Investigating soundscapes perception through acoustic scenes simulation

G. Lafay[1)], M. Rossignol[2)], N. Misdariis[2)], M. Lagrange[1)], J-F. Petiot[1)]

[1)] *Laboratoire des Sciences du Numérique de Nantes-CNRS-École Centrale de Nantes, Nantes, France.*
mathieu.lagrange@cnrs.fr
[2)] *STMS Ircam-CNRS-UPMC Institut de Recherche et Coordination Acoustique/Musique, Paris, France*

Abstract

This paper introduces a new experimental protocol to study mental representations of urban soundscapes through a simulation process. Subjects are asked to create a full soundscape by means of a dedicated software tool, coupled with a structured sound data set. This paradigm is used to characterize urban sound environment representations by analyzing the sound classes that were used to simulate the auditory scenes. A rating experiment of the soundscape pleasantness using a 7-point bipolar semantic scale is conducted to further refine the analysis of the simulated urban acoustic scenes. Results show that 1) a semantic characterization in terms of presence / absence of sound sources is an effective way to characterize urban soundscapes pleasantness, and 2) physical descriptors computed for specific sound sources better characterize the appraisal than global descriptors.

Keywords: cognitive psychology, soundscape perception, soundscape simulator

1. Introduction

One of the main goals of soundscape studies is to identify which components of the soundscape influence human perception [1]. Establishing a link between an environment (urban sound scene) and the induced human sensation (calmness) is equivalent to instantiating the mental representation of a specific sound scene (a calm urban sound scene). Several means have been

considered in the literature to perform this task.

First, a subject can be asked to assess a given sound following a given perceptual scale (*e. g.* calmness, pleasantness) [2, 3, 4]. The amount of information that can be gathered strongly depends on the nature of the available stimuli, be they recorded sound scenes as part of a within laboratory experiment, or *in situ*. Working with actual stimuli makes it possible to analyze them, thus allowing the experimenter to gather a coarse-grained physical description of them.

Second, a subject can be asked to describe a given sound environment [5, 6]. A large amount of quantitative and semantic information is then collected about the subject's representation of this type of sound environment. Unfortunately, without any reference to sound data, this representation can hardly be characterized physically.

We propose in this paper to consider the use of a soundscape simulator that the subject can employ to objectify his/her representation of a given sound environment. We believe that the use of such a device allows us to gain the benefits of the two above mentioned approaches. As the subject is asked to produce audio data (the signal of the simulated scene), it allows the experimenter to study a precise modal version of the subject's mental representation that is characterized both semantically (the nature of sound sources) and physically (*e. g.* the levels of sound sources).

We believe that the availability of a fine grain description of the sound stimuli is of great interest, because recent studies demonstrate that the sound sources do not contribute equally to the perception of the sound scene [7, 8, 5, 9, 10]. Thus, much attention is given to the specific contributions of the different sources on the notion of emotional quality of the scene [11, 12].

For several reasons that are detailed in this paper, the use of a soundscape simulator such as the one proposed here can lead to interesting outcomes as it allows the experimenter to separately study the influence of the sound sources that compose a sound environment. With such material, not only is the type of sound sources available for study, but also the exact level and audio waveform for each source together with the structural properties of the scene, that is, the temporal distribution of the sound events.

To demonstrate the potential of the proposed approach in its ability to question the human perception of sound environments, we study in this paper the notion of perceptual pleasantness of urban environmental scenes. Results and outcomes of a series of three experiments that build upon the use of the simulator are studied in order to better comprehend how different

sound sources typically present in a urban scene impact pleasantness:

1. *experiment 1.a, simulation*: the subjects use the soundscape simulator to produce ideal / non ideal soundscapes that are considered as material for the following experiments;
2. *experiment 1.b, pleasantness evaluation*: the subjects judge the pleasantness of the simulated scenes on a semantic scale;
3. *experiment 2, pleasantness evaluation after modification of the scenes*: the subjects judge the pleasantness of the simulated scenes on a semantic scale as in 1.b, but some scenes are modified beforehand, *i. e.* some specific sounds classes that are identified as having a significant impact on perceived pleasantness are removed.

To the best of our knowledge, only Bruce [13, 14] considered the use of a simulator to question soundscape perception. They propose a tool that allows the user to modify a given soundscape by adding or removing specific sound sources, by changing the acoustic level of the sources as well as their spatial location. The authors shows that the addition or removal of the sources globally follows more social or semantic matters than their acoustical characteristics. A lack of diversity in terms of sound sources is nonetheless mentioned by the authors that limit the strength of the outcomes given in this study.

In our approach, the simulator developed for this study[1] only yields a monophonic representation of the scene, but that simplification comes with the benefit of a wider range of available sound sources and scheduling parameters in order to provide outputs which are both expressive and useful for analysis.

The remaining of the paper is organized as follows: the soundscape simulator *simScene* is introduced in Section 2. Experiments 1.a *simulation* and 1.b *pleasantness evaluation* are presented in Section 3. Experiment 2 *pleasantness evaluation after modification of the scenes* is presented in Section 4. Conclusions and discussion about future work follow in Section 5.

[1] simScene, available online `https://bitbucket.org/mlagrange/simscene`

2. The simulator

Simscene is an online digital audio tool whose first version has been developed as part of the HOULE project[2]. It has been designed to run on the popular web browsers *Chrome* and *Firefox*. It is fully written in javascript using the *angular.js* library[3] and the *Web Audio* standard that allows the manipulation of digital audio data within the browser[4]. The interface for selecting sound sources (cf. Section 2.3) uses the popular *D3.js* [15] visualization library.

Simscene is designed as a simplified audio sequencer, with sequencing parameters specifically chosen for the generation of realistic soundscapes. To do so, the user first selects a sound source using a non verbal selection interface presented in Section 2.3. A track is then created for this sound source within the simulator interface. The user can then manipulate some parameters detailed in Section 2.2 to control the time and magnitude distribution of the occurrences of the sound sources. Text fields are also available for the user to 1) name each track, 2) name the entire scene, 3) provide free comments about the simulated scene.

2.1. Sound database

In order to provide the user with a sound database that is well organized and covers as much as possible the variety of sound sources that are present in urban areas, a typology of urban sounds is first defined.

Typology

The chosen typology is established based on the category/classes of sounds found while reviewing several articles or thesis manuscripts [16, 17, 18, 7, 19, 6, 20, 5, 21, 16, 22, 23, 24, 25] that study how humans discriminate different kind of urban soundscapes.

We choose not to include any musical content in the sound database, as the study of the pleasantness of a given style or genre of music is beyond the scope of this study.

[2] *Projet HOULE* : `houle.ircam.fr`
[3] *angular.js* : `angularjs.org`
[4] *Web Audio*: `www.w3.org/TR/webaudio`

Events and Textures

A commonly accepted distinction consists in separating:

- sound events: isolated sounds, precisely located in time, whose acoustical characteristics may change with respect to time;
- sound textures: isolated sounds of long duration whose acoustical characteristics are stable with respect to time [26].

In order to account for the possibility that the morphological differences between those two classes may have some important consequences on the perception of the scenes, the simulation tool *Simscene* follows this distinction and considers two distinct sound databases: one with classes of events only and the other with classes of textures only. These two types of sound classes also have specific scheduling procedures during the simulation process.

According to the Auditory Scene Analysis (ASA) theory, the brain processes separately events emerging from the different sources of a sound environment [27]. Several auditory streams are created, one for each sequence of events emitted by the same source [28], while events that cannot be isolated are grouped into a same stream.

Maffiolo [16] distinguishes two separated categorization processes, either of which is triggered depending on the listener's ability to identify sound events. She shows that those processes lead to two abstract cognitive categories respectively termed "event sequences" and "amorphous sequences".

Event sequences are composed of salient events that can easily be recognized, such as *car start* or *male speech*. They arise from a descriptive analysis based on the identification of the sound sources. On the contrary, amorphous sequences are sound environments where distinct events cannot be readily identified, such as *traffic hubhub*. They result from a holistic analysis based on global acoustic features.

Concerning sound textures, *i. e.* sounds that have stable characteristics over time, [29, 30] demonstrates that the human brain can opt for an abstract, statistical representation of the perceived information, discarding precise physical properties of the sound [31].

In the light of those results, it appears that the processing of auditory information comprises some sort of decision concerning the nature of the stimuli [31, 30]. We thus consider in this study the soundscape as a "skeleton of events on a bed of textures" as coined in [31].

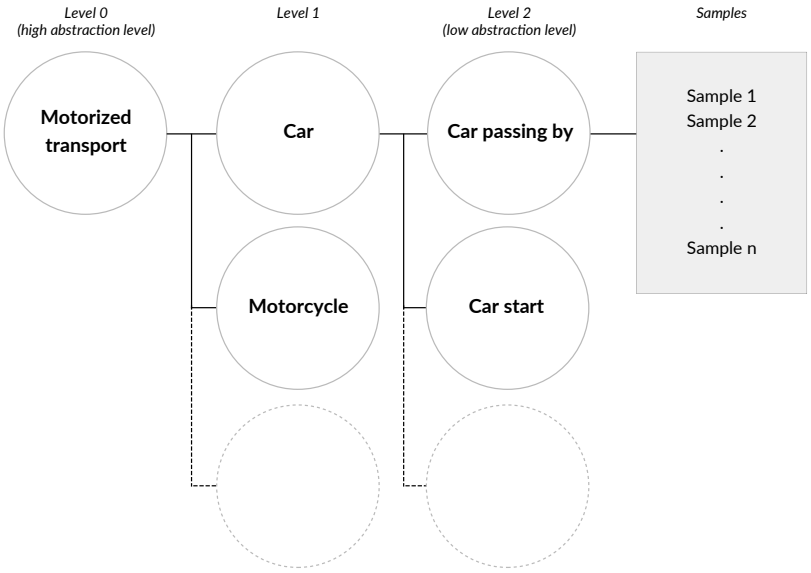


Figure 1: Hierarchical organisation of the isolated sounds used in the simulation.

Taxonomy

We call "sample" a recording of an isolated sound, be it an event or a texture. Each sound class is implemented as a collection of samples judged to be perceptually equivalent.

The sound classes are organized hierarchically (cf. Figure 1) according to a structure similar to the vertical axis of the categorical organization proposed by E. Rosch [32]. The lower the level of abstraction, the more precise the description of the class and the more perceptually similar the sound sources. For classes with a high level of abstraction that have sub classes, their collection of samples is the union of the collections of the sub classes.

Accounting for the previously detailed perceptual matters, two taxonomies are built, one for event sounds and one for texture sounds. Four levels of abstraction are considered from the most generic (level 0) to the most specific classes (level 2), leading to a taxonomy close to the one used in [33]. Only three levels of abstraction are considered for the texture sounds.

Sound samples collection

483 isolated sounds are collected and organized with the two typologies discussed above, 381 events and 102 textures. Among those samples, 332 have been recorded and 151 have been taken from two sound libraries: *SoundIdeas*[5] and *Universal SoundBank*[6].

Original sounds have been recorded using a shotgun microphone *AT8035*[7] plugged to a *ZOOM H4n*[8] recorder. The use of such a microphone allowed us to isolate as much as possible sound events of interest from the urban background. It also allowed us to avoid dominant sound sources while recording texture sounds by targeting distant areas with no dominant sound sources.

All samples are normalized to the same RMS level of -12 dB FS, *i. e.* relative to Full Scale. In our case, the full scale level is set arbitrarily to 1 Volt.

2.2. Parameters

By a deliberate design choice, the simulation tool does not allow the user to interact with and control directly a specific sample. Interaction is done at the track level, a track being a sequence of samples. Several parameters are available to the subject to control the track:

- *sound level* (dB): for each sample, the sound levels are drawn randomly following a normal distribution parameterized by the subject in terms of mean value and variance;
- *inter-onset spacing* (second): for event tracks only, and for each sound event sample, the inter-onset spacings are drawn randomly following a normal distribution parameterized by the subject in terms of mean value and variance;
- *start and end time* (second): the subject sets the start and end times between which the texture or sequence of repeated events occurs.

To improve simulation quality, two parameters are also proposed:

[5] *SoundIdeas*: `www.sound-ideas.com`
[6] *Universal SoundBank*: `www.universal-soundbank.com`
[7] *AT8035 shotgun microphone*: `eu.audio-technica.com`
[8] *ZOOM H4n* : `www.zoom.co.jp/english/products/h4n`

- *event fades* (seconds): for the event tracks only, the subject can set a fade in / fade out duration applied to each sample;
- *global fades* (seconds): the subject can set global fade in and fade out durations applied to the entire track.

Texture samples are sequenced without time spacing, therefore the parameters *event fade* and *inter-onset spacing* are not available for this kind of track.

2.3. Selection interface

Once the typology and the set of sounds are available, an important design issue is the need of a suitable way to display the sound dataset to the user. Most browsing tools are based on keyword indexing; however, for sensory experiments that study the objectivation of a subject's mental representations, this means is problematic: indeed, the availability of a verbal description of the sound can influence the subject's choice, and potentially induce biases in the analysis conclusions. For example, a subject may automatically select sounds referenced as belonging to a *park* environment to build a *calm* soundscape, rather than focusing on their perception.

Therefore, the selection interface considered in this study is text free and designed so as to force the user to rely on listening.

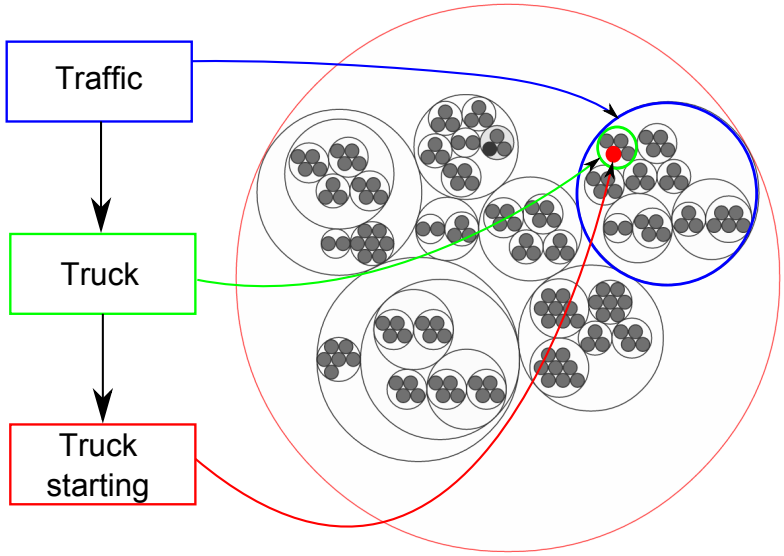
Figure 2a shows the interface used for the selection of events. Each circle corresponds to a sound class, the lowest level of abstraction (leaves) colored in grey. The spatial location of those circles is chosen according to the hierarchical organization of the sound database: sub-classes belonging to the same class are close to each others, and so on until the user reaches the leaf classes, which are directly linked to a collection of samples.

Each of those classes has a representative sound chosen arbitrarily by the authors in order to provide the same sound each time the user clicks on the circle. The subject can browse the database by listening to those prototype sounds. The efficiency of this interface compared to several others designs have been evaluated and the outcomes are discussed in [34].

2.4. Simulation interface

As shown on Figure 2b, the simulation interface displays a schematic of the scene under creation. Each track is represented as a horizontal strip with a

(a)



(b)

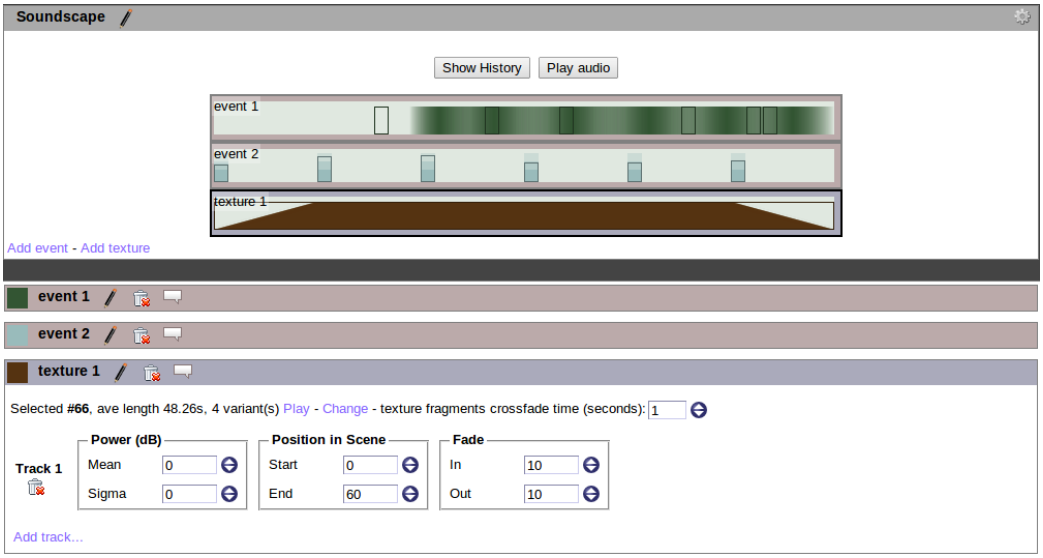


Figure 2: *SimScene* graphical interfaces for the selection of sound classes (a) and their sequencing (b).

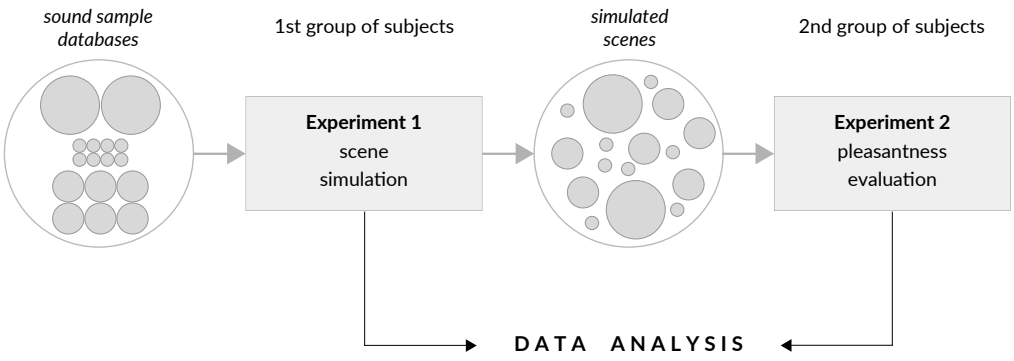


Figure 3: Experimental protocol of the simulation experiment (1.a) and pleasantness evaluation experiment (1.b).

temporal axis. Each sample of this track is displayed as a rectangle whose height is proportional to the amplitude of the sample. For event tracks, the horizontal spacing between those rectangles is a function of the time delay between their onsets. For texture tracks, a unique rectangle is displayed as this kind of sounds does not allow spacing with silence. As the actual amplitude and spacing values are drawn from random variables, each time the subject changes the value of a parameter, the location and height of the rectangles are updated to reflect the changes in the sequencing of the samples. The subject can listen to the resulting sound scene at any time.

As such, the underlying model of the scene is a sum of sound sources. Further details about the simulation interface can be found in [35].

3. Experiment 1

3.1. Objective

Experiment 1 aims at using a simulation paradigm to investigate the specific influences of the various sound sources constituting urban soundscapes on the perceived pleasantness. For that, the two first experiments are planned as follows (cf. Figure 3):

- *experiment 1.a (simulation)*: during this experiment, subjects are asked to simulate urban sound environments using *Simscene* (see Section 2).

Each of them has to create two sound environments: one ideal / pleasant, and the other non ideal / unpleasant.

- *experiment 2.a (evaluation)*: after the simulation phase, only a binary information on the pleasantness property of the respective scenes is available: respectively ideal (i) or non ideal (ni). Furthermore, this information is given by the creator of the scene. The second experimental step aims at investigating more deeply and more broadly our knowledge on the pleasantness of the simulated scenes. For that, a second group of subjects is asked to evaluate the pleasantness of each scene produced during (1.a), on a semantic scale. This experiment has two goals:

  1. to evaluate more precisely the respective influence of the various sources composing the scenes on the pleasantness (i or ni) thanks to a finer quantification of the pleasantness of the scene;
  2. to detect the presence of *outliers* or ambiguous scenes. Indeed, throughout our analyses, the predefined hedonic properties of the scenes (i and ni) are used as reference. We thus need to ensure beforehand that no ambiguity exists between extreme cases of i and ni scenes, *i. e.* that the least pleasant i-scene remains more highly rated on average than the most pleasant ni-scene.

The data collected by these two experiments (1.a and 1.b) are analyzed conjointly.

3.2. Planning of Experiment 1.a

The design of this experiment has been validated with a pilot study described in [36].

Procedure

The subjects are asked to simulate two urban sound environments of one minute each, following these instructions:

- first simulation: create a **plausible urban soundscape** which is ideal, according to you (where you would like to live);

- second simulation: create a **plausible urban soundscape** which is non ideal, according to you (where you would not like to live);

All the subjects start by designing the ideal environment; they read the second set of instructions at the end of the first experiment. Subjects are completely free of their choices concerning sounds and synthesis parameters. The created sound environments must nevertheless fulfill the two following constraints :

- the listening point of view is that of a fixed listener;
- the soundscape must be realistic, *i. e.* physically plausible. For instance, subjects are free to insert ten dogs in the soundscape but they cannot insert one dog barking every 10 milliseconds.

These constraints are notified in the instructions; no control is done *a priori* in the simulation interface.

Each simulation process is decomposed into several steps:

1. Simulation, where the user is asked to:
   - select sound classes,
   - give each of them a name,
   - set the parameters of the tracks related to the selected sound classes of sounds, see Section 2.2.
2. Feedback : writing of a free form comment about the composed soundscape.

In addition, once the two sound scenes are completed, subjects are invited to:

- point out the sound sources that were missing for the composition;
- give a comment about the ergonomics of the simulation environment;
- give a comment about the ergonomics of the selection tool.

Before starting the first simulation, a 20-minute tutorial is given to the subjects in order to familiarize them with the simulation interface and the sound database. The experiment is planned to last about two hours and a half, including breaks that the subjects are allowed to take.

Apparatus

All the subjects performed the experiment on standard desktop computers with the same hardware and software configurations. The audio files were played in diotic conditions using headphones. During the tutorial, subjects were asked to adjust the sound volume to a comfortable level. Once set, they were not allowed to modify it during the remaining of the experiment.

All the subjects performed the experiment at the same time. They were equally distributed in three identical quiet rooms, and are not allowed to talk to each other during the experiment.

Three experimenters (one in each room) were available during the whole duration of the experiment in order to assist subjects with potential hardware and software issues, and to answer questions.

Subjects

44 students (30 male, 14 female; averaging 21.6 years of age, s.d. of 2.0 years) from *Ecole Centrale de Nantes* (a french engineering school) took part in the experiment. All the subjects had been living in Nantes, France, for at least two years at the time of the experiment.

Among the 44 subjects, 40 succeeded, producing at the end 80 simulated sound scenes (40 ideal scenes, 40 non ideal scenes). the 4 other subjects were excluded from the process due to a lack of understanding of the instructions or failure to respect them, or for exceeding the maximum duration allowed to perform the experiment. The software platform used for the experiment, the parametrization of the software platform for each generated scene, as well as a 2 dimensional projection of the resulting scenes are available `http://soundthings.org/research/urbanSoundScape/XP2014`. The resulting waveforms are available for download `https://archive.org/details/soundSimulatedUrbanScene`.

3.3. Planning of experiment 1.b

Procedure

Due to temporal constraints, subjects only assess 30 seconds of the initial 1-minute simulated scenes generated in experiment 1.a (from second 15 to sec. 45).

The assessment is done with a 7-point bipolar semantic scale going from -3 (non ideal / unpleasant) to +3 (ideal / pleasant). Before evaluating a scene, subjects must listen to the first 20 seconds of the stimuli. After the evaluation, they are free to continue to the next scene.

For each participant, sound scenes are played in a quasi random order. 5 ideal scenes and 5 non ideal scenes are first sequenced to allow the subjects to calibrate their scores. These first 10 scenes are played back again at the end of the experiment. Only the last evaluations are taken into account. Each participant evaluates all the sound scenes.

The experiment is planned to last 30 minutes. The subjects do not know anything beforehand about the nature of the sound scenes.

Apparatus

All the subjects perform the experiment on computers with equivalent hardware and software configurations. The audio files are played in diotic conditions by semi-open headphones *Beyerdynamic DT 990 Pro*. The stimuli are the scenes obtained in experiment 1.a. The output sound level is the same for all the subjects.

All the subjects perform the experiment simultaneously. They are not allowed to talk to each other during the experiment.

An experimenter is available during the whole duration of the experiment in order to assist subjects and answer questions any questions the subjects may have.

Subjects

10 students (8 male, 2 female; averaging 23.1 years of age, s.d. of 1.8 years) from *Ecole Centrale de Nantes* took part to the experiment. All the subjects had been living in Nantes, France, for at least two years at the time of the experiment. None of them took part to the previous simulation experiment (experiment 1.a).

All the subjects succeeded in doing the experiment.

3.4. Data and statistical analysis

A set of features, upon which the analysis is conducted, is attached to each sound scene. A summary of those features (and the corresponding acronyms)

Terms	Acronyms
Sound level	L
Sound level (events)	$L(E)$
Sound level (textures)	$L(T)$
Average pleasantness (per scene)	$\mathcal{A}_{scene}$
Average pleasantness (per subject)	$\mathcal{A}_{subject}$
ideal/pleasant	i
non-ideal/unpleasant	ni
scene ideal/pleasant	i-scene
scene non-ideal/unpleasant	ni-scene

Table 1: Acronyms of features used in the analysis of the sensory experiments.

is presented in Table 1. In order to be consistent with the evaluation of experiment 1.b, features are not computed on the whole duration of the sequences but only on their 30-second reduced version used as stimuli for experiment 1.b (Section 3.3).

For each sound scene, three types of features are considered :

- *perceptual features*: the perceived pleasantness of the composed scene, assessed on a 7-point bipolar semantic scale. $\mathcal{A}_{scene}$ is the average pleasantness for one scene, as the average of the marks of all the subjects. $\mathcal{A}_{subject}$ is the average pleasantness for each participant, computed as the average of all the score given by each participant. Given the low number of subjects in experiment 1.b, we choose not to normalize the pleasantness scores. $\mathcal{A}_{subject}$ is computed for i-scenes and ni-scenes.

- *semantic features*: we use a boolean vector $S = (x_1, x_2, \ldots, x_n)$ that indicates the class of sounds involved in the scene, *i.e.* the sound classes that are present / absent from the scene. Each boolean x of this vector corresponds to a specific class of sounds: $x = 1$ if the class is present in the scene, and $x = 0$ otherwise. The vector dimension (n) depends on the level of abstraction that is considered for the analysis. For instance, for the abstraction level 1, that includes 44 classes of sounds, the dimension is thus 44 ($n = 44$).

- *structural features*: while *SimScene* allows us to acces a variety of information about the scene structure (such as the density of events), we

only focus in this first study on the sound levels. To figure those out, we draw inspiration from the L_{Aeq}. In our case, it is a scalar computed from the signal (in Volt and not in Pascal), and converted in decibels, with a reference of 1 Volt (full scale). The level is obtained by computing the quadratic mean of the signal every second and averaging the results over the total duration of the scene. An A-filtering module processes the data before the quadratic means are computed. We note L, $L(E)$ and $L(T)$ the computed levels by respectively considering the whole set of samples, only the set of event samples, and only the set of texture samples.

In order to evaluate the specific impact of the various sound sources on the perceived pleasantness, we run the data through the five following significance tests:

- *Analysis of perceived pleasantness*: the goal is to evaluate whether the perceived pleasantness is in accordance with the pleasantness label given by the creators of the i- and ni-scenes during the experiment 1.a. To do so, we consider if there exist significant differences for the i- and ni-scenes of the average $\mathcal{A}_{scene}$ and the average $\mathcal{A}_{subject}$. The significance is evaluated by a two-sample Student test for $\mathcal{A}_{scene}$ and by a paired-sample Student test for $\mathcal{A}_{subject}$.

- *Analysis of sound levels*: the goal is to evaluate whether the sound levels differ between the i- and ni-scenes. The significance is measured with a two-sample Student test.

- *Influence of sound levels on perceived pleasantness*: the goal is to evaluate if the sound levels impact the perceived pleasantness. To do so, we consider linear correlations between these features and $\mathcal{A}_{scene}$. The Pearson correlation coefficient is used for that purpose.

- *Analysis of semantic features*: the goal is to evaluate if specific classes are more frequently used to simulate a given type of environment (i or ni). To do so, a V-test is considered [37]. With c being the total number of classes used for both types of environment, c_k the number of classes used for a given type of scene k ($k = \{i, ni\}$), c_j the number of times a class j has been used, and c_{jk} the number of time a class j has been used for a given type of scene k, the V-test evaluates the null

hypothesis that the ratio $\frac{c_{jk}}{c}$ is not significantly different from the ratio $\frac{c_{jk}}{c_k}$. For each class j, and each environment type k, an approximation of the statistical value V_{jk} is computed as follows:

$$V_{jk} = \frac{c_{jk} - c_k \frac{c_j}{c}}{\sqrt{c_k \frac{c-c_k}{c-1} \frac{c_j}{c} (1 - \frac{c_j}{c})}} \tag{1}$$

If the null hypothesis is rejected, the class j is said to be typical with respect to the type of scene k. Such typical classes are called **sound markers**. Testing is done for each class, at each level of abstraction, and separately for texture and event classes.

- *Representation space induced by the semantic features*: the goal is to determine if a representation space of the scenes solely based on the presence / absence of sound sources allows us to distinguish between the two types of scene. Denoting as S_i the semantic features of scene i, we compute the distances between all S_i vectors. A Hamming distance is used: considering two n-dimension vectors $S_1 = (x_{1,1}, x_{1,2}, \ldots, x_{1,n})$ and $S_2 = (x_{2,1}, x_{2,2}, \ldots, x_{2,n})$, with $x \in \{0, 1\}$, the Hamming distance d_{ham} measures the proportion of coordinates that differ between the two vectors. It is defined as follows:

$$d_{ham}(S_1, S_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} \bigoplus x_{2,i}) \tag{2}$$

where $\bigoplus$ is the *exclusive-or* operator. Two scenes having similar source compositions will be close in such space. Using the Hamming distance allows us to take into account equally the presence and absence of classes. In order to measure the intrinsic ability of the space to discriminate between i- and ni-scenes, we use a ranking metric named the precision at rank k ($P@k$). The $P@k$ computes the precision obtained after the k closest items with respect to a given seed item have been found. Formally, for each s_i scene (considered as seed), we compute the proportion of s_j scenes in the k nearest neighbors of s_i that share the same label as s_i. The $P@k$ is then the average of this ratio for all the items considered as search seeds.

- *Influence of the sound markers on the perceived pleasantness*: in order

to assess the specific contributions of some sound sources, we again estimate the impact of the sound levels on the perceived pleasantness by taking into account only the sound markers for the computation of those features.

All statistical significance tests are conducted with a critical threshold of $\alpha = 0.05$. For the V-test, considering that a large number of classes is tested, a Bonferroni correction is applied. For the p-value, if $p \geq 0.05$, the value is reported. If $0.01 \leq p < 0.05$, we only report $p < 0.05$, otherwise we report $p < 0.01$.

3.5. Results

Analysis of perceived pleasantness

First, in order to ensure the coherence of the data, we check that none of the ni-scenes gets a $\mathcal{A}_{scene}$ higher than one of a i-scene. 4 ni-scenes do not fulfill that constraint: they are thus removed, together with their corresponding i-scenes. As a consequence, 36 i-scenes and 36 ni-scenes remain for analysis. Second, we verify that subjects really perceived a difference in terms of pleasantness between i- and ni-scenes. For that, we investigate the mean pleasantness score for each participant $\mathcal{A}_{subject}$, computed separately for each type of environment. It indeed appears that the i-scenes were perceived significantly more pleasant ($p < 0.01$) than the ni-scenes.

Analysis of sound levels

First, our analysis focuses on the sound levels. Figures 4a, 4b and 4c respectively depict the distributions of levels L, $L(E)$ and $L(T)$. There is a significant difference in terms of sound levels between i- and ni-scenes (L: $p < 0.01$). This difference is significant for events ($L(E)$: $p < 0.01$, mean deviation: -7 dB) and for textures ($L(T)$: $p < 0.01$, mean deviation: -6 dB).

As expected, the sound level of the sources is indeed a pleasantness indicator, as the ni-scenes tend to be louder. This results is also an outcome of a large number of related works. We also notice that this difference of sound levels is significant for both events and textures.

It appears that the biggest influence on the global sound levels comes from the events, the difference between L and $L(E)$ being only 1 dB for i- and ni-scenes. This observation is in agreement with the results obtained by

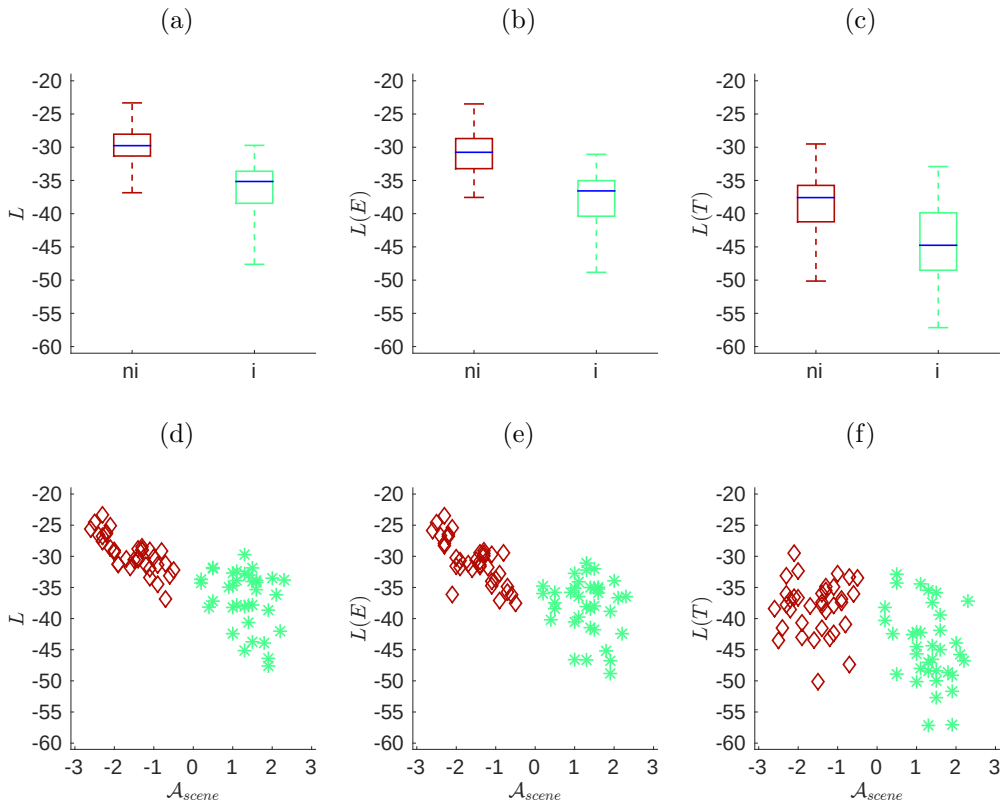


Figure 4: Distributions of the sound levels L (a, d), $L(E)$ (b, e) et $L(T)$ (c, f), with respect to scene type (a, b, c) and perceived pleasantness $\mathcal{A}_{scene}$ of experiments 1.b (d, e, f).

Kuwano *et al.* [38]. During their experiment, the authors ask their subjects to assess a set of soundscapes at a global level and then to do the same judgment at the time when they detect a sound source. The study shows that there is no significant difference between global and averaged instantaneous judgments. In our case, the result can be interpreted as if the subjects had unconsciously integrated this perceptual reality when composing the scenes, by allocating most of the global sound levels to well identified and relatively short sounds, *i. e.* the events.

However, level alone is not sufficient to fully differentiate between i- and ni-scenes. In fact, 20% of the i-scenes have a sound levels higher than the lowest level of the ni-scenes, while there is no overlap when considering the perceived pleasantness $\mathcal{A}_{scene}$.

	all scenes	i-scenes	ni-scenes
L	**-0.77** ($p < 0.01$)	-0.32 ($p = 0.06$)	**-0.78** ($p < 0.01$)
$L(E)$	**-0.75** ($p < 0.01$)	-0.20 ($p = 0.24$)	**-0.75** ($p < 0.01$)
$L(T)$	**-0.53** ($p < 0.01$)	-0.33 ($p = 0.05$)	-0.00 ($p = 0.99$)

Table 2: Linear correlation coefficients computed between mean perceived pleasantness $\mathcal{A}_{scene}$ of experiment 1.b and sound levels.

Influence of sound levels on the perceived pleasantness

In this section, more detailed relationships that could exist between sound levels and perceived pleasantness are investigated. Contrary to the previous test, we do not limit ourselves to a binary i-scenes *vs.* ni-scenes disctinction: we consider here the mean pleasantness $\mathcal{A}_{scene}$ as the perceptual feature. The aim is to investigate the level of correlation between sound levels and $\mathcal{A}_{scene}$. The linear correlation coefficients computed between $\mathcal{A}_{scene}$ *vs.* L, $L(E)$, $L(T)$ are shown in Table 2. Relationships between $\mathcal{A}_{scene}$ and the sound levels are depicted in Figure 4d, 4e and 4f.

Concerning L, a strong negative correlation with $\mathcal{A}_{scene}$ is measured ($r = -0.77$, $p < 0.01$), indicating that the higher the sound level is, the more unpleasant the scene is perceived. Nevertheless, Figure 4d suggests that this relationship does not occur in the same way for i- and ni-scenes. In fact, the correlation between L and $\mathcal{A}_{scene}$, remains high for ni-scenes ($r = -0.78$, $p < 0.01$), but is not significant ($p = 0.06$) for i-scenes.

When considering the whole set of scenes, the fact that the level is indeed a good indicator of pleasantness can be explained by the fact that the i-scenes tend to be softer than the ni-scenes, thus allowing us to extend artificially to the i-scenes the negative correlation observed for the ni-scenes.

We thus conclude that L:

- allows us to differentiate between i- and ni-scenes,
- characterizes precisely the perceived pleasantness for ni-scenes (unsurprisingly, an unpleasant scene gets all the more unpleasant as it gets louder),
- is not a relevant feature for modeling the perceived pleasantness of an *a priori* pleasant soundscape (i-scene).

The same conclusions can be drawn for $L(E)$, see Figure 4e. For $L(T)$, as shown on Figure 4f, the moderate correlation observed for the whole set of scenes disappears when separate scenes are considered (i-scenes: $r = -0.33$, $p = 0.05$, ni-scenes: $r = -0.00$, $p = 0.99$). Again, we believe that the negative correlation coming from the whole is an artifact due to the level difference between the two types of scenes (i-scenes tend to be softer than ni-scenes). Thus, while sound event levels maintain a relative ability to predict the pleasantness of the ni-scenes, texture levels do not bring much information, whatever the type of environment is.

To sum up, for an unpleasant environment, sound levels – especially those of events – negatively influence the perceived pleasantness. On the contrary, for a pleasant environment, none of the sound levels considered in the study seem to influence the perceived pleasantness.

Those first outcomes tend to show that 1) two modes of perception exist depending on the nature of the environment (i or ni), and 2) each involves distinct, independent features. The fact that L is not sufficient to characterize the pleasantness of the i-scenes can lead us to conclude that all sound sources do not equally contribute to the perception of pleasantness. We thus put forward the hypothesis that only the level of some of them can influence this perception. In order to investigate further in that direction, we analyze in the next section the scenes from a semantic point of view, *i. e.*we take an interest in the nature of the sources they are made of.

Analysis of the semantic features

We study the composition of the scenes by counting the number of subjects who used a given class of sounds to simulate a given type of environment (i or ni). For the 36 i and ni-scenes considered, results are shown on Figure 5a for events and on Figure 5b for textures. For the sake of clarity, a transitional level of abstraction between level 0 and 1, named 0+, is used to depict classes, see Figures A.9, A.10 and A.11.

We observe a noticable difference in terms of class choices between the i- and ni-scenes. The distribution of the classes is very similar to the one obtained in a related work on ideal urban soundscapes [5], *i. e.* on one hand, classes involving human presence and nature are prevailing in the i-scenes, and on the other hand, classes involving mechanical sounds and/or public works are prevailing in the ni-scenes. These results confirm a previously observed fact: the semantic nature of the sound sources play an important

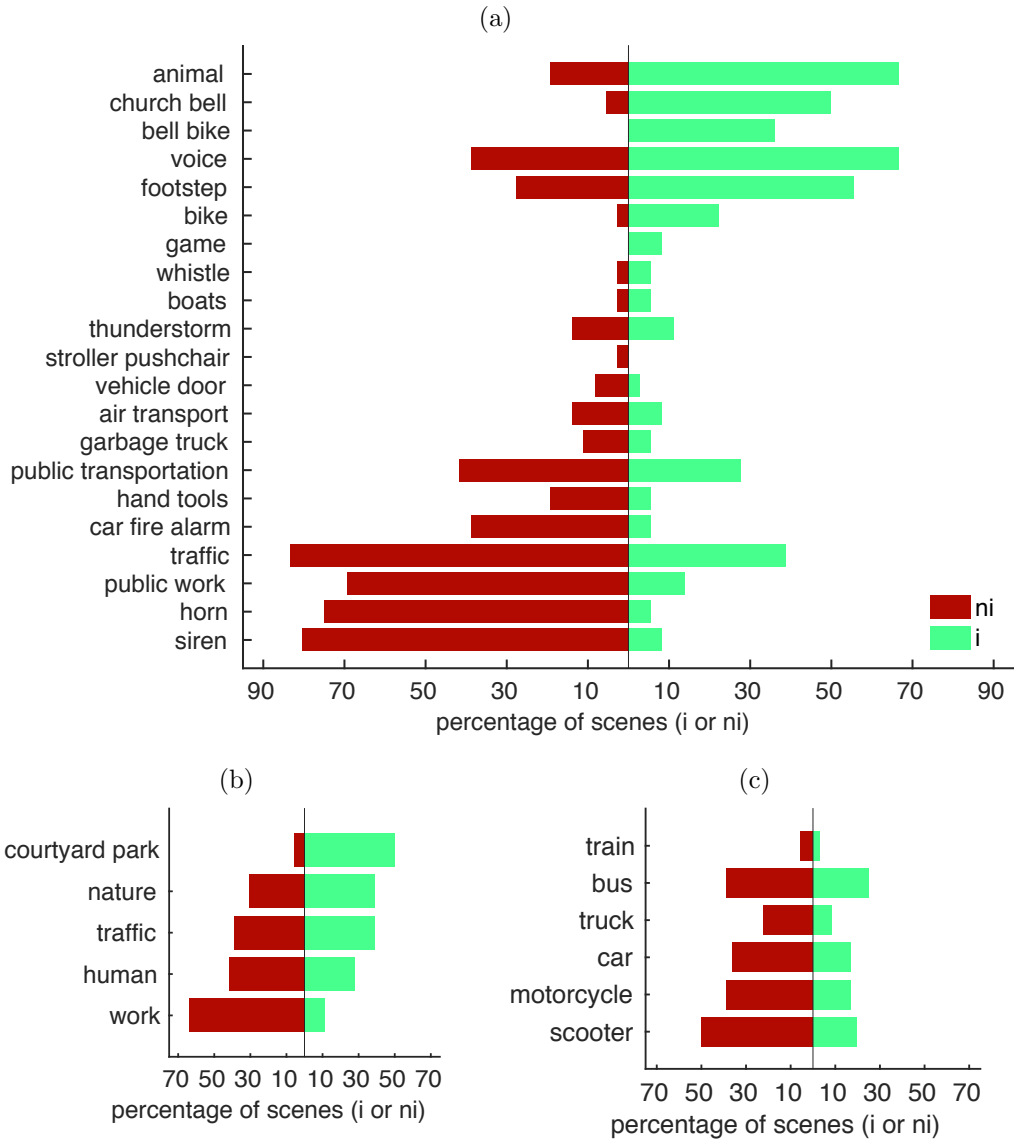


Figure 5: Proportion of simulated scenes (i or ni) involving a specific class of sounds: (a) event classes at an abstraction level 0+, (b) texture classes at an abstraction level 0, (c) event sub-classes at an abstraction level 1 belonging to *traffic* and *public transportation* classes of the abstraction level 0.

role in the assessment of the environment [19, 6].

Nevertheless, we notice some differences with [5]: Guastavino's results show that sounds of *public transportation* are specific of ideal urban soundscapes. The authors interpret this by the fact that the perception of pleasantness is, among other things, due to socio-cultural factors. Thus, in our representation of the world, sounds of public transportation are positively connoted and tend to be more accepted than sounds of personal vehicles.

To a certain extent, our results contradict this result. In fact, Figure 5a shows that *public transportation* classes (*bus* and *train*, cf. Figure 5c) have been used by the subjects for 28% of the i-scenes, and for 42% of the ni-scenes. Those results do not question the fact that sounds of *public transportation* are well accepted: 25% of the subjects used *bus* for the i-scenes, a level that is comparable to the *bike* class, and much higher than all the personal vehicles classes. However, *public transport* classes are also strongly present in the ni-scenes, for instance more than *light vehicle* or *truck* classes. On the basis of our results, the public transportation class cannot be considered as typical of an ideal urban soundscape.

This difference may be explained by the nature of the experimental protocol used in the two studies. As in our study, Guastavino asks the subjects to describe an environment. She asks them to perform this task starting only from their memories, whereas in our case subjects perform the same task using actual sound samples that they can listen to. The fact that subjects in our experiment are faced to the acoustic reality of the sounds for composing the environment may have decreased the socio-cultural impact. Other studies that considered sounds as stimuli have shown that the *bus* class can have a negative influence on the assessment of the environment [8].

Sound markers

We have shown that, from a qualitative point of view, the composition of the scenes in terms of sound sources differs between i- or ni-scenes. We now investigate whether some of the sound classes are specific to a given environment. For that purpose, the V-test detailed in Section 3.4 is considered separately for each abstraction level. Results are presented in Table 3.

Regarding the sound events, 9 markers are identified for all abstraction levels. As shown on Figure 5, classes related to human presence (*male footsteps on concrete*, *bicycle bell*), and of nature (*animals, bird*, and *bird song*) are i-scenes markers as well as the *church bell* class. This latter result

Abstraction level	Event sound markers	
	i-scenes	ni-scenes
0		construction work (3.78)
1	church bell (4.5) bicycle bell (4.3) animal (4.2)	car horn (3.9) siren (3.9)
2	bird (4.8) church bell (4.4) bicycle bell (4.2)	car horn (4.0) siren (4.0)
3	bird song (4.8) church bell (4.3) bicycle bell (4.2) foot steps (3.6)	klaxon (4.1) siren (4.0)

	Texture sound markers	
	i-scenes	ni-scenes
0	courtyard / park (4.1)	construction (3.9)
1	park (3.65)	crossroad (3.6) construction vehicule (3.3)
2	park (3.64)	crossroad (3.56)

Table 3: Event and texture classes identified as sound markers. In each cell, markers are ranked as decreasing order of V-test value, shown between parenthesis. $p \leq 0.01$ for all sound markers.

may be due to the socio-cultural background of the subjects who are mostly European citizens. In fact, according to Schafer, a sound that is identified by a person as being an important element of his/her environment, is well accepted. Sound markers of ni-scenes are classes related to construction site (*construction works*), or suggesting a strong traffic (*horn*, *siren*).

Regarding the textures, 5 markers are identified. For the i-scenes, those classes related to subdued or quiet ambiances (*courtyard*, *park*). The marker classes for the ni-scenes are, as for the events, related to construction site (*construction*, *construction vehicle*), together with a class related to traffic (*crossroads*).

Although the whole set of identified markers are rather intuitive, none of the event classes related to the noise of motor vehicles are identified as markers, except for the texture class *crossroads*. To generate an unpleasant traffic, subjects chose the classes *horn* or *siren*. We thus conclude that isolated motor vehicle sounds are understood as being part of the urban en-

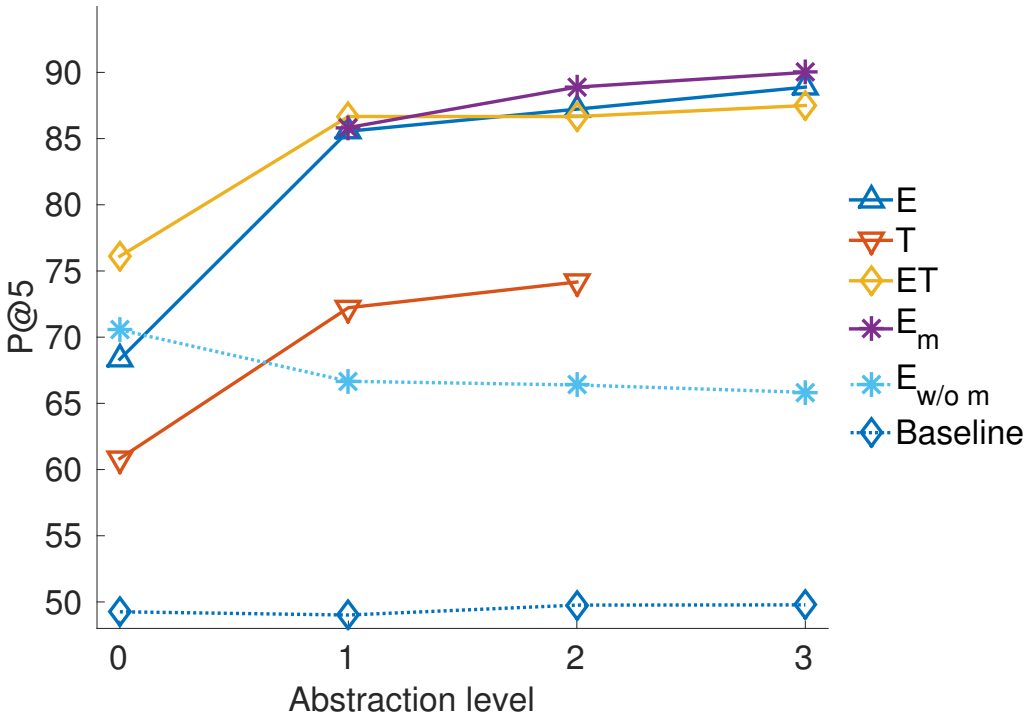


Figure 6: P@5 obtained by considering the dissimilarity matrix computed from the paired Hamming distances of the semantic features vectors as a function of the abstraction level. The vectors are built by using all the classes (ET), only the event classes (E), only the texture classes (T), only the event classes corresponding to sound markers (E_m), or only the event classes excluding sound markers $E_{w/o,m}$. Baseline results are achieved by considering random vectors as input.

vironment, and thus their nature is not necessarily linked to an unpleasant soundscape.

Representation space induced by semantic features

In this section, we evaluate at which level a semantic representation of the scenes allows us to discriminate between the two types of environment. For this purpose, a rank 5-precision (p@5) is computed on the space induced by the semantic features S, and for each abstraction level (see Section 3.4). The vectors S are built by using all the classes (ET), only the event classes (E), only the texture classes (T), only the event classes corresponding to sound markers (E_m), or only the event classes excluding sound markers ($E_{w/o,m}$).

Texture classes corresponding to sound markers are not numerous enough to reliably compute the metric, and are thus not considered. For the same reasons, event classes corresponding to sound markers at abstraction level 0 are also discarded. Results are shown on Figure 6.

Concerning ET, the p@5 is 76% at abstraction level 0 (the most abstract), and remains above 86% for higher abstraction levels. Considering only the presence / absence of sound classes thus allows us to nicely discriminate between the two types of environments. We also notice that the less abstract (and therefore more precise) the description is, the more effective it is to predict agreement.

Considering E and T separately, it appears that: 1) the p@5 obtained with E is similar to the one obtained with ET ; 2) the p@5 obtained with T is always lower than the one obtained with E, by 10% to 15%. Those results indicate that the semantic information that is discriminative is mostly carried by the events. Those results are in line with [16]. As discussed in Section 2.1, it seems that humans analyze the event scenes which are composed of several sound events in a causal manner, *i. e.* by identifying the sources.

The dimension of the vectors S for E_m is lower than the dimension of vector S for E, itself lower than the dimension of vector S obtained when all the classes are considered (ET). S being a boolean vector, the smaller the dimension, the lower the amount of information it can carry. Despite this, it appears that the p@5 obtained with E_m is equal – or superior – to the ones obtained with E or ET, although only a partial information is used in that case to describe the scenes. Reciprocally, if the sound markers are not taken into account for the description ($E_{w/o,m}$), the performance is below the one achieved when considering only the textures as features. Thus, most of the semantic information allowing to differentiate between i- and ni-scenes is included in the markers.

To sum up, the outcomes of this analysis are:

1. unlike what we outlined with the sound levels, a semantic description of the scenes composition in terms of presence / absence of sound sources allows us to reliably differentiate between the two types of environments (i or ni);
2. the semantic information is mainly contained in the event sound classes;
3. only a part of the event classes, i.e. the sound markers, are useful to differentiate between the i- and ni-scenes.

	i-scenes	ni-scenes
L_m	0.03 ($p = 0.88$)	**-0.75** ($p < 0.01$)
$L(E)_m$	0.08 ($p = 0.66$)	**-0.71** ($p < 0.01$)
$L(T)_m$	-0.11 ($p = 0.66$)	-0.17 ($p = 0.37$)
L_b	**-0.52** ($p < 0.01$)	-0.32 ($p = 0.06$)
$L(E)_b$	**-0.51** ($p < 0.01$)	-0.30 ($p = 0.07$)
$L(T)_b$	-0.32 ($p = 0.05$)	**-0.73** ($p < 0.01$)
$L_m - L_b$	**0.67** ($p < 0.01$)	-0.31 ($p = 0.07$)
$L(E)_m - L(E)_b$	**0.66** ($p < 0.01$)	-0.28 ($p = 0.10$)
$L(T)_m - L(T)_b$	0.16 ($p = 0.54$)	0.21 ($p = 0.28$)

Table 4: Linear correlation coefficients computed between mean perceived pleasantness $\mathcal{A}_{scene}$ (Exp. 1.b) and sound levels related to the presence of sound markers.

Since we have extracted the typical classes of the i- and ni-scenes and verified that the distinction between them was largely dependent on the presence of these classes, we shall now investigate whether a description of the scenes only based on the sound pressure level of sound markers could characterize the perceived pleasantness, perhaps better than a globally computed sound level.

Influence of sound marker levels on the perceived pleasantness

To do so, the correlations between $\mathcal{A}_{scene}$ and the sound levels are evaluated. In this section, the sound levels are computed by taking into account only the previously identified sound markers. We define X_m, the X feature computed by taking into account only the sound markers, and X_b, the X feature computed by taking into account all the sound classes, except the sound markers. When the feature characterizes an i-scene (resp. ni-scene), only the markers identified for the i-scenes (resp. ni-scenes) are considered. We henceforth call *i-markers* and *ni-markers* the two types of markers. Results are shown on Table 4.

Concerning the sound levels, the same trends are measured between L_m, $L(E)_m$ and $L(T)_m$ on one side, and L, $L(E)$ and $L(T)$ on the other side, see Figures 7a and 7d, respectively. No matter whether all the classes or only the markers are considered, it appears that:

1. a significant difference between levels of i- and ni-scenes exists (L_m, $L(E)_m$ et $L(T)_m$: $p < 0.01$);

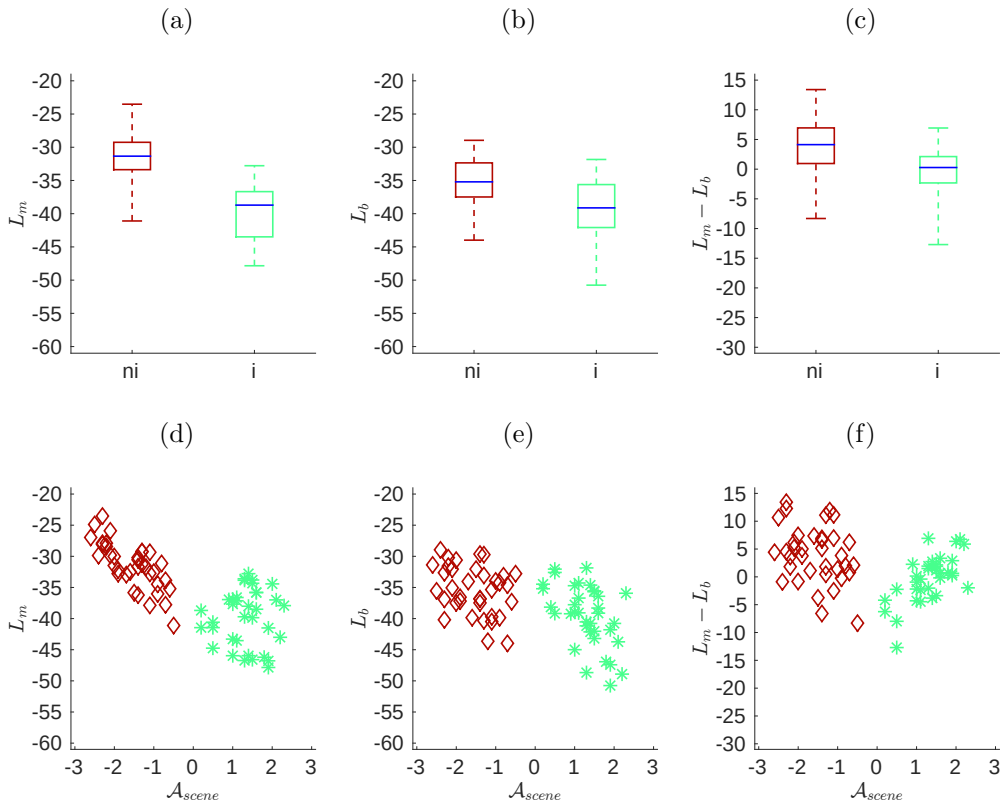


Figure 7: Distribution of the relative sound levels related to the presence of markers L_m (a, d), L_b (b, e) and $L_m - L_b$ (c, f), versus scene type (a, b, c) and perceived pleasantness $\mathcal{A}_{scene}$ of experiment 1.b (d, e, f).

2. the sound level of scenes is mainly related to the sound events, compared to the textures;
3. the sound level of events has an influence on the perception of pleasantness for ni-scenes, but not for i-scenes;
4. the sound level of textures does not play any role in the perception of the pleasantness.

To conclude, the level of ni- markers has a negative influence on the pleasantness for the ni-scenes. On the other hand, the level of i- markers does not influence the perceived pleasantness for the i-scenes.

Considering the non markers classes, we can observe on the i-scenes results, a weak negative correlation for L_b ($r = -52$, $p < 0.01$) et $L(E)_b$

($r = -51$, $p < 0.01$), see Figures 7b and 7e. It is the first time that an objective feature allows us to define the pleasantness of i-scenes. This leads us to conclude that the level of non-typical sound classes of a pleasant environment has a negative influence on the pleasantness.

Moreover, whereas $L(T)$ did not show any significant correlation for ni-scenes, a strong negative correlation is observed for $L(T)_b$ ($r = -0.73$, $p < 0.01$). This indicates that the level of non-marker texture classes does not influence the perceived pleasantness in the same way for i- and ni-scenes. Sound level of textures seems to have a negative effect for the ni-scenes, and no significant effect for the i-scenes.

A last group of features is now considered, namely $L_m - L_b$, $L(E)_m - L(E)_b$ and $L(T)_m - L(T)_b$. These features describe the difference between the markers level and those of the other sound classes, see Figures 7c and 7f. They express the saliency of the markers with respect to the sound mixture.

For the i-scenes, a moderate positive correlation is observed for $L_m - L_b$ ($r = 0.67$, $p < 0.01$) and $L(E)_m - L(E)_b$ ($r = 0.66$, $p < 0.01$). for the ni-scenes no correlation is observed. Thus, for the i-scenes, it is not the absolute markers level that is important, but their relative level with respect to the other sounds composing the scene. A double perceptual mechanism for the ideal environments can thus be observed:

- the higher the absolute level of sounds not being i-markers, the weaker the pleasantness,
- the higher the relative level of i-markers is, with respect to the remaining sounds, the higher the pleasantness.

On the contrary, the fact that we observe significant correlations for L_m and $L(E)_m$ and no correlation for $L_m - L_b$ and $L(E)_m$ - $L(E)_b$ for the ni-scenes, shows that this is indeed the absolute level that matters.

3.6. Discussion

From this analysis, the following points can be outlined:

- *differentiating i- and ni-scenes*: the semantic features, and the global sound levels (L, $L(E)$ et $L(T)$), allow us to differentiate reliably between i- and ni-scenes. The semantic description seems to be more powerful;

- *events or textures*: whatever the feature type, be it semantic or related to the sound pressure level, events are the most useful components of the scene to differentiate the two types of scenes; textures bring a limited amount of information;

- *pleasantness prediction*: considering the correlation between sound levels and pleasantness, it seems that the way subjects perceived the quality of a given environment depends on its very nature (i or ni). From the data gathered in those experiments, the same set of features cannot be considered to predict the pleasantness of i- and ni-scenes:

  - *for ni-scenes*, the global levels (L et $L(E)$), or the level of sound markers (L_m et $L(E)_m$), have a negative influence on pleasantness. Taking into account the contribution of each of the different sources of the scene does not improve the prediction performance compared to a global analysis of the environment.
  - *for i-scenes*, on the contrary, the sound markers characteristics and those of the other sounds have to be considered separately to predict the pleasantness. The markers level relative to the background noise ($L(E)_m - L(E)_b$ and $L_m - L_b$) is positively correlated to the pleasantness, whereas the noise level (L_b and $L(E)_b$) is negatively correlated.

The fact that the pleasantness of the i-scenes is not correlated to global physical features, contrary to the pleasantness of ni-scenes, has also been studied recently [11].

We can assume two perceptual modes of operation that involve different types of features and rely on the hedonic nature of the stimuli. It thus appears that the features considered for the pleasantness judgment also depends on a preliminary judgment of the global hedonic nature of the environment (ideal or non ideal).

A similar phenomenon is observed for the perception of textures, see Section 2.1. It seems that the brain selectively adapts the way it encodes the information (statistic summary for textures, finer description for events) following a previous decision making process based on the nature of the stimuli, *i. e.* is it an event or a texture ?

Another hypothesis would be that the volume somehow act as an hedonic "gain" factor. If the volume of a negative marker is high, it lowers the

overall pleasantness. If the volume of a positive marker is high, it highers the overall pleasantness. Evidently, the positive gain is expected to saturate at a given level and will quickly decreases as the level raises above a given threshold.

4. Experiment 2: modification of the semantic content

4.1. Objective

The previous experiment demonstrated that, among the classes of sounds occurring in urban soundscapes, those gathering markers are specific to some environments. Those sound markers seem to have a great impact on perception. This impact is studied here in more detail using an added benefit of the simulation paradigm proposed in this paper, the ability to manipulate and modify the generated scenes.

In order to investigate deeper into the relation between the pleasantness of the ideal and non ideal scenes and the sound markers, the audio waveform of the simulated scenes are regenerated with or without the classes identified as markers. To do so, i-markers are removed from the i-scenes, and ni-markers are removed from the ni-scenes. To evaluate the impact on the perception of pleasantness caused by those modifications, a perceptual test is conducted with a protocol close to the one considered in experiment 1.b.

The objective of this experiment is to study if the removal of the previously identified markers have an impact on the perceived pleasantness. Two hypothesis are thus formulated:

- *for the ni-scenes*, we hypothesize that the absence of ni-markers will **increase** the pleasantness score;
- *for the i-scenes*, we hypothesize that the absence of the i-markers will **decrease** the pleasantness score.

If the first hypothesis is rather intuitive, the second is less. Indeed, it is not obvious that the removal of the i-markers, though perceptively positively connoted, will decrease the global quality of a soundscape, since this removal also decreases the global sound level of the scene. However, as discussed before, the global amplitude level can only be considered as a partial indicator

of pleasantness of the ideal soundscapes. Furthermore, the level of i-markers positively impact the pleasantness. For those reasons, the validation of the second hypothesis is of high interest.

4.2. Planning of experiment 2

Stimuli

There are 144 stimuli of 30 seconds duration. More precisely:

- *72 km-scenes*: the 72 scenes originally simulated by the subject of experiment 1.a where the sound classes identified as markers are kept. The 36 ideal scenes with markers are noted i/km-scenes, and the 36 non ideal scenes with markers are noted ni/km-scenes (km for kept markers).
- *72 rm-scenes*: 72 scenes where the sound classes identified as markers are removed. The 36 ideal scenes without markers are noted i/rm-scenes, and the 36 non ideal scenes without markers are noted ni/rm-scenes (r for removed markers).

Notwithstanding the presence or absence of markers, the km-scenes and rm-scenes are exactly the same.

In order to create rm-scenes with still some sound diversity and no absence of sound activity within a long period of the scene, only the sound classes of events of the first level of abstraction are removed, see Table 3. Those classes are:

- *church bell*, *bicycle bell*, and *animals* for the i/rm-scenes;
- *siren*, *car horn* for the ni/rm-scenes.

Thus, only part of the i-markers and ni-markers are removed from the rm-scenes.

Experiment

The subjects evaluate the 144 scenes. The evaluation is done on a 11-point bipolar semantic scale ranging from -5 (non-ideal / very unpleasant) to +5

(ideal / very pleasant). Before rating a scene, subjects must listen to the first 20 seconds. After scoring, they are free to move on to the next scene.

For each subject, the scenes are presented in a random order. The first 10 scenes allow the subject to calibrate their scores. Those calibration scenes are 5 i/km-scenes and 5 ni/km-scenes. These first 10 scenes are replayed at the end of the experiment, and only the scores given at the second occurrence are taken into account.

The experiment is scheduled to last 1 hour. The subjects do not know the nature of the scenes.

Experimental protocol

All the subjects perform the test at the same time, on the same type of computers, in a quiet environment. They listen to the stimuli with semi-open *Beyer-Dynamic DT 990 Pro* headphones, with the sound level set to be the same for all the subjects. They are not allowed to talk to each other during the experiment.

A supervisor is available during the whole duration of the experiment to monitor it and answer to any question the subjects may have.

Subjects

12 subjects perform the test (8 male, 4 female; averaging 29.5 years of age, s.d. of 14 years). None of them have performed experiments 1.a and 1.b. All the subjects had been living in Nantes, France, for at least two years at the time of the experiment. All the subjects performed the test successfully.

4.3. Data and analysis method

The type of data analyzed in this experiment have been considered for experiment 1.a, see Section 3.4 for details.

The aim here is to validate the hypothesis that the removal of i-markers and ni-markers have an impact on the perceived pleasantness. To do so, we perform a variance analysis. We consider $\mathcal{A}_{subject}$ as a dependent variable, and as independent variables, the type of environment (i/ni) and the presence/absence of markers (km/rm). As each subject evaluated the whole set of scenes, a two factors repeated measure ANOVA is used to evaluate whether there exist a significant difference of perceived pleasantness between

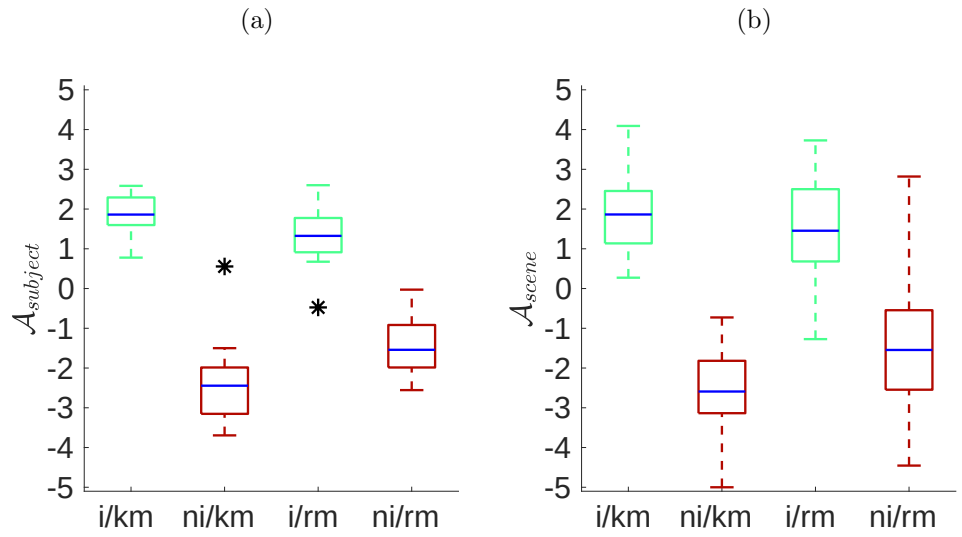


Figure 8: Distribution of the mean perceived pleasantness per subject $\mathcal{A}_{subject}$ (a) and the mean perceived pleasantness per scene $\mathcal{A}_{scene}$ (b) versus scene type. Black stars in subfigure (a) stand for the detected outlier, *i. e.* subject 7.

km-scenes and rm-scenes. The two independent variables are considered as *within-subject* factors. The factors being of only two levels each (type: i/ni, marker: km/rm), the sphericity hypothesis does not need to be checked. *Post hoc* analyses are done using the Tukey-Kramer procedure.

All the significance tests are performed with critical threshold $\alpha = 0.05$.

4.4. Results

4.4.1. Outliers detection

Let us consider $\mathcal{A}_{subject}$ for the km-scenes (Figure 8a). Close inspection of the answers show that one subject's judgments strongly differ from the others'. This subject evaluated positively close to half of the ni/km-scenes (see Annex Appendix B, subject 7), and gave a score above 0 for 58% of the ni/km-scenes, contrary to an average of 11% for the other subjects.

Furthermore, this subject used the whole scale (from -5 to 5) to score both the i-scenes and the ni-scenes. Those behaviors strongly differ with the ones of the other subjects, be they from this experiment or the previous ones. Subject 7 is thus discarded from the analysis.

4.4.2. Influence of the markers on perceived pleasantness

In this section, we study the scores given by the subject while listening to the several types of scenes, namely the i/km-scenes, ni/km-scenes, i/rm-scenes and ni/rm-scenes (Figure 8b). The repeated measure ANOVA applied to $\mathcal{A}_{subject}$ shows a significant effect of the type environment (i/ni: $F[1, 10] = 175$, $p < 0.01$), of the presence/absence of markers (km/rm: $F[1, 10] = 7$, $p < 0.05$), and of the interaction between those two factors ($F[1, 10] = 67$, $p < 0.01$).

The *post hoc* analysis exhibits significant differences between all groups of observations, notably between the i/km- and the i/rm scenes ($p < 0.05$) as well as between the ni/km- and the ni/rm-scenes ($p < 0.01$).

These results indicate that the removal of the markers indeed modify the perception of the scenes by the subjects. Our two hypotheses are thus verified :

- the removal of ni-markers improved the pleasantness of the ni-scenes;
- the removal of i-markers reduced the pleasantness of the i-scenes.

The significant interaction shows that the type of environment influences the effect of the presence/absence of the markers. Indeed, the average difference between km-scenes and rm-scenes is larger for the ni-scenes (1.1) than for the i-scenes (0.5).

4.5. Discussion

This experiment shows that the presence of the markers identified during the analysis of experiment 1 does have an impact on the perceived pleasantness. The removal of the ni-markers has a positive effect on the perception of the ni-scenes. Perhaps more surprisingly the removal of the i-markers slightly decreases the perception of the i-scenes: this is a more striking observation since, due to the removal of the markers, the acoustic pressure level of the i/km-scenes is higher than the one of the i/rm-scenes.

This strongly confirms that i-markers do have a positive impact on the perception of an environment. The fact that their removal decreases $\mathcal{A}_{scene}$ indicate that it should be possible to improve the perceived pleasantness of a given urban area by the addition of sounds commonly considered as pleasant such as *bird calls*. Those conclusions are in line with the positive approach introduced by Schafer in [39].

5. Outcomes for soundscape perception

This series of three experiments showed that most of the descriptors used in this study, be they of a semantic or acoustic nature, allow us to distinguish between an ideal scene and a non ideal one.

That being said, we observe that the physical characteristics correlated with the perceived pleasantness clearly differ depending on the type of scenes. In the case of ideal scenes, it is above all the emergence of sound markers that determines the perceived quality, whereas in the case of non-ideal scenes, it is the overall sound level that prominently influences it.

These results show that the perception of the qualities of a scene indeed depends primarily on its identifiable sound sources. The characteristics that are taken into account during the perceptual process appear to vary from one source to the other, from one type of environment to the other. This fact leads the authors to believe that there is little hope to find a holistic physical descriptor that can adequately account for the affective qualities of all types of sound environment.

Those results may have an impact on the relevant strategies to adopt while trying to improve the quality of a sound environment:

- in the case of non ideal scenes, one should focus on reducing the acoustic pressure level, whether globally, or by discarding specific sources such as *sirens* or *car horns*

- in the case of ideal scenes, one should first identify which sources are pleasant to the targeted community, second lower the volume of the other sound sources, and, if possible, raise the contribution or add positive sound markers.

Finally, this work allows us to conjecture as to the nature of the mental representations of the concepts "pleasant (urban sound) environment" (PE) and "unpleasant (urban sound) environment" (UE).

First, the fact that the semantic information (which sound source is present) and compositional information (at which level) are different for ideal and non-ideal scenes leads us to believe that these two types of information characterize the PE and UE concepts.

Second, the fact that the removal of sound markers changes the perceived pleasantness leads us to believe that the abstract concept related to pleasantness depends on the activation of a network of concepts strongly

linked to the sources which are in the case of this study: *bird*, *church bell* and *bicycle bell*.

6. Conclusion

The outcomes of the three experiments described in this paper demonstrate the usefulness of considering a dedicated simulation tool such as *simScene* in order to scientifically question the perception of soundscapes in an innovative way. We also believe that its wider usage could enable urban planning decision-makers to question an entire community about its own representations of the sound environment to which they are exposed, and about the representations of the sound environments to which they would like to be exposed to.

Future work should consider several other simulation experiments by changing the emotional qualities (quiet, comfortable, troublesome, etc.), but also by specifying specific urban locations (park, square, street, etc.), in order to provide to the scientific community an entire corpora of cognitively informed soundscapes.

There are also many more avenues of research to fully explore the capabilities of the proposed paradigm; first by taking into account a wider range of structural features (for example the density or regularity of events), and second by studying further the effects caused by the voluntary modification of scene composition, as during the suppression of the sound markers practiced in experiment 2.

One interesting avenue in this direction would be to study the impact of adding positively appreciated sound markers to a non ideal scene in order to study the hypothesis that this type of addition would improve the perceived quality of the scene.

Finally, one should study the influence of socio-cultural contexts on perception. Indeed, if the sound of the church bell is most often a quality environment marker for a Westerner, this does not necessarily hold true for subjects of Eastern, Middle Eastern or other cultures.

Once again, besides the interesting possibilities already mentioned, the simulation protocol presented here as well as its implementation brings in this case two important advantages:

- The simulator can be deployed on a large scale via the Internet thanks to the web-based software architecture;

- Simulated scenes can be analyzed without the need to take into account the different mother tongues of the subjects, the semantic nature of the classes of sounds used being known *a priori* by the experimenter and without the need to analyze and annotate the sound scene to identify the sources, their occurrences in the simulated scene being directly available.

Acknowledgements

Research project partly funded by ANR-11-JS03-005-01. The authors would like to thank the students of the Ecole Centrale de Nantes for their willing participation.

References

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

[2] Ö. Axelsson, B. Berglund, M. E. Nilsson, Soundscape assessment, The Journal of the Acoustical Society of America 117 (4) (2005) 2591–2592.

[3] W. J. Davies, M. D. Adams, N. S. Bruce, R. Cain, A. Carlyle, P. Cusack, D. A. Hall, K. I. Hume, A. Irwin, P. Jennings, et al., Perception of soundscapes: An interdisciplinary approach, Applied acoustics 74 (2) (2013) 224–231.

[4] R. Cain, P. Jennings, J. E. Poxon, The development and application of the emotional dimensions of a soundscape, Applied Acoustics 74 (2) (2013) 232–239.

[5] C. Guastavino, The ideal urban soundscape: Investigating the sound quality of french cities, Acta Acustica united with Acustica 92 (6) (2006) 945–951.

[6] D. Dubois, C. Guastavino, M. Raimbault, A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories, Acta acustica united with acustica 92 (6) (2006) 865–874.

[7] B. Defréville, C. Lavandier, M. Laniray, Activity of urban sound sources, in: Proceedings of the 18th International Congress in Acoustics (ICA), Kyoto, Japan, 2004.

[8] C. Lavandier, B. Defréville, The contribution of sound source characteristics in the assessment of urban soundscapes, Acta Acustica united with Acustica 92 (6) (2006) 912–921.

[9] M. E. Nilsson, Soundscape quality in urban open spaces, in: Proceedings of the 36th International Congress and Exposition on Noise Control Engineering (InterNoise), Istanbul, Turkey, 2007.

[10] B. Szeremeta, P. H. T. Zannin, Analysis and evaluation of soundscapes in public parks through interviews and measurement of noise, Science of the total environment 407 (24) (2009) 6143–6149.

[11] G. R. Gozalo, J. T. Carmona, J. B. Morillas, R. Vílchez-Gómez, V. G. Escobar, Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes, Applied Acoustics 97 (2015) 1–10.

[12] P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, P. Aumond, Sound quality indicators for urban places in paris cross-validated by milan data, The Journal of the Acoustical Society of America 138 (4) (2015) 2337–2348.

[13] N. S. Bruce, W. J. Davies, M. D. Adams, Development of a soundscape simulator tool, in: Proceedings of the 38th International Congress and Exposition on Noise Control Engineering (InterNoise), Ottawa, Canada, 2009.

[14] N. S. Bruce, W. J. Davies, The effects of expectation on the perception of soundscapes, Applied Acoustics 85 (2014) 1–11.

[15] M. Bostock, V. Ogievetsky, J. Heer, D^3 data-driven documents, IEEE Transactions on Visualization and Computer Graphics 17 (12) (2011) 2301–2309.

[16] V. Maffiolo, De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain, (*Semantic and acoustical characterisation of the sound quality of urban environment*), Ph.D. thesis, Université du Mans, Le Mans, France (1999).

[17] M. Raimbault, Simulation des ambiances sonores urbaines: intégration des aspects qualitatifs, *Urban soundscape simulation: focusing on qualitative aspect*), Ph.D. thesis, Unversité de Nantes - Ecole polytechnique de Nantes, Nantes, France (2002).

[18] C. Guastavino, Etude sémantique et acoustique de la perception des basses fréquences dans l'environnement sonore urbain, (*Semantic and acoustic study of lowfrequency noises perception in urban sound environment*), Ph.D. thesis, Unversité Paris VI UPMC, Paris, France (2003).

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

[20] A. Devergie, Relations entre perception globale et composition de séquences sonores, Master's thesis, IRCAM, Paris VI UPMC (2006).

[21] M. E. Niessen, C. Cance, D. Dubois, Categories for soundscape: toward a hybrid classification, in: Proceedings of the 39th International Congress and Exposition on Noise Control Engineering (InterNoise), Vol. 2010, Lisbon, Portugal, 2010, pp. 5816–5829.

[22] J. Beaumont, S. Lesaux, B. Robin, J.-D. Polack, C. Pronello, C. Arras, L. Droin, Pertinence des descripteurs d'ambiance sonore urbaine, Acoustique et techniques.

[23] J.-D. Polack, J. Beaumont, C. Arras, M. Zekri, B. Robin, Perceptive relevance of soundscape descriptors: a morpho-typological approach, Journal of the Acoustical Society of America 123 (5) (2008) 3810.

[24] A. Leobon, Analyse psycho-acoustique du paysage sonore urbain, (*Psychoacoustic analysis of urban soundscape*), Ph.D. thesis, Unversité Louis Pasteur, Strasbourg, France (1986).

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

[26] N. Saint-Arnaud, Classification of sound textures, Master's thesis, Massachusetts Institute of Technology (1995).

[27] A. S. Bregman, Auditory scene analysis: The perceptual organization of sound, MIT press, Cambridge, MA, 1994.

[28] R. P. Carlyon, How the brain separates sounds, Trends in cognitive sciences 8 (10) (2004) 465–471.

[29] J. H. McDermott, E. P. Simoncelli, Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis, Neuron 71 (5) (2011) 926–940.

[30] J. H. McDermott, M. Schemitsch, E. P. Simoncelli, Summary statistics in auditory perception, Nature neuroscience 16 (4) (2013) 493–498.

[31] I. Nelken, A. de Cheveigné, An ear for statistics, Nature neuroscience 16 (4) (2013) 381–382.

[32] E. Rosch, B. B. Lloyd, Cognition and categorization, Hillsdale, New Jersey, 1978.

[33] J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: Proceedings of the 22st ACM International Conference on Multimedia, Orlando, FL, USA, 2014.

[34] G. Lafay, N. Misdariis, M. Lagrange, M. Rossignol, Semantic browsing of sound databases without keywords, Journal of the Audio Engineering Society 64 (9) (2016) 628–635.

[35] M. Rossignol, G. Lafay, M. Lagrange, N. Misdariis, Simscene: a web-based acoustic scenes simulator, in: Proceedings of the Web Audio Conference (WAC), IRCAM, Paris, France, 2015.

[36] G. Lafay, M. Rossignol, N. Misdariis, M. Lagrange, J.-F. Petiot, A new experimental approach for urban soundscape characterization based on sound manipulation: A pilot study, in: Proceedings of the International Symposium on Musical Acoustics (ISMA), SFA, Le Mans, France, 2014.

[37] R. Rakotomalala, A. Morineau, The tvpercent principle for the counterexamples statistic, in: Statistical Implicative Analysis, Springer, 2008, pp. 449–462.

[38] S. Kuwano, S. Namba, T. Kato, J. Hellbrück, Memory of the loudness of sounds in relation to overall impression, Acoustics Science and Technics 4 (24).

[39] R. Schafer, The Tuning of the World, Borzoi book, Knopf, New York, 1977, (Reprinted as *Our Sonic Environment and the Soundscape: The Tuning of the World*. Destiny Books, 1994).

Appendix A. Taxonomy of sound classes

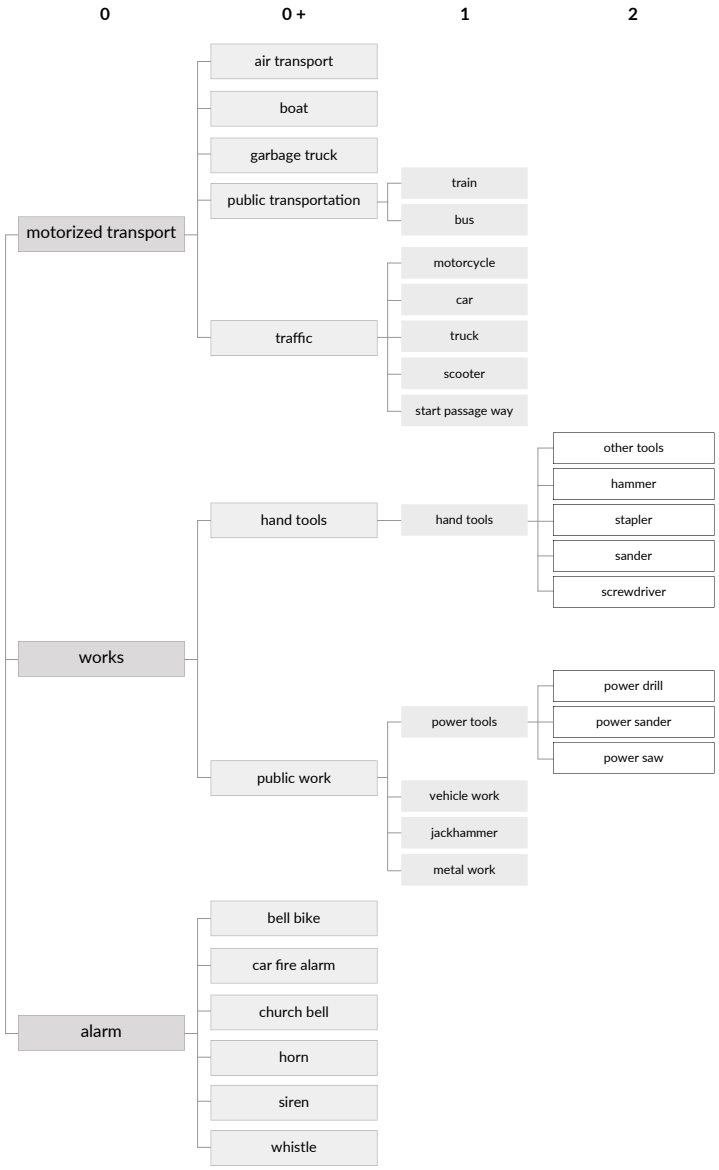


Figure A.9: Taxonomy of sound classes of mechanical events used for the simulation urban soundscapes with level of abstraction from left to right.

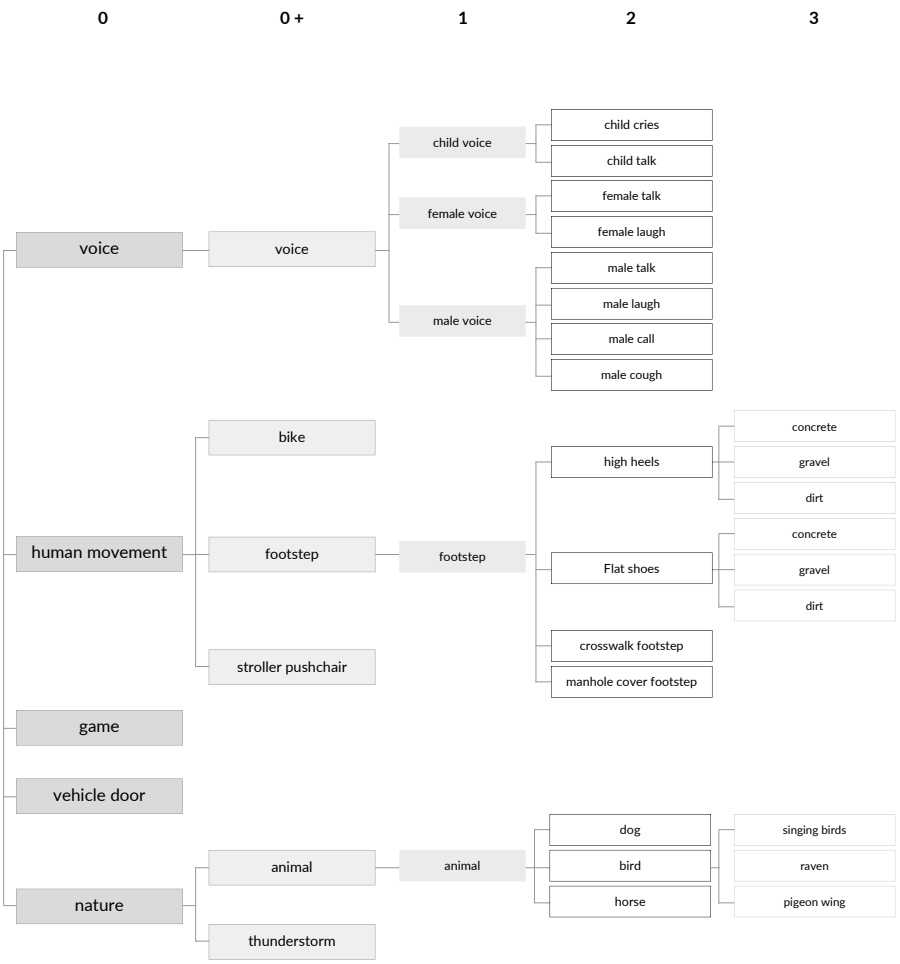


Figure A.10: Taxonomy of sound classes of non-mechanical events used for the simulation urban soundscapes with level of abstraction from left to right.

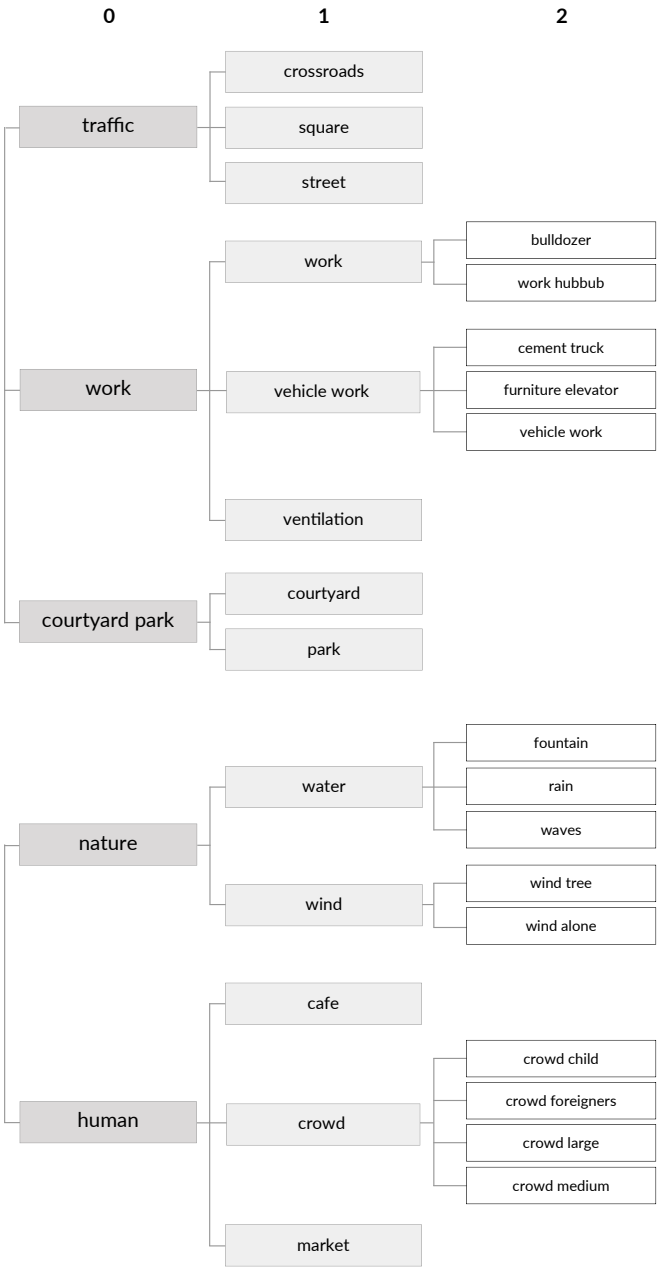


Figure A.11: Taxonomy of sound classes of textures used for the simulation urban soundscapes with level of abstraction from left to right.

Appendix B. Distribution of pleasantness scores given by the subjects during experiment 2

Figure B.12: Distribution of scores given by the subjects during experiment 2 for i/am-scenes (green) et ni/am-scenes (red). The horizontal axis is, from left to right, the 11-point bipolar semantic scale ranging from -5 (non-ideal / very unpleasant) to +5 (ideal / very pleasant).