1. INTRODUCTION

The urban sound environment is studied both from the perceptual[1,2] and from the physical point of view[3].[4] These studies can be based on real sound environments either by acoustic listening through a *soundwalk*[5] or with recordings made in city,[6] or in a laboratory by using sound mixtures resulting from a simulation process.[7] In the first case, since it has an undeniable ecological validity, it doesn't make it possible to have a repeatable experience nor to propose a controlled framework where the presence and the level of the different sound sources can be adjusted. Then the use of a simulation tool to create specific urban sound environments is a useful way to control the degree of contribution of each class of acoustical sources. It allows us to create urban sound environments by adjusting some parameters[8] or to evaluate performances of some classification or detection tools.[9]

A critical question is then to obtain scenes that are realistic enough to be considered simulated scenes as similar to some reference recordings. To do so, various methods have been proposed. The first approach would be to simulate completely a neighborhood taking into account the architecture of the buildings, the sound dynamics of the different sources present (car, voice, bird, bell ...) and their propagation to a receptor, as proposed by.[10] If the first renderings are interesting, this method remains complex to implement. A more straightforward approach composes sound mixtures from isolated sounds mixed together. The idea is to consider the urban sound environment as the sum of acoustic events, i.e. events with a sufficient sound level to be discernible, superposed to a sound background, i.e a sound constant on the sound mixture whose properties vary slowly during the scene. The difficulty here is to have a representative database of isolated sounds with a sufficient quality to not compromise the resulting sound quality and to sequence the acoustic events wisely.

The method proposed in[11] enables to resolve the first issue by extracting the acoustical events directly in real recordings and by manipulating them to re-use them in new mixtures. Since the tool has a sufficient number of parameters to control the sound event, their method is limited by the extraction phase: the overlapping between the acoustic events deteriorates the sounds and then the final rendering. Furthermore, the manipulations (duration, frequency range) can create artifacts that can deacrease the realism. In a similar way, the tool proposed by Davis and Bruce[8] enables to compose sound mixtures with an isolated sound database composed of recordings made specially for their application. The creation of the urban sound mixtures is made possible with the perceptual evaluation of a panel of listeners of the different sound environments. This method enables to determine which sounds are the most representative or the most linked to an urban sound environment from a perceptual point of view.

However, as our objective here is to get simulated scenes that are as realistic as possible, this method can not be considered because if this method proposes a perceptual validity, it can not provide an ecological one. Thus, this study proposes to create urban sound mixtures from the listening of urban audio recordings and the estimation of some high level sequencing parameters extracted from the recordings to tune an audio simulation tool. The realism of the simulated scenes is then evaluated using a perceptual test.

The remaining of the paper is organized as follows: Section 2 focuses on the study of recordings made in Paris, Section 3 deals with the simulation tool and Section 4 summarizes the results obtained in the perspective test.

2. STUDY OF REAL SCENES

Real urban noise recordings are listened in details to determine the composition of typical urban sound environments in terms of sound sources, presence and level. The recordings considered in this study have been made on the 13th district of Paris (France) as part of the GRAFIC project[12] during a soundwalk which was designed to cover different types of urban sound environments.

Figure 1: Map of the soundwalk with the 19 stop points

The walk was 2.1 km long, consisting of 19 stop points with a recording duration between 3 to 5 minutes (see figure 1). It was roamed on two days (03/23/2015 and 03/30/2015), twice a day (on the morning and on the afternoon) and in one direction (from West to East, WE) and then in the other (from East to West, EW). The acquisition system used was a ASAsense and was carried on the backpack of the operator which was walking with the participants. In the end, 76 (4×19) audio files (sampled at 44,1 kHz) were available and serve as a basis for the study. More details can be found on.[13] Each file is fully annotated in terms of sound classes that are present in the urban area and their recurrence. It is labeled using the following types of sound environments (*park*, *quiet street*, *noisy street*, *very noisy street*) as proposed by[14] or[3] (Table 1).

day	**journey**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	**EW**																			
1	**WE**																			
2	**EW**																			
2	**WE**																			

park | quiet street | noisy street | very noisy street | unclassified

Table 1: Classification of the scenes according to the sound environment

Most of the scenes of the recordings belong to *quiet street* and *noisy street*. 2 scenes are removed form the database: a noisy cleaning truck is present inside the *1-WE-3* heavily polluting the sound scene and the *1-WE-19* record is too short to be considered in this study. From the annotations, each sound environment is characterized by the sound classes present in the recordings. The most recurrent sound classes are presented in Table 2.

As the calibration audio file was not available, the sound level cannot be considered as absolute but the relative difference between each sound environment can be considered. The road traffic, the voice and the bird components are the most discriminant classes of the sound environment: with exclusively road traffic, the *noisy* and *very noisy* environments sound environments are different to the *park* and *quiet street* where the voice and the bird's whistles are predominant. Furthermore, multiple sound classes can be heard whatever the urban sound environments : dog barks, church bell ringing, car horn, sirens ... with a density that can be very low (¡ 0.1/min). Finally, a lot of brief sounds with unknown origin can be listened in almost all the files. Consequently, all these indefinite sounds are annotated in one unique class sound, *street noise*. From this information (density, level, sound classes present in each sound environment), it is now possible

Sound environment	Sound level (dB)	Background	Event	number events/min
Park	69.0	voice, bird's whistles	road traffic	0.5
			voices	0.5
			bird's whistle	0.5
			street noise	0.5
			foot step	0.3
Quiet street	70.2	road traffic bird	road traffic	1.0
			voices	0.7
			street noises	0.7
			foot step	0.5
			bird	0.2
			construction site noise	0.1
			door house	0.2
			door car	0.2
Noisy street	73.5	road traffic	traffic	9.0
			foot step	0.5
			voice	0.6
			street noise	0.4
			bell	0.1
			bird	0.2
			car horn	0.3
			car's door	0.2
			siren	0.1
Very noisy street	76.0	road traffic	traffic	40
			voice	0.3
			siren	0.2
			car horn	0.3
			bird	0.2
			foot step	0.3
			car's door	0.2
			street noise	0.3

Table 2: Sound level and description of the most recurrent sound classes (backgrounds and events) present in the sound environments

to compose urban sound mixtures.

3. SIMULATION OF REALISTIC AUDITORY URBAN SCENES

A. PRESENTATION OF THE WEB-SIMULATOR *SIMSCENE*

simScene[15] is a simulator that creates sound mixtures in a .wav format by superposing audio samples that come from an isolated sound database composed of two categories: the brief sounds (from 1 to 20 seconds) that are considered as salient belong to the *event* category whereas all the sound that are of long duration and whose acoustic properties do not vary with respect to time belong to the *background* category. Inside each category, the sound samples are then grouped in sound classes (bird, car, foot steps ...). The software enables the user to control for each sound class some parameters as the number of events of each class that appears in the mixture, the elapsed time between each sample of a same class or the existence of a fade in and a fade out .... Each parameter is completed with a standard deviation that may brings some random behavior between the scenes.

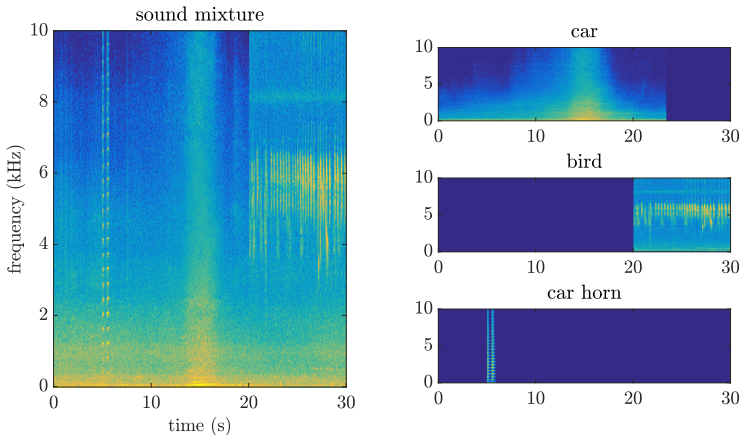

Figure 2: Example of a scene composed of three sound classes

The sound mixtures can be created following two modes: the *abstract* mode allows to create mixtures from the specified parameters while the *replicate* mode reproduces an existing scene by following as much as possible its annotation file. With the audio file of the global sound mixture, each sound class can be resumed in a separate audio file, which allows to know their exact contribution in the scene.

B. CREATION OF THE URBAN SOUND DATABASE

Based on the annotations of the 74 real scenes, a sound database is built using the *replicate* mode of *simScene* with sounds found online on the *freesound* project and with the help of the urbanSound8k database.[16] This database consists in more than 8000 files with a 4 seconds or less duration, found on the web site freesound.org too, and classified in 10 sound classes : ventilation, car horn, children playing, dog barking, bell, engine idling, gunshot, jackhammer, siren and street music. All the audio files have been sorted and those who presented the best signal to noise ratio have been selected.

In addition, as road traffic is a prime audio source in an urban sound environment, it seemed interesting to have, in the database, a part composed of well-recorded car sounds. As a result, passages of 4 different cars (Renault Scenic, Clio, Megane and Dacia Sandero) were recorded on the Ifsttar-Nantes runway at different speeds and gear ratios for steady, acceleration, braking (Table 3) and stopped phases. Among the 108 recordings planned, those for the acceleration and braking phase for the Renault Senic could not be done

whereas some recordings have been made twice for the other vehicles. In all, 103 car passages have been recorded.

	Trans.	1	2	3	4	5
stabilized speed (km/h)	20	×				
	30		×	×		
	40		×	×	×	
	50			×	×	
	60				×	×
	70				×	×
	80					×
	90					×

braking		acceleration	
speed (km/h)	gear r.	speed (km/h)	gear r.
$50 \rightarrow 0$	$3 \rightarrow 2$	$0 \rightarrow 30$	$1 \rightarrow 2$
$40 \rightarrow 0$	$2 \rightarrow 2$	$0 \rightarrow 40$	$1 \rightarrow 2$
$50 \rightarrow 30$	$3 \rightarrow 2$	$20 \rightarrow 40$	$1 \rightarrow 3$
$60 \rightarrow 40$	$4 \rightarrow 3$	$30 \rightarrow 50$	$2 \rightarrow 3$
$70 \rightarrow 50$	$4 \rightarrow 3$	$40 \rightarrow 60$	$3 \rightarrow 4$
$80 \rightarrow 50$	4 or 5 $\rightarrow 3$	$50 \rightarrow 70$	$3 \rightarrow$ 4 or 5

Table 3: Description of the recordings set on runway with vehicle passages with stabilized speed (left) and in acceleration and braking phase(right) with different gear ratios

To obtain the cleanest possible sound samples, a median filter[17] has to be applied to filter out bird's whistles present in the recordings. The filter consists in taking a window around a center point and to attribute to it the median value of this window. This method allows to attenuate the brief variations both in the spectral and the temporal plan. The size of the filter window is chosen with a 5 points width and a 9 points height in the frequency range $[2500 - 6500]$ Hz, so that the effect of the filter is sufficient without deteriorating the quality of the audio (see Figure 3).

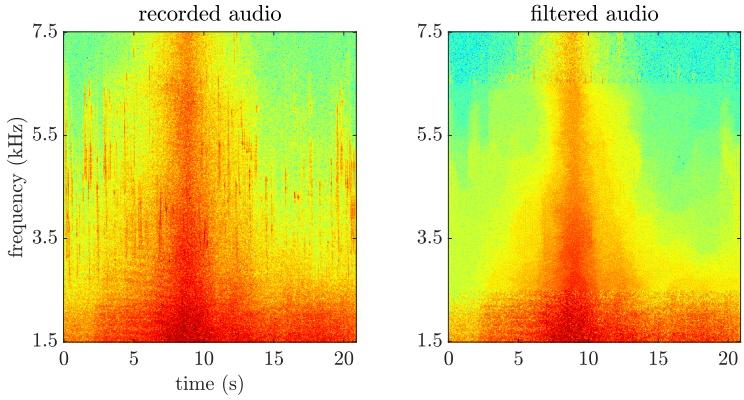

Figure 3: Spectrogram in the frequency range $[2500 - 6500]$ ***Hz of a recorded passage of the Renault Megane car at 40 km/h with the 3rd gear ratio (*** $N_w = 2^{12}$ ***with 50*** $\%$ ***overlapping,*** $N_{fft} = 2^{12}$***,*** $sr = 44.1$ ***kHz). On the left, the original recording, on the right, the filtered audio with the filtered part.***

Finally, the resulting database is composed of 245 sound events (bell, whistle bird, sweeping broom, car horn, car, hammer and drill, coughing, dog barking, car and house door slamming, plane, siren, foot step, thunder, street noise, suitcase rolling, train and tramway passing, truck and voice) and of 154 background sounds (birds, construction site, crowd, park, rain, schoolyard, traffic, ventilation, wind). With this built-up database, the 74 real sound scenes are replicated. In order to obtain a corpus of urban sound scenes, resulting from a simulated process, sufficiently realistic, a perceptual test is set up to evaluate the realism on several

of them.

4. PERCEPTUAL TEST

A. DESIGN OF THE TEST

A perceptual test is conducted with a panel of listeners that are asked to evaluate the level of realism on a 7-point scale (1 is *not realistic at all*, 7 is *very realistic*) of an auditory scenes, recorded and simulated. The total number of sound scenes to be tested is set at 40. It consists of 20 30-seconds audio files, including 5 scenes that belong to the sound environment *Park*, 6 from *Quiet street*, 4 from *Noisy street* and 5 from *Very noisy street* chosen randomly among the 74 recorded scenes, and the same 30 seconds from the replicated scenes. In order to limit the duration of the test and to preserve the concentration of the subjects, each subject listens a subset of 20 sound scenes (10 real sounds scenes and 10 replicated). Furthermore, to prevent the listeners from changing the sound level of their speakers, all the scenes are normalized to the same sound level, chosen at 65 dB.

The experimental design is elaborated following a Balanced Incomplete Block Design (BIBD)[18] where J participants make K evaluations from B sound scenes, each of them are tested R times and each couple of sound sources are presented λ times in one block. In order to have BIBD, a 3 rules govern these parameters :

$$B \geq K, \tag{1a}$$

$$JK = BR, \tag{1b}$$

$$\lambda = R\frac{K-1}{B-1}. \tag{1c}$$

with $[J, B, K, R, \lambda] \in \mathbb{N}$. A reasonable number of listeners is chosen at $J = 50$. With theses parameters fixed, $J = 50$, $K = 20$, $B = 40$ and $R = 25$, the BIBD is not fully balanced (condition (1c) $\notin \mathbb{N}$). It is finally with a partially BIBD that the listening plan is elaborated.[19] It consists, with the parameters J, K and R fixed, in determining the optimal plan that would be the most balanced. The package *sensoMineR* on the *R* software provides the plan where the listening order per participant is defined inside.[20] As the distribution of the simulated and real samples for each participant is almost identical, the obtained listening plan can be considered as balanced (Figure 4).

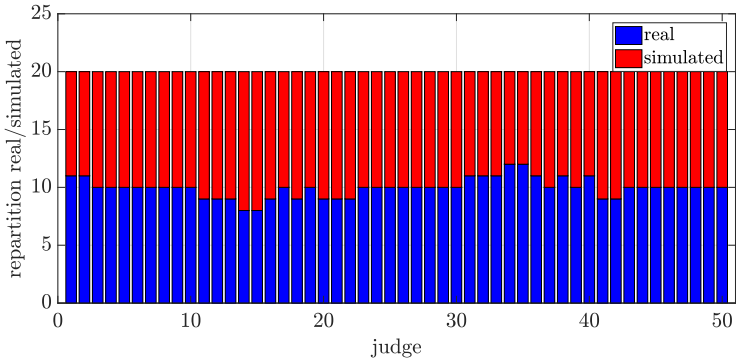

Figure 4: Distribution between the real and simulated scenes for each judge. The cumulative sum is equal to the number of tested elements K**.**

The test was administered online[1] on the 8 February 2017 and the number of 50 participants has been reached 12 days later. During the test, the participant had the possibility to listen to each scene as many times as wanted before evaluating it, without being able to change his/her judgment afterwards. The participant could also leave a comment on each audio to explain the rating. Finally, the age, gender and experience on listening to urban sound mixtures were asked at the end of the test. The panel of 50 listeners was made of 31 males and 18 females (one not documented) with an average age of 36 ($\pm$ 12) years old. 62% of the participants declared having no experience in the listening of urban sound mixtures.

B. RESULTS

First, the note distributions are represented by a box-and-whiskers plot according to whether they belongs to the type 'real' or 'simulated'.

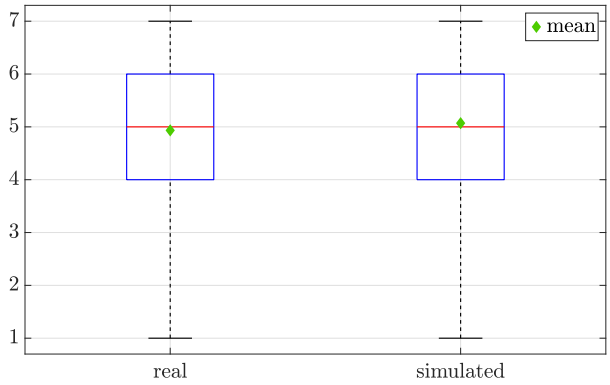

Figure 5: Box-and-whiskers plot of the rating of realism according to the type of scene

The distribution allow to considered that the notation given the subjects are extremely similar. One can note that the mean for the simulated scene is even superior to the real one ($m_{simul.} = 5.1(\pm 1.6)$, $m_{real} = 4.9(\pm 1.6)$).

A paired samples *t-test* is then performed to validate the similarity between the two types. It consists in validate (or not) an H_0 hypothesis which considers the similarity between the distribution of the average scores for the recorded and the simulated scenes of each judge with a Student test. This test establishes a *p-value* that is compared to a threshold value α of 5%. The H_0 hypothesis is then considered if $p-value > \alpha$. The results of the *t-test* (degrees of freedom (DOF), the absolute value *t* and *p-value*) are sum up in the table 4.

	DOF	$\mid t \mid$	*p-value*
type	49	1.37	0.17

Table 4: T-test performed on the distribution of the average scores for the recorded and simulated scenes

The $p - value$ is calculated at $0.17 > \alpha$, which confirms that all the real and simulated scenes are perceived in a similar way by the panel.

[1] `http://soundthings.org/research/xpRealism`

From this global result, a first analyze of variance (ANOVA) is performed to determine if the experience in the listening of urban sound mixtures is an influential factor to distinguish real and replicated scenes. In a similar way than the *t-test*, but based on a Fischer statistical test, an ANOVA allows to determine if the distribution of different level of multiple factors are similar. It allows too to take into account multiple factors and their interactions. Therefore, a two-way ANOVA with interaction with the factors 'experience of the panelist in the listening of urban sound scene' (yes or no) and 'type of scene' (real or simulated) is carried out. The DOF, the Fisher statistics and the associated *p-value* are given in Table 5.

	DOF	*F-statistic*	*p-value*
type	1	0.59	0.44
experience	1	1.97	0.16
type/exp.	1	2.14	0.14

Table 5: Two-ways ANOVA with the factors 'type of sound' and 'listener experience'.

The effect of each factors and the interaction between them are not significant ($p\text{-}value > 0.05$). This means that even listeners trained to hear urban sound environments do not dissociate the real scenes from the simulated scenes better than the non-experienced listeners.

The effect of the 'scene type' and the 'sound environment' factors and the interaction are then studied in a two-ways ANOVA (data distributions displayed on Figure 6 and results in Table 6).

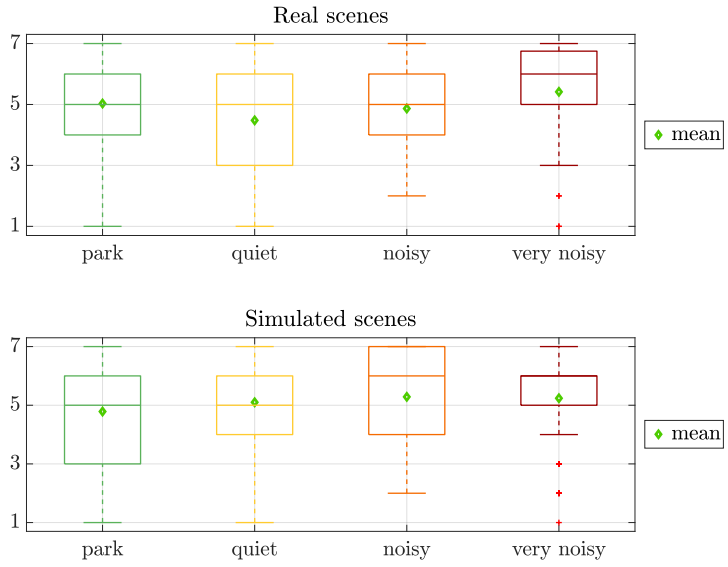

Figure 6: Data distribution for the 'scene type' and 'sound environment' factors

Again, the 'type' factor is not influential in the distribution of the notes ($p - value > \alpha$) whereas the 'judge' and the 'sound environment' factors are. But most of all, interactions effects occur for the sound environment with respectively the 'type' factor and the 'judge' factors. The interaction effect can be easily illustrated for the case of the type and sound environment factors (figure 7). It reveals that the perception of

	DOF	*F-statistic*	*p-value*
type	1	1.38	0.24
sound env.	3	4.69	3.60×10^{-3}
judge.	49	5	0
type/sound env.	3	6.80	0.20×10^{-3}
type/judge	49	1.07	0.35
sound env./judge	147	1.43	1.60×10^{-3}

Table 6: Two-ways ANOVA for the 'type of sound' and 'sound environment' factors with the interaction effect

realism of the simulated and real scenes is not the same depending on the sound environment.

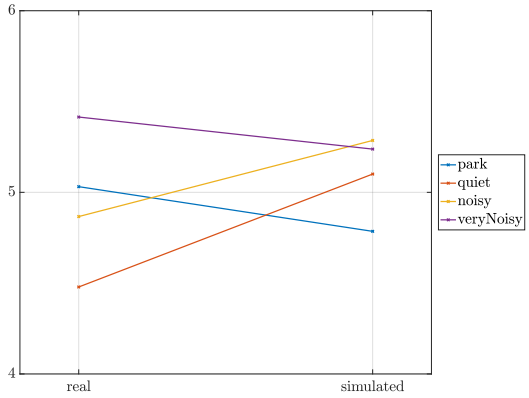

Figure 7: Interaction of the two factors 'type of sound' and 'sound environment'

For the *quiet* and *noisy* sound environments, the mean scores increase when the scenes are simulated whereas for the *park* and *very noisy* sound environment it decreases. As the effect of the *type* factor is not the same according to the *sound environment* levels, there is therefore an interaction effect. One notices that, on the simulated scenes, as the average scores of the *noisy* and *very noisy* atmospheres are higher for both, it seems that the presence of cars has been a preponderant element to increase the perceived realism of an audio mixture. On the contrary, for the *quiet* and *park* atmospheres, the evaluation is more complex. From the comments given by the subjects, it is mainly the sound class *foot step* and a sound background composed of birds that have decreased the scene scores. In a similar way, in the real scenes, it is the presence of too loud birds and some street noises with unknown origins that have been remarked by the panelist and have degraded the evaluation.

5. CONCLUSION

In order to study the acoustic properties of the urban environment, the research community designs tools that have to be evaluated to demonstrate their merit. The availability of simulated scenes realistic enough with precise annotation in terms of content and level will foster research in this direction.

To do so, realistic urban sound mixtures have been composed from the analysis of urban recordings. A

work of annotation has allowed to extract, for different acoustic atmospheres, some useful information as the sound classes presence, the traffic flow rates, the sound level. From these observations, a urban sound database has been created including multiple sound classes both for the sound events and backgrounds. As the road traffic is one of the most noise source, to have a full control on this, multiple car passages recordings have been made on a runway. A perceptual test has been set up to quantify the level of realism of the simulated scenes and has demonstrated that they are comparable to the recorded scenes. Then, the different ANOVA performed have demonstrated that, according to the judge and, most of all, to the sound environment factors, they may still be significant differences in the distribution of notes. In the *park* and *quiet street* environment, for the simulated scenes, the evaluations are more spread and the mean notes lower than the *noisy* and *very noisy* atmospheres. Theses differences come mainly from the sound level of some events which are found to be too loud to be realistic enough.

This problem solved, databases of sound sources and urban scenes can be shared with communities interested in methods of recognizing or detecting sound sources in order to help them to develop methods based on urban sound environments.

6. ACKNOWLEDGEMENTS

We would like to thank Pierre Aumond and Catherine Lavandier from the University of Cergy-Pontoise for transmitting us the data of the *Grafic* project.

REFERENCES

[1] J. Jin Yong, H. Joo Young, and L. Pyoung Jik. Soundwalk approach to identify urban soundscapes individually. *The Journal of the Acoustical Society of America*, 134(1):803–812, 2013.

[2] D. Botteldooren, C. Lavandier, A. Preis, D. Dubois, I. Aspuru, C. Guastavino, L. Brown, M. Nilsson, and T. Andringa. Understanding urban and natural soundscapes. In *Forum Acusticum*, pages 2047–2052. European Accoustics Association (EAA), 2011.

[3] A. Can and B. Gauvreau. Describing and classifying urban sound environments with a relevant set of physical indicators. *The Journal of the Acoustical Society of America*, 137(1):208–218, January 2015.

[4] M. Raimbault, C. Lavandier, and M. Brengier. Ambient sound assessment of urban environments: field studies in two French cities. *Applied Acoustics*, 64(12):1241–1256, 2003.

[5] M. D. Adams, N. S. Bruce, W. J. Davies, R. Cain, P. Jennings, A. Carlyle, P. Cusack, K. Hume, and C. Plack. Soundwalking as a methodology for understanding soundscapes. volume 30, Reading, U.K., 2008.

[6] D. Botteldooren, B. De Coensel, and T. De Muer. The temporal structure of urban soundscapes. *Journal of Sound and Vibration*, 292(12):105–123, 2006.

[7] G Lafay, M Rossignol, N Misdariis, M Lagrange, and J. Petiot. A New Experimental Approach for Urban Soundscape Characterization Based on Sound Manipulation : A Pilot Study. In *International Symposium on Musical Acoustics*, Le Mans, France, 2014.

[8] N.S. Bruce, W.J. Davies, and M.D. Adams. Development of a soundscape simulator tool. In *Proceedings of the INTERNOISE Congress*, Ottawa, Canada, August 2009.

[9] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, October 2013.

[10] CSTB. *Simulation de bruit dynamique*. October 2015.

[11] A. Misra, G. Wang, and P. Cook. Musical Tapestry: Re-composing Natural Sounds. *Journal of New Music Research*, 36(4):241–250, 2007.

[12] P. Aumond, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, and C. Lavandier. Sound pleasantness evaluation of pedestrian walks in urban sound environments. In *Proceedings of the 22nd International Congress on Acoustics*, 2016.

[13] P. Aumond, A. Can, B. De Coensel, D Botteldooren, C. Ribeiro, and C. Lavandier. Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(11), 2016.

[14] M. Rychtrikov and G. Vermeir. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74(2):240–247, 2013.

[15] M. Rossignol, G. Lafay, M. Lagrange, and N. Misdariis. SimScene: a web-based acoustic scenes simulator. In *1st Web Audio Conference (WAC)*, 2015.

[16] J. Salamon, C. Jacoby, and J. Bello. A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the ACM International Conference on Multimedia - MM '14*.

[17] D. Fitzgerald. Harmonic/Percussive Separation Using Median Filtering. *Conference papers*, January 2010.

[18] P. Dagnelie. *Principes d'exprimentation: planification des expriences et analyse de leurs rsultats*. Presses Agronomiques de Gembloux, 2003.

[19] J. Pags and E. Prinel. Blocs incomplets quilibrs versus plans optimaux. *Journal de la Socit Franaise de Statistique*, 148(2):99–112, 2007.

[20] S. Ł and F. Husson. SensoMineR: A package for sensory data analysis (PDF Download Available). *Journal of Sensory Studies*, pages 14 – 25, February 2008.