Investigating soundscapes perception through acoustic scenes simulation

Grégoire Lafay, Mathieu Lagrange, Jean-François Petiot, Mathias Rossignol, Nicolas Misdariis

Approaching mental representations of urban soundscape using auditory scenes simulation

Grégoire Lafay, Mathieu Lagrange, and Jean François Petiot
Institut de Recherche en Communications et Cybernétique de Nantes (IRCCYN), Ecole centrale de Nantes, France

Mathias Rossignol and Nicolas Misdariis
Institut de Recherche et Coordination Acoustique/Musique (IRCAM), France

This paper introduces a new subject-centered experimental protocol to study mental representations of urban soundscapes through a simulation process. Subjects are asked to recreate a full sound environment by means of a structured sound data set and a software dedicated to sound manipulation. This paradigm is used to characterize urban sound environment representations, by analyzing the sound classes that were used to simulate the auditory scenes. Results show that a semantic characterization in terms of presence / absence of sound sources is an effective way to characterize urban sound environments.

PACS numbers: 43.50.Lj, 43.50.Rq, 43.50.Qp, 43.66.Lj, 43.66.Ba

I. INTRODUCTION

The notion of soundscape has been introduced by Schafer[31,32] in the 1970s as the auditory equivalent to landscape. Following this paradigm, a sonic environment is described by focusing on the listener's evaluation, rather than only taking into account the acoustic parameters of the sound. Focusing on subjective criteria, Schafer proposed to decompose the soundscape between three main components called *keynote sounds*, *signals* and *soundmarks*. The notion of soundmark refers to sounds with subjective qualities that bring them into focus for a certain category of people, whereas keynote sounds and signals may roughly be considered as respectively background and foreground sounds. Schafer also claims that the improvement of the sonic environment requires a "positive" approach to the soundscape, involving the identification and reinforcement of pleasant, important or meaningful sounds in the environment. The soundscape thus appears as a powerful tool to develop perceptively motivated acoustical indicators[34].

The main goals of studies addressing soundscape perception may be summarized as follow:

1. *Describing the soundscape*: How does the human brain identify different types of soundscapes? What are the elements composing each type of soundscape? How do they differ from one soundscape type to another?

2. *Evaluating the soundscape*: How do those elements influence the qualitative evaluations of a soundscape, such as "noise annoyance" or "pleasantness", to name but a few?

The second question finds its origin in the growing need to improve the quality of artificial sonic environments, such as the urban soundscape. It is only during the 1980s that policy-makers started taking into account the link between *noise* and pollution, considering noise as a significant degradation of the quality of life. To fight this pollution, the first approach consisted in identifying unwanted sounds, and lowering their intensities. Thus, in the wake of this realization, several regulations have emerged, often limited to enforcing sound level thresholds. However, several studies showed that urban "noise" perception and evaluation is a complex phenomenon that cannot be described solely with the help of objective acoustical measures such as A-weighted levels and L_{Aeq}[10–12,28,33,35]. More precisely, Yang and Kang[38] as well as Kang and Zhang[17] found significant differences between objective acoustic comfort measurements and subjective evaluations of sonic environments, thus confirming that "noise" is a cognitive object which depends on a listener's appreciation and the context in which the noise is heard. Many urban "noises" such as that of a *siren*, can annoy as well as warn of a danger. Many town districts are appreciated because of their lively and animated atmosphere which often results in higher sound levels.

To summarize, we believe that even if acoustical regulations are to some extent effective, there is a need to describe urban soundscapes not only through objective acoustical measures. Qualitative attributes also need to be considered[7,36] in order to understand better the relationship between humans and their acoustic environment[31].

II. BACKGROUND

The strength of the soundscape approach is, to some extent, also its weakness, as it implies to fully appreciate the interaction of many factors. Perception of soundscape is thus an interdisciplinary field of investigation, each discipline coming with its well identified experimental protocol. In each one, integration of the results can be difficult[9,34]. Recently, ambitious projects have been undertaken as the European Cooperation in Science and Technology Action TD0804-Soundscape of European Cities and Landscapes to standardize soundscape assessment and indicators.

We believe that studies addressing soundscape perception may be roughly divided into two approaches according to their methodologies[9].

The first approach is the dimensional approach, which tends to derive relevant emotional dimensions using semantic differential analysis[7,16,17], and to link those dimensions with acoustical descriptors[36]. The idea is to investigate the soundscape using both acoustical measurements and semantic data. In order to establish the semantic axis[9] prior to the dimensional analysis, those semantic data usually come from qualitative surveys. Using a Principal Component Analysis (PCA) on descriptors of urban soundscapes, Cain and al.[7,9] found two independent emotional dimensions named "Calmness" and "Vibrancy". Using a hierarchical cluster analysis on both semantic differential attributes and acoustical descriptors, Torija and al.[36] derive 15 soundscape typologies and show that crest factor and sound level at 125Hz are relevant acoustical variables to recognize types of soundscape. Although the dimensional approach provides important cues to describe soundscape perception, it remains a purely holistic approach as it bypasses the way in which the human brain decomposes soundscapes into perceptual auditory objects[3,37] and how each of those objects may influence the qualitative judgment.

The second approach is the categorical approach, which investigates mental representations of soundscape. Mental representations can be seen as mirrors of the perceived external realities (see Dubois et al.[11] p.869 for a definition). They are the memory of the knowledge acquired by a subject, and act as the base of *top-down* cognitive systems[21]. As explained by Dubois et al.[11] "semantic categories can be seen as mediating individual sensory experiences to collective representations by means of a shared language". As they cannot be observed, experimenters must use objectivation methods to reach them. Traditional methods used to objectify human mental representations attempt to derive mental categories and perceptual components, either by relying on verbal descriptions coming from questionnaires or interviews[2,13,28] (categories of sound sources) or by sorting tasks[14,15,20] (categories of sound sources and soundscapes). To identify the mental categories, these methods require psycho-linguistic and lexical analyses. Results from those studies tend to show that the soundscape appreciation depends upon the identification and the assessment of the sound sources which compose the soundscape[11,13,14,20,28]. The categorical approach thus agrees with the Auditory Scene Analysis (ASA) theory[3], which stands that auditory stimuli are decomposed into "auditory objects" called "streams", which may be regarded as sequences of auditory events emitted by putative sound sources[8,37]. This view is in line with recent cognitive neuroscience studies suggesting that the primary auditory cortex (A1) produces representations of such auditory objects that act as bases for high level cognitive processing[23].

As underlined by Davies and al.[9], while the categorical approach provides useful information to understand how a soundscape is composed, it does not provide information on how the relations (acoustic or semantic) between the soundscape components (sequences of audio events emitted by sound sources) are perceived and influenced the qualitative evaluation. Furthermore, as category "names" may occur at different semantic levels, categorical approaches based on linguistic analysis are not amenable to facilitating comparison between studies[4,25]. To go further in the categorical approach, there is a need for methods providing 1) a finer characterization of sound source categories with semantic and numeric data, 2) a description of the inter-relations existing between these sound sources, and 3) subjective data that may be used as basis for inter-studies comparisons.

III. METHOD OVERVIEW

With those considerations in mind, we propose a new subject-centered experimental protocol to characterize sound environments with qualitative data related to quantitative data. In this protocol, subjects are asked to simulate complex sound environments with pre-imposed associated qualitative appreciations. To do so, they have access to a soundscape simulator that uses classes of sounds as base elements. The soundscape simulator is thus bound to a sound data set of urban environmental sounds, organized hierarchically around sound categories found in previous studies[4,11,13,14,20,25]. The proposed protocol was previously tested and validated thanks to a pilot study[19] with 10 subjects.

Previous works done by Bruce et al. also use a soundscape simulator to study soundscape perception[5,6]. But in their approach, subjects may only add or remove a short set of long recordings of foreground sounds, and adjust their levels. In our approach, subjects are presented with a full data set of classes of urban environmental sounds, that they may manipulate in terms of sound intensity and time positioning.

The proposed protocol is used to characterize two antagonistic soundscapes: subjects are asked to simulate two full urban sonic environments, one "ideal" and the other "non-ideal". The names "ideal" and "non-ideal" have been chosen in order to compare the results with those of a previous study addressing ideal urban sound environment representations[13]. In this paper, the focus is put on the acoustical analysis of the simulated scenes to figure out what the characteristics of an ideal (resp. non-ideal) urban environment are, in terms of sound source composition, sound event densities and sound levels. To refine the analysis, sound sources categories and their related quantitative data are investigated according to different semantic levels of categorization (*urban transport*>*car*>*car-passing*)[4,25].

The simulated scenes as well as the annotations are available on the archive Data Repository: `https://archive.org/details/soundSimulatedUrbanScene` The experiment web-page is available via the link `http://soundthings.org/simScene/`. It should be run on the Chrome browser or the Mozilla Firefox browser.

IV. EXPERIMENTAL PARADIGM

The proposed experimental protocol is largely inspired by hypotheses coming from cognitive psychology[10,11] and categorization theory[29]. A sound is considered both for its physical properties and for its semantic value which both depend on the subject and the context (urban environment). The purpose of our protocol is to objectify human mental representations of the sonic world without the need for a linguistic analysis, by characterizing them with combined semantic and numerical data.

To that end, subjects are asked to recreate a complex sound environment, making use of an environmental sound data set that they may explore without any written textual help thanks to a selection interface designed for this study. The selection process has been designed to rely only on the listening of sounds themselves, in order not to influence subjects with *a priori* associated semantic values. Once a sound element is selected, subjects must name it. They may then modify some of its physical parameters, thanks to a set of audio controllers. This framework is depicted on Figure 2. It exposes three types of data to objectify mental representations:

1. *Generic semantic data*: the "tags" of the sounds chosen by the subject. A tag is relative to the nomenclature of our typology, and thus pre-defined by experimenters. (ex: *male-voice*)

2. *Quantitative data*: the set of "audio parameters" attributed to the sounds by the subject (ex: *sound level* dB).

3. *Non-generic semantic data*: all the subjective verbal data including the *names* given by the subject to the sounds they selected (ex: *the cries of a man*), a general *title* of the simulated scene, and a *free comment* concerning the creation process.

The proposed paradigm can thus be seen as the inverse of that of a description task (questionnaire, interviews) that uses audio data as input (see Figure 1). In contrast with interviews and questionnaires, which both require subjects to describe a soundscape, *i.e.* to decompose the soundscape in elements, subjects are asked to recompose the soundscape from a sound data set of urban environmental sounds. This sound data set represents the "sonic world": a semantic discretization of it in term of "sound sources". It ideally provides the subject with all the sound diversity he may desire.

The proposed approach questions the following: "What sounds have been *used*?", "how have they been *used*?" and "how have they been *named*?", whereas descriptions tasks respond to: "What sounds have been *named*?" and "how have they been *described*?".

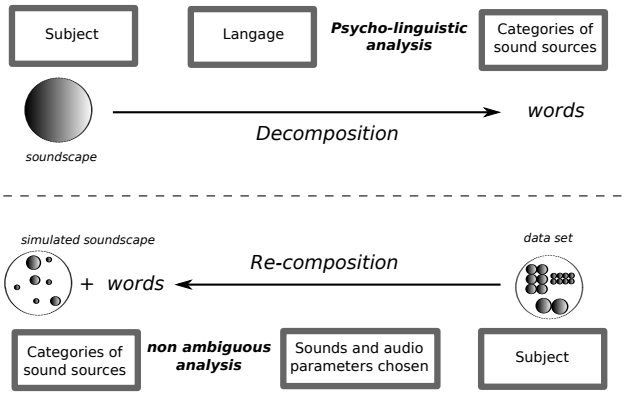

FIG. 1. The psycho-linguistic analysis paradigm (top) and the proposed approach (bottom)

V. SIMULATING THE SOUNDSCAPE

A. Soundscape model

In order to allow the creation of a complex sound environment, a model of soundscape has been designed whose key aspects are now presented. A soundscape is assumed to be a sum of tracks. Each track is modeled by a temporal sequence of sound samples that belong to the same sound class. To this sense, each track is related to one specific sound class (leaf sound class more specifically, see V.B). To generate a soundscape, subjects may interact with the tracks, but not with specific sound samples. In other words, if a subject wishes to put a sequence of car sounds, he may choose the sound class *car passing* and manipulate its associated track, but he cannot interact with individual sound samples of *car passing*.

Concerning the sound classes, the commonly accepted distinction made between "sound events" and "sound textures" is taken into account. A soundscape is thus regarded as "a skeleton of events on a bed of texture"[24]. Several studies point out the fact that textures and events drive two distinct cognitive processes[20,22,30]. If there is to some extend an analogy between the notion of "sound event", and the Schaferian notion of "signal", the analogy is not obvious between the notions of "texture" and keynote. Schafer's definition of keynote is perceptively motivated whereas most definitions of texture adopt a morphological approach (see[30] for a definition of texture, which is the one adopted in this paper).

Following this distinction (events *vs* textures), two separate treatments are handled by the soundscape simulator:

- *For the sound event classes* the temporal sequence is made of randomly selected sound samples belonging to the considered sound class of events. the temporal sequence is scaled by the mean/average spacing time between the sound events, the beginning and the end of the sequence.

- *For the sound texture classes* the temporal sequence is made of randomly selected texture samples belonging to the considered texture class, which are sequenced without time space (crossfading is used to guarantee that the transition is seamless). Only the beginning and the end of the sam-

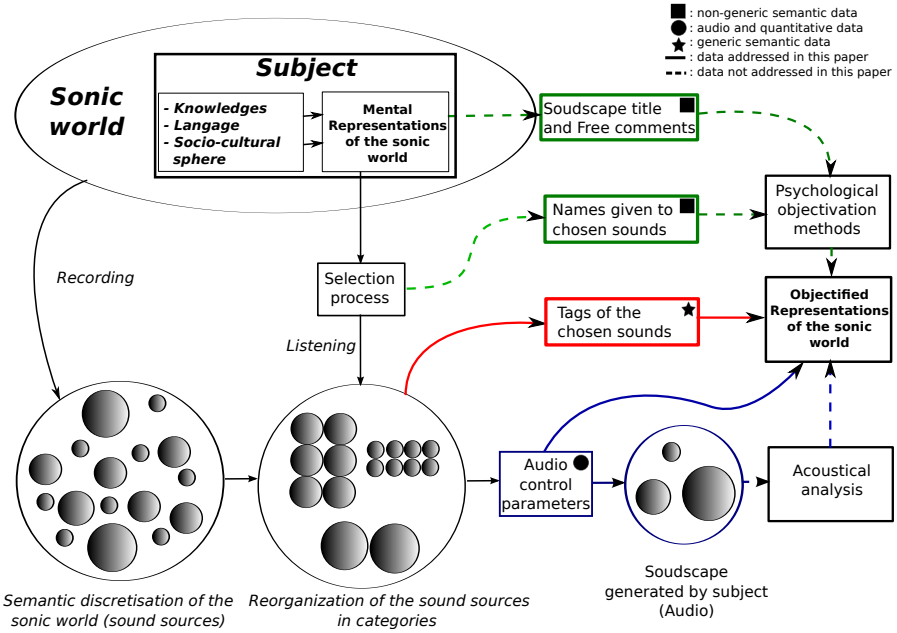

FIG. 2. Paradigm of the proposed experimental protocol. *square*: non-generic semantic data; *star*: generic semantic data; *circle*: audio and quantitative data; *Dashed-line*: data not addressed in this paper; *Line*: data addressed in this paper

ple sequence may be set by the subject. Furthermore, in order to avoid any unnatural effect due to the repetition of identical sounds[1], the soundscape simulator guaranties that a sample will never transition to itself.

Six audio parameters allow the subject to manipulate sound sequences of each selected class, in terms of sound levels and time positioning, see Table I. The parameters have an effect on all the samples of a sound sequence. To this sense, sound levels, and time spaces between events (only for events classes) are controlled in terms of average and variance between all the sound samples. Global parameters as global fade in/out and start/stop positions act on all the sequence. A last parameters called "sample fade in/out" allows subject to set a same fade effect separately on each sample of an events sequence (only for events classes).

B. Sound data set

Segmenting the sonic world to provide the subject with a reasonable and representative number of sound events and textures is a crucial step of our experiment. In order to create our sound data set, we rely on previous studies addressing urban sound source categories[4,11,13,14,20,25]. Based on the names of those categories, we build two hierarchical structures of event and texture sound classes, respectively. Top classes of those structures represent concepts as "urban transport" grouping sound classes such as "boat" or "cars", which are themselves grouped in sub-classes. The deeper the class level, the lower the variability between the exemplars lying under the class. In this paper, the class levels will be referred as **semantic levels**, the top classes being at the semantic level 0, the first subclasses at the semantic level 1 and so forth until the leaf classes are reached. Leaf classes are collections of recorded sound samples. Subjects may only interact with the leaf classes. The Figure 3 illustrates the hierarchical structure.

C. Simulation process

The simulation process involves two steps (see Figure 4), each of them relying on particular data resources and software interfaces. The two steps are:

1. *Sound Class selection*: where the subject has to select a leaf class of sound ("car passing", "heavy rain", ...). The selection is made without any written verbal help thanks to a particular selection interface.
2. *Sound Class modification*: where the subject may tune the parameters (time and intensity) of the temporal sequence formed by sounds belonging to the selected sound class.

VI. EXPERIMENT

A. Participants

44 post-graduate students of the École Centrale de Nantes (French engineering school), were asked to take part in the experiment. They were 30 males and 14 females and were about the same age (M: 21.6, STD: 2). All of them have been living in the same large French city (Nantes) for at least two years.

Audio Parameters	Description
Sound intensity (dB)	Average and variance
Spacing time (sec)	Average and variance
(only for event classes)	
Position in Scene (sec)	Start and Stop
Fade In/out (sec)	Global
Sample fade In/out (sec)	For each events
(only for event classes)	

TABLE I. Description of the audio parameters

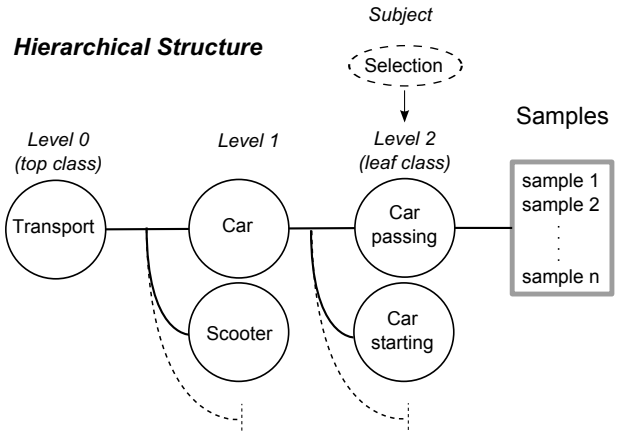

FIG. 3. Scheme of the hierarchical structure used both for the event and texture data sets. Depending on the considered top class, there could be more than 3 semantic levels.

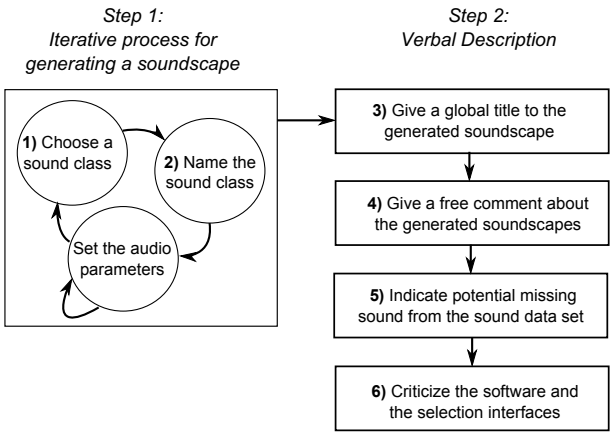

FIG. 4. The simulation process

B. Apparatus

483 urban environmental sounds were collected, of which 381 sound events and 102 textures. Among them, 260 events and 72 textures were recorded by the authors. The rest, which proved to be particularly difficult to record, came from existing sound banks. All recordings were performed using the shotgun microphone *Audio Technica AT8035* connected to a *ZOOM H4n* recorder. The use of a shotgun microphone allowed us to isolate as much as possible sound recordings from undesired events. All the sound were normalized to the same root mean square (RMS) level. The experiment was run simultaneously with the 44 subjects spread in three identical rooms with calm environment. Subjects were forbidden to talk to one another. Audio was presented diotically to each participant via *BeyerDynamic DT 990 Pro* semi open headphones. Three experimenters were always present (one in each room) to give instructions and answer queries if needed, including explanations, software installation and warm up session.

C. Task

Subjects are asked to successively create two urban soundscapes. The first must be ideal (*ie* the favorite urban soundscape of the subject in which they would like to live), the second non-ideal (*ie* the worst urban soundscape of the subject in which they would not like to live). They are asked to mimic a static listener. Subjects are absolutely not restricted in their design choices, but are forbidden to create physically impossible situations such as "a dog barking every 10 milliseconds". Each simulation involves six steps which are exposed in Figure 4. At the start of the experiment, a small tutorial of 20 minutes has been planed in order to familiarize the subject with the software environment. The experiment is scheduled to last about two hours.

D. Data collection and analysis

Four subjects were excluded from the analysis due to a misunderstanding of the instructions or a failure to comply with the time limits. 40 subjects completed the experiment successfully, giving us 40 ideal urban auditory scenes (i-scenes) and 40 non-ideal urban auditory scenes (ni-scenes).

The proposed experimental protocol generated a large range of data suggesting numerous avenues of investigations. We here choose to restrict ourselves to 1) the analysis of the holistic properties of the created scenes such as sound levels, event density and diversity, and 2) a class wise analysis, *i.e.* the analysis of the properties attached to each specific sound class, and a study of the distribution of the classes among the ideal and non ideal scenes. Other aspects are intentionally left for future research. Thus, among the collected data, we focus on 2 types of data: 1) the generic semantic data being the "tags" of the chosen sound classes, and 2) the quantitative data being the sound levels attached to the sound classes. For the holistic analysis, the data are averaged over the scenes, without looking at particular sound classes. For the class-wise analysis, the scene content is investigated with respect to the different sound classes used.

	Event classes	Texture classes
i-scenes	-6.8 (5.4)	-2.6 (3.9)
ni-scenes	-2.4 (3.2)	-1.6 (2.6)

TABLE II. Sound levels: mean sound levels in dB averaged over the subjects

Semantic	i-scenes		ni-scenes	
level	Coverage	Diversity	Coverage	Diversity
0	100 %	5.7 (1.4)	100 %	5.7 (1.4)
1	95 %	7.6 (2.7)	95 %	9.5 (2.5)
2	85 %	8.1 (3.2)	89 %	9.7 (2.7)
3	86 %	8.3 (3.4)	89 %	9.7 (2.7)

TABLE III. Diversity and coverage: the coverage is the percentage of the sound classes (events and Textures) of the data set that have been used by all the subjects. The diversity is the mean number of distinct sound classes chosen by each subject, averaged over the subjects

VII. HOLISTIC ANALYSIS

For the holistic analysis, all the statistical tests are performed using a Wilcoxon signed ranks test at the 5% significance level. Observations are obtained by averaging the data related to each scene. Results are presented in the tables using the convention *average(standard deviation)*.

A. Sound levels

Table II shows the results for the mean sound levels averaged over all subjects. For each scene, the mean sound level is obtained by averaging the sample levels. Samples are the elements which composed a temporal sequence of a particular sound class: if a subject selects the event sound class *car passing* and creates a sequence composed of 10 audio events, 10 samples are considered. This measure is thus not equivalent to the global sound level, but can reasonably be considered a good indicator of the overall sound level of a scene. Considering the event samples of the i- and ni-scenes, results show that the sound levels are significantly higher ($p = 2.08{*}10^{-5}$) for the ni-scenes, indicating that sound levels are indeed on average higher for unpleasant sound environments. But if there is a blatant difference between the event sound levels of the i- and ni-scenes, the deviation between the texture sound levels is not significant($p = 0.14$). This result suggests that textures have less direct influence on the sound level perception. This could be due to the fact that textures are sounds with low semantic weight[30], in other words, sounds which are less easy to identify. This is in line with early results of Kuwano et al.[18] showing that overall judgment of loudness of a sound environment sequence is not statistically different with average instantaneous judgment of the recalled sound events of the sequence.

	Density of the sound events
i-scenes	53 (65)
ni-scenes	63 (64)

TABLE IV. Density of the sound events: mean number of sound events of each scene averaged over the subjects

Semantic level	Event and texture	Event	Texture
0	81 %	76 %	70 %
1	90 %	91 %	78 %
2	92 %	89 %	80 %
3	93 %	91 %	—

TABLE V. Precision at rank 5 ($P@5$) computed from the *Jaccard* distances between the scenes for different semantic levels

B. Diversity and coverage

The notion of diversity is addressed by observing the number of distinct sound classes (events and textures) used in each i- and ni-scenes. This number depends on the semantic level considered. For example, let us consider two ni-scenes, one having a sound class *car-passing* of the semantic level 2, and the other a sound class *car-starting*, also of the semantic level 2. Both classes belong to the hierarchical structure *transport* (level 0) > *car* (level 1). We will count 2 distinct sound classes for the semantic levels 2 and 3 (as *car-starting* and *car-passing* have no subclass), and only 1 sound class for the semantic levels 0 and 1.

Averaged results are shown in Table III. Except for the semantic level 0, the diversity is significantly higher for the ni-scenes (level 1: $p = 2.5 * 10^{-4}$; level 2: $p = 0.006$; level 3: $p = 0.011$;). This reveals that ni-scenes are composed of a larger variety of sounds than i-scenes, suggesting that a non ideal urban environment contains more distinct sound sources than an ideal urban environment.

To qualify this measure of diversity, and verify that it is not biased by the selection interface, we look at the coverage. The coverage is defined by the percentage of the sound classes (events and textures) of the data set that have been used by all the subjects. For example, if all the classes of the semantic level 1 have been chosen by at least one subject, the coverage for the semantic level 1 will be of 100%. A low coverage would suggest that some parts of the data set have not been explored. Results are shown in Table III. The coverage remains superior to 85% regardless the semantic level, indicating that most of the sounds have been used for each type of scenes.

C. Event density

The event density is defined as the number of event sound samples used to generate a scene, regardless of the classes. We can see that the numbers of sound samples vary widely with the subjects. There is no statistical

difference between the densities of the i- and ni-scenes ($p = 0.14$). This suggests that the global sound events density does not influence the qualitative evaluation of a urban soundscape.

VIII. CLASS-WISE ANALYSIS

For the class-wise analysis, all the statistical analyses are done using a Wilcoxon rank-sum test (Mann-Whitney-Wilcoxon test) at the 5% significance level. Observations are obtained by averaging the data related to each class.

A. Semantic features: an effective way to evaluate soundscape quality

In order to describe the scenes in terms of sound sources, the "tags" of the sound classes (event and texture) chosen by the subjects are now investigated. Figures 5 and 6 display the "tags" of the semantic levels 1 or 2. For ease of reading, the tags *horn*, *fire alarm* and *car alarm* are grouped into a class named *alarm/horn* the classes *bus* and *train* are grouped into a class named *public transportation*, and the classes *truck*, *car*, *scooter*, *motorcycle* and *start passage-way* are grouped into a class named *traffic*. Classes with a selection percentage less than 2 % are not displayed. We only count the selected sound classes and not the number of samples of their respective sequences.

There is considerable difference between the "tags" of the i- and ni-scenes, confirming that sound semantics play an important role in soundscape evaluation. To address this, each simulated scene is represented by a boolean vector of n dimensions $S_i = (x_1, x_2, \ldots, x_n)$, $i \in [1, 80]$. Each dimension corresponds to a sound class (event and texture) of a particular semantic level (for examples $n = 44$ classes for the semantic level 1). Thus $x_1 = 1$ indicates that the sound class x_1 is present in the scenes S_i (not present if $x_1 = 0$ resp.). A *Jaccard* distance[26] is then computed between the vectors S_i. To quantify how semantic features characterize both the i- and ni-scenes, the precision at rank 5 ($P@5$) metric is used, *i.e.* the average number of items of the same class among the 5 closest items to a given seed item. For each scene S_i, $P@5$ is thus obtained by averaging the number of scenes having the same label as S_i (i or ni) among the five scenes S_j ($j = 1, ..., 5$) that are closest to S_i. Results are then averaged over the scenes S_i ($i = 1, \ldots, 80$).

Results are shown in Table V. For the semantic level 1, a $P@5$ of 90% is found, with a random threshold of 58%. The deeper the hierarchical level, the higher the $P@5$. We find 92.25% for semantic level 2 and 92.75% for semantic level 3. This is a strong indicator that semantic values of sounds are good descriptors to distinguish between the i- and ni-scenes. To refine the analysis, the same test is run considering separately the sound textures and the events to describe the scenes (for the semantic level 1, events: $S_i \in S_i = (x_1, x_2, \ldots, x_{31})$ and texture: $S_i \in S_i = (x_{32}, x_{33}, \ldots, x_{44})$). The $P@5$ achieved for the semantic level 1 of 91% for the events and 78.5 % for the textures demonstrate that sound events contribute more to the perception of soundscape quality than sound textures.

B. Hierarchical analysis of sound categories

The sound classes showed in the Figures 5 and 6 are close to those found by Guastavino[13] in a psycho-linguistic study also addressing ideal urban soundscape perception. Classes that suggest human presence and nature are the most present in i-scenes while classes referring to mechanical and construction work sounds are used for ni-scenes. This fact has been observed in other psycho-linguistic studies[11,14,28]. It confirms the *biophilia* hypothesis that "humans are attracted to nature" (Wilson quoted by Guastavino[13]). However some differences are to be noted. In her study, Guastavino[13] point out that "public transports" are typical sounds of an ideal urban environment (also mentioned by Dubois et al.[11]). She attributes this observation to the fact that urban environment perception is driven by the meaning attributed to the identified sources. As this meaning is influenced by social values, public transport sounds are better accepted than those of private vehicles. Our results tend to qualify this claim by showing that sounds of public transports (*bus* and *train*) were chosen for both the i-scenes (3.97% of the selected sound classes) and the ni-scenes (5.1% of the selected sound classes). Moreover, the sound densities (i-scenes: 1.1(2.1); ni-scenes: 1.4(2.2); $p = 0.2$) as well as the sample levels (i-scenes(dB): -1.5(3.1); ni-scenes(dB): -1.6(3.3); $p = 0.47$) do not differ significantly. Thus even if the idea of public transport sounds are well accepted due to societal considerations, the sound itself remains similar to that of any vehicles and appears in the ni-scenes more than *car* or *truck* sounds. Although semantic features influence the qualitative evaluation of public transport sounds (*bus* sounds correspond indeed to 3.64% of the event classes used in the i-scenes, as much as *bicycle* sounds and more than any other vehicle sounds), it seems that, for this sound category, it is not predominant over the role played by the physical attributes of the sound.

Figures 5 and 6 indicate a counter-intuitive result showing that the event class *traffic* (composed of the event subclasses *truck*, *car*, *scooter*, *motorcycle* and *start passage-way*) is well represented in the i-scenes (10.3% of the selected sound classes) as well as in the ni-scenes (21.6%), thus suggesting that traffic sounds are an integral part of an ideal urban environment. If we look at the sample levels, there is no statistical difference between the i- and ni-scenes (i-scenes (dB): -1.8(2.1); ni-scenes (dB): -1.7(3.6); $p = 0.17$). The difference occurs only if we look at the sample densities (i-scenes(dB): 2.7(4.5); ni-scenes (dB): 10.4(11.9); $p = 3.6 * 10^{-5}$) which is significantly superior for the ni-scenes. These results tend to indicate that traffic sounds are not absent from the representation of an ideal urban environment provided their densities are not excessive. If we refine the analysis by assessing separately each event subclass of *traffic*, we

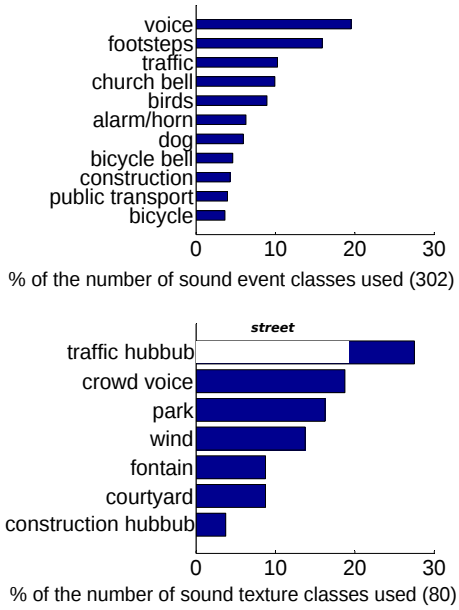

FIG. 5. Percentage of the number of sound classes of events (*top*) and textures (*bottom*) used by the subjects for the i-scenes at the semantic level 1 or 2 depending of the sounds

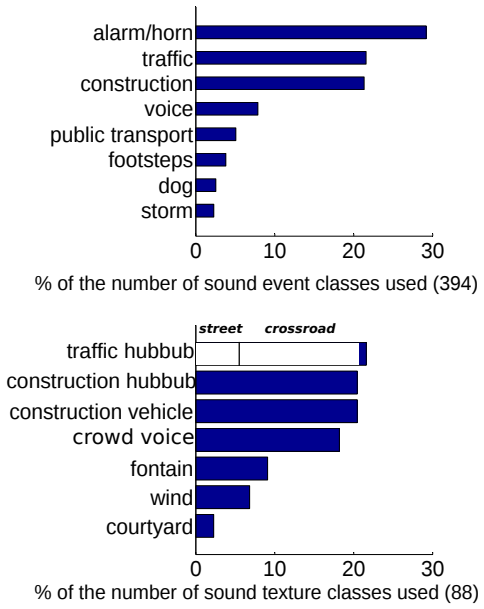

FIG. 6. Percentage of the number of sound classes of events (*top*) and textures (*bottom*) used by the subjects for the ni-scenes at the semantic level 1 or 2 depending of the sounds

find significant differences only for the subclasses *scooter* and *motorcycle*. The former has a significantly higher density ($p = 5.6 * 10^{-4}$) for the ni-scenes, and the latter has both significantly higher sample levels ($p = 0.03$) and density ($p = 0.02$) for the ni-scenes. Putting these

Semantic level	Markers	
	i-scenes	ni-scenes
0		construction work
1	church bell	klaxon
	bicycle bell	siren
	animal	vehicle work
	footsteps	
2	church bell	klaxon
	birds	siren
	bicycle bell	vehicle work
	female laugh	
	male laugh	
3	church bell	klaxon
	birds singing	siren
	bicycle bell	vehicle work
	female laugh	
	male footsteps concrete	

TABLE VI. Event classes found to be markers. In each cell, markers are ordered using descending order of V-test values

results with the urban problematic of "traffic noise", it appears that 1) it is mostly the sounds of two-wheeled vehicles that are the cause of the annoyance, and 2) that in this case, sound levels are not a relevant indicator of quality compared to sound density.

The texture class *traffic hubbub* is well present in the i-scenes. There is no statistical difference between the i- and ni-scenes for both the sample levels ($p = 0.11$) and the densities ($p = 0.77$). Considering now the subclasses of *traffic hubbub* (see Figures 5 and 5), we see that the *Traffic hubbub of streets* class has been mostly used for the i-scenes (18.75 % of the textures used in i-scenes) whereas the *Traffic hubbub of crossroad* class has been only used for the ni-scenes. If we compare those two subclasses, we find again no statistical differences for both the sample levels ($p = 0.19$) and the densities ($p = 0.37$). These observations show the importance of context and expectation in soundscape evaluation. The fact that *traffic hubbub* is not depreciated for ideal urban environments shows that "traffic sounds" are understood as being an inherent element of urban environment. This is in contradiction with an intuitive idea that traffic background are indeed "noises". Similar findings are presented by Guastavino[12]: asking subjects to describe urban background, she found that appreciation of the *traffic background* depends on the subject, as it can be reassuring or even appreciated, providing it is not too loud.

C. Markers

This section investigates the existence of potential event sound markers of an ideal urban environment (resp. non-ideal urban environment), *ie* an event class which has been mostly used in one type of soundscape. To

identify markers, the V-test statistical value is used. For each semantic level, considering the population as being the total number n of event classes used for both i-scenes and ni-scenes, n_k the number of event classes used for one type of scene k (ni-scenes or i-scenes), n_j the number of event classes j used for both i-scenes and ni-scenes, and n_{kj} the number of event classes j used for one type of scene k (ni-scene or i-scene), the V-test of the event class j (modality) in the scene type k (group) can be computed as follow:

$$\text{V-test}_{jk} = \frac{n_{jk} - n_k \frac{n_j}{n}}{\sqrt{n_k \frac{n-n_k}{n-1} \frac{n_j}{n}(1 - \frac{n_j}{n})}}$$

The V-test tests the null hypothesis that the n_k individuals of the group k are randomly drawn from the population of n individuals. Usually the V-test value is assessed at the 5% significance level, that is, if the null hypothesis is true, the V-test value has a 95% chance to fall within the confidence interval $[-1.96, 1.96]$. Thus the V-test is considered as statistically significant if its absolute value remains superior to 2. It is safer to correct the usual 5% significance level as we are testing many modalities (49 event classes for the third semantic level). To do so, we use the Bonferroni adjustment[27]. The Bonferroni adjustment is approximated by dividing the 5% significance level by the number of modalities j (event classes) to be tested. For example, if we test the 49 event classes of the semantic level 3, the corrected significance level will be 0.001%, and the V-test value will be statistically significant if its absolute value is superior to 3.29.

Results are shown on the table VI. Eleven markers are found across all the semantic levels. Confirming again the *Biophilia* hypothesis, sounds of human (*footsteps*, *female laugh*, *male laugh* and *male footsteps concrete*) and nature (*animal*, *birds* and *birds singing*) are markers of the i-scenes. The presence of *Church Bell* as a strong marker of the i-scenes can be due to the socio-cultural background of the subjects, in great majority French citizens. It confirms Schafer's claim that sounds that are recognized by a community as integral part of its sonic environment are popular[31].

For the ni-scenes, the markers (*Klaxon*, *Siren*, *vehicle work* and *construction work*) are rather intuitive. Interestingly, none of the event classes related to traffic sounds is a marker of the i- or ni-scenes. It confirms the trend observed in VIII.A that traffic sounds are understood to be intrinsically part of a urban environment. In other words, although the event densities of traffic sound classes differ between the i- and ni-scenes, traffic sound classes presence is still not a characteristic of a non-ideal urban sound environment.

IX. DISCUSSION

A. About protocol validity

One way to validate the protocol is to look at the subject comments. Looking at the database criticisms, 28 subjects stated that they couldn't find one or several desired sounds, with a maximum of 4 sounds for one subject. From all the missing sounds we identified 26 sound classes at different semantic levels. Among those classes, 16 were effectively present in the database, 1 referred to musical sounds which we had chosen to exclude from the database and only 9 were effectively absent. Among the 16 sound classes, they have all been used by at least one subject, except for one sound class (*stroller/trolley*). Among the 9 missing ones, we find very specific sounds as *sport car* or *teenager voice*. We believe that those results show that the database diversity was sufficient for the purpose of the study. If we look at the interface criticisms, 32.5 % of all the subject clearly stood that the selection interface was a "straightforward yet effective way" to find a sound whereas only 10 % indicated that they encountered difficulties. The remaining 57.5 % did not report difficulties. Those results tend to confirm that both the data set and the selection interface are well suited for the experiment.

B. Outcomes

We believe that four main benefits may derive from using the proposed simulation approach: 1) Other studies based on description tasks use full soundscapes recordings as input. Those recordings are specific exemplars of a very many potential soundscapes which could occur at the same location, at the same time. The re-composition process prevents experimenters from this bias as the simulated soundscape is directly related to the subject representation. Although the simulation is limited by the diversity of the sound data set, we believe that this bias is more controllable, as the expressiveness of the subjects' responses will be less constrained by the size of the data set than by the exposure to imposed exemplars of a soundscape. 2) The presence of a structured sound data set allows us to focus directly on the sounds chosen by subjects instead of verbal data only. Thus it may reduce the potential bias of a psycho-linguistic analysis, the first being the mastering of the subject's language by the subject, the second being the lack of definite terms to fully describe sounds or complex sound environments[13], and the third being the high inter-subject variability concerning the description of a same sound. 3) The simulation protocol provides sound signals of simulated urban environments, annotated in term of sound sources, sound intensity, time positioning and qualitative appreciations. It allows us to appreciate new soundscapes descriptors as sample density, which otherwise would be tedious to obtain from recordings.

To some extend, those simulated scenes may be regarded as cognitive summaries of mental representations and may be a useful data set for all Computational Auditory Scene Analysis (CASA) research or cognitive neuroscience studies investigating 'auditory object' decomposition[23] or regularities representation impact on high level perceptive processes[37].

X. CONCLUSIONS AND PERSPECTIVES

In this paper we introduced a new experimental protocol to study soundscape perception based on a simulation paradigm. Application of this protocol allows us to gain knowledge about the perception of urban soundscape by asking human subjects to simulate two urban sound environments: one ideal and the other non-ideal. The results of the presented study show that:

- The simulation paradigm provides the same high level sound categories as those observed by questionnaire based studies[13], and allows us to refine the hierarchical analysis of those sound categories.
- A semantic characterization of the soundscape in term of presence / absence of sound sources is an effective way to characterize ideal or non-ideal urban environment.
- Global sound level is indeed a good indicator of pleasantness, but not for specific sound categories.
- Structural features such as sound event density are found to be relevant descriptors to characterize traffic sounds.

These results indicate that associated semantic values of events is an effective information to categorize the simulated scenes between ideal and non ideal scenes. More precisely, the study shows that some event categories may be considered as markers of a specific type of soundscape. We believe that those results could be useful to guide the design of computational categorization paradigms based on automatic sound event detection.

We believe that this protocol allows us to directly objectify important aspects of the mental representations of the subjects by being able to look at 1) the sound classes chosen by the subjects to simulate the scenes and 2) their associated quantitative data. Lastly, the use of a pre-fixed structured data set may facilitate the inter-subjects or inter-studies comparison as all subjects have access to the same decontextualised material to recreate their own vision (contextualisation) of a particular sound environment.

We believe that performing this experiment with subjects of different socio-cultural backgrounds would be interesting for future work as it would allow us to study the cultural impact on the hearing cognitive processes involved.

XI. ACKNOWLEDGEMENTS

Research project partly funded by ANR-11-JS03-005-01. Thanks to the 44 students of the Ecole Centrale de Nantes for their willing participation.

[1] Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (**2010**). "Rapid formation of robust auditory memories: Insights from noise", Neuron **66**, 610–618.

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

[3] Bregman, A. S. (**1994**). *Auditory scene analysis: The perceptual organization of sound* (MIT press).

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

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

[6] Bruce, N. S., Davies, W. J., and Adams, M. D. (**2009**). "Development of a soundscape simulator tool", proceedings of Internoise 2009 .

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

[8] Ciocca, V. (**2007**). "The auditory organization of complex sounds.", Frontiers in bioscience: a journal and virtual library **13**, 148–169.

[9] Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., Hall, D. A., Hume, K. I., Irwin, A., Jennings, P., Marselle, M., Plack, C. J., and Poxon, J. (**2013**). "Perception of soundscapes: An interdisciplinary approach", Applied Acoustics **74**, 224–231.

[10] Dubois, D. (**2000**). "Categories as acts of meaning: The case of categories in olfaction and audition", Cognitive Science Quarterly **1**, 35–68.

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

[12] Guastavino, C. (**2003**). "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. dissertation, Université Paris **6**.

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

[14] Guastavino, C. (**2007**). "Categorization of environmental sounds", Canadian Journal of Experimental Psychology .

[15] Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (**2012**). "A lexical analysis of environmental sound categories.", Journal of Experimental Psychology: Applied **18**, 52–80.

[16] Jeon, J. Y., Hong, J. Y., and Lee, P. J. (**2013**). "Soundwalk approach to identify urban soundscapes individually", The Journal of the Acoustical Society of America **134**, 803–812.

[17] Kang, J. and Zhang, M. (**2010**). "Semantic differential analysis of the soundscape in urban open public spaces", Building and Environment **45**, 150–157.

[18] Kuwano, S., Namba, S., Kato, T., and Hellbrück, J. (**2003**). "Memory of the loudness of sounds in relation to overall impression", Acoustics Science and Technics **4**.

[19] Lafay, G., Rossignol, M., Misdariis, N., Lagrange, M., and Petiot, J.-F. (**2014**). "A new experimental approach for urban soundscape characterization based on sound manipulation : A pilot study", in *Proceedings of the 26th International Symposium on Musical Acoustics*.

[20] Maffiolo, V. (**1999**). "De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain", Semantic and acoustic characterization of urban environmental sound quality") Ph. D. dissertation, Université du Maine, France .

[21] McAdams, S. and Bigand, E. (**1994**). *Penser Les Sons :*

Psychologie cognitive de l'audition, Psychologie et sciences de la pensée, Presses Universitaire de France.

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

[23] Nelken, I. and Bar-Yosef, O. (**2008**). "Neurons and objects: the case of auditory cortex", Frontiers in neuroscience **2**, 107.

[24] Nelken, I. and de Cheveigné, A. (**2013**). "An ear for statistics", Nature neuroscience **16**, 381–382.

[25] Niessen, M., Cance, C., and Dubois, D. (**2010**). "Categories for soundscape: toward a hybrid classification", in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2010, 5816–5829.

[26] Pang-Ning, T., Steinbach, M., Kumar, V., *et al.* (**2006**). "Introduction to data mining", in *Library of Congress*.

[27] Perneger, T. V. (**1998**). "What's wrong with bonferroni adjustments", Bmj **316**, 1236–1238.

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

[29] Roach, E. and Lloyd, B. B. (**1978**). "Cognition and categorization", Hillsdale, New Jersey .

[30] Saint-Arnaud, N. (**1995**). "Classification of sound textures", Master thesis, Massachusetts Institute of Technology.

[31] Schafer, R. (**1977**). *The Tuning of the World*, Borzoi book (Knopf).

[32] Schafer, R. M. (**1969**). *The new soudscape* (Universale Edition, Vienna).

[33] Schulte-Fortkamp, B. (**2013**). "Soundscape-focusing on resources", in *Proceedings of Meetings on Acoustics*, volume 19, 040117 (Acoustical Society of America).

[34] Schulte-Fortkamp, B., Brooks, B. M., and Bray, W. R. (**2007**). "Soundscape: An approach to rely on human perception and expertise in the post-modern community noise era", Acoustics Today **3**, 7–15.

[35] Schulte-Fortkamp, B. and Fiebig, A. (**2006**). "Soundscape analysis in a residential area: An evaluation of noise and people's mind", Acta acustica united with acustica **92**, 875–880.

[36] Torija, A. J., Ruiz, D. P., and Ramos-Ridao, A. (**2013**). "Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes", The Journal of the Acoustical Society of America **134**, 791–802.

[37] Winkler, I., Denham, S. L., and Nelken, I. (**2009**). "Modeling the auditory scene: predictive regularity representations and perceptual objects", Trends in cognitive sciences **13**, 532–540.

[38] Yang, W. and Kang, J. (**2005**). "Acoustic comfort evaluation in urban open public spaces", Applied Acoustics **66**, 211 – 229, urban Acoustics Urban Acoustics.