On the Effectiveness of Facial Expression Recognition for Evaluation of Urban Sound Perception

Qi Meng,[a,b] Xuejun Hu,[a] Jian Kang[a,b*], Yue Wu[a*]

[a] Key Laboratory of Cold Region Urban and Rural Human Settlement Environment Science and Technology, Ministry of Industry and Information Technology, School of Architecture, Harbin Institute of Technology, Harbin 150001, 66 West Dazhi Street, Nan Gang District, Harbin, China

[b] UCL Institute for Environmental Design and Engineering, University College London (UCL), London WC1H 0NN, UK

*Corresponding Author
Prof. Jian Kang, j.kang@ucl.ac.uk
Dr. Yue Wu, wuyuehit@hit.edu.cn;

Abstract

Sound perception studies mostly depend on questionnaires with fixed indicators. Therefore, it is desirable to explore methods with dynamic outputs. The present study aims to explore the effects of sound perception in the urban environment on facial expressions using software named FaceReader based on facial expression recognition (FER). The experiment involved three typical urban sound recordings, namely, traffic noise, natural sound, and community sound. A questionnaire on the evaluation of sound perception was also used, for comparison. The results show that, first, FER is an effective tool for sound perception research, since it is capable of detecting differences in participants' reactions to different sounds and how their facial expressions change over time in response to those sounds, with mean difference of valence between recordings from 0.019 to 0.059 ($p<0.05$or $p<0.01$). In a natural sound environment, for example, facial expression increased by 0.04 in the first 15 seconds and then went down steadily at 0.004 every 20 s. Second, the expression indices, namely, happy, sad, and surprised, change significantly under the effect of sound perception. In the traffic sound environment, for example, happy decreased by 0.012, sad increased by 0.032, and surprised decreased by 0.018. Furthermore, social characteristics such as distance from living place to natural environment ($r=0.313$), inclination to communicate ($r=0.253$), and preference for crowd ($r=0.296$) have effects on facial expression. Finally, the comparison of FER and questionnaire survey results showed that in the traffic noise recording, valence in the first 20 s best represents acoustic comfort and eventfulness; for natural sound, valence in the first 40 s best represents pleasantness; and for community sound, valence in the first 20 s of the recording best represents acoustic comfort, subjective loudness, and calmness.

Keywords: facial expression recognition; sound perception; urban soundscape; FaceReader

2019 Science of the Total Environment
Date received: 28 Septmeber 2019, Date accepted: 10 November 2019
Publish online: 10 November 2019

1. *Introduction*

Sound quality is considered to be a key part of the sustainable development of urban open spaces (Zhang et al., 2006; Hardy, 2012). Reducing sound levels from certain sound sources (reducing noise) may not necessarily result in an acoustic environment of high quality; one reason is that sound perception is equally important as absolute sound level (Axelssonet al., 2014; European Environmental Agency, 2014). To reflect this fact, the term 'soundscape' was put forward by Schafer, with the definition "a sonic environment with emphasis on the way it is perceived and understood by the individuals or by a society" (Brown et al., 2011). Ever since this concept emerged, researchers have been studying the effects of acoustic environments on the sound quality and how sounds can be used in urban planning and design (Pijanowski, 2011). In 2014, the International Organization of Standardization (ISO, 2014) gave a broader definition of soundscape: "acoustic environment as perceived or experienced and/or understood by a person or people, in context." That is, the soundscape is an evaluation of the acoustic quality of a space, and such an evaluation cannot neglect cultural factors and the lived experiences of people (Brown et al., 2011).

In sound perception studies, the typical data collection methods are: (1) questionnaires, (2) semantic scales, (3) interviews (Alettaet al., 2018; Pérez et al., 2018), (4) physiological measurements, and (5) observation protocols (Schulte, 2002). Of these tools, questionnaire surveys are the most common in both on-site studies and laboratory studies (Alettaet al., 2016). In previous soundscape studies, questionnaires have been used to evaluate pleasantness, sound preference and subjective, loudness (Liu et al., 2014) as well as audio–video and sound–odour interaction in urban environments, in laboratory settings (Ba and Kang, 2019).Semantic scales are usually used to find out the extent of certain characteristics of the sound environment. For instance, Pérez et al. (2018) studied the soundscapes of monumental places through on-site questionnaires, asking about perceived sounds' sources, loudness, and a set of 12 semantic attributes (such as 'pleasant', 'natural', 'comfortable'). Mackrill et al. (2013) conducted laboratory experiments to study hospital sounds, having participants semantically (such as 'relaxed', 'reassured', 'at ease' and 'intrigued) rate a range of sound clips representative of a ward soundscape.

However, questionnaire surveys have several limitations, as follows: For one thing, questionnaires are subjective, and an 'experimenter effect' might occur if the questionnaire is not well designed (Brown et al., 2011). For another, the results of the questionnaire for a participant's experience about a recording are a fixed value and cannot show the trend over time.

Another method of data collecting in sound perception is physiological measurement. Physiological indicators such as heart rate (HR) (Medvedev et al., 2015; Blood and Zatorre, 2001), SCL (Medvedev et al., 2015), electrodermal activity, respiratory rate, and facial electromyography (Lee et al., 2018) are also used in sound perception studies. Compared with questionnaire surveys, physiological measurement has some advantages: it is objective and can show change over time. However, physiological methods cannot directly reflect the emotions of research participants (Alettaet al., 2016).

As a environment evaluation tool, FaceReader (Noldus, 2014), a software based on facial expression recognition (FER), has been applied in psychological evaluations (Zarbakhsh and Demirel, 2018; Bartlett et al., 2005; Amor et al., 2014). Video cameras have been the predominant method of measuring facial expressions in this context (Oliver et al., 2000). In FaceReader, facial expressions are divided to reflect six types of emotions: happiness, surprise, fear, sadness, anger, and disgust, as per a universally recognized classification of expressions (Ekman, 1982). The validity of FER has been proven in many previous studies (Sato et al., 2019). FaceReader is capable of measuring emotions with an efficacy of over 87% according to a self-assessment test (Christos et al., 2010). The validity of FaceReader for East Asian people in particular has been shown to be 71% (Yang and Li, 2015). The efficiency of this method has been tested in many domains of study, for example in relation to tourism advertisements, where FaceReader proved useful for collecting and analysing real-time data concerning seven discrete emotions plus valence and arousal (Hadinejadet al., 2017), or for characterizing customers' response to different kinds of sweeteners (Leitch et al., 2014). The effects of external factors on people are usually a cumulative process. Leitch et al. (2014) indicated the length of time after tasting sweeteners has influence on valence and arousal of facial expression. Hadinejad et al. (2017) proved that arousal and positive emotions keep decreasing as participants were watching tourism commercials. Therefore, FER has been proven to be effective in getting the accumulated changes of under the effect of external factors. However, whether it can be used for sound perception has not been tested by any study. Furthermore, people's demographic characteristics, such as their gender (Yang et al., 2018; Ma and Wong, 2017), age (Yi and Kang, 2019), and professions and educational levels (Meng and Kang, 2013), are expected to correlate with their evaluation of sound perception. Whether these characteristics will influence the results of facial expression recognition, however, is unknown.

Thus, the present study explores the effectiveness of facial expression recognition for evaluating urban sound perception, focusing on the following research questions: (1) whether this facial expression analysis system can be applied to sound perception studies; (2) what the mechanics are of the effects of different sound environment on valence and the different facial expression indices; (3) whether the characteristics of participants influence their facial expression; and (4) compared with questionnaires and semantic scales, what the advantages FER may have for urban sound perception study. A laboratory experiment using FER with 32 participants was conducted. The experiment involved three typical urban sound recordings, namely, traffic noise, natural sound, and community sound.

2. Method

2.1 Materials

The recordings were collected in typical urban public spaces along the banks of the Majiagou River in Harbin, which is located in northeast China, as shown in Figure 1. Three typical recordings representative of sounds described in a number of studies from the soundscape literature were chosen (Gygiet al., 2007; Yang et al., 2013; Kang et al., 2019), namely, the traffic sounds on the main roads of the city (Location 1), the natural sounds in the city parks (Location 2), and the sounds of human activity in the public spaces of the community (Location 3). The HEAD multi-channel acoustic data acquisition front end was used to collect binaural recordings, and the osmo+ panoramic camera was used to collect the corresponding scene video. The recordings were collected from 1 pm to 3 pm (Liu et al., 2014). Both the audio and video devices were placed at the height of 1.5m (Kang et al., 2019) to simulate the average height of the human ear. A BSWA 801 sound-level meter was placed at the same height to measure sound pressure level. In addition, the recordings were then played using high-fidelity headphones to play back to 30 volunteers randomly selected from students in Harbin Institute of Technology. They were asked to record sound sources they heard in each recording, and sort by significance. From the record, two most heard sound sources are the dominant source of sound for this survey site. The A-weighted equivalent SPL for 3-min of the traffic sound recording was 71.79 dBA in Location 1, and the main sound sources were car engines (55.4%), and horns (28.3%); the A-weighted equivalent SPL for 3-min of the natural sound recording was 43.83 dBA in Location 2, and the main sound sources included birdsong (66.4%), and wind (20.9%); and the A-weighted equivalent SPL for 3-min of community recording was 57.96 dBA in Location 3, mainly including people's speech (72.6%) and footsteps (20.8%). The same recording equipment was used to record a background soundscape with no significant sound in a quiet night environment. A BSWA 801 sound-level meter was used to record the A weighted equivalent SPL for 3mins, which is 38.42 dBA, which represents the average sound level. It was used to adapt the participants to the outdoor environment before playing each recording.

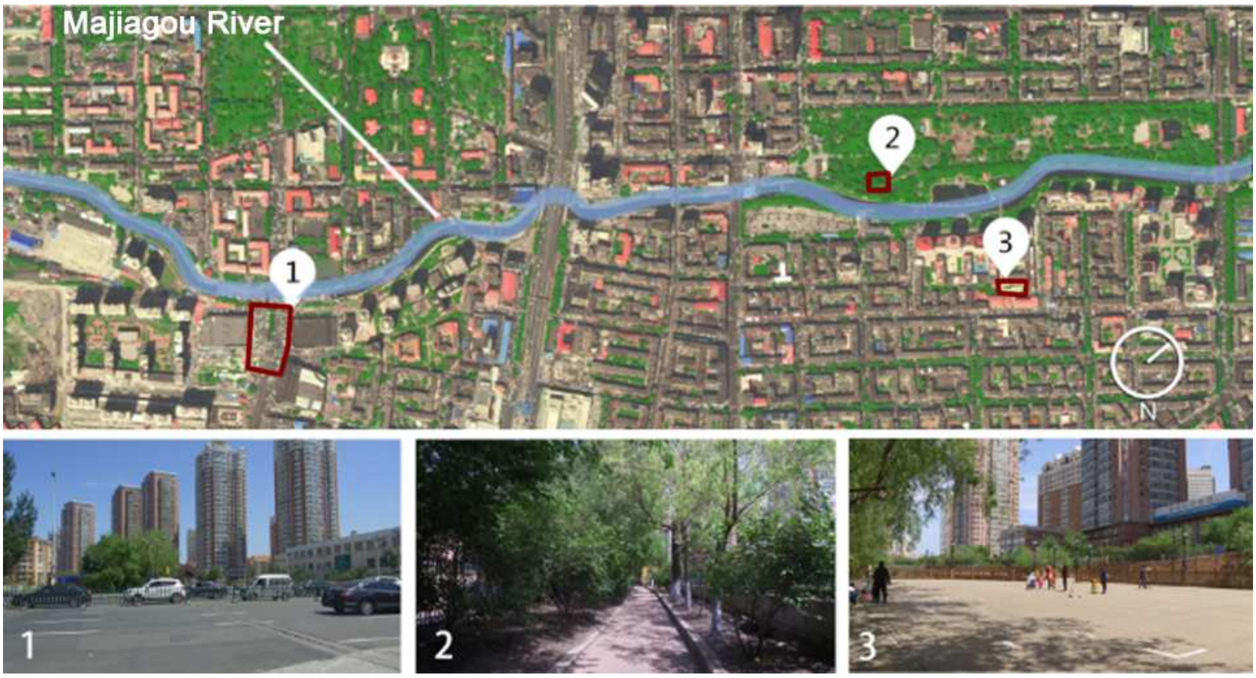
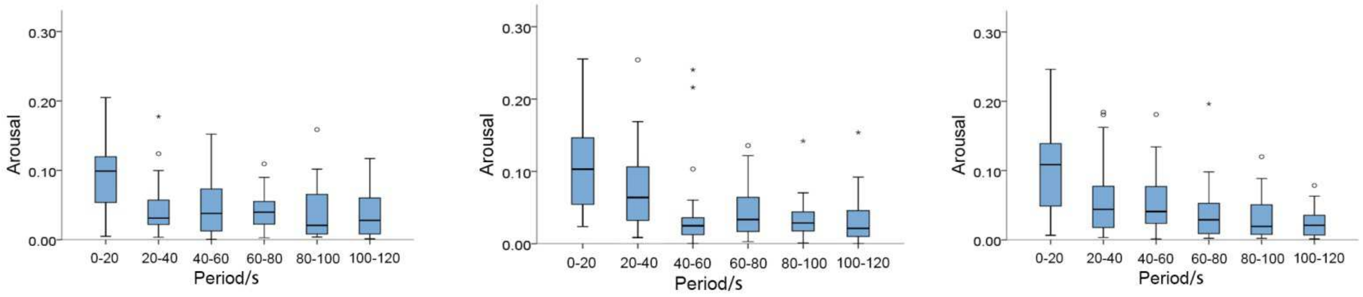


Figure 1. The survey sites and survey locations. 1-traffic sounds of the main roads of the city, mainly car engine, horns, and other mechanical noises;2-natural sounds in the city parks, mainly birdsong, wind and human footsteps; 3-sounds of human activity in the public spaces of the community, mainly people's speech and footsteps.

2.2 Stimuli

The recordings and videos were merged using Adobe Premiere software (Botello, 2006) into three video files, each lasting 3 minutes, composed of panoramic video and sound. In order to eliminate the impact of the video on the results, the three sections of the corresponding video are the same—scenes of a path with trees in the park—and there is no human activity in the video. Before formal experiment, a pilot experiment was carried out to test the effect of visual factors on facial expression. The participants were asked to listen to background sound with the same visual environment, the results showed that the facial expressions of the participants got into a stable state within 1 minute. Therefore, a 1 minute background sound with video were set up before the formal experiment to avoid the effect of visual factors. After the background sound, the main recording then plays after a 2-second transition, for 2 minutes (Zarbakhsh and Demirel, 2018).

In order to decide on a suitable duration of recordings, a pilot study was conducted. In the pilot study, the sampling rate of FER was set as 15 per second. Arousal, which indicates whether the test participant is active (+1) or not active (0), was used for the determination of duration. Figure 2 shows the absolute value of change in arousal every 20 seconds for the three recordings. The trends for the three recordings are similar (these results do not include the one-minute background sounds heard by the participants): in the first 20 seconds, arousal changes the most; comparing the period from 0 s to 20 s and the period from 20s to 40s, the amount of change in arousal significantly decreases (from 0.091 to 0.126 compared with 0.044 to 0.052), and then it remains relatively stable until the end of the recordings. However, as the length of time listening to the recording increases, the amount of change in arousal starts to increase slightly, in particular after 80 s, which may be due to distraction among the participants. These rises may not be caused by the recordings, indicating that when the length of the recording exceeds 80 s, the resulting facial expression analysis results may be inaccurate. This shows that when using FaceReader to study the acoustic landscape in a laboratory environment, recording duration should be from 60 s to 80 s.



(a) (b) (c)

Figure 2. The trends of arousal change in different sound sources.
traffic noise, (b) natural sound, and (c) community sound.

2.3 Participants

To determine the required number of, taking the traffic noise recording as an example, according to the difference between the average value of the 20 s valence and the initial value for facial expression recognition results, a one-sample t-test with 0 as test value was performed with different sample sizes, as shown in Table 1. It is worth noting that, unless otherwise stated, “40 s” indicates the ranges of 0s - 40s accumulated, and it is the same for 60 s, 80 s and so on. It can be seen from the value of t in Table 1 that the impact of traffic noise on facial expression is negative: when the recording duration is 20 s, to confirm the negative effect of traffic noise, the number of participants has to be 30 or more (p=0.038), but when the recording time is 40 s to 60 s, as few as 10 people (p=0.046) can prove the effect. Thus, when the recording duration is more than 40 seconds, at least 10 participants should be recruited to study the effect of the recording on facial expressions through FaceReader. In the present study, 32 participants (Hong and Jeon, 2014; Ba and Kang, 2019) were used—17 women and 15 men, all of whom were students at the Harbin Institute of Technology aged 19 to 26 (mean=23.32, SD=1.452). Power analyses based on g-power software have been used to calculate the sample size for paired sample t-test, ANOVA and linear regression, the power are 81%, 82% and 89% respectively, proving that the sample size meets the requirements for statistical analysis (Faul et al., 2007). The analysis was conducted in a laboratory with a panoramic screen, and participants were informed about the precautions and scheduled for a time to enter the laboratory the day before the analysis. The precautions included: (1) The experiment involves evaluation of sound under three types of environment using questionnaires, which will take approximately 30 minutes; (2) remember to take their smartphone for filling out online questionnaires; (3) if someone else is experimenting while entering the lab, please wait outside to prevent influence. In order to prevent the participants from knowing that this is an experiment of the influence of sound on facial expressions, which may lead to subjective errors, the participants were told before the experiment that it was just a subjective evaluation experiment of sound. On the day of the analysis, after a participant entered the laboratory, he or she was given five minutes to adapt to the laboratory environment.

Table1 Mean difference between valence of different time and test value of facial expressions with different sample size

	Sample size			
Listening time	5	10	20	30
20s	-0.002	-0.005	-0.006	-0.009 *
40s	-0.021*	-0.031*	-0.020 *	-0.021 *
60s	-0.019	-0.046 **	-0.028 *	-0.023 *
80s	-0.021	-0.030	-0.027	-0.025**

** means p<0.01; * means p<0.05

2.4 Questionnaire design

The questionnaire data were compared with the results of FaceReader, and were divided into three parts as shown in Table 2. The first part covered basic demographic information on the participants. The second part asked the participants to evaluate the subjective loudness and acoustic comfort of the recordings on a scale of 0 to 10 from ‘completely silent’ to ‘extremely loud’ and from ‘extremely uncomfortable’ to ‘extremely comfortable’. The third part asked the participants to rate the recordings on four semantic scales, namely, eventful, vibrant, calm, and pleasant. The questionnaire, and especially the semantic attributes chosen, were based on previous studies (Harris et al., 2005, Axelssonet al., 2010; Jennings et al., 2010; Cain et al., 2013). The fourth part was filled out after the experiment. In the community recording, the participants’ facial expressions showed different trends, which may be related to the participant’s preference for language sounds. Therefore, after the experiment, the participants were asked to fill out a supplementary online questionnaire (part 4), asking the extent to which their living place was near to the natural landscape, their inclination to communication, and their preference for crowds on a five-point scale (Pérez et al., 2018).

Table 2 Questions and scales in the questionnaire survey

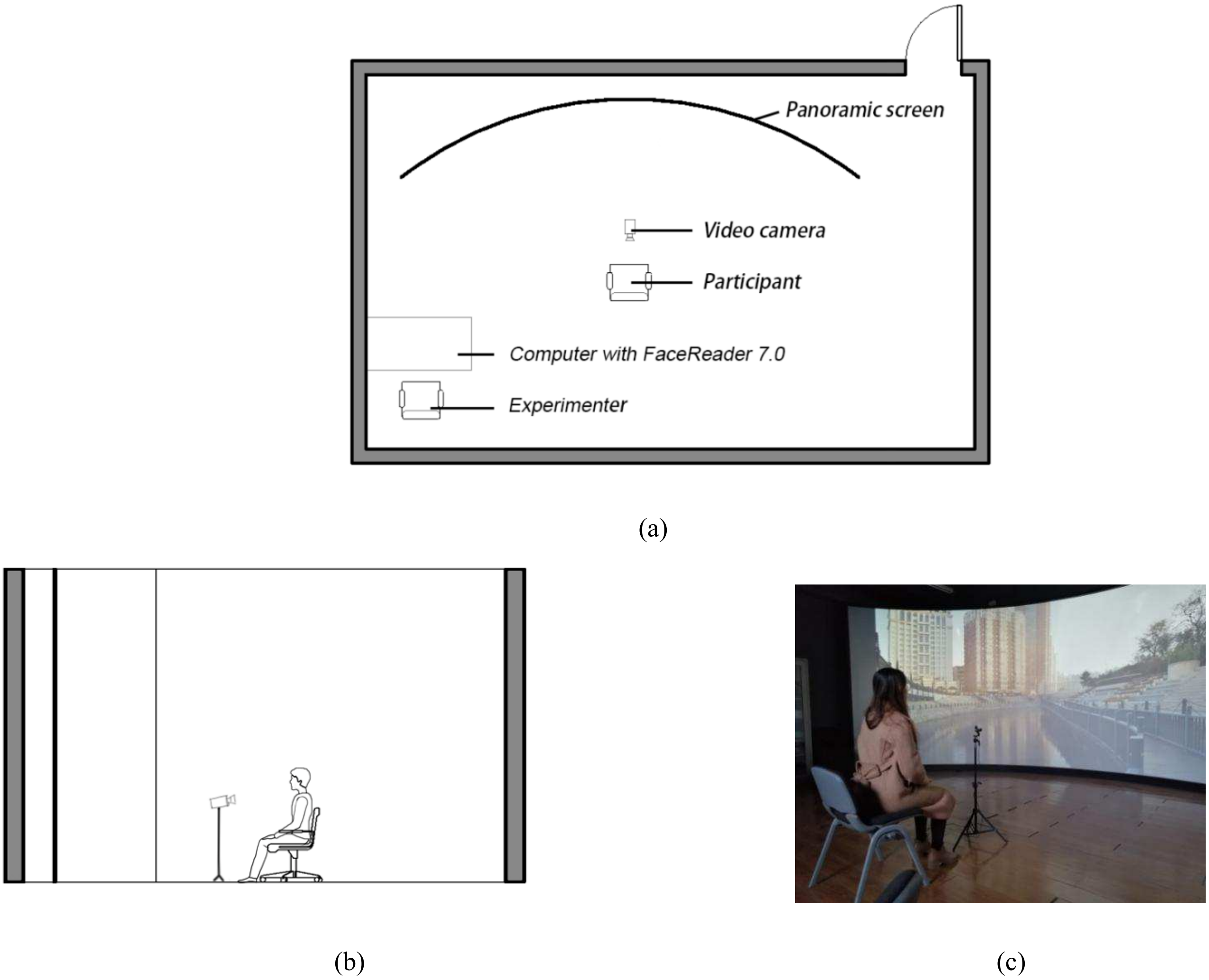
Parts	Questions	Scales
Part I Basic information	Name, gender, age	Nominal and ordinal
Part II Subjective loudness and acoustic comfort	Subjective loudness and acoustic comfort of the recordings	Scale of 0 to 10 from “completely silent” to “extremely loud” and from “extremely uncomfortable” to “extremely comfortable”
Part III Evaluation of sounds using semantic dimensions	The degree to which the recording was eventful/ vibrant/ calm/ pleasant	Scale of 0 to 10 from “completely silent” (extremely uncomfortable)
Part IV Preference for speech-related activities	Distance from their living place to the nature, inclination to communicate with others and preference for crowd	Scale of 1 to 5 from “completely silent” (extremely uncomfortable)

2.5 Facial expression recognition

The input of FaceReader is photos or videos of the participant’s face. To recognize facial expressions, FaceReader works in three steps. The first step is detecting the face (Viola and Jones, 2001). The next step is accurate 3D modelling of the face using an algorithmic approach based on the Active Appearance Method (AAM) described by Cootes and Taylor (2000). In the last step, the actual classification of facial expressions is done by training an artificial neural network; the AAM is used to compute scores of probability and intensity of six facial expressions (happiness, surprise, fear, sadness, anger, and disgust) on a continuous scale from 0 to 1 indicating intensity. ‘0’ means that the expression is absent, ‘1’ means that it is fully present (Lewinski et al., 2014).

Besides the intensity of individual facial expressions, FaceReader also calculates their valence, that is, whether the emotional state of the participant is positive or negative (Frijda, 1986). ‘Happy’ is the only positive expression; ‘sad’, ‘angry’, ‘scared’, and ‘disgusted’ are considered to be negative expressions; and ‘surprised’ is not used in the calculation of valence. FaceReader also calculates arousal, that is, whether the test participant is active (+1) or not active (0). Arousal is based on the activation of 20 Action Units (AUs) of the Facial Action Coding System (FACS) (Ekman et al., 2002).

Participants’ facial expression data were gathered in a laboratory (Figure 3). A panoramic screen was placed in the laboratory, on which the panoramic video taken of the recording area was displayed. The participants sat in a fixed seat in front of the screen, and a video camera connected to a computer running FaceReader was placed in front of the participant. Each participant listened to 3 recordings using earphones; FaceReader computed facial expression data as the recording went on. At the end of each recording, the QR code for the questionnaire was displayed on the screen, and the participants were asked to fill out a questionnaire related to their subjective feelings about the sound environment they had just experienced. The experimenter reminded the participants to fill out the questionnaire based on the 2 minutes of the recording, without including the previous 1 minute of background sound. The participants were given 5minutes to adapt to the laboratory environment and, after each recording, were asked to rest for 3 minutes before the next recording. After the three recordings, each participant was presented with a small gift.



(a)

(b)

(c)

Figure 3. The laboratory experiment for facial expression recognition.
(a)plan, (b)section, and (c) view from participant.

2.6 Data Analysis

SPSS (Feeney, 2012) was used to perform the analyses of survey data. First, a dispersion test was operated to delete the data with a large degree of dispersion. Then, a one-sample t-test was used to confirm the number of participants. An analysis of variance (ANOVA) was used to test the difference in valence between recordings. Paired-samples t-test and independent-samples t-test at $p<0.01$ and $p<0.05$ were conducted to compare difference of valence between recordings and test the significance of the influence of gender on facial expression. Pearson correlation tests were used to calculate the relationship between the results from FaceReader and the questionnaire. Further, linear and nonlinear regression analyses were performed to model change of valence and the six types of facial expression over time.

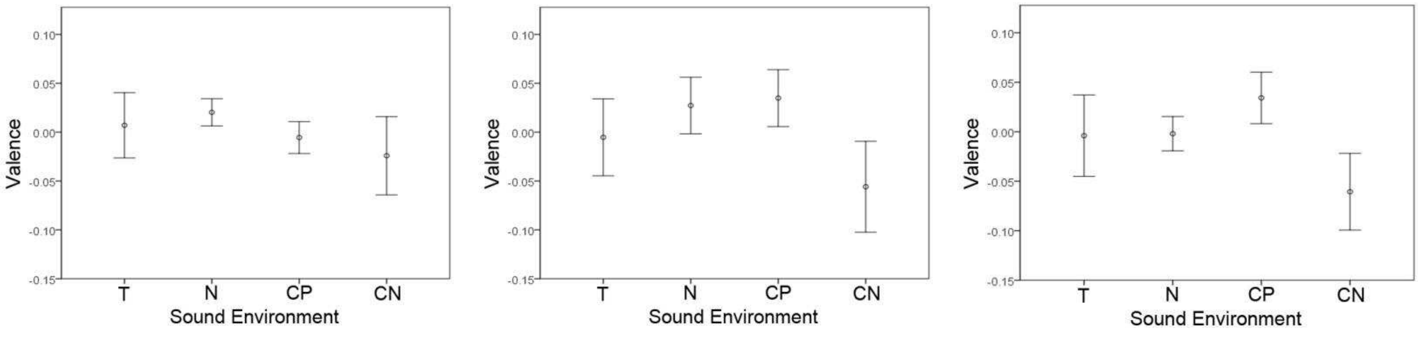
For individual difference in facial expressions and the comparison between questionnaire survey and FaceReader data, effect sizes were also reported using an effect size calculator (Lipsey and Wilson, 2012).Effect size is a quantitative measure for the strength of a statistical claim;

reporting of effect sizes facilitates the interpretation of the substantive magnitude of a phenomenon, while the significance of a research result reflects its statistical chance of being meaningful (Kelley and Preacher, 2012). The measurements of effect size are as follows: for Pearson correlation, 0.10 or above means a small effect, 0.30 or above means a medium effect, and 0.50 or above means a large effect (Cohen, 1988, 1992); for independent-samples t-test, d-value greater than 0.20 means a small effect, greater than 0.50means a medium effect, and greater than 0.80 means a large effect (Cohen, 1988).

3 Results

3.1 Effect of sound perception on facial expression

In FaceReader, the valence, which indicates whether the emotional state of the participant is positive or negative, is calculated for the facial expression in each frame. In this experiment, since the facial expression of each participant has different initial values in its natural state, the average value of all the data when listening to the background sound are taken as the initial value for each participant; measurement every 20 seconds after the start of the recording is then taken, and the average value is compared with the initial value to obtain the amount of change every 20 seconds. As mentioned in 2.4, there were two valence change models reflecting the features of the community sounds (the third recording); therefore, in the analyses below, the results of the third recording are separated and abbreviated into CP (community-positive) and CN (community-negative).



(a) (b) (c)

Figure 4 The changes of valence with different sound types from 20s to 80s with error bar of 95% confidence interval; (a) shows20s, (b) 40s and (c) 80s
T-traffic noise, N-natural sound, CP-community sound(positive), CN-community sound(negative)

Figure 4 shows the average valence change (with error bar of 95% confidence interval) from 20 s to 80 s, based on valences after listening to the recordings for 20 s, 40 s, and 80 s. In general, CN showed the greatest valence drop (from -0.061 to -0.024), followed by traffic noise (from -0.007 to -0.004), where the valence changed positively after listening to the recording for 20 s (0.020) and 40 s (0.027), but became below zero at 80 s (-0.002). The valence for CP was relatively low for the first 20 s (-0.006) but highest of the four at 40 s (0.035) and 80 s (0.034). To confirm the differences among sound environments, a one-way ANOVA was conducted; significance at 20 s, 40 s and 80 s was 0.157, 0.013 and 0.010, respectively, which indicates that after listening to recordings for 40 s and 80 s, the effect of sound source on facial expression is significant ($p<0.05$).

To test the differences between recordings, a paired-sampled t-test was conducted (Table 3). The greatest differences are between traffic noise and natural sounds, with mean difference reaching 0.028 at 40 seconds (p=0.025), and between natural sounds and CN, with mean difference reaching 0.059(p=0.034). The valence in the first 40 seconds between natural sounds and CP is similar, but deviation appears at 80 s, with a mean difference of 0.022(p=0.021).From valence results in Figure 4 and Table 3, we can see that there are differences between recordings and between measurement times, indicating that FaceReader is capable of detecting differences in participants' reactions to different sounds.

Table 3 Mean differences between different recordings of valence change from initial value to 20s, 40s, and 80s, where T means traffic noise, N means natural sound, CN means community (negative), CP means community (positive)

Listening time	T&N	T&CP	T&CN	N& CP	N&CN
20s	-0.019 *	0.007	0.006	0.043	0.019 **
40s	-0.028*	-0.012	0.014	0.008	0.059*
80s	0.004	-0.009	0.008	0.022*	-0.029 **

** means p<0.01; * means p<0.05

A linear or quadratic regression was performed for each recording to determine the trends in valence of the four recordings (Figure 5). With p<0.05, the valences of the four recordings all changed significantly over time. In the traffic noise recording, the valence went down at around 80 s by 0.023 and then recovered slightly. In terms of natural sounds, the valence increased by 0.038 immediately (within 15s) and then steadily declined (by 0.004 every 20 s) to the initial value. For CP, the valence rose until 60 s, by 0.033, and then fell back to initial value at approximately the same pace. For CN, the valence dropped by 0.020 as soon as the recording started and kept going down (by 0.037) up to 80 s before recovering slightly, similar to traffic noise. As in most previous studies (Hu et al., 2019; Kang et al., 2012), natural sounds had positive effects on facial expression, and traffic noise, negative. In addition, although CP and natural sounds both influence valence positively, and CN and traffic noise negatively, the curves are different. From the results above, FaceReader as a tool for urban soundscape study can be seen not only to show to what extent an acoustic environment is subjectively positive or negative but also to display how facial expression changes over time.

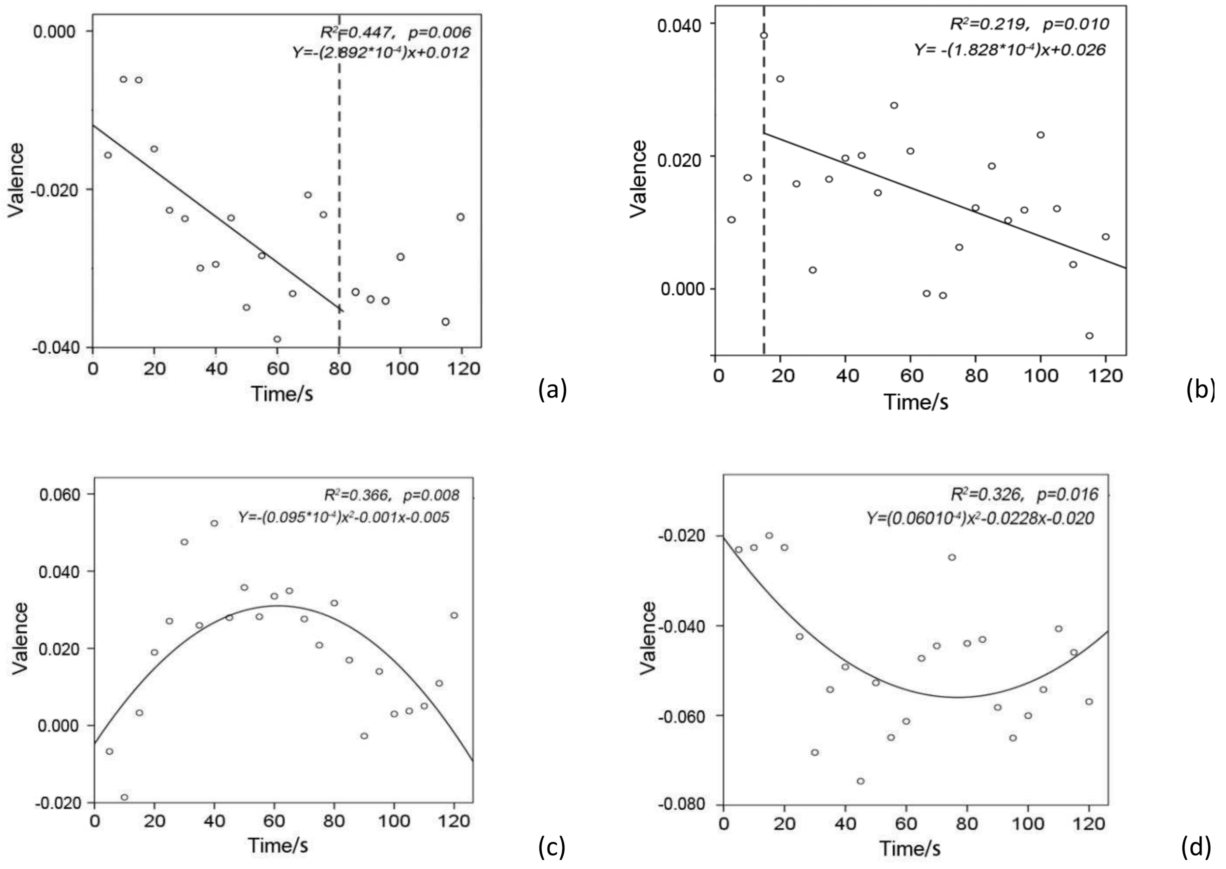
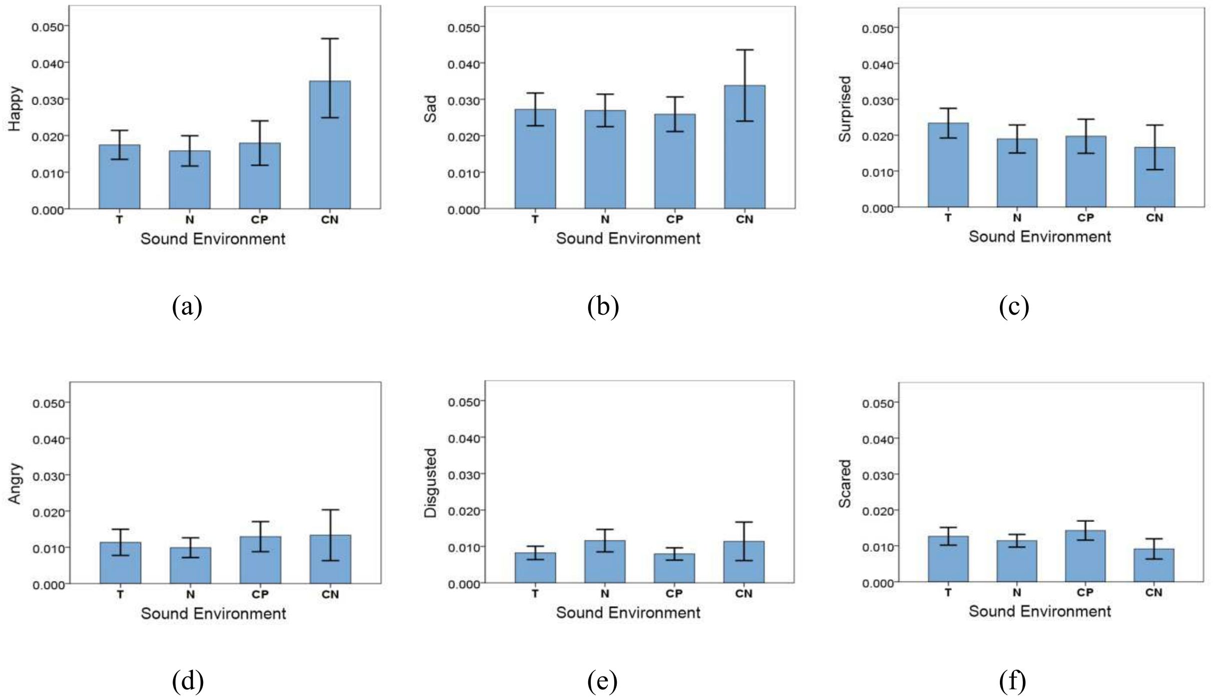


Figure 5. Relationships between listening time and valence in traffic sound (a), natural sound (b), community sound (positive) (c), and community sound (negative) (d).

3.2 Effects of sound perception on facial expression indices

Facial expression can be evaluated based on six indices—happiness, surprise, fear, sadness, anger, and disgust—using a scale from 0 (absent) to 1 (fully present). This section considers the effects of sound perception on these indices.Figure6 shows the average change (absolute value, with 95% confidence interval error bar) in the scores for the respective expressions over the four recordings. Except happy in CN, which is higher than that for the other three recordings, by 0.015, in the other five expressions there is no significant difference between recordings. Expressions that show significant change in all the four recordings are sad (mean=0.028, SD=0.031), happy (mean=0.028, SD=0.037), and surprised (mean=0.020, SD=0.026), while angry (mean=0.011, SD=0.023), disgusted (mean=0.009, SD=0.017) and scared (mean=0.012, SD=0.014) show relatively small, non-significant effects. Therefore, in this section, happy, sad and surprised are chosen for analysing their effect of sound perception.



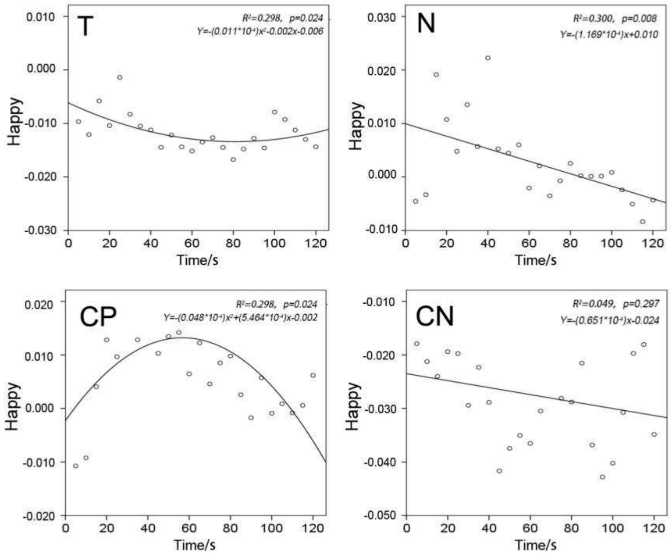
(a) (b) (c)

(d) (e) (f)

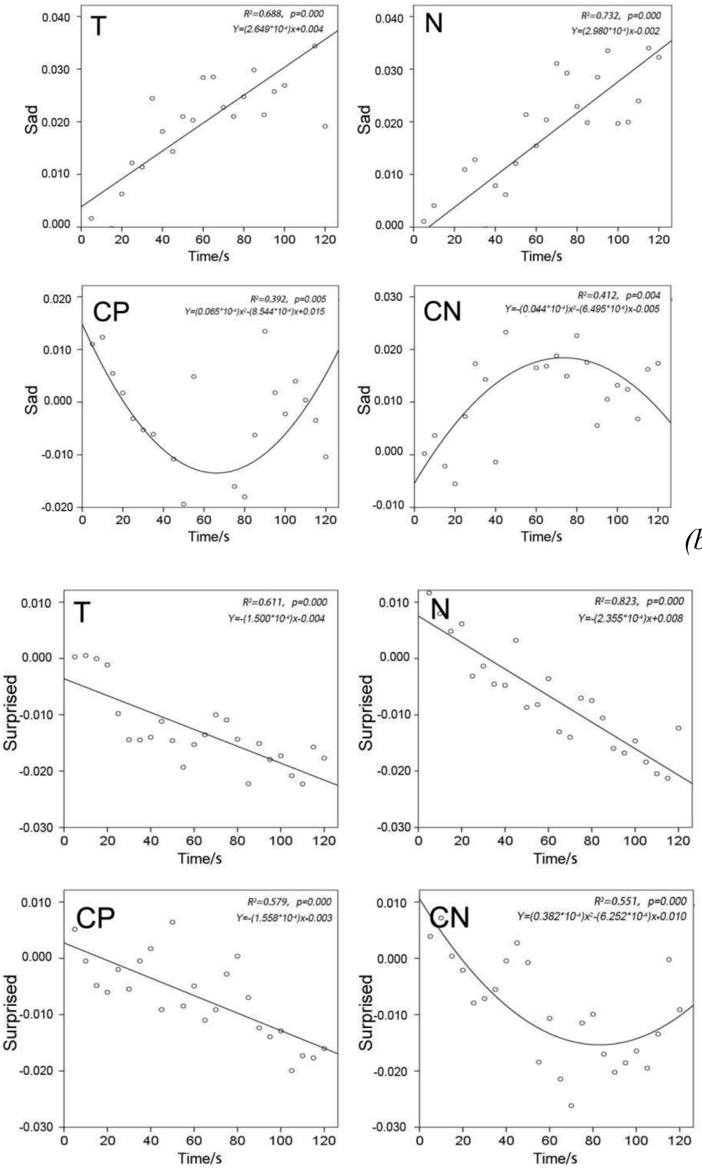
Figure 6. Effect of sound source on different expressions with error bar of 95% confidence interval.

T-traffic noise, N-natural sound, CP-community sound (positive), CN-community sound (negative)

Happy, (b) Sad, (c) Surprised, (d) Angry, (e) Disgusted, and (f) Scared



(a)



(b)

(c)

Figure 7. Relationships between listening time and valence indices for different sound sources

Happy, (b) Sad, and (c) Surprised, T-traffic noise, N-natural sound, CP-community sound (positive), CN-community sound (negative)

Figure 7 shows the results of linear or quadratic regressions of happy, sad, and surprised. In all 12 graphs, the scores of these three expressions are significantly affected by time ($p<0.05$ or $p<0.001$), except happy in CN ($p=0.297$).

It can be seen, in Figure 7a, that The happy expression goes down in traffic noise by 0.012 in the first 80 seconds and then starts to rise slightly. For natural sounds, similar to the trend for valence with this recording, happy expression displays an immediate rise of 0.010 at the beginning and a downward trend (about 0.002 every 20 seconds) thereafter, being still greater than the initial value until 80 s. In the community-positive results for the community recording(CP), happy rises by 0.015 in the first 60 s and then goes down by 0.025 by the end of the recording, while for the negative type, happiness does not change significantly over time ($p=0.297$).

In terms of sad, in Figure 7b, in both traffic noise and natural sound, the sad expression increases with time (0.005 every 20 s) through the whole recordings. In community (positive) sad decreases by 0.029 in the first 60 s and started to rise by 0.024 until the end of the recording. In community (negative), sad rises by 0.025 in the first 60 s and starts to go down by 0.012 until the end of the recording. In community (positive) and community (negative), curves of sadness are both quadratic and the trends are opposite.

In terms of surprised, in Figure 7c, in traffic noise, natural sound and community(positive), surprise goes down linearly from the beginning, at a rate of 0.003,0.005,0.003 every 20 s, respectively. While community (negative) sees an increase after going down by0.026 for the first 80 seconds. The surprised expression decline in all the four cases, it could be implied that with the time of recording goes on, the participants generally feel less and less surprised.

From the results above, happy, sad and surprise are significantly influenced by listening time (p<0.05) in all the sound environments. Therefore, in soundscape studies, thesefacial expression indices could be used for the evaluation of acoustic perception.

3.3 Individual differences

Through independent-samples t-test, gender does not significantly affect the change in valence and the six components of facial expressions, which is consistent with some previous studies in urban soundscape using questionnaire evaluations. In the questionnaire designed for assisting the present study, there are three questions that are about the social characteristics of the participants (2.4).

The relationship between the questionnaire questions and average change in different expressions using Pearson's correlation, reporting Cohen's effect size, is shown in Table 4. Sad expression and arousal are not influenced by social characteristics. Living place has small effects on happy (r=0.127), angry (r=0.116), disgusted (r=0.243), and valence (r=0.104)when listening to the recordings and a medium effect on surprised (r=-0.313); inclination to communication has small effects on surprised (r=-0.132), scared (r=0.123) disgusted (r=0.188), and valence (r=-0.253); and preference for crowds has small effects on happy (r=-0.153), angry (r=-0.151), surprised (r=0.296), scared (r=0.126), and disgusted (r=0.136). From these results, we see that the selected social characteristics generally have small effects on different expressions and valence and do not influence arousal.

Table 4 Relationships between the Social characters and average change in different expressions

Social characters	Happy	Sad	Angry	Surprised	Scared	Disgusted	Arousal	Valence
Distance to nature	0.127*	-0.075	0.116*	-0.313**	-0.077	0.243*	-0.025	-0.104*
Inclination to communication	-0.019	0.041	-0.002	-0.132*	0.123*	0.188*	-0.088	-0.253*
Preference for crowd	-0.153*	-0.030	-0.151*	0.296*	0.126*	0.136*	-0.076	-0.076

*** means large effect, r>0.5; **means medium effect, r>0.3; * means small effect, r>0.1

As mentioned in chapter 2.3, there are two types of curve for the community recording, thus an independent-samples t-test is also performed to see what factors influence which type a participant belongs to (Table 5). In T-test, Cohen's d was calculated to determine effect size (Cohen, 1988); distance from living place to natural landscape and inclination to communication had medium effects on valence (d=0.673 and 0.705, respectively), while preference for crowds had a small effect (d=0.284). From the result, the inclination of the participants to communicate the distance from their living place to natural environment causes the preference for human speech sound, and the preference is also affected by their preference for crowd.

Table 5 The difference in social characters between community (positive) and community (negative).

Social characters	Mean difference	ES(d)
Distance to nature	1.217	0.673**
Inclination of communication	1.167	0.705**
Preference to crowd	0.450	0.284*

** means medium effect, d>0.4; * means small effect, d>0.2

4 Discussion

In order to decide whether and how FaceReader can take the place of questionnaires as a tool in sound perception research, the results of these two methods should be compared. Here, a Pearson's correlation test between participants' evaluations and valence change, with effect size, was conducted (Table 6). Based on the sign of r, valence is negatively correlated with subjective loudness and eventfulness and positively correlated with acoustic comfort, vibrancy, pleasantness, and calmness. In the recording of traffic noise, acoustic comfort (r ranging from 0.351 to 0.547) and eventfulness (r ranging from 0.459 to 0.463) had significant correlations with valence at all three time points, while vibrancy (r from 0.213 to 0.273), pleasantness (r from 0.243 to 0.352), and calmness (r from 0.295 to 0.226) had small correlations with it. Therefore, when studying traffic noise, change in valence can be used to predict acoustic comfort and eventfulness, and can to a certain extent reflect the evaluation of vibrancy, pleasantness, and calmness. As for natural sound, acoustic comfort (r from 0. 029 to 0.164), subjective loudness (r from 0.090 to 0.158), eventfulness (r from 0.067 to 0.136), vibrancy (r from 0.152 to 0.218), and pleasantness (r from 0.157 to 0.315) can all be reflected based on valence from FaceReader. Finally, with community sounds, valence in the first 20 seconds can be used to predict acoustic comfort (r=0.472), subjective loudness (r from 0.212 to 0.580) and pleasantness (r from 0.116 to 0.179), while calmness can be reflected by valence change in the first 20 s (r=0.378).

Table 6 Relationship between the results of semantic evaluations and valence change through 20s, 40s, and 80s of the three recordings.

Sound environment	Listening time	Comfort	Loudness	Eventful-ness	Vibrancy	Pleasant-ness	Calmness
Traffic noise	20 s	0.547***	-0.010	-0.462**	0.273*	0.352**	0.204*
	40 s	0.351**	0.029	-0.463**	0.213*	0.276*	0.195*
	80 s	0.380**	0.084	-0.459**	0.250*	0.243*	0.226*
Natural sound	20s	0.164*	-0.090	-0.161*	0.152*	0.266*	0.030
	40 s	0.029	-0.129*	-0.136*	0.218*	0.315**	0.059
	80 s	0.134*	-0.158*	0.067	0.158*	0.157*	0.070
Community sound	20 s	0.472**	-0.580***	-0.066	0.044	0.179*	0.378**
	40 s	0.008	-0.212**	-0.035	-0.060	0.121*	0.018
	80 s	0.027	-0.002	-0.206*	-0.123*	0.116*	0.167*

*** means large effect, r> 0.5; ** means medium effect, r>0.3; * means small effect, r>0.1

Table 7 The difference in semantic evaluations between community (positive) and community (negative).

Semantic evaluations	Mean difference	ES(d)
Comfort	0.416	0.237^{*}
Loudness	-0.083	-0.065
Eventfulness	-0.361	-0.205^{*}
Vibrancy	1.083	0.649^{**}
Pleasantness	1.594	0.594^{**}
Calmness	1.200	0.447^{*}

** means medium effect, d>0.4; * means small effect, d>0.2

An independent-samples t-test was conducted for the difference in questionnaire evaluations between participants whose facial expression changes showed different trends (positive and negative), and the effect size was also calculated (Table 7). It can be seen that whether people react positively or negatively to community sound has medium effects on their evaluation of vibrancy (d=0.649) and pleasantness (d=0.594) and small effects on acoustic comfort, eventfulness, and calmness (d=0.237, 0.205, and 0.447). From the result, the preference for human speech sound is significantly correlated with acoustic comfort, eventfulness, vibrancy, pleasantness and calmness.

5 Conclusions

This study put forward facial expression recognition as a potential method for sound perception study. Based on a laboratory experiment that involved 32 people, the following conclusions can be drawn.

First, FaceReader is capable of detecting differences in the participants' reactions to different sounds. The valence for natural sounds and community (positive) is significantly different up to 80 s, with mean difference of 0.022 (p<0.05). The valence for traffic sounds and natural sounds decreases with increase of listening time(p<0.01). The valence for community exhibits a parabolic change, which increase followed by decrease in community (positive) and decrease followed by increase in community (negative).

Second, expression indices can also change under the effect of sound perception. Expressions that change significantly under four typical sound environments are sad, happy and surprised, but not angry, disgusted, or scared. Social characteristics generally significantly affect expressions and valence but not arousal. Living place significantly affects happy (r=0.127), angry (r=0.116), disgusted (r=0.243), valence (r=0.104), and surprised (r=-0.313); distance from their living place to natural landscape, inclination to communication, and preference for crowds significantly affect (d=0.673, 0.705, and 0.284 respectively) whether they belong to the community-positive or -negative group.

In terms of comparison between FaceReader and questionnaire survey data, the results show that valence in the first 20 s best represents acoustic comfort (r=0.547) and eventfulness (r=0.463) in the traffic noise condition; valence in the first 40 s best represents pleasantness (r=0.315) in the natural condition; and valence in the first 20 s best represents acoustic comfort (r=0.472), subjective loudness (r=0.580), and calmness (r=0.179) in the community condition.

Since many factors may bring effects on facial expression such as vision, odour and people's emotional status in a real situation, it seems difficult to identify the causal link between the sound with facial expression only. The aim of the present study is just to verify the Effectiveness of FER for Evaluation of Urban Sound Perception. Therefore, to avoid the effects of other factors, the experiment was carried out in the laboratory after collecting sound environment materials. In future study, the possibility and procedures of field study using FER will be explored.

Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) [grant numbers 51878210, 51678180, and 51608147].

Reference

Aletta, F., Lepore, F., Kostara-Konstantinou, E., Kang, J. &Astolfi, A. (2016). An Experimental Study on the Influence of Soundscapes on People's Behaviour in an Open Public Space. *Applied Sciences*. 6(10), 276.

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., Renterghem, T.V., Botteldooren, D. (2018). Can attitude towards greenery improve road traffic noise perception? A case study of a highly-noise exposed cycling path. *Euronoise 2018 - Conference Proceedings.*

Amor, B., Drira, H., Berretti, S. (2014). 4-D Facial Expression Recognition by Learning Geometric Deformations. *IEEE TRANSACTIONS ON CYBERNETICS.*44(12), 2443-2457.

Axelsson, Ö., Nilsson, M. E., Berglund, B. (2010). A Principal Components Model of Soundscape Perception. *The Journal of the Acoustical Society of America.*128 (5), 2836–2846.

Axelsson, Ö., Nilsson, M. E., Hellström, B., &Lundén, P. (2014). A field experiment on the impact of sounds from a jet-and-basin fountain on soundscape quality in an urban park. *Landscape and Urban Planning,* 123(1), 49-60.

Ba, M., Kang, J. (2019). A laboratory study of the sound-odour interaction in urban environments. *Building and Environment.*147, 314–326.

Bartlett, M., Littlewort, G., Frank, M. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 568-573.

Blood, A.J., Zatorre, R.J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences.* 98(20), 11818-11823.

Brown, A.L., Kang,J., Gjestland, T. (2011). Towards standardization in soundscape preference assessment .*Applied Acoustics.* 72(6), 387–392.

Botello, C. (2006). *Adobe Premiere Pro 2.0 Revealed*. Thomson/Course Technology.

Cain, R., Jennings, P., J. Poxon. (2013). The Development and Application of the Emotional Dimensions of a Soundscape. *Applied Acoustics.*74 (2), 232-239.

Christos, V., Moridis, N., Anastasios, A. (2010).Measuring instant emotions during a self-assessment test: the use of FaceReader. *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research. Eindhoven, The Netherlands. August 24 - 27, 2010.*

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Routledge.

Cohen, J. (1992). "A power primer".*Psychological Bulletin.*112 (1) 155–159.

Cootes, T. and Taylor, C. (2000). Statistical models of appearance for computer vision. *Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering, 2000.*

Ekman, P. (1982). Emotion in the human face (2nd ed.). *Cambridge, MA: Cambridge University Press.*

Ekman, P.W., Friesen, V., and Hager, J. C. (2002).*"FACS manual."A Human Face.*

European Environmental Agency.(2014). *Good Practice Guide on Quiet Areas.*

Faul, F., Erdfelder, E., Lang, A., Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods.* 39(2). 175-191.

Feeney, B.C. (2012). A Simple Guide to IBM SPSS Statistics for Version 20.0.*Cengage Learning: Boston, MA, USA,2012.*

Fields, J.M. (1993). Effect of personal and situational variables on noise annoyance in residential areas. *JAcoustSoc Am.* 93, 2753–2763.

Frijda, N.H. (1986). *"The Emotions", Cambridge (UK): Cambridge University Press.*

Gygi, B., Kidd, G.R., Watson, C.S. (2007). Similarity and categorization of environmental sounds *Perception & Psychophysics* 69(6), 839-855.

Hadinejad, A., Moyle, B.D., Scott, N., Kralj, A. (2019). Emotional responses to tourism advertisements: the application of FaceReader. *TOURISM RECREATION RESEARCH.*44, 131-135.

Hardy, E. G. (2012).On the Lex Iulia Municipalis.*The Journal of Philology (Cambridge Library Collection - Classic Journals).*(Volume 35). Cambridge: Cabridge University Press.

Harris, D., Chan-Pensley, J., McGarry, S.(2005). The Development of a Multidimensional Scale to Evaluate Motor Vehicle Dynamic Qualities. *Ergonomics.*48 (8), 964– 982.

Hong, J.Y., Jeon, J.Y. (2013). Designing sound and visual components for enhancement of urban soundscapes. *The Journal of the Acoustical Society of America.* 134 (3), 2026.

Hong, J.Y., Jeon, J.Y. (2014). The effects of audio–visual factors on perceptions of environmental noise barrier performance.*Landsc.Urban Plann.*125 (6), 28–37.

Hu, X., Meng, Q., Kang, J., Han, Y. (2019). Psychological assessment of an urban soundscape using facial expression analysis. *The 49th international congress and exposition on noise control engineering (Internoise 2019).*

ISO 12913-1 (2014). *Acoustics–Soundscape–Part 1: Definition and conceptual framework.*

Jennings, P., Dunne, G., Williams, R., Giudice, S. (2010). Tools and Techniques for Understanding the Fundamentals of Automotive Sound Quality. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering.* 224 (d10), 1263–1278.

Jeon, J.Y., Lee, P.J., Hong, J.Y., Cabrera D., (2011).Non-auditory factors affecting urban soundscape evaluation.*J.Acoust.Soc. Am.* 130 (6), 3761.

Kang, J., Meng, Q., Jin, H.(2012). Effects of Individual Sound Sources on the Subjective Loudness and Acoustic Comfort in Underground Shopping Streets. *Science of The Total Environment*.435,80-89.

Kang, S., De Coensel, B., Filipan, K., Botteldooren, D., (2019). Classification of soundscapes of urban public open spaces. *Landscape and Urban Planning* 189(C), 139-155.

Kelley, K., Preacher, K.J. (2012). "On Effect Size". *Psychological Methods.*17 (2) 137–152.

Lee,P.J., Hong, J.Y., Jeon, J.Y.(2014). Assessment of rural soundscapes with high-speed train noise.*Science of The Total Environment*. 482(1),432-439.

Lee, P.J., Park, S.H., Jung, T., Swenson, A. (2018). Effects of exposure to rural soundscape on psychological restoration. *Euronoise 2018 - Conference Proceedings.*

Leitch, K.A., Duncan, S.E., O'Keefe, S., Rudd, R., Gallagher, D.L. (2015). Characterizing consumer emotional response to sweeteners using an emotion terminology 2 questionnaire and facial expression analysis. *Food Research International.* 76(2), 283-292.

Lewinski, P., Uyl T.M., Butler, C. (2014). Automatic facial coding: validation of basic emotions and facsaus recognition in noldusfacereader. *Journal of Neuroscience, Psychology, and Economics.*7(4) 227-236.

Lipsey, M.W., Wilson, D. (2000). "Practical Meta-Analysis (Applied Social Research Methods)". *SAGE Publications, Inc, 1st edition (August 18, 2000).*

Liu, J., Kang, J., Behm, H. (2014). Birdsong As an Element of the Urban Sound Environment: A Case Study Concerning the Area of Warnemünde in Germany. *ActaAcustica united with Acustica.* 100(3)

Ma, K.W., Wong, H.M, Mak, C. M. (2017). Dental Environmental Noise Evaluation and Health Risk Model Construction to Dental Professionals. *International Journal of Environmental Research and Public Health.* 14(9), 1084.

Mackrill, J.B., Jennings, P.A., Cain, R. (2013). Improving the hospital 'soundscape': a framework to measure individual perceptual response to hospital sounds. *Ergonomics,* 56(11), 1687–1697.

Medvedev, O., Shepherd, D., Hautus, M.J. (2015). The restorative potential of soundscapes: A physiological investigation. *Applied Acoustics.*96, 20-26.

Meng, Q., Kang, J. (2013). Influence of social and behavioural characteristics of users on their evaluation of subjective loudness and acoustic comfort in shopping malls. *PLoSOne.*8(1).

Meng, Q., Kang, J. (2015).The influence of crowd density on the sound environment of commercial pedestrian streets.*Science of The Total Environment*. 511C,249-258.

Miedema, H.M.E., Vos, H. (1998). Exposure-response relationships for transportation noise. *J Acoust Soc Am.* 104, 3432–3445.

Oliver, N., Pentland, A., Berard, F. (2000). LAFTER: A realtime face and lips tracker with facial expression recognition. *Pattern Recognition* 33(8), 1369-1382.

Pérez-Martínez, G., Torija, A.J., Ruiz D.P. (2018). Soundscape assessment of a monumental place: A methodology based on the perception of dominant sounds. *Landscape and Urban Planning.*169, 12–21.

Pheasant, R., Horoshenkov, K., Watts, G., Barrett, B. (2008). The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places? *The Journal of the Acoustical Society of America.* 123 (3), 1446-1457.

Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Dumyahn, Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N. (2011). Soundscape Ecology: The Science of Sound in the. Landscape *Bioscience* 61(3), 250.

Rådsten-Ekman, M., Axelsson, Ö., Nilsson, M. E. (2013). Effects of sounds from water on perception of acoustic environments dominated by road-traffic noise. *ActaAcustica united with Acustica,* 99(2), 218-225.

Sato, W., Hyniewska, S., Minemoto, K., Yoshikawa, S. (2019). Facial Expressions of Basic Emotions in Japanese Laypeople. *Frontiers in psychology.* 259(10).

Schulte-Fortkamp, B. (2002). How to measure soundscapes: A theoretical and practical approach. *The Journal of the Acoustical Society of America.* 112(5), 2434-2434.

Viola, P., Jones, M.J. (2001).Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, U.S.A., December, 2001.*

Yang, C., Li, H. (2015). Validity Study on Facereader's Images Recognition from Chinese Facial Expression Database. *Chinese Journal of Ergonomics.*21(1), 38-41.

Yang, M., Kang, J. (2013).Psychoacoustical evaluation of natural and urban. *The Journal of the Acoustical Society of America.* 134(1), 840-851.

Yang, W., Moon, H.J., Kim, M. (2018). Perceptual assessment of indoor water sounds over environmental noise through windows. *Applied Acoustics.*135, 60–69.

Yi, F., Kang, J.(2019). Effect of background and foreground music on satisfaction, behavior, and emotional responses in public spaces of shopping malls. *Applied Acoustics.*145, 408–419.

Zarbakhsh, P., Demirel, H. (2018). Low-rank sparse coding and region of interest pooling for dynamic 3D facial expression recognition. *Signal Image and Video Processing.* 12(8), 1611-1618.

Zhang, Y., Yang, Z., Yu, X. (2006). Measurement and evaluation of interactions in complex urban ecosystem. *Ecological Modelling.* 196(1), 77–89.