

Davies WJ, Adams MD, Bruce NS, Cain R, Carlyle A, Cusack P, Hall DA, Hume KI, Irwin A, Jennings P, Marselle M. (2013). Perception of soundscapes: An interdisciplinary approach. Applied Acoustics, 74(2), 224-231.

De Coensel B, Botteldooren D, De Muer T, Berglund B, Nilsson ME, Lercher P. (2009). A model for the perception of environmental sound based on notice-events. The Journal of the Acoustical Society of America, 126(2), 656-665.

De Coensel B, Sun K, Botteldooren D. (2017). Urban Soundscapes of the World: selection and reproduction of urban acoustic environments with soundscape in mind. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 255(2), 5407-5413. Institute of Noise Control Engineering.

Dubois D, Guastavino C, Raimbault M. (2006). A cognitive approach to soundscapes: using verbal data to access auditory categories. Acta Acust United Acust, 92(6), 865-874.

Echevarria Sanchez GM, Van Renterghem T, Sun K, De Coensel B, Botteldooren D. (2017). Using Virtual Reality for assessing the role of noise in the audio-visual design of an urban public space. Landscape and Urban Planning, 167, 98-107.

Ekkekakis P. (2008). Affect circumplex redux: the discussion on its utility as a measurement framework in exercise psychology continues. International Review of Sport and Exercise Psychology, 1(2), 139-159.

Fields JM, De Jong RG, Gjestland T, Flindell IH, Job RFS, Kurra S, Lercher P, Vallet M, Yano T, Guski R, Felscher-Suhr U. (2001). Standardized general-purpose noise reaction questions for community noise surveys: Research and a recommendation. Journal of sound and vibration, 242(4), 641-679.

Filipan K, De Coensel B, Aumond P, Can A, Lavandier C, Botteldooren D. (2019). Auditory sensory saliency as a better predictor of change than sound amplitude in pleasantness assessment of reproduced urban soundscapes. Building and Environment, 148, 730-741..

Gidlöf-Gunnarsson A, Öhrström E. (2007). Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas. Landscape and Urban Planning, 83(2-3), 115-126.

Gillespie MAK, Baude M, Biesmeijer J, Boatman N, Budge GE, Crowe A, Memmott J, Morton DR, Pietravalle S, Potts SG, Senapathi D, Smart SM, Kunin WE. (2017). A method for the objective selection of landscape-scale study regions and sites at the national level. Methods in Ecology and Evolution, 8(11), 1468-1476.

Guastavino C. (2006). The ideal urban soundscape: Investigating the sound quality of French cities. Acta Acustica united with Acustica, 92(6), 945-951.

Hong JY, He J, Lam B, Gupta R, Gan WS. (2017). Spatial Audio for Soundscape Design: Recording and Reproduction. Applied Sciences, 7(6), 627.

Hong JY, Jeon JY. (2015). Influence of urban contexts on soundscape perceptions: A structural equation modeling approach. Landscape and Urban Planning, 141, 78-87.

Ishihara S. (1957). Test for Colour Deficiency – 24 Plates Edition. Tokyo: Kanehara Shuppan, 24.

ISO (2014). ISO 12913-1:2014 Acoustics — Soundscape — Part 1: Definition and Conceptual Framework. Geneva: International Organization for Standardization.

ISO (2018). ISO/PRF TS 12913-2, "Acoustics—Soundscape—Part 2: Data collection and reporting requirements", ISO Technical Specification, Geneva, Switzerland.

Jennings P, Cain R. (2013). A framework for improving urban soundscapes. Applied Acoustics, 74(2), 293-299.

Kamenický M. (2018). Enhanced sound source composition methods for qualitative mapping of urban sound environment. In 11th European Congress and Exposition on Noise Control Engineering (Euronoise 2018).

Kang J, Aletta F, Gjestland TT, Brown LA, Botteldooren D, Schulte-Fortkamp B, Lercher P, van Kamp I, Genuit K, Fiebig A, Coelho JL. (2016). Ten questions on the soundscapes of the built environment. Building and Environment, 108, 284-294.

Kaya EM, Elhilali M. (2017). Modelling auditory attention. Phil. Trans. R. Soc. B, 372(1714), p.20160101.

Krause B, Márquez-Ruiz J, Cohen Kadosh R. (2013). The effect of transcranial direct current stimulation: a role for cortical excitation/inhibition balance?. Frontiers in human neuroscience, 7, p.602.

Latinjak AT. (2012). The underlying structure of emotions: A tri-dimensional model of core affect and emotion concepts for sports. Revista Iberoamericana de Psicología del Ejercicio y el Deporte, 7(1), 71-87.

Léobon A. (1995). La qualification des ambiance sonores urbaines. Natures Sciences Société, 3(1), 26-41.

Lindau A, Weinzierl S. (2012). Assessing the plausibility of virtual acoustic environments. Acta Acustica united with Acustica, 98(5), 804-810.

Liu J, Kang J. (2015). Soundscape design in city parks: exploring the relationships between soundscape composition parameters and physical and psychoacoustic parameters. Journal of Environmental Engineering and Landscape Management, 23(2), 102-112.

Longstreth R. (ed.). (2008). Cultural landscapes: balancing nature and heritage in preservation practice. Minneapolis: University of Minnesota Press.

Macdonald JS, Lavie N. (2011). Visual perceptual load induces inattentional deafness. Attention, Perception, & Psychophysics, 73(6), 1780-1789.

Maris E, Stalen PJ, Vermunt R, Steensma H. (2007). Noise within the social context: annoyance reduction through fair procedures. Journal of the Acoustical Society of America, 121(4), 2000-2010.

Nielbo FL, Steele D, Guastavino C. (2013). Investigating soundscape affordances through activity appropriateness. In Proceedings of Meetings on Acoustics ICA2013 (Vol. 19, No. 1, p. 040059). ASA.

Nilsson M, Botteldooren D, De Coensel B. (2007). Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas. In Proceedings of the 19th International Congress on Acoustics.

Oldoni D, De Coensel B, Boes M, Rademaker M, De Baets B, Van Renterghem T, Botteldooren D. (2013). A computational model of auditory attention for use in soundscape research. The Journal of the Acoustical Society of America, 134(1), 852-861.

Payne SR. (2013). The production of a perceived restorativeness soundscape scale. Applied Acoustics, 74 (2), 255-263.

Puyana-Romero V, Lopez-Segura LS, Maffei L, Hernández-Molina R, MasulloM. (2017). Interactive Soundscapes: 360°-Video Based Immersive Virtual Reality in a Tool for the Participatory Acoustic Environment Evaluation of Urban Areas. Acta Acustica united with Acustica, 103(4), 574-588.

Raimbault M, Dubois D. (2005). Urban soundscapes: Experiences and knowledge. Cities, 22(5), 339-350.

Richardson AD, Jenkins JP, Braswell BH, Hollinger DY, Ollinger SV, Smith M. (2007). Use of digital webcam images to track spring green-up in a deciduous broadleaf forest. Oecologia, 152, 323-334.

Russell JA. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161-1178.

Rychtáriková M, Vermeir G. (2013). Soundscape categorization on the basis of objective acoustical parameters. Applied Acoustics, 74(2), 240-247.

Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, Yacoub E, Formisano E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proceedings of the National Academy of Sciences, 114(18), 4799-4804.

Schafer RM. (1994). The Soundscape: Our Sonic Environment and the Tuning of the World. Destiny Books, Rochester, Vermont.

Schisterman EF, Perkins NJ, Liu A, Bondell H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology, 16(1), 73–81.

Schönwiesner M, Zatorre RJ. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proceedings of the National Academy of Sciences, 106(34), 14611-14616.

Simons DJ, Chabris CF. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception, 28(9), 1059-1074.

Southworth M. (1969). The sonic environment of cities. Environment and Behavior, 1(1), 49-70.

Sun K, De Coensel B, Echevarria Sanchez GM, Van Renterghem T, Botteldooren D. (2018a). Effect of interaction between attention focusing capability and visual factors on road traffic noise annoyance. Applied Acoustics, 134, 16-24.

Sun K, Echevarria Sanchez GM, De Coensel B, Van Renterghem T, Talsma D, Botteldooren D. (2018b). Personal audiovisual aptitude influences the interaction between landscape and soundscape appraisal. Frontiers in Psychology, 9:780.

Terroir J, De Coensel B, Botteldooren D, Lavandier C. (2013). Activity interference caused by traffic noise: Experimental determination and modeling of the number of noticed sound events. Acta Acustica united with Acustica, 99(3), 389-398.

Tress B, Tress G, Fry G, Opdam P (eds.). (2006). From Landscape Research to Landscape Planning – Aspects of Integration, Education and Application. Dordrecht, The Netherlands: Springer.

van den Bosch KA, Andringa TC, Post WJ, Ruijssenaars WA, Vlaskamp C. (2018). The relationship between soundscapes and challenging behavior: A small-scale intervention study in a healthcare organization for individuals with severe or profound intellectual disabilities. Building Acoustics, 25(2), 123-135.

Van Renterghem T. (2018). Towards explaining the positive effect of vegetation on the perception of environmental noise. Urban Forestry & Urban Greening.

Van Renterghem T, Botteldooren D. (2016). View on outdoor vegetation reduces noise annoyance for dwellers near busy roads. Landscape and Urban Planning, 148, 203-215.

Vorländer M. (2008). Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, Springer, Berlin.

Walker AJ, Ryan RL. (2008). Place attachment and landscape preservation in rural New England: A Maine case study. Landscape and Urban Planning, 86(2):141-152.

Yu L, Kang J. (2015). Using ANN to study sound preference evaluation in urban open spaces. Journal of Environmental Engineering and Landscape Management, 23(3), 163-171.

Xue M, Atallah BV, Scanziani M. (2014). Equalizing excitation–inhibition ratios across visual cortical neurons. Nature, 511(7511), p.596.

Zannin PHT, Calixto A, Diniz FB, Ferreira JAC. (2003). A survey of urban noise annoyance in a large Brazilian city: the importance of a subjective analysis in conjunction with an objective analysis. Environmental Impact Assessment Review, 23(2), 245-255.

Appendix I

Methods for objective data collection – Recording protocol

Site selection protocol

Sampling of urban sites for performing soundscape evaluation studies is most often performed in an *ad hoc* manner. Systematic site selection methods for landscape studies, conservation and planning are often based on objective factors such as land cover (Gillespie et al., 2017), as well as perception, visual preference and emotional attachment of local residents (Longstreth, 2008; Walker and Ryan, 2008). The latter are typically evaluated through surveys or interviews, in order to select a sample of sites covering a wide range of landscapes (Tress et al., 2006).

A similar approach for site selection was also applied at the early stage of this study. An online questionnaire survey was conducted among 30 to 50 inhabitants (depending on the city), in which they were asked to pinpoint outdoor public spaces within their city that they perceive along the soundscape perception dimensions of pleasantness and eventfulness. Locations obtained from the online survey were then spatially clustered using the Google MapClusterer API, which allows extracting a shortlist of prototypical locations. This approach was designed to lead to a range of urban sites with a large variety in soundscapes, more or less uniformly covering each of the four quadrants of the 2D core affect perceptual space (Axelsson et al., 2010; Cain et al., 2013). In each city, participants were recruited among local students, and through calls for participation on relevant Facebook pages and with local guide associations. Details of the site selection protocol can be found in De Coensel et al. (2017).

Audio-visual recording

Combined and simultaneous audio and video recordings were performed at the selected locations within each city, using a portable, stationary recording setup (Figure I). The setup consists of the following components: binaural audio (HEAD acoustics HSU III.2 artificial head with windshield and SQobold 2-channel recording device), first-order ambisonics (Core Sound TetraMic microphone with windshield and Tascam DR-680 MkII 4-channel recording device) and 360-degree video camera (GoPro Omni spherical camera system, consisting of 6 synchronized GoPro HERO 4 Black cameras). The ears of the artificial head, the video camera system and the ambisonics microphone are located at heights of about 1.50m, 1.70m and 1.90m, respectively. It was chosen to stack the audio and video recording devices vertically, such that no horizontal displacement between devices is introduced, which could otherwise result into an angular mismatch for the localization of sound sources in the horizontal plane. A minimal separation distance of about 20cm between the camera and both the binaural and ambisonics microphones is required, such that these do not show up prominently on the recorded video, and can be masked easily using video processing software. All audio was recorded with a sample rate of 48 kHz and a bit depth of 24 bits, and were stored in uncompressed .wav format; moreover, the binaural recordings were performed according to the specifications set forth in ISO/TS 12913-2:2018 (ISO, 2018). Note that the recording setup is highly portable: when disassembled, all components can be carried by a single person. Assembling the setup takes about 10 minutes, and batteries and memory of all recording devices allow for about a full day of recording.

At each location, the recording system is oriented towards the most important sound source and/or the most prominent visual scene—this orientation defines the initial frontal viewing direction for the 360-degree video and ambisonics recordings, and the fixed orientation for the binaural recordings. Time synchronization is performed at the start of each recording by clapping hands directly in front of the system; this also allows checking correct 360-degree alignment of all components when post-processing. At each location, at least 10 minutes of continuous recordings were performed, such that 1-minute or 3-

minute fragments containing no disturbances can be extracted easily. During recording, the person handling the recording equipment was either hiding (in order not to show up on the 360-degree video) or, in case hiding was not possible, blended in the environment (e.g. performing the same activities as the other people around).

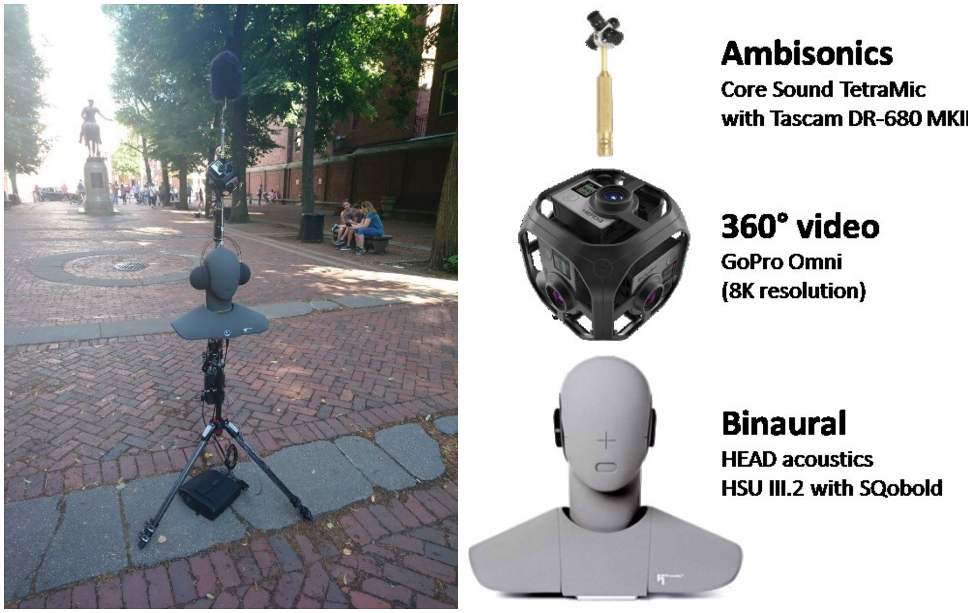

Figure I – Audio-visual recording setup (*Left*: photo on location (Boston); *Right*: position diagrammatic sketch of the recording equipments).

Post-processing for Virtual Reality

Since the six cameras from GoPro Omni use a parallel program, the six individual videos are automatically synchronized. The stitching work that combines these six videos together as a single 360-degree video is achieved with Autopano Video and Autopano Giga from Kolor software team. It gives the postproduction a stable, color-balanced and sustained 360-degree view. Since the postproduction captures the full surroundings, it is impossible to know what the viewer will eventually be focusing on (within the 360-degree sphere) at any given moment. In this study, only the opening scene of each recording (the coordinates of the image) was fixed, which ensures all the participants receive the same view at the beginning. With this setting, it also sets a reference for the audio-spatial synchronization.

Since the GoPro Omni cameras stand between the tripod stand, the HEAD and the Tascam (Figure 2), the videos will also record these devices, shown in zenith and nadir (top and bottom) in the postproduction, respectively. These were carefully camouflaged with a patch created in Photoshop, ensuring that no recording equipment appears in the final playback. Also, a color equalization has been applied to the postproduction by using ffmpeg (saturation=2), which highlights the color vividness in the video. All videos were exported in 4k quality. Together with the presentation by an Oculus Virtual Reality device, it gives a visually realistic and immersive experience as if the participants were in the place standing right on the recording position.

These 360-degree video is paired with ambisonics audio recording. The reason why first-order ambisonics audio can be used is explained in Appendix II. Video and audio synchronization was conducted by ffmpeg. Google Spatial Media Metadata Injector was used to achieve the spatial audio effect following head rotations.

Appendix II

Preliminary study – Validation of the recording and playback protocol

Overview

With the virtual reality device presents the video, it is expected to pair with corresponding audio recording, that ensures a high quality and spatial effect. Note that the audio recording by GoPro Omni cameras itself was not used in this study. As the recording contains both ambisonics and binaural audio (Figure 2), it is essential to decide which audio recording performs better through headphone playback when combined with virtual reality. A preliminary experiment was designed for this purpose.

Binaural audio recordings, performed using an artificial head, are generally considered to provide the highest degree of realism. Using an artificial head, the sound is recorded as if a human listener is present in the original sound field, preserving all spatial information in the audio recording. The main disadvantage of binaural audio recordings is that the frontal direction, and as such the acoustic viewpoint of the listener, is fixed by the orientation of the artificial head during the recording. This drawback could in theory be solved using ambisonics audio recording (Gerzon, 1985), a multichannel recording technique that allows for unrestricted rotation of the listening direction after recording. In principle, this technique could therefore provide an alternative to binaural recordings in the context of soundscape studies. However, the ambisonics technique has its own disadvantages, such as the more complex process of playback level calibration and equalization as compared to the binaural technique, the necessity of head tracking and real-time HRTF updates in case of playback through headphones, and the limited spatial resolution that can be achieved with lower-order ambisonics recordings—to date, there are no truly portable higher-order ambisonics recording systems available. Nevertheless, (first-order) ambisonics has become the de facto standard for spatial audio in VR games and platforms providing 360 video playback such as YouTube or Facebook.

Material & Experiment setup

Five 1-minute recordings were chosen for experiment 1 (Table I). The stimuli contain a fixed HD video, cut out from the original video in the frontal viewing direction, and padded with black in order to obtain again a 360-degree spherical video that can be viewed through a head-mounted display. This creates a “window” effect, forcing the participant to watch only in the frontal direction (Supplement 3). Furthermore, these stimuli are created in two flavors: with first-order ambisonics spatial audio track (allowing for head rotation) and with binaural audio track (which provides a fixed, i.e. head-locked, listening direction).

Table I – Stimuli used in the validation experiment.

Label	City	Date	Time	Location	Longitude	Latitude	$L_{Aeq,1min}$
R0001	Montreal	2017/6/22	8:02	Palais des congrès	45.503457	-73.561461	65.8
R0012	Boston	2017/6/28	9:36	Boston Public Garden	42.353478	-71.070151	62.5
R0030	Tianjin	2017/8/24	16:00	Century Clock	39.13262	117.198314	63.2
R0038	Hong Kong	2017/8/29	17:07	Taikoo Shing	22.286715	114.218385	64.6
R0055	Berlin	2017/9/10	12:08	Checkpoint Charlie	52.507796	13.390011	66.5

The experiment setup is the same as described in Section 2.2.2. During the experiment, participants were seated inside a soundproof booth. Recordings are played back using a PC (placed outside the booth), equipped with the GoPro VR Player 3.0 software, which allows to play back video with spatial audio. The 360-degree video is presented through an Oculus Rift head-mounted display, and the participant could freely move the head and look around in all directions. The audio is played back through Sennheiser HD 650 headphones, driven by a HEAD acoustics LabP2 calibrated headphone amplifier. Stimuli with binaural audio track are automatically played back at the correct level, as the headphone amplifier and headphones are calibrated and equalized for the artificial head that made the recordings. The gain of the ambisonics audio tracks have been adjusted such that their level is as close as possible to that of the corresponding binaural audio tracks.

Procedure & Participants

Since 5 stimuli paired with 2 audio recordings were involved, these 10 videos were played randomly to participants (20 participants, 6 female, Age_{mean}=28.9 yr, standard deviation 2.8 yr, range: 25-35 yr). After each video, 6 questions were shown in the VR screen (Table II, Guastavino et al., 2007). Participants needed to answer each question on a 5-point scale by verbal talking.

Table II – Questions asked to the participants in the validation experiment.

Question:	Answer (5-point scale):
1. The sonic environment sounds __ enveloping.	little – very
2. I feel __ immersed on the sonic environment.	little – very
3. Representation of the sonic environment:	poor – good
4. Readability of this scene:	poor – good
5. Naturalness, true to life:	not truthful – truthful
6. The quality of the reproduction is __.	poor – good

Results

Table III shows the results of the comparison between ambisonics (allowing head rotation) and binaural (head-locked) audio playback. The table shows, on a scale from 1 to 5, the median scores on the questions asked (similar results are obtained with average scores). When there is a difference in median between the binaural and ambisonics playback cases, the higher value is underlined.

Table III – Median score of five pairs of soundscapes in the second stage of the validation experiment: a) ambisonics, b) binaural.

Label	Envelopment		Immersion		Representation		Readability		Realism		Overall quality	
	a	b	a	b	a	b	a	b	a	b	a	b
R0001	4.0	4.0	3.5	4.0	4.0	3.5	4.0	3.0	3.5	4.0	4.0	4.0
R0012	3.5	4.0	3.0	3.5	3.0	3.0	3.0	3.5	3.0	3.0	3.0	3.0
R0030	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
R0038	4.0	3.5	4.0	3.0	4.0	4.0	4.0	3.5	4.0	4.0	4.0	4.0
R0055	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	3.0

Earlier research (Guastavino et al., 2007) showed that ambisonics audio results in a high degree of envelopment and immersion. Intuitively, one would expect that the possibility of rotating one's head during playback would result in a higher degree of envelopment and immersion, as compared to the case when one's listening direction is locked. On the other hand, due to the limited spatial resolution offered by first-order ambisonics, one would expect the binaural reproduction to result in a higher degree of readability and realism. The results shown in Table III do not allow to draw these conclusions; using a two-sample *t*-test with significance level 0.05, no significant difference is found between both sound reproduction methods, for any of the perceptual dimensions considered. Moreover, the difference between soundscapes is found to be larger than between the audio reproduction methods; some differences are significant, e.g. between R0012 and R0030 regarding representation (both ambisonics and binaural) and realism (binaural), or between R0012 and R0055 regarding immersion (ambisonics), readability (ambisonics) and representation (both ambisonics and binaural). This pilot test therefore justifies the use of ambisonics in the first stage of the experiment; either reproduction method could have been used.

Reference

Gerzon MA. (1985). Ambisonics in multichannel broadcasting and video. Journal of the Audio Engineering Society, 33(11), 859-871.

Guastavino C, Larcher V, Catusseau G, Boussard P. (2007). Spatial audio quality evaluation: comparing transaural, ambisonics and stereo, In Proceedings of the 13th International Conference on Auditory Display (ICAD), Montréal, Canada.

Appendix III

Overview of the basic characteristics of the recordings used for the VR experiment.

Table IV – Overview of the stimuli presented in the two repetitions of the soundscape classification experiment: (above division line) collection 1, (below division line) collection 2.

Label	City	Date	Time	Location	Longitude	Latitude	$L_{Aeq,1min}$/dB
R0002	Montreal	2017/6/22	8:43	Place d'Armes	45.504683	-73.55715	66.5
R0003	Montreal	2017/6/22	9:43	Tour de l'horloge	45.511973	-73.545911	55
R0007	Montreal	2017/6/22	15:26	Chalet du Mont-Royal	45.503405	-73.587005	54.8
R0010	Montreal	2017/6/22	17:53	Square Phillips	45.503807	-73.568543	67.5
R0011	Montreal	2017/6/22	19:10	Place Jacques Cartier	45.50768	-73.552625	66.1
R0015	Boston	2017/6/28	12:41	Old State House	42.359039	-71.057139	69.5
R0016	Boston	2017/6/28	13:11	Quincy Market	42.35986	-71.055825	74.6
R0017	Boston	2017/6/28	13:47	Post Office Square	42.35623	-71.0556	65.8
R0018	Boston	2017/6/28	14:23	R. F. Kennedy Greenway	42.354721	-71.052073	66.1
R0020	Boston	2017/6/28	16:31	Paul Revere Mall	42.365687	-71.053446	57.4
R0022	Tianjin	2017/8/24	8:54	Peiyang Square (TJU campus)	39.107327	117.170222	62.2
R0026	Tianjin	2017/8/24	11:46	Water Park North	39.090986	117.163317	60.4
R0029	Tianjin	2017/8/24	15:29	Haihe Culture Square	39.130202	117.193256	73.5
R0031	Tianjin	2017/8/24	16:26	Tianjin Railway Station	39.133779	117.203206	65.2
R0033	Tianjin	2017/8/24	17:59	Nanjing Road	39.118566	117.185557	65.3
R0036	Hong Kong	2017/8/29	15:43	Wanchai Tower	22.279705	114.17245	68.7
R0040	Hong Kong	2017/8/30	7:44	Hong Kong Park	22.277824	114.161488	64.1
R0041	Hong Kong	2017/8/30	8:50	Wong Tai Sin Temple	22.342062	114.194042	69.7
R0047	Hong Kong	2017/8/30	13:36	Peking Road	22.296512	114.171813	77
R0048	Hong Kong	2017/8/30	14:30	Ap Lei Chau Waterfront	22.245093	114.155663	62.2
R0050	Berlin	2017/9/9	16:57	Breitscheidplatz	52.504926	13.336556	72.4
R0054	Berlin	2017/9/10	11:32	Gendarmenmarkt	52.513517	13.3929	60.8
R0058	Berlin	2017/9/10	14:18	Lustgarten	52.518604	13.399195	65.2
R0060	Berlin	2017/9/10	15:39	James-Simon Park	52.521787	13.399158	65.9
R0061	Berlin	2017/9/10	16:32	Pariser Platz	52.516145	13.378545	67.7
R0001	Montreal	2017/6/22	8:02	Palais des congrès	45.503457	-73.561461	65.8
R0004	Montreal	2017/6/22	10:39	Place Marguerite-Bourgeoys	45.507368	-73.555006	62.1
R0005	Montreal	2017/6/22	12:21	Parc La Fontaine	45.523279	-73.568341	53.7
R0006	Montreal	2017/6/22	14:22	Monument à Sir George-Étienne Cartier	45.514488	-73.586564	58.7
R0008	Montreal	2017/6/22	16:26	McGill University campus	45.504202	-73.576833	54.7

R0012	Boston	2017/6/28	9:36	Boston Public Garden	42.353478	-71.070151	62.5
R0013	Boston	2017/6/28	10:12	Boston Common	42.353705	-71.065063	62.3
R0023	Tianjin	2017/8/24	9:23	Jingye Lake (TJU campus)	39.107495	117.166476	57.4
R0027	Tianjin	2017/8/24	12:14	Water Park Center	39.087846	117.162092	58.5
R0030	Tianjin	2017/8/24	16:00	Century Clock	39.13262	117.198314	63.2
R0032	Tianjin	2017/8/24	16:55	Jinwan Plaza	39.131835	117.202969	60.7
R0034	Tianjin	2017/8/24	18:44	Drum Tower	39.140833	117.174355	54.5
R0037	Hong Kong	2017/8/29	16:14	Johnston Road	22.277781	114.176621	71.6
R0038	Hong Kong	2017/8/29	17:07	Taikoo Shing	22.286715	114.218385	64.6
R0039	Hong Kong	2017/8/29	17:55	Victoria Park	22.281835	114.187832	57.0
R0042	Hong Kong	2017/8/30	9:44	Nelson Street	22.318352	114.170164	67.2
R0043	Hong Kong	2017/8/30	10:32	Signal Hill Garden	22.296008	114.174859	62.1
R0045	Hong Kong	2017/8/30	12:45	Hong Kong Cultural Centre	22.29343	114.170038	60.7
R0049	Hong Kong	2017/8/30	15:53	The Peak	22.270879	114.150917	55.6
R0052	Berlin	2017/9/10	9:28	Tiergarten	52.512166	13.347172	53.3
R0053	Berlin	2017/9/10	10:48	Leipziger Platz	52.509296	13.37818	68.8
R0055	Berlin	2017/9/10	12:08	Checkpoint Charlie	52.507796	13.390011	66.5
R0057	Berlin	2017/9/10	13:43	Neptunbrunnen	52.519829	13.406623	66.2
R0062	Berlin	2017/9/10	18:06	Sony Center	52.510166	13.373572	66.9
R0063	Berlin	2017/9/10	18:31	Potsdamer Platz	52.509192	13.376332	67.4