Auditory sensory saliency as a better predictor of change than sound amplitude in pleasantness assessment of reproduced urban soundscapes

Karlo Filipan, Bert de Coensel, Pierre Aumond, Arnaud Can, Catherine Lavandier, Dick Botteldooren

Auditory Sensory Saliency as a Better Predictor of Change than Sound Amplitude in Pleasantness Assessment of Reproduced Urban Soundscapes

Karlo Filipan [a,*], Bert De Coensel [a], Pierre Aumond [b,c], Arnaud Can [b], Catherine Lavandier [c], Dick Botteldooren [a]

[a]Waves Research Group, Department of Information Technology, Ghent University, Technologiepark-Zwijnaarde 15, 9052 Ghent, Belgium; karlo.filipan@ugent.be (K.F.); bert.decoensel@ugent.be (B.D.C.); dick.botteldooren@ugent.be (D.B.)

[b]IFSTTAR, CEREMA, UMRAE, F-44344 Bouguenais, France; pierre.aumond@ifsttar.fr (P.A.); arnaud.can@ifsttar.fr (A.C.)

[c]ETIS, UMR 8051,University of Paris Seine, University of Cergy-Pontoise, ENSEA, CNRS, F-95000, France; catherine.lavandier@u-cergy.fr (C.L.)

[*]Corresponding author: Karlo Filipan, karlo.filipan@ugent.be; Ghent University, Department of Information Technology, Research group WAVES, iGent Technologiepark-Zwijnaarde 15, 9052 Gent, Belgium; Tel: +32-9-264-3325, Fax: +32-9-264-3593

Abstract

The sonic environment of the urban public space is often experienced while walking through it. Nevertheless, city dwellers are usually not actively listening to the environment when traversing the city. Therefore, sound events that are salient, i.e. stand out of the sonic environment, are the ones that trigger attention and contribute highly to the perception of the soundscape. In a previously reported audiovisual perception experiment, the pleasantness of a recorded urban sound walk was continuously evaluated by a group of participants. To detect salient events in the soundscape, a biologically-inspired computational model for auditory sensory saliency based on spectrotemporal modulations is proposed. Using the data from a sound walk, the present study validates the hypothesis that salient events detected by the model contribute to changes in soundscape rating and are therefore important when evaluating the urban soundscape. Finally, when using the data from an additional ex-

periment without a strong visual component, the importance of auditory sensory saliency as a predictor for change in pleasantness assessment is found to be even more pronounced.

Keywords: Auditory Saliency; Sensory Saliency; Listening Experiment; Computational Model; Granger Causality

1 Introduction

People often experience an urban public space while walking through it. The perceived quality of walking routes through the urban environment may affect their usability and thereby promote an active lifestyle [1]. Pleasant routes may also promote the choice of walking as a travel mode [2] and thereby help reducing inner city car traffic and its negative influence on the living environment. Walking through an agreeable environment may even become a mentally restoring activity [3]. The soundscape—the sonic environment as perceived or experienced and/or understood by a person or people, in context [4]—is part of this experience [5] and is important for the perceived quality of the walking routes. The sonic environment itself relates to the physical (acoustic) environment which constitutes the sound at the receiver from all sound sources as modified by the environment [4].

It has been shown that sounds that are noticed influence soundscape perception [6]. While walking through an urban environment, people in general pay little attention to details in their surroundings unless asked to do so [7]. Most environmental sounds may therefore remain unnoticed and hence would not contribute to the cognitive appraisal of the sonic environment [8]. Subliminal environmental sound might still contribute to the overall affect, emotion, and stress but would not trigger conscious changes in pleasantness rating [9].

However, some sound events have higher probability to be noticed depending on how much they stand out of the sonic environment. The term saliency is used to refer to the degree to which an event stands out of the environment [10]; therefore, such sounds are deemed salient. Correspondingly, salient sound events trigger people's attention and evoke a reaction depending on cognitive appraisal: from the fight-or-flight response to fast approaching car honking to the appreciation of the bird singing in the tree [11].

Auditory saliency could be divided into two non-excluding dimensions: sensory and semantic saliency. Sensory saliency is determined by the enhanced sensitivity or tuning of the human hearing system to specific sound features [12]. On the other hand, semantic saliency requires recognition of the sound and incongruency within the environment [13]. Sensory saliency has been investigated by explicitly identifying features that alter behavior [12] or by inspection of the spectrogram using methods similar to the ones used to model visual saliency [14].

The tuning of the auditory system can nowadays be measured using several brain imaging techniques. Therefore, responses and findings obtained using brain imaging can serve as a basis for the features used for calculating auditory

saliency. In [15] it was shown that tonotopically-localized regions of the brain respond to spectrotemporal modulations, i.e. ripple sounds that have simultaneous modulation in amplitude and frequency domain. In this study we therefore use a biologically-inspired computational model for auditory sensory saliency which evaluates the similarity of the input to spectrotemporal modulations.

This paper explores the hypothesis that sensory salient events trigger changes in the appraisal of the sonic environment. Two laboratory experiments, an audiovisual and an audio-only experiment, in which people continuously rated the pleasantness of the sonic environment recorded during a city walk [16], are used for verifying this hypothesis. More specifically, this paper evaluates if the sensory saliency, computed using the proposed model, increases the probability that the participants in the experiments will change their pleasantness rating of the sonic environment.

In Section 2, the audiovisual listening experiment carried out in a laboratory context is briefly presented and the obtained dataset and calculated metrics are discussed. In Section 3, the layout and the implementation of the computational saliency model are discussed. Section 4 outlines the metrics and the statistical methods used in the analysis and presents the results for the audiovisual experimental dataset. Finally, using the same methodology, the results for the audio-only experimental dataset are presented and discussed in Section 5.

2 Evaluating pleasantness of a sound walk

A listening experiment was conducted where participants were asked to continuously assess the pleasantness of the sound environment which was a recorded soundwalk. Five paths covering Paris boulevards, streets, passageways and parks were chosen as a representative of the urban sound environment. The walks were recorded in both directions to capture the transition between the environments. For the complete explanation of the experiment we refer to [16].

There were 30 participants in the experiment. They were recruited inside the university (students and university staff) with no relation to any soundscape study. The group consisted of 18 women and 12 men, with a mean age of 33 years (SD = 14). The participants were naive to the tested hypotheses, and received a small monetary compensation for participation. Prior to the experiment, all of the participants gave their informed written consent.

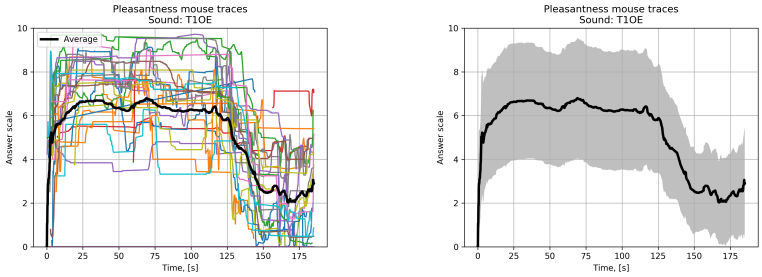
The same experiment was also carried out in situ, walking on the same paths, asking for pleasantness at different locations [17]. The use of audiovisual stimuli was preferred above the use of only sound stimuli, in order to be able to compare the real pleasantness (measured in situ) with the pleasantness measured in laboratory. The results were found to be comparable, and the continuous pleasantness measured in laboratory has been considered ecologically valid.

Ten recorded sound scenes corresponding to urban walks in Paris through different types of areas (boulevards, streets and park) were played back inside a semi-anechoic room through a transaural system. The sampling frequency used was 48 kHz and the duration of each scene was 185 s. The recorded sounds were

reproduced with the corresponding videos on a large screen in the laboratory. Participants evaluated the sound pleasantness on a continuous scale by moving a marker bar with the mouse. The sampling frequency of the mouse movement measurement was 8 Hz (1480 samples for 185 seconds sound duration).

2.1 Perceptual dataset of pleasantness rating

The perceptual dataset consisted of 300 collected mouse traces from each combination of participant and sound. In order to create a single trace for each sound sample from 30 participants, the data was averaged. As it can be seen in Figure 1, the reaction time response varied highly between participants; therefore, an average metric would not provide a relevant reference and the data of each single participant would need to be considered. Furthermore, it was found that some of the traces were broken when the mouse was not positioned above the scale. This was later accounted for by retaining only the participant-sound combinations that had less than 25% of missing data.



(a) (b)

Figure 1: Recorded trace of answers for sound sample T1OE: (a) raw data from all 30 participants and (b) average data with standard deviation. Missing data from the traces was not taken into account when calculating the average.

2.2 Probability of change in pleasantness rating

The perceptual evaluation data consisted of slider traces as shown in Figure 1. As it can be seen, these included long periods without participant's reaction. As hypothesized, the change of the pleasantness rating would happen when the participant's attention is triggered by the salient event [18]. Therefore, a metric was created representing the probability of the change in the pleasantness rating.

This metric was calculated by taking the absolute value of the finite difference of the pleasantness rating ψ. This difference was calculated for each of the samples in the signal of the length D equal to 1480 samples (Equation 1). Additionally, the difference signal was also rectified with threshold T equal to 0.001, i.e. very small changes were not taken into account.

Due to the nature of the numerical difference and in order to maintain the same length as the original signal, the initial sample was set to have a difference equal to zero (Equation 1). Such calculation process produced a non-avoidable 62.5 ms delay (half a period of the 8 Hz sampling frequency) between the change signal and the original pleasantness rating.

$$d\psi|_{n\in\{1,2,\ldots,D\}}[n] = \begin{cases} 1, \ |\psi[n] - \psi[n-1]| > T \\ 0, \ otherwise \end{cases} \\ d\psi[0] = 0 \tag{1}$$

The probability of the pleasantness rating change ($P_{d\psi}$) was subsequently calculated with a sliding window w across the signal $d\psi$ (Equation 2). The duration of the sliding window was 2 seconds (L equal to 16 samples) with a step of 250 ms (S totaling 2 samples), thus making the sampling frequency of the final output equal to 4 Hz.

$$P_w(d\psi) = \frac{\sum_{n=w}^{w+L} d\psi[n]}{L} \\ w \in \{0, S, 2S, \ldots, D-L\} \tag{2}$$

Furthermore, if the signal inside a sliding window consisted of more than 50% of missing data, this portion of the output was labeled as missing. Finally, the mask (0 – perceptual data present, 1 – no data recorded) was stored in order to remove the same portions when comparing probability of change to other metrics.

2.3 Recordings of the sonic environment

The sound samples used in the listening experiment were 10 recorded urban sequences of three minute duration, recorded with a binaural system (two omnidirectional microphones inserted in the ears of the experimenter) during five trips traveled in both directions in the 13^{th} district of Paris, in April 2015.

The left and right channels were recorded with slightly different sensitivities which needed to be accounted for. As the absolute value was not important in the analysis, a relative (mutual) calibration was performed. Therefore, the left and right channel of the sound file were equalized by fixing the lower level channel (left in this case) and adjusting (lowering) the higher level channel by the ratio of the root-mean-square values of the calibration signals.

2.4 Sound pressure amplitude

The amplitude of the sound pressure was calculated on the mutually calibrated left and right channels separately to emulate the saliency calculation procedure (Section 3.5). No calibration of the absolute sound pressure was performed since the relative pressure was sufficient for the performed statistical analyses. The

extraction of the sound pressure signal started from the monaural signal which was firstly weighted with an A-weighting digital filter. The A-weighting was chosen as it is a standardized modification of the sound signal based on human perception that is used extensively in environmental noise and soundscape studies [19, 20, 21].

In the next step, the equivalent level was calculated using a time window with a length of 6000 samples to achieve the same 8 Hz sampling frequency as the pleasantness rating (Section 2.1). Left and right levels were then logarithmically summed to a single level and converted back to sound pressure amplitude. Finally, a windowing procedure was applied—an average sound pressure was calculated inside a 250 ms window which was shifted with a 250 ms time step until the end of the signal.

3 Computational model for auditory saliency

A biologically-inspired computational model for auditory saliency is created for calculating the sensory saliency of the input sound. The model is comprised of two stages: the auditory periphery stage and the brain (central) processing stage. The input sound is fed first to a simplified auditory periphery model that uses Gammatone periphery [22] and simulates the peripheral processing up to the level of the auditory brainstem [23]. The output of the periphery model at each tonotopic region, i.e. central frequency, is then used as an input to the simulation of brain processing: spectrotemporal modulation content reacted to in the auditory cortex [24] followed by a sensory activation stage based on leaky integration [25].

Although most recent research employs spectrotemporal modulation features in models that focus exclusively on speech [26, 27], previous models also investigated sensory saliency of mixtures of environmental sounds including noise, animal sounds, music, sirens, etc. [28, 29]. Moreover, studies that investigated the perception of spectrotemporal modulations have analyzed complex mixtures of sounds (harmonic complexes, tones in noise, amplitude and frequency modulated tones, etc.) [30, 31, 32], although often with a focus on the analysis and reconstruction of speech signals. However, most urban sounds from sources such as road traffic, birds or airco units, have their dominant frequency contributions in the same frequency range as speech.

The proposed model evaluates the spectrotemporal content in all frequency bands evaluated by the periphery model, therefore, not focusing only on the range specific to human vocalizations. Similarly to previously proposed models that extract indicators of temporal structure of the soundscape in relation to music [33], this model extracts the spectrotemporal modulation content from the acoustic input as a relevant soundscape indicator of the tuning of a human brain to spectrotemporal modulations [24, 15].

3.1 Spectrotemporal modulations

The created saliency model is inspired by the observation of the sensitivity of the human auditory cortex to spectrotemporal modulations [15, 30, 32]. Ripples, i.e. sound signals that have simultaneous sinusoidal modulation in time and frequency domain, can be considered as prototypes of this spectrotemporal modulation. Using functional magnetic resonance imaging (fMRI), it was shown in [15] that they excite particular, spatially separate regions in the auditory cortex.

The modulation function $M(t, x)$ of such ripple sound is shown in Equation 3. There, modulations are given as amplitude modulation (AM) ω on time axis t, frequency modulation (FM) Ω on octave band axis x with the modulation depth Δm usually set to 1.

$$M(t, x) = 1 + \Delta m \cdot sin(\omega t + \Omega x) \tag{3}$$

For the creation of ripple sounds in the digital domain, the frequency axis f is discretized with N_f discrete frequencies, according to the number of octaves N_{oct} calculated from the start and end frequency, with the number of separate divisions of octave band axis x as shown in Equation 4.

$$\begin{aligned} x &\in \left\{0, \frac{1}{N_x}, \frac{2}{N_x}, \ldots, N_{oct}\right\} \\ N_{oct} &= \left\lceil log_2\left(\frac{f_{end}}{f_{start}}\right)\right\rceil ;\ N_x = \frac{N_f}{N_{oct}} \end{aligned} \tag{4}$$

With this, the term Ωx in Equation 3 becomes a constant and the carrier on a frequency band c is only an amplitude modulated function $S_c(t)$ shown in Equation 5. There, x_c is given as an octave band number on a frequency band f_c, corresponding to one value from the discretized octave band axis in Equation 4, and β_c is a randomly selected phase. Finally, the complete ripple sound $S(t)$ is calculated by summing over all the AM modulated frequency band signals as displayed in Equation 6.

$$S_c(t) = M(t, x_c) \cdot sin(2\pi f_c t + \beta_c) \tag{5}$$

$$S(t) = \sum_{c=1}^{N_f} S_c(t) \tag{6}$$

3.2 Auditory periphery stage

The saliency model is created to be able to separate the spectrotemporal content from the input sound into ripples. Firstly, the input sound (sampled at 48 kHz) is passed through a filter simulating the outer/middle ear transfer function. This is modeled using a band-pass filter between 0.6 and 4 kHz simplifying the transfer functions of the human middle ear [34]. The next step, which represents

cochlear processing, is modeled using a filterbank of N_f filters. This filterbank consists of 120 Gammatone filters [35] created across the hearing range with central frequency division according to [36]. The central frequencies of these filters provide the basis f_c values for the Equation 5 as well as the later stages of the model.

The final step of the auditory periphery model is a simplification of the brainstem response. The input signal to the brainstem stage is demodulated by squaring the content from each frequency band, in a manner similar to simple nonlinear functions used in other periphery models [37, 30]. According to [38], the demodulation procedure positions the AM content around the DC value and therefore, in combination with low-pass filtering, it could be separated from the rest of the signal contents. Finally, the output signal is also downsampled which is allowed according to the Nyquist theorem. The parameters used for demodulation and downsampling in this study were a cut-off frequency of 152 Hz for the low-pass filter and an output sampling frequency equal to 320 Hz. This cut-off frequency is high enough to cover the main peak in the frequency response of the inferior colliculus neurons [39].

3.3 Auditory cortex stage

The auditory cortex stage is modeled based on the analysis of the amplitude and frequency modulation content from the signal according to basic ripples (Equations 3-5). Since the ripple can be expressed with a separate AM and FM portion, the auditory cortex stage is based on two simulation stages: detection of AM using resonator filters and FM using carrier frequency dependent time delays.

Firstly, the AM content is extracted using a filterbank of constant-gain single-pole resonator filters. The transfer function of such filters in discrete-time Z-transform domain is presented in Equation 7. The parameters of the filter are the resonator pole radius R_p and the resonance frequency F_r, where F_s equals the sampling frequency of the input. In this study, a filterbank was created with 10 resonator filters all having the pole radius R_p equal to 0.0001. This assured that the actual pole radius R_r was close to the minimum possible which maximally reduces the duration of the impulse response. The resonance frequencies of the filters were logarithmically spaced from 1 to 10 Hz, i.e. ω was spaced from 2π to 20π cycles.

$$\begin{aligned} H(z) &= G\frac{1-z^{-2}}{1-2R_r\cos(\Theta_r)z^{-1}+R_r^2z^{-2}} \\ G &= \frac{1-R_r^2}{2} \\ R_r &= 1-\frac{2\pi}{F_s}(1-R_p) \\ \Theta_r &= \frac{2\pi}{F_s}F_r \end{aligned} \tag{7}$$

For each AM/FM combination the time delay $\delta_c(\omega, \Omega)$ on frequency band c is calculated according to Equation 8. This enables calculating the first step of FM using buffers with a length equal to the time delay in each of the frequency bands f_c.

$$\delta_c(\omega, \Omega) = \frac{\Omega}{\omega} \cdot x_c \tag{8}$$

Now, consider a ripple sound with AM ω_0 cycles and FM Ω_0 cycles/octave (Section 3.1) as an input to the auditory periphery stage of the saliency model (Section 3.2). The (perfect) output of the auditory periphery would then consist of a sinusoid on each frequency band f_c with angular frequency ω_0 and delay equal to $\delta_{c,0}(\omega_0, \Omega_0)$ (Equation 8).

We then analyze every signal on f_c only with a single resonator (Equation 7) which has $2\pi F_r$ equal to ω_0 and R_p equal to 1. The output is then fed to the N_f buffers each with a delay corresponding to $\delta_{c,0}$. The output consists of sinusoids with angular frequency ω_0 and delays $\delta_{c,0}$ over frequencies f_c. These delayed sines are then fed to the overlapping summation which is shown in Equation 9. As it can be seen, for a single output band q a window of length U is applied over the frequency band axis c and the time signals (sines in this case) are summed together. Afterwards, the window is shifted to the next frequency band with step T and the summation is repeated.

$$\begin{aligned} V_q &= \sum_{c=q+1}^{q+1+U} sin(\omega t + \delta_c) \\ q &\in \{0, T, 2T, \ldots, N_f - U + 1\} \end{aligned} \tag{9}$$

This procedure reduces the number of output bands to N_{out} calculated from the number of frequency bands N_f, window size U and step T. In this study, an overlapping summation over one octave band with a half-octave step was used. As there were 120 frequency bands coming from the auditory periphery stage with 12 bands inside each octave, the number of output bands amounted to 19.

$$N_{out} = \left\lceil \frac{N_f - U + 1}{T} \right\rceil \tag{10}$$

It can be proven mathematically that the summation of any number of sinusoidal signals of the same angular frequency gives either another sinusoid with this angular frequency or no response because of the cancellation due to opposite phases. This is the theoretical baseline of the final calculation step of the auditory cortex stage—calculation of the amplitude across the buffer with the length corresponding to ω. In turn, this step ensures that the rippling effect in the output is smoothed to a constant value. When the modulation content of the input coincides with the delay, as in our example with ω_0 and Ω_0 corresponding to delay $\delta_{c,0}$, the output produces the highest value.

In this study, five frequency modulations were selected with linear spacing from 0 to 1 Hz/octave, i.e. Ω was spaced from 0 to π cycles/octave. To reduce

the long delays that can occur for high values of Ω/ω, the modulation axis x in Equation 4 was wrapped. Since the delays in the first octave are the smallest due to the lowest values of x_c, they were repeated for all other octaves according to Equation 11.

$$\begin{aligned} x_{oct} &\in \left\{0, \frac{1}{N_x}, \frac{2}{N_x}, \ldots, 1\right\} \\ N_{oct} &= \left\lceil log_2\left(\frac{f_{end}}{f_{start}}\right)\right\rceil ;\ N_x = \frac{N_f}{N_{oct}} \\ x &\in \{x_{oct}, \ldots, x_{oct}\} \end{aligned} \tag{11}$$

To summarize, in the auditory cortex stage the input, i.e. the modulation content of the signal, is expanded across ripple-like functions while the output of any of the AM/FM combinations peaks if the input is a matching spectrotemporal modulation.

3.4 Sensory activation stage

The sensory activation stage simulates the excitation and inhibition processes in the brain which are found to be important for human attention and gating [40, 41, 42]. To model the excitation and inhibition in the sensory activation stage of the model [43], a leaky integrator implementation is used [44]. The mathematical expression of leaky integration in discrete-time domain is provided in Equation 12. As it can be seen, the current output signal z_{out} on time step n depends on the previous output on step $n-1$ and the α_i portion of the difference between the current input z_{in} and the previous output.

$$z_{out}[n] = z_{out}[n-1] + \alpha_i \left(z_{in}[n] - z_{out}[n-1]\right) \tag{12}$$

Equation 13 shows the mathematical expression for calculating the value α_i. It is calculated based on the time constant for integration time τ_i and sampling frequency F_s which is a measure of time spacing between the two samples in discrete-time domain. Additionally, different time constants τ_i could be given for rise and fall of the signal, therefore α_i would change if the difference between the current input and previous output in Equation 12 is greater or lower than zero.

$$\alpha_i = 1 - e^{-\frac{1}{F_s \tau_i}} \tag{13}$$

In the model, excitation and inhibition are determined according to the expressions shown in Equation 14. Firstly, excitation e is calculated on the input a_{in}. The current excitation therefore depends on the current input evaluated with leaky integration (Equation 12). Afterwards, inhibition g is evaluated on the excitation signal also with the leaky integration function, however with different values of α_i (Equation 13). Corresponding to the previous work [45, 46], the values of τ_i used for excitation were 0.05 s and 2 s for rise and fall time

respectively. Additionally, inhibition time constants were given as 1.8 s and 10 s for rise and fall time respectively.

$$\begin{aligned} e[n] &= leaky(a_{in}[n]) \\ g[n] &= leaky(e[n]) \end{aligned} \tag{14}$$

The effect of the sensory activation stage a_{out} is determined as an interplay between the excitation and inhibition and calculated according to Equation 15. Firstly, the difference between the excitation signal e and the inhibition signal g delayed by m samples and multiplied with a constant K is calculated [46]. Secondly, the output is rectified using maximum rectification, an approach found in neuronal circuits of the neocortex [47] and widely used in neural network research [48]. In this study, the buffer length m was set to be 3 samples (9.375 ms, i.e. an approximation of 0.01 s up to a sampling frequency step) and no multiplication was applied (K equal to 1) [46].

$$a_{out}[n] = \max(e[n] - K \cdot g[n - m], 0) \tag{15}$$

The computational saliency model generates several output values, i.e. on each time sample $N_{AM} \times N_{out} \times N_{FM}$ values are calculated. As noted previously, we used 10 amplitude and 5 frequency modulations, as well as the 19 output bands in the model. These 950 separate outputs were compressed into a single value on each time sample using a simple summation. This ensured that, when evaluating the continuous input to the model, the output saliency becomes a single-valued time signal Y.

3.5 Saliency computed from the recordings

The saliency model does not have an implementation of binaural hearing characteristics, therefore, left and right channel for each sound were evaluated separately. Consequently, the saliency signal Y was computed for the left Y_{left} and right Y_{right} channels separately which needed to be combined into a single saliency value. The saliency computed from the model implementation discussed in previous paragraphs can be in the range $[0, \infty)$. Therefore, a sigmoid function [49] was used to confine the saliency value into $[0, 0.5)$ as shown in Equation 16. With the implementation and parameters listed in Sections 3.2, 3.3 and 3.4, the calculations were performed faster than real time on a conventional PC.

$$Y = \frac{1}{1 + e^{-(Y_{left} + Y_{right})}} - 0.5 \tag{16}$$

As discussed, the output of the computational model was produced with a sampling frequency of 320 Hz. This enabled capturing the complete range of the amplitude modulations as well as the transients in the response over time. However, the computed saliency was additionally downsampled to 8 Hz to be directly comparable to the dataset of pleasantness ratings (Section 2.1). Finally, the average saliency was calculated within a moving window to enable a direct comparison to the probability of the change in pleasantness rating (Section

2.2). Conversely to the probability which predicts the response in the next two seconds (2 s time window), saliency was calculated within a window of 250 ms while the step between the windows was the same 250 ms, i.e. a sampling frequency of 4 Hz. This windowing procedure also directly corresponded to the calculation of the sound amplitude (Section 2.4).

4 Predicting change in pleasantness assessment

This study investigates whether salient sounds in the environment trigger changes in the appraisal of the sonic environment. On the one hand, people's appraisal of a walk in the city was measured using a continuously monitored pleasantness rating within the laboratory experiment (Section 2). On the other hand, the saliency of the sound recorded during the walk was calculated using a proposed biologically-inspired computational model for auditory saliency (Section 3).

In order to test the hypothesis that the salient events as evaluated by the computational model are the ones that trigger changes of pleasantness rating, an additional metric was used: A-weighted sound pressure amplitude (Section 2.4). Therefore, the analysis involved three types of signals: the probability of change in pleasantness assessment, the computed saliency and the sound pressure amplitude (Figure 2). Finally, using the same approach, the relationship between sound pressure and computed saliency was also investigated.

The first comparison was between the computed saliency and the probability that people change their pleasantness rating when listening to the same sound. Additionally, sound pressure amplitude as a predictor for the change was analyzed in order to reject the hypothesis in case it was found to be a better predictor for the change in rating than the saliency itself. Finally, the potential existence of a prediction relationship between sound amplitude and saliency was investigated.

Each combination of three signals was evaluated separately for each participant-sound combination and an example of these data is shown in Figure 3.a and b. It should be noted that the saliency and amplitude signals were the same for each sound, nevertheless, the probability of change varied across participant and sound. For example, sometimes a participant changed his/her answer constantly over a period of time as shown by the probability of change equal to one between 75 and 100 seconds in Figure 3.a. On the other hand, data for a different participant displays only a small probability of change over time (Figure 3.b between 25 and 100 seconds) since the participant was changing the rating only sometimes and during a long period of time.

Two additional datasets are represented in Figure 3 to compare the sound events in the recordings. Firstly, the spectrograms of the left and right channel of binaural recordings which were used to calculate the sound amplitude trace (Section 2.4) are shown in Figure 3.c and d. The other dataset comprises the sound events which were labeled by listening to the binaural recordings. Additionally, the transition between the recorded environments is also shown in the top banners in Figure 3.e and f.

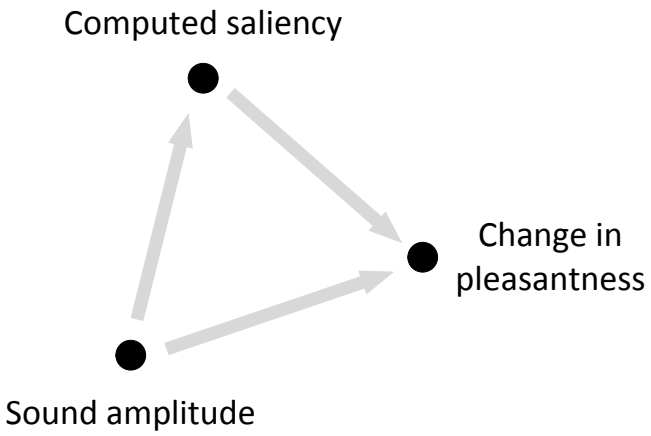


Figure 2: Three signal types used in the prediction analysis: output of the computational saliency model (Computed saliency), probability of the change in participant's reaction (Change in pleasantness rating) and time evolution of the A-weighted sound pressure amplitude (Sound amplitude).

For the sound named T3OE, i.e. the recording that starts in a small street and finishes in a large boulevard, there are several labeled events that can be seen in the calculated saliency trace. In the first place, the sounds of a meal (clanking of the cutlery and plates, voices, etc.) coming from the building above the recorded path are recognized in the model while the participant 7 also changed the pleasantness rating in this time frame. Furthermore, the squeaking around 50 second time stamp is also recognized by the model as a salient event, however, this participant did not change the pleasantness rating then. The highest saliency comes from a loud car pass-by before the 100 second mark which is also easily recognizable in the spectrogram. Finally, the difference between the calculated saliency and the level could be seen from the second 140 onward where the level stays up while the saliency model is reacting to individual sound events (speech, motorbike and squeaking brakes).

Sound T5OE covers the transition between the three environments: the recording started in a boulevard, then continued in a small pedestrian passageway until the end in another boulevard. When listening to the sound recording, the first clearly noticeable event is a sound of a woman walking in high heels. This is also reflected in the change of the rating of participant 16. Similar to the recording T3OE, in the sound T5OE between seconds 25 and 75, loud traffic noise is also visible in the sound amplitude and the spectrogram but not in the calculated saliency trace. Several events are recognized by the saliency model in this time frame—one of which is a bird around the second 70. Moreover, a honk of a car is a clearly dominant event in the calculated saliency trace around

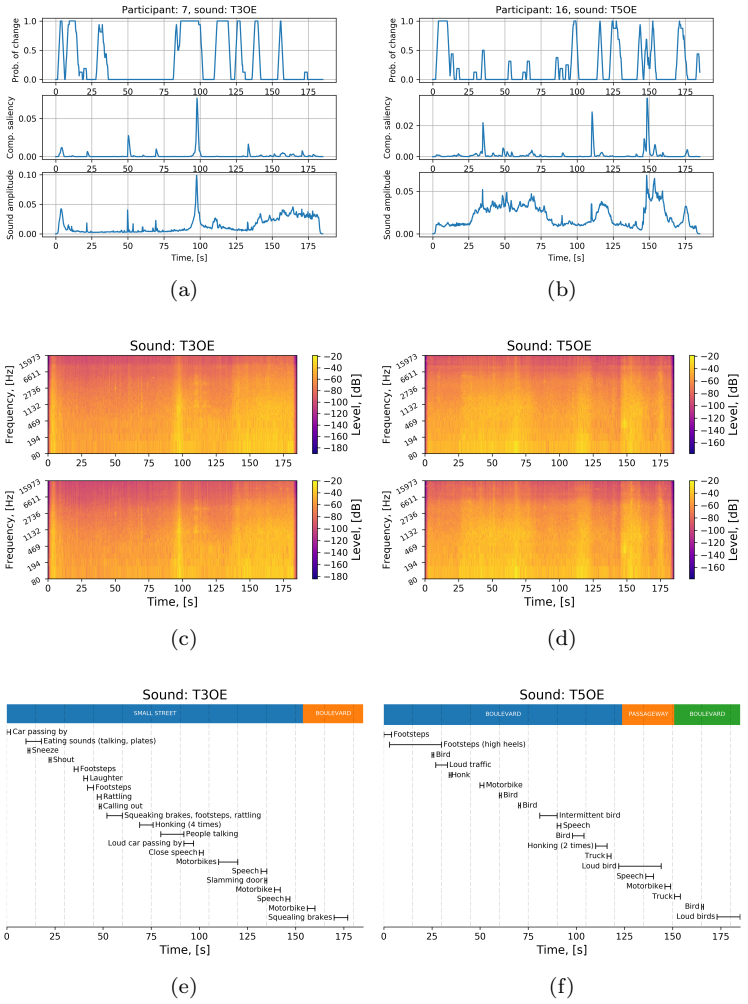


Figure 3: (a) and (b): Example of signal traces for the three types of data (probability of change in pleasantness rating, saliency calculated from the computational model and sound pressure amplitude) shown for two participant-sound combinations. (c) and (d): Spectrogram of the binaural recordings (left channel in top graph, right channel in bottom graph) corresponding to the sound amplitude signal trace above. (e) and (f): Markings of the recorded environment (filled banner) and sound events labeled from listening to the binaural recordings.

the second 110. However, loud bird around the second 125 is not recognized by the model as a salient event even though this event produced the change in the

pleasantness rating of participant 16. Contrary to this, birds singing around the second 175 are marked as salient by the computational model.

4.1 Granger causality

To assess the hypothesis that the auditory saliency computed by the proposed saliency model can be used as a predictor for changes in the appraisal of the sonic environment, a Granger causality analysis was performed [50]. Granger causality is a measure of one signal being predicted by another: if a signal X_1 "Granger-causes" a signal X_2, then past values of X_1 should contain information that helps to predict X_2 above and beyond the information contained in past values of X_2 alone [51].

Granger causality has been used extensively in econometric studies [52] and in the recent years to analyze brain imagining datasets [53]. Additionally, some studies related to sound (music) also analyzed other types of data using Granger causality: D'Ausilio et al. investigated causal relationship among musicians using the recorded movement kinematics during an execution of a musical piece [54], while Dean et al. studied continuously rated perception of arousal in relation to the varying intensity change of a musical piece [55]. Similarly, this study relates the perceptual data from a listening experiment to the varying indicators (saliency and sound amplitude) extracted from the listened sounds.

In this study, a Granger causality analysis was performed for each participant-sound combination between the metrics shown in Figure 2. Granger causality lag (LAG) was set to be up to 500 ms, a representative of delays found in brain imaging studies [56, 57]. As the exact reaction time of the participants was unknown, an additional time shift between the signals (SHIFT) was included in the analysis. To remove the risk of not covering the minimum possible reaction time which could be below 250 ms [58], a zero time shift was also included. On the other hand, the largest time shift was selected to be 1.5 seconds, in accordance with the largest reaction time found for multi-sensory stimuli [59].

The time shift and Granger lag could also be related to the previous analysis of the reaction and integration time when evaluating the connection between the continuous level and evaluated pleasantness [16]. However, in the previous study, the assessed time constants were on a larger scale up to several seconds while here only the values around one second were investigated. Furthermore, correlations for the previously reported results were found to be the highest around lower values of SHIFT-LAG space which coincides with the ranges investigated in this study.

4.2 Unidirectional Granger causality

Three types of data signals (probability of change in pleasantness rating, saliency calculated from the computational model and sound pressure amplitude) were analyzed using a Granger causality analysis for the combinations shown in Figure 2. Granger causality is assessed by creating two vector autoregression (VAR) models [51], the second one of which fits the prediction of the signal with the

past values of itself and those of another signal. Accordingly, for this study, the Akaike Information Criterion (AIC) of the second model was selected to be a distinguishing factor when selecting the best model.

An example of the analysis process of one combination of data signals, i.e. computed saliency and probability of change in rating, is shown in Figure 4. Several VAR models across TIME-LAG space were evaluated, however, the top graphs in Figure 4 indicate only the AIC values where Granger causality is confirmed with asymptotic significance of p-value less than 0.05. Furthermore, in the same figure the result of the selection process of unidirectional Granger causality (UGC) is represented in the bottom graph.

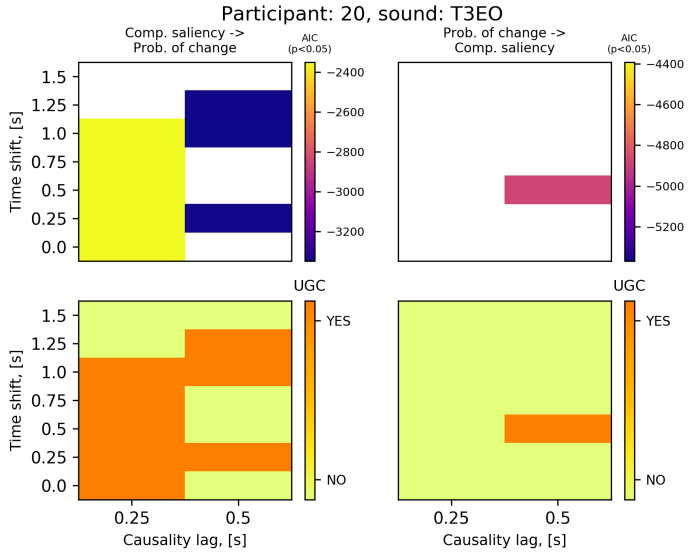


Figure 4: Granger causality measures calculated for relationship between the computed saliency and probability of change in rating for participant 20 and sound sample T3EO. Evaluated time shifts (SHIFT) included delays up to 1.5 seconds and number of assessed lags (LAG) up to 500 ms. The Akaike information criterion is displayed in the top part of the graph for Granger causality confirmed with $p < 0.05$. The lower part of the graph represents unidirectional Granger causality, i.e. indication of SHIFT-LAG combinations where causality was confirmed in one direction and rejected in another (Equation 17).

UGC is calculated using Equation 17 where k denotes a single SHIFT-LAG combination while $A \rightarrow B$ denotes the direction of analysis between signals A and B. Consequently, UGC for one SHIFT-LAG combination between two signals is confirmed only if Granger causality is confirmed in one direction and rejected in another.

$$UGC_{k,A\to B} = \begin{cases} 1, & (p_{k,A\to B} < 0.05) \wedge (p_{k,B\to A} \geq 0.05) \\ 0, & otherwise \end{cases} \tag{17}$$

4.3 Stationarity of the analyzed signals

The properties of Granger causality assume the stationarity of the analyzed signals. Time signals are deemed stationary if the shift in time does not produce a change in the shape of its probability distribution. In turn, statistical values of mean, variance and covariance are constant over the length of the portion of the signal and its position in time.

Since the data traces used do not appear immediately as such (Figure 3), it was important to check their stationarity. An Augmented Dickey-Fuller (ADF) test, i.e. a unit root test for stationarity, was used for this purpose [60]. Regression models in the ADF test were built by adding a constant (assuming no zero-mean of the signals) while simultaneously no trend was included. Furthermore, maximum checked lag was four samples, a reasonably higher number than the maximum Granger causality lag of two samples. Finally, the AIC value of the ADF test (not to be confused with the AIC for Granger causality) was used to select the optimal lag for the significance check.

The results show that the computed saliency was stationary for all 10 sounds. Similarly, the stationarity was confirmed for sound pressure amplitude of all sounds except T2EO and T2OE, as well as T5OE. Moreover, the perceptual data was stationary for all participants-sound combinations, except for participant 14 and sound T3EO, participant 21 and sounds T1OE and T2OE, and participant 22 and sound T1EO.

Although the majority of the signals were confirmed stationary, the main limitation of the study is on the sound amplitude which is not stationary for three out of the 10 evaluated sounds. It should be noted, however, that the stationarity of the sound amplitude is confirmed when the difference of the amplitude is used. Nevertheless, to keep a direct reference to this widely used acoustical metric, it was decided to make the analysis on the sound pressure amplitude.

4.4 Determining Granger causality across combinations

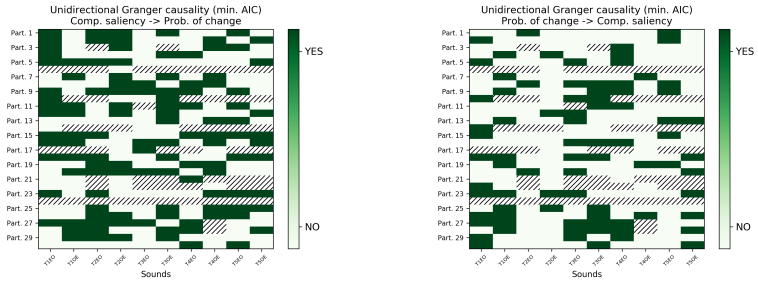
In order to summarize the large amount of data obtained for each participant-sound-SHIFT-LAG combination, a single statistical measure was created. Firstly, for each participant and sound, an unidirectional Granger causality in SHIFT-LAG space was determined as explained in Section 4.2 and shown on Figure 4. Next, a single model with the lowest AIC value was selected across SHIFT-LAG combinations where UGC was confirmed. In case no UGC was found for this participant-sound combination, the output was marked as negative. This procedure was repeated for all the signal types used in prediction analysis (Figure 2).

The results for the computed saliency and the probability of change in rating are shown in Figure 5. Counting the amount of times UGC was confirmed across participants shows that the largest difference between the original and the reverse direction of saliency predicting probability of change is found for sounds T2EO and T2OE. For both sounds, UGC was confirmed for 14 participants in the original, while only 3 and 4 participants respectively had the UGC confirmed in the reverse direction. Looking at the recorded environment, it could be observed that both sounds were recorded on a path which featured a transition between a park, a passageway and a boulevard.

However, sounds T4EO and T4OE were also recorded on a path that included a park environment. Nonetheless, for sound T4OE, UGC was confirmed for the computed saliency predicting the change for 12 participants and in the other direction for 3 participants. For sound T4EO, there were 10 participants with UGC confirmed in both directions. As the sounds that people hear matter for their perception in the parks [61], the observed difference between the results for the same environment could come from the difference in the recorded sonic environment: for sounds T2EO and T2OE, a park was located in a shielded space between the buildings which allowed more prominent park sounds (birds in this case) to be more noticeable in the recording, while for sounds T4EO and T4OE, a park was an open area with large amount of visitors and a high level of traffic noise.

The environment featured in recordings T1EO and T1OE, T3EO and T3OE, T5EO and T5OE included boulevards, streets and passageways. Counting the amount of participants for each sound, computed saliency was found to be a better or equally good predictor in the original direction of UGC compared to the reverse direction for five sounds. Only outlier was sound T3EO which had a notable difference of 6 participants with the confirmed UGC in the original direction and 12 in the reverse direction. Although for this sound the environments are the same but recorded in reverse, there is a less amount of noticeable events in sound T3EO than for sound T3OE (Figure 3.e). Moreover, this recording had a notable difference in the visual stimulus, i.e. a clouded sky and thus a darker video, which could also influence the importance that the auditory sensory saliency had on the rating of the participants [62].

As it can be seen from Figure 5, for almost 50% of the assessed combinations the saliency is confirmed as a predictor of the change in soundscape rating. However, this representation needs to be contrasted with the same analysis but in the opposite direction (i.e. change in rating predicts the saliency). To obtain such contrasting measure, a ratio of confirmed Granger causality across participant-sound space was calculated by counting the participant-sound combinations where UGC is confirmed and by dividing by the total number of cases. Moreover, the uncertainty on this measure was calculated using Equation 18. There, A and B represent the data signals, r is the ratio of confirmed UGC and ρ represents the uncertainty of the measure. On the one hand, when calculating the relationship with change in pleasantness, the degrees of freedom N_{df} equals 300, i.e. the total number of combinations that was used. On the other hand, N_{df} equals 10 when evaluating the 10 sound recordings between sound



(a) (b)

Figure 5: **Audiovisual experiment:** Unidirectional Granger causality (Equation 17) for both directions between computed saliency and probability of the change in participants' rating. Participant-sound combinations are marked as confirmed where the model evaluated by the AIC value in the SHIFT-LAG combination space exists for UGC (Figure 4). Dashed lines denote the participant-sound combinations with missing data which were excluded from the analysis.

amplitude and computed saliency.

$$\rho_{A\to B} = \sqrt{\frac{r_{A\to B}(1 - r_{A\to B})}{N_{df,A\to B}}} \tag{18}$$

The results shown in Table 1 demonstrate that there is a larger amount of cases confirming Granger causality in the original direction for both computed saliency and sound pressure amplitude. However, the sound amplitude is better predicted by the computed saliency than vice versa. This could be a surprising result since the saliency is evaluated later in time than the amplitude of the sound itself. However, the saliency model is created to react better to the change rather than the level itself, a fact exhibited in, for example, the audibility of impulsive sounds [63]. It should also be noted that as there are only 10 sounds, the UGC percentage in the last row of Table 1 could only be calculated in steps of 10%, which is much more discrete than the other two relationships.

Table 1: **Audiovisual experiment:** Percentage of confirmed unidirectional Granger causality across participant-sound combinations for three data signals: Computed saliency (Y), Probability of change in pleasantness rating (P), Sound pressure amplitude (E). The confidence intervals are calculated using Equation 18.

	Unidirectional Granger causality, [%]	
	Original direction	Reverse direction
$Y \to P$	47.54 ± 2.88	32.79 ± 2.71
$E \to P$	42.62 ± 2.86	38.11 ± 2.80
$E \to Y$	0.00 ± 0.00	30.00 ± 14.49

Furthermore, when comparing the values including their uncertainty, the percentages are distinctly separated for the saliency as a predictor of the change in pleasantness. For the sound amplitude, the intervals of uncertainty are overlapping and the percentages are not significantly different.

Table 1 also shows that there is a relatively low amount of combinations where Granger causality is confirmed for the prediction of probability of change. This could be due to various reasons, however, most notably the influence could come from the fact that the experiment was performed in an audiovisual setting. Therefore, the saliency of the event would not only be present in the sound but also in the visual scene [64]. What is more, the non-explained portion of the results could arise from the responses of the participants that are determined by top-down attention [65], i.e. when the participants focused on the non-salient portions of the stimuli.

Finally, the results show that for the prediction of the change in pleasantness as evaluated by this dataset, the saliency of the signal computed by the proposed model is a better predictor than the sound amplitude. Consequently, this validates the hypothesis that the saliency of the sound predicts change better than the sound amplitude in the appraisal of the perceived soundscape.

5 Prediction without a visual component

The experimental data that was used to investigate the relevance of the computed saliency came from an audiovisual experiment where participants listened to the sound recordings and at the same time watched the matching videos. This setting provides the most difficult one for auditory sensory saliency, due to the influence of the visual component [62, 5]. In order to evaluate the influence of sensory saliency for the change in pleasantness rating in an easier setting, a Granger causality analysis was also performed on a dataset from an audio-only experiment.

The experimental data came from the same study as the audiovisual experiment [16]. The group of participants in the audio-only experiment was, however, different. The recruitment procedure was the same as explained in Section 2. Initially, there were 11 women and 19 men with a mean age of 33 years (SD = 14). However, seven participants were eliminated from the analysis due to measured hearing loss and/or incoherent responses (very incomplete, constant or random ratings). Therefore, only 23 participants were included in the analysis.

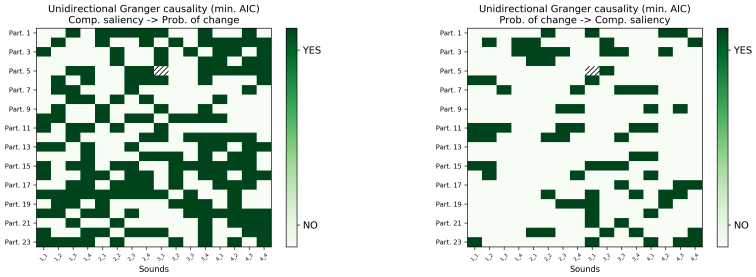
Contrary to the audiovisual experiment conducted in the same laboratory setting, in the audio-only case, participants listened to 16 recordings with the visual component reduced to a minimum by presenting only a blurred stationary image of an urban environment on a screen. It should also be noted that the 16 recordings used in the audio-only experiment were specifically constructed from the two audio files and therefore different from the ones used in the audiovisual experiment. Therefore, comparison on a sound-by-sound instance between the obtained results was impossible, however, the evaluation of the cumulative results of the Granger causality analysis was attainable.

The audio-only experimental dataset was evaluated using the same procedure as explained in Sections 2-4. In particular, from the perceptual dataset, the change in pleasantness assessment was calculated using Equation 2. Moreover, the sound amplitude was extracted from the recordings as explained in Section 2.4. Finally, the saliency of the sound recordings was calculated using the same proposed saliency model presented in Section 3. The dataset of three signal types was then analyzed using the procedure explained in Section 4.

In order to confirm the applicability of the Granger causality analysis to the new dataset, the obtained data traces were firstly tested for their stationarity. The results of the Augmented Dicky-Fuller test confirmed that all sound amplitude and computed saliency signals are stationary. For the perceptual data traces, however, the stationarity was not confirmed for 12 out of the 368 participant-sound combinations. Therefore, although this presents a limitation of the analysis, all the combinations were kept in the dataset, similar to the analysis of the audiovisual data (Section 4.3).

The same SHIFT-LAG combinations were assessed in the analysis of the audio-only experimental dataset. Therefore, the time shifts between the signals included delays up to 1.5 seconds, while the assessed lags for Granger causality were up to 500 ms. Although the multi-sensory reaction from the visual part [59] was minimized in the audio-only case, it was decided to keep the analyzed SHIFT-LAG space the same in order to obtain more comparable results.

The results from the unidirectional Granger causality analysis of the audio-only dataset are shown in Figure 6 and Table 2. When comparing the same representation (Figure 5) for the audiovisual experiment, it can be seen that the reverse prediction for the Granger causality is confirmed in more cases than for the audio-only experiment.



(a) (b)

Figure 6: **Audio-only experiment:** Unidirectional Granger causality (Equation 17) for both directions between computed saliency and probability of the change in participants' rating. Participant-sound combinations are marked as confirmed where the model evaluated by the AIC value in the SHIFT-LAG combination space exists for UGC (Figure 4). Dashed lines denote the participant-sound combinations with missing data, which were excluded from the analysis.

The larger relative difference between the confirmed UGC in both directions

Table 2: **Audio-only experiment:** Percentage of confirmed unidirectional Granger causality across participant-sound combinations for three data signals: Computed saliency (Y), Probability of change in pleasantness rating (P), Sound pressure amplitude (E). The confidence intervals are calculated using Equation 18.

	Unidirectional Granger causality, [%]	
	Original direction	Reverse direction
$Y \rightarrow P$	48.77 ± 2.89	21.53 ± 2.37
$E \rightarrow P$	38.15 ± 2.80	18.53 ± 2.24
$E \rightarrow Y$	0.00 ± 0.00	43.75 ± 15.69

is further substantiated using the values from Table 2. For instance, the relative difference in the original and reverse direction of the prediction of the computed saliency and the probability of change in pleasantness rating is 27.24%. On the other hand, the difference between the original and reverse direction for the sound pressure amplitude and the change in pleasantness rating is 19.62%. Contrary to the previous results (Table 1), both these differences are inside the confidence intervals, therefore it can be concluded that the prediction is confirmed in both cases.

The most important comparison between the experiments, however, comes from the confirmed UGC for the original direction between the computed saliency and the sound amplitude respectively and the predicted change in pleasantness. In particular, for the audio-only experiment, there is a 10.62% difference of the confirmed instances between the saliency and the sound amplitude as predictors. On the other hand, for the audiovisual experiment, this difference falls to 4.92%. This result is in line with the idea that auditory sensory saliency should be more relevant when assessing the environment using the acoustic stimulus alone. In turn, this finding also shows the applicability of the proposed computational model for calculating the sensory saliency of the sound environment.

6 Conclusions

In this paper we evaluated the hypothesis that auditory saliency triggers change in pleasantness assessment of the soundscape. Recordings of walking trips through urban environments were assessed in a previous audiovisual experiment by their pleasantness [16]. In this study, the continuous rating obtained from this experiment was used as a basis for determining the probability of change in pleasantness rating over time.

The recordings from the audiovisual experiment were analyzed by the proposed biologically-inspired auditory sensory saliency model. The model is based on the fact that the human auditory cortex is sensitive to a range of spectrotemporal modulations [15, 30, 32]. Thus, the model evaluates the similarity of the input to the spectrotemporal modulation content. Finally, at the last stage,

the model utilizes sensory activation to interplay the excitation and inhibition processes taking place in the neural circuits [40, 41, 42].

To test the hypothesis that the saliency of the sound as determined by the saliency model is indeed a predictor of changes in pleasantness, the A-weighted sound pressure amplitude, a common indicator used in soundscape studies [19, 20, 21], was also calculated. The prediction between the signals was then evaluated using a Granger causality analysis with a unidirectional causality constraint.

It was found that saliency better predicts the probability of change than sound amplitude. In particular, for 47.54% of combinations, the computed sensory saliency predicts change in pleasantness while the opposite is confirmed in 32.79% of cases. Sound amplitude was found to predict change in rating in 42.62% and 38.11% of combinations in each direction respectively, thus having a smaller number of cases predicted and a smaller difference between the two directions than with computed saliency.

Finally, to account for the effect of audiovisual interaction, which happens even at the lowest stages of attention processing [62, 5], the data from the audiovisual experiment was compared to the data from an audio-only experiment conducted in the same study [16]. The results show that computed sensory saliency becomes an even better predictor for the change in pleasantness rating in comparison to sound amplitude, as shown by the 48.77% of confirmed cases for the computed sensory saliency and only 38.15% of confirmed cases for sound amplitude. This result also shows the applicability of sensory saliency as evaluated by the proposed computational model in assessment of sound environments.

To conclude, the proposed model could serve as an evaluation tool in other urban soundscape studies. One of the studies, in which the model is currently used, is the categorization of urban soundscapes. Other future studies could also include evaluation of large datasets of environmental sound and comparison with the indicators currently established in soundscape research. Finally, the proposed saliency model could be extended with binaural hearing traits, to better represent the processes of auditory perception taking place in the brain.

Acknowledgements

The research leading to these results has received funding from the Research Foundation Flanders (FWO-Vlaanderen) under Grant G0D5215N, ERC Runner-up project MAESTRO. Furthermore, this work has also been carried out in the framework of the GRAFIC project, supported by the French Environment and Energy Management Agency (ADEME) under contract No. 1317C0028.

References

[1] C. Carlson, S. Aytur, K. Gardner, and S. Rogers, "Complexity in built environment, health, and destination walking: a neighborhood-scale analysis," *Journal of Urban Health*, vol. 89, no. 2, pp. 270–284, 2012.

[2] S. Park, K. Choi, and J. S. Lee, "To walk or not to walk: Testing the effect of path walkability on transit users' access mode choices to the station," *International Journal of Sustainable Transportation*, vol. 9, no. 8, pp. 529–541, 2015.

[3] C. J. Gidlow, M. V. Jones, G. Hurst, D. Masterson, D. Clark-Carter, M. P. Tarvainen, G. Smith, and M. Nieuwenhuijsen, "Where to put your best foot forward: Psycho-physiological responses to walking in natural and urban environments," *Journal of Environmental Psychology*, vol. 45, pp. 22–29, 2016.

[4] International Organization for Standardization, "ISO 12913-1:2014 Acoustics – Soundscape – Part 1: Definition and conceptual framework," 2014.

[5] G. M. E. Sanchez, T. Van Renterghem, K. Sun, B. De Coensel, and D. Botteldooren, "Using Virtual Reality for assessing the role of noise in the audio-visual design of an urban public space," *Landscape and Urban Planning*, vol. 167, pp. 98–107, 2017.

[6] J. Kang, F. Aletta, T. T. Gjestland, L. A. Brown, D. Botteldooren, B. Schulte-Fortkamp, P. Lercher, I. van Kamp, K. Genuit, A. Fiebig, J. L. B. Coelho, L. Maffei, and L. Lavia, "Ten questions on the soundscapes of the built environment," *Building and Environment*, vol. 108, pp. 284–294, 2016.

[7] J. A. Droll and M. P. Eckstein, "Gaze control and memory for objects while walking in a real world environment," *Visual Cognition*, vol. 17, no. 6-7, pp. 1159–1184, 2009.

[8] K. Filipan, M. Boes, B. De Coensel, H. Domitrović, and D. Botteldooren, "Identifying and recognizing noticeable sounds from physical measurements and their effect on soundscape," in *10th European Congress and Exposition on Noise Control Engineering (Euronoise 2015)*, pp. 1559–1564, 2015.

[9] B. Gatersleben and I. Griffin, "Environmental Stress," in *Handbook of Environmental Psychology and Quality of Life Research*, pp. 469–485, Springer, 2017.

[10] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, pp. 115–141, Springer, 1987.

[11] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[12] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160101, 2017.

[13] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos, "COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 54, 2017.

[14] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[15] M. Schönwiesner and R. J. Zatorre, "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*, vol. 106, no. 34, pp. 14611–14616, 2009.

[16] P. Aumond, A. Can, B. De Coensel, C. Ribeiro, D. Botteldooren, and C. Lavandier, "Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments," *Applied Sciences*, vol. 7, no. 2, p. 144, 2017.

[17] P. Aumond, F. Masson, L. Beron, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, and C. Lavandier, "Influence of experimental conditions on sound pleasantness evaluations," in *22nd International Congress on Acoustics (ICA 2016)*, 2016.

[18] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS biology*, vol. 7, no. 6, p. e1000129, 2009.

[19] G. Watts, A. Khan, and R. Pheasant, "Influence of soundscape and interior design on anxiety and perceived tranquillity of patients in a healthcare setting," *Applied Acoustics*, vol. 104, pp. 135–141, 2016.

[20] R. Cain, P. Jennings, and J. Poxon, "The development and application of the emotional dimensions of a soundscape," *Applied Acoustics*, vol. 74, no. 2, pp. 232–239, 2013.

[21] Ö. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2836–2846, 2010.

[22] A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, and S. Verhulst, "A comparative study of seven human cochlear filter models," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1618–1634, 2016.

[23] S. Verhulst, A. Altoè, and V. Vasilkov, "Computational modeling of the human auditory periphery: auditory-nerve responses, evoked potentials and hearing loss," *Hearing Research*, 2018.

[24] P. Lakatos, G. Musacchia, M. N. OConnel, A. Y. Falchier, D. C. Javitt, and C. E. Schroeder, "The spectrotemporal filter mechanism of auditory selective attention," *Neuron*, vol. 77, no. 4, pp. 750–761, 2013.

[25] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[26] A. Bellur and M. Elhilali, "Speech processing using adaptive auditory receptive fields," in *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 6, pp. 63–73, 2018.

[27] A. Bellur and M. Elhilali, "Feedback-driven sensory mapping adaptation for robust speech activity detection," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 3, pp. 481–492, 2017.

[28] V. Duangudom and D. V. Anderson, "Identifying salient sounds using dual-task experiments," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pp. 1–4, IEEE, 2013.

[29] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *Signal Processing Conference, 2007 15th European*, pp. 1206–1210, IEEE, 2007.

[30] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[31] R. P. Carlyon and S. Shamma, "An account of monaural phase sensitivity," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 333–348, 2003.

[32] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[33] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," *Journal of Sound and Vibration*, vol. 292, pp. 105–123, 2006.

[34] S. Puria, "Measurements of human middle ear forward and reverse acoustics: implications for otoacoustic emissions," *The Journal of the Acoustical Society of America*, vol. 113, no. 5, pp. 2773–2789, 2003.

[35] M. Slaney *et al.*, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[36] American National Standards Institute, "Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters," 2004.

[37] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 422–438, 2008.

[38] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on signal processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[39] L. H. Carney, T. Li, and J. M. McDonough, "Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations," *Eneuro*, vol. 2, no. 4, pp. ENEURO–0004, 2015.

[40] M. Xue, B. V. Atallah, and M. Scanziani, "Equalizing excitation–inhibition ratios across visual cortical neurons," *Nature*, vol. 511, no. 7511, p. 596, 2014.

[41] G. Deco, A. Ponce-Alvarez, P. Hagmann, G. L. Romani, D. Mantini, and M. Corbetta, "How local excitation–inhibition ratio impacts the whole brain dynamics," *Journal of Neuroscience*, vol. 34, no. 23, pp. 7886–7898, 2014.

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

[43] S. A. Neymotin, H. Lee, E. Park, A. A. Fenton, and W. W. Lytton, "Emergence of physiological oscillation frequencies in a computer model of neocortex," *Frontiers in computational neuroscience*, vol. 5, p. 19, 2011.

[44] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[45] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "Attention-driven auditory stream segregation using a SOM coupled with an excitatory-inhibitory ANN," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 516–523, IEEE, 2012.

[46] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *20th International Congress on Acoustics (ICA-2010)*, pp. 1–8, 2010.

[47] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, p. 947, 2000.

[48] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, 2013.

[49] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[50] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[51] A. Seth, "Granger causality," *Scholarpedia*, vol. 2, no. 7, p. 1667, 2007. revision #91329.

[52] C. W. J. Granger, *Modelling economic series: readings in econometric methodology*. Oxford University Press, 1991.

[53] M. Ding, Y. Chen, and S. L. Bressler, "Granger causality: basic theory and application to neuroscience," *Handbook of time series analysis: recent theoretical developments and applications*, pp. 437–460, 2006.

[54] A. D'Ausilio, L. Badino, Y. Li, S. Tokay, L. Craighero, R. Canto, Y. Aloimonos, and L. Fadiga, "Leadership in orchestra emerges from the causal relationships of movement kinematics," *PLoS one*, vol. 7, no. 5, p. e35757, 2012.

[55] R. T. Dean, F. Bailes, and E. Schubert, "Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation," *PloS one*, vol. 6, no. 4, p. e18591, 2011.

[56] G. Deshpande and X. Hu, "Investigating effective brain connectivity from fMRI data: past findings and current issues with reference to Granger causality analysis," *Brain connectivity*, vol. 2, no. 5, pp. 235–245, 2012.

[57] R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping," *Magnetic resonance imaging*, vol. 21, no. 10, pp. 1251–1261, 2003.

[58] H. Colonius and A. Diederich, "The optimal time window of visual-auditory integration: a reaction time analysis," *Frontiers in integrative neuroscience*, vol. 4, p. 11, 2010.

[59] J. Bigelow and A. Poremba, "Achilles ear? Inferior human short-term and recognition memory in the auditory modality," *PloS one*, vol. 9, no. 2, p. e89914, 2014.

[60] S. M. Miller, "Are saving and investment co-integrated?," *Economics Letters*, vol. 27, no. 1, pp. 31–34, 1988.

[61] K. Filipan, M. Boes, B. De Coensel, C. Lavandier, P. Delaitre, H. Domitrović, and D. Botteldooren, "The personal viewpoint on the meaning of tranquility affects the appraisal of the urban park soundscape," *Applied Sciences*, vol. 7, no. 1, p. 91, 2017.

[62] K. Sun, G. M. Echevarria Sanchez, B. De Coensel, T. Van Renterghem, D. Talsma, and D. Botteldooren, "Personal audiovisual aptitude influences the interaction between landscape and soundscape appraisal," *Frontiers in psychology*, vol. 9, p. 780, 2018.

[63] T. Pedersen, "Audibility of impulsive sounds in environmental noise," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 2000, pp. 4158–4164, Institute of Noise Control Engineering, 2000.

[64] O. Rummukainen, J. Radun, T. Virtanen, and V. Pulkki, "Categorization of natural dynamic audiovisual scenes," *PloS one*, vol. 9, no. 5, p. e95848, 2014.

[65] D. L. Strait, N. Kraus, A. Parbery-Clark, and R. Ashley, "Musical experience shapes top-down auditory mechanisms: evidence from masking and auditory attention performance," *Hearing research*, vol. 261, no. 1-2, pp. 22–29, 2010.