

# Chapitre 1

## Création de corpus de mixtures sonores urbaines

L'utilisation d'un estimateur permet d'obtenir un niveau sonore du trafic approximé. Mais son utilisation directe sur des enregistrements sonores (où le niveau sonore du trafic réel est inconnu) ne permet pas de connaître son efficacité. En effet, comment connaître avec justesse le niveau sonore estimé puisque que le niveau sonore réel exacte est lui-même inconnu ? La solution proposée est ici d'appliquer cet estimateur sur des scènes sonores urbaines simulées où la contribution du trafic routier sera connue et où les estimations des niveaux sonores pourront être comparé aux solutions exactes. Plusieurs questions sont alors soulevées : comment composer des mixtures sonores urbaines aussi réaliste que des enregistrements sonores ? Comment s'assurer que les corpus de sons sur lesquels sont appliqués les algorithmes permettent de tester les limites des estimateurs ?

### 1.1 Création de scènes sonores : une revue de l'état de l'art

Créer des environnements sonores urbains dépasse le cadre de celui de l'estimation du niveau sonore du trafic. Dans le cadre de l'étude des environnement sonores urbains ou la perception des citadins est étudié, des phases d'écoutes sont réalisées. Ces écoutes peuvent être fait directement *in situ* en ville [Adams *et al.*, 2008] [Raimbault *et al.*, 2003], dans la rue ou bien en laboratoire. Dans ce dernier cas, l'auditeur peut écouter soit des enregistrements audio [Guastavino *et al.*, 2005] soit des mixtures sonores issues d'un processus de simulation [Lafay *et al.*, 2014]. Si la réalisation de *soundwalks* ou l'écoute d'enregistrements audio en laboratoire permettent indéniablement d'avoir une validité écologique, elles n'offrent pas un cadre contrôlé où la présence des sources sonores, leur niveaux sonores pourraient être choisis et modifiés. Il est donc utile de savoir modéliser de tels environnements malgré sa complexité. En effet, l'environnement sonore urbain est un milieu extrêmement variables à la fois temporellement (à un endroit donné, les sources sonores varient constamment) et spatialement (d'un quartier à un

## 1.1. CRÉATION DE SCÈNES SONORES : UNE REVUE DE L'ÉTAT DE L'ART

---

autre, les sources ne sont pas les mêmes). Simuler ces ambiances de façon suffisamment réaliste pour être assimilable à des enregistrements faits en ville n'est donc pas trivial.

### 1.1.1 Simulation totale d'ambiances sonores urbaines

Une des premières approches possible est d'utiliser les techniques d'auralisation pour restituer un environnement sonore urbain [Forssén *et al.*, 2009]. Cette méthode vise à restituer un signal sonore en un point en prenant en compte l'environnement spatial et les modifications qu'il apporte sur ce signal sonore. Cette méthode est couramment utilisée en acoustique du bâtiment. Dans ce domaine, on réalise la convolution entre la réponse impulsionnelle de la salle, obtenue par des mesures réalisées directement dedans si la dite-salle est déjà existante ou bien encore à partir de la modélisation de celle-ci par un logiciel (CATT-acoustics, Odeon), avec un signal sonore, enregistré dans une salle anéchoïque ou bien synthétisé. L'effet de la pièce (réverbération, diffusion) sur la restitution du son en un point donné peut alors être écouté [Vorländer, 2007]. Dans le cas d'un environnement sonore urbain, si l'approche reste la même que dans le cas de l'acoustique du bâtiment, la tâche est plus complexe pour obtenir les réponses impulsionales des rues [Picaut *et al.*, 2005]. En effet, leur mesure est une tâche complexe à réaliser puisqu'elle nécessite tout un dispositif expérimental avec des conditions les plus neutres possibles (faible bruit de fond, conditions météorologiques neutres). C'est pourquoi cette réponse impulsionnelle est le plus souvent simulée à l'aide d'une modélisation numérique. Ce choix pratique nécessite toutefois de savoir correctement modéliser les phénomènes de propagations du son dans un milieu urbain ; tâche complexe car de nombreux phénomènes interviennent : dispersion géométrique, effets météorologiques, effets des sols, des façades et des objets urbains (réflexion, absorption et diffusion). La modélisation de ces phénomènes de propagation acoustique dans un milieu urbain, si elle a fait l'objet de nombreux travaux [Embleton *et al.*, 1976] [Lihoreau *et al.*, 2006], reste à l'heure actuelle une thématique toujours à l'étude afin d'offrir de meilleurs outils prédictifs [Leroy *et al.*, 2010] [Guillaume *et al.*, 2015] et de prendre en compte l'évolution architecturale des villes (végétalisation des bâtiments par exemple [Guillaume *et al.*, 2014]).

Pour obtenir l'auralisation d'une rue ou d'un quartier, il faut également pouvoir prendre en compte le dynamisme de cet espace et les différentes sources sonores qui peuvent s'y trouver. Il est alors possible soit de modéliser certaines sources sonores, comme le trafic routier ou ferroviaire en utilisant des modèles dynamiques pour simuler leur déplacement, soit d'utiliser des enregistrements audio qui, si elle est une approche plus simple, est plus restreinte en possibilité là où la modélisation des sources sonores et le contrôle des paramètres par l'utilisateur permet de réaliser un plus grand nombres de scénarios. Dans [Stienen et Vorländer, 2015], les auteurs résument ces différents aspects, les questions soulevés et les champs d'applications que permet l'auralisation des environnements sonore urbains. Le logiciel *MithraSON* du CSTB propose de générer des auralisations d'environnements sonore<sup>1</sup>. Les sources sonores liées au trafic sont générées en temps réel à partir d'une synthèse granulaire là où l'ensemble des autres sources sonores sont basées sur des enregistrements audio. La propagation des signaux est générée à l'aide d'une

---

1. extrait sonore <https://www.youtube.com/watch?v=ACCV2mi81j8>

méthode de tirs de rayons. Même si les résultats permettent une forte immersion, grâce à la spatialisation du son, cette méthode reste complexe à implémenter et nécessitent des ressources numériques importantes.

### 1.1.2 Composition de scènes sonores

Une autre approche pour simuler les environnements sonores urbains, proposée par M. Schaffer [Schafer, 1993], consiste à les considérer comme la superposition d'événements sonores distinctifs sur un bruit de fonds sonore continu. Ce processus additif permet alors de créer des mixtures sonores en combinant des sons brefs (de 1 à 20 secondes) à des sons plus long (plusieurs minutes) dont les propriétés acoustiques ne varient pas dans le temps. Le défi est alors de disposer de signaux sonores suffisamment différents pour pouvoir recréer la diversité de cet environnement. Plusieurs outils existent comme le logiciel TAPESTRA [Misra *et al.*, 2007] qui se base sur l'extraction de signaux sonores issu d'enregistrements et de leur modulation afin de les insérer dans des mixtures sonores. Les scènes sont alors créées par un processus en trois parties :

- un analyse de phase où des événements sinusoïdaux, transitoires et le bruit de fond sont séparés d'un enregistrement audio. Les événements sinusoïdaux sont sélectionnés à partir d'une représentation temps-fréquence du signal. En fixant des fréquences limites et une amplitude seuil, les événements sont extraits du signal ; les régimes transitoires sont extraits à partir des variations d'énergies brusques du signal dans le domaine temporel. Enfin le bruit de fond est le signal résiduel restant après l'extraction des événements sonores.
- Une phase de synthèse où chaque signal extrait est modifié. Pour les sons sinusoïdaux, ces modifications peuvent être fréquentielles, en multipliant les fréquences des spectres par un facteur, ou bien temporelles en changeant sa durée (allongement, troncature...). Les signaux en régime transitoire peuvent aussi modifiés en hauteur et en durée à l'aide d'un vocoder de phase. Quant au bruit de fond, le choix est fait de générer un nouvel audio similaire à un des audio extraits, à partir d'un algorithme d'apprentissage en arbres d'ondelettes.

Les audio modifiés peuvent alors être placés dans une scène sonore soit de manière bouclée, c'est-à-dire qu'un événement sonore sera placé  $n$  fois dans un intervalle de temps, soit plus précisément en situant temporellement son emplacement superposé à un bruit de fond.

Ces techniques présentent l'avantage de s'appuyer sur des sons réelles issus directement d'enregistrements sonores, et non des sons synthétisés, de pouvoir modifier à l'infini les sons extraits ainsi que d'avoir une grande maîtrise dans la construction des scènes sonores. La limite de cette technique est la phase d'extraction où les événements sonores doivent soit avoir un rapport *signal/bruit* élevé, soit ne pas présenter de recouvrement temporel et fréquentiel avec d'autres sources sonores. Sans cela, l'extraction des signaux est moins performante. Dans le cas d'un milieu sonore urbain, de nombreuses sources sonores présentent du recouvrement et rendent donc

## 1.2. PRÉSENTATION DE *SIMSCENE*

---

l'utilisation de cette méthode difficile. De plus, si pour la création de contenus musicaux savoir modifier des sons est utile, dans le cas de mixtures sonores urbaines, cette modification peut générer des artefacts qui rendraient les mixtures sonores peu réalistes et dénaturés.

D'autre simulateurs se base sur des bases de données de sons pré-existantes comme chez Davies [Bruce *et al.*, 2009]. Leur outil fonctionne sur la superposition d'évènements sonores sur un bruit de fond. Dans leur étude, la base de son utilisée est constituée d'enregistrements des classes de sons qui ont été définies par un panel d'auditeurs comme prépondérantes à la création des ambiances sonores urbaines. Cette approche était justifiée par l'objectif de leur étude qui visait à étudier l'environnement sonore urbain et l'influence de la présence des classes de sons. Ici, puisqu'on souhaite simuler des enregistrements sonores réalisés en ville, on s'attarde à établir un ensemble plus exhaustif des classes de sons présentes. Pour cela, des enregistrements réels sont étudiés afin de savoir quelles sont les sources sonores présentes qui permettront de simuler correctement des mixtures sonores urbaines. Afin de conserver la présence de sons réels dans les mixtures sonores tout en s'affranchissant de ces contraintes, le choix est fait d'utiliser le simulateur *SimScene*.

### 1.2 Présentation de *SimScene*

Le logiciel *SimScene* [Rossignol *et al.*, 2015] est un simulateur de scènes sonores<sup>2</sup> qui consiste à superposer des *évenement* sonore, issus d'une base de données de sons isolés, à une signal *bruit de fond* qui dure tout le long de l'échantillon. À la différence de l'outil TAPESTRA, la base de données est constitué de sons isolés et non plus à partir d'une phase d'extraction. Cette particularité permet d'avoir une grande liberté quant aux sources sonores qu'on peut intégrer. *SimScene* permet de renseigner plusieurs paramètres de hauts niveaux pour réaliser des mixtures sonores :

- le rapport *évenement/bruit de fond* (abrégé ebr pour *Event Background Ratio*),
- le temps de présence moyen d'une classe de son,
- l'occurrence moyenne d'une classe de son dans une scène,
- l'intervalle temporel entre chaque audio d'une même classe de son,
- la présence d'un *fade in* et d'un *fade out* pour chaque échantillon.

Chaque paramètre est également complété par un écart-type permettant d'instaurer de la variabilité entre les scènes simulées. En plus d'un audio pour la mixture sonore globale, un audio pour chaque classe de son présent dans la scène est généré permettant de connaître sa contribution exacte. Dans notre cas, ce sont toutes les classes de sons relatifs au trafic routier qui nous intéressent et qui permettent d'estimer son niveau sonore exact dans la scène.

En parallèle, *SimScene* génère 3 fichiers images (l'évolution temporelle du niveau sonore,

---

2. projet open-source disponible à <https://bitbucket.org/mlagrange/simscene>

le spectrogramme et un *piano Roll* pour visualiser la répartition dans le fichier de chacune des classes, Figure 1.1), un fichier .txt résumant les temps de présence de l'ensemble des sons présents dans la scène et un fichier .mat où se trouve la totalité des résultats et des paramètres de la scène.

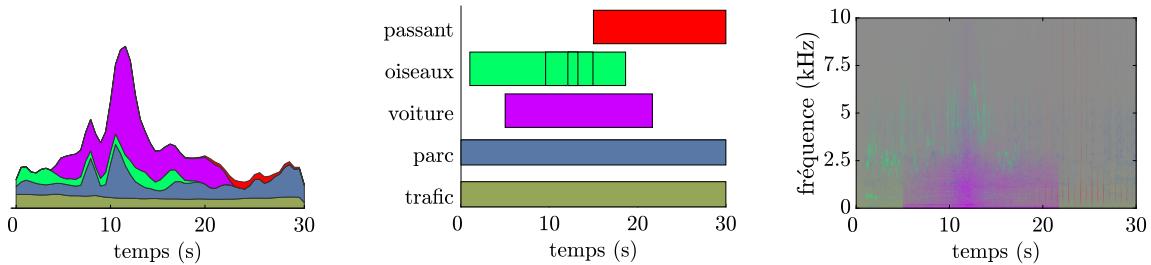


FIGURE 1.1 – Représentation temporelle (à gauche), *Piano Roll* (au centre) et spectrogramme (à droite) générés par *SimScene* d'une scènes composé d'un bruit de fond *trafic*(en vert foncé) et *parc* (en gris) et d'évènements *oiseaux* (en vert), *voiture* (en magenta) et *passant* (en rouge).

La génération de scènes sous *SimScene* peut se faire selon 2 modes. Dans le mode *abstract*, l'utilisateur renseigne lui-même les échantillons sonores présents dans la scène et chaque paramètre permettant de créer des scènes complètement artificielles. À l'inverse, dans le mode *replicate*, le schéma de la scène s'appuie sur un fichier texte où la position d'évènements sonore (début et fin) et leur classe de son correspondante sont détaillées. Ce mode permet de reproduire des scènes réelles annotées.

Si *SimScene* offre de nombreux paramètres pour créer de multiples scènes sonores variées, il nécessite d'avoir une base de données de sons isolés, appelée corpus élémentaire, devant être suffisamment exhaustive. De plus, la qualité de chaque audio (rapport Signal/Bruit élevé, échantillonnage à 44,1 kHz) doit être suffisante pour que leur juxtapositions ne viennent pas détériorer le rendu final.

## 1.3 Crédit d'un corpus élémentaire d'échantillons audio

### 1.3.1 Constitution des évènements et des bruits de fond sonores

La base de données de sons pour *SimScene* comprend un ensemble de classes de sons isolés (oiseaux, voiture, klaxon ...) qui contiennent chacune plusieurs échantillons (*oiseaux01.wav*, *oiseaux02.wav* ...) pour permettre une grande variabilité dans les mixtures sonores créées. La plupart des échantillons sont trouvés sur des sites en ligne de sons<sup>3</sup><sup>4</sup> et à l'aide de la base de données constituée par J. Salamon et al. [Salamon *et al.*, 2014]. Leur base de données comprend en tout plus de 8000 fichiers audio, collectés également sur le site *freesound.org*, d'une durée inférieure à 4 secondes répartit en 10 classes de sons : ventilation, klaxon de voiture, enfants qui

3. [www.freesound.org](http://www.freesound.org)

4. [www.universalsoundbank.com](http://www.universalsoundbank.com)

### 1.3. CRÉATION D'UN CORPUS ÉLÉMENTAIRE D'ÉCHANTILLONS AUDIO

---

joue, chien qui aboie, sonnerie, moteur en fonctionnement, coup de feu, marteau-piqueur, sirène et musique dans la rue. L'ensemble des échantillons a été trié afin de ne conserver que les audio ayant un rapport signal à bruit élevé et un échantillonnage de 44,1 kHz. À partir de la liste des noms des fichiers originaux fournis avec cette base de données, les fichiers audio sont récupérés dans leur intégralité sur le site internet et intégrés dans la base de données.

Afin d'obtenir un rapport signal à bruit acceptable, certains audio ont été filtrés à l'aide du logiciel d'Audacity par un filtre de Wiener. D'autres signaux ont, quant à eux, été tronqués ou bien divisés en plusieurs fichiers afin d'obtenir des durées convenables.

#### 1.3.2 Enregistrements de passages de véhicules

S'il est possible de trouver l'ensemble des classes de son dans une qualité suffisante en ligne, dans le cas de la classe *voiture*, il nous a semblé utile de réaliser des enregistrements de passages de véhicules sur une piste d'essai afin de posséder un ensemble varié et maîtrisé de vitesses et de modèles de véhicules. Pour cela, 4 voitures ont été enregistrées (Renault Mégane, Renault Clio, Renault Sénic et Dacia Sandero) en suivant un plan de mesure défini comprenant plusieurs vitesses stabilisées à différents rapports de vitesses ainsi que des phases d'accélération et de freinage du véhicule. Photo et moteur ??

	Rapport	1	2	3	4	5			
Vitesse stabilisée (km/h)	20	x							
	30		x	x					
	40		x	x	x				
	50			x	x				
	60				x	x			
	70				x	x			
	80					x			
	90					x			
	Total	14							

Freinage		Accélération	
Vitesse (km/h)	Rapport	Vitesse (km/h)	Rapport
50 → 0	3 → 2	0 → 30	1 → 2
40 → 0	2 → 2	0 → 40	1 → 2
50 → 30	3 → 2	20 → 40	1 → 3
60 → 40	4 → 3	30 → 50	2 → 3
70 → 50	4 → 3	40 → 60	3 → 4
80 → 50	4 ou 5 → 3	50 → 70	3 → 4 ou 5

TABLEAU 1.1 – Ensemble de mesures réalisées sur pistes avec des passages de véhicules à vitesses stabilisée (à gauche) et en accélération et freinage (à droite)

Les enregistrements ont été réalisés sur la piste d'essais de l'Ifsttar de Nantes le 7 et 8 juillet 2016 à l'aide du système d'acquisition A COMPLETER, la position du microphone a respecté la norme A COMPLETER et fut donc situé à 7 m de la piste à une hauteur de 1m50. Enfin, les conditions météorologiques étaient satisfaisantes (temps clair et dégagé, température à l'ombre de 25° C , vitesse moyenne du vent inférieure à 2 m/s). Les enregistrements sont ensuite extraits en fichiers audio en format .wav échantillonnés à 44,1 kHz.

Afin d'obtenir des échantillons suffisamment propres, la présence d'oiseaux dans les enregistrements a été atténuée à l'aide d'un filtre médian [Fitzgerald, 2010] appliqué dans la bande de fréquence [2500 – 6500] Hz, correspondante aux fréquences d'émission des oiseaux. Ce filtre consiste à définir une fenêtre et à attribuer la valeur médiane de cette fenêtre à l'élément central. Puisque les aspects à la fois temporel et fréquentiel sont à prendre en compte, le fenêtre du filtre est de forme rectangulaire de dimension 5 × 9 (96 Hz × 230 ms). Un exemple de l'application

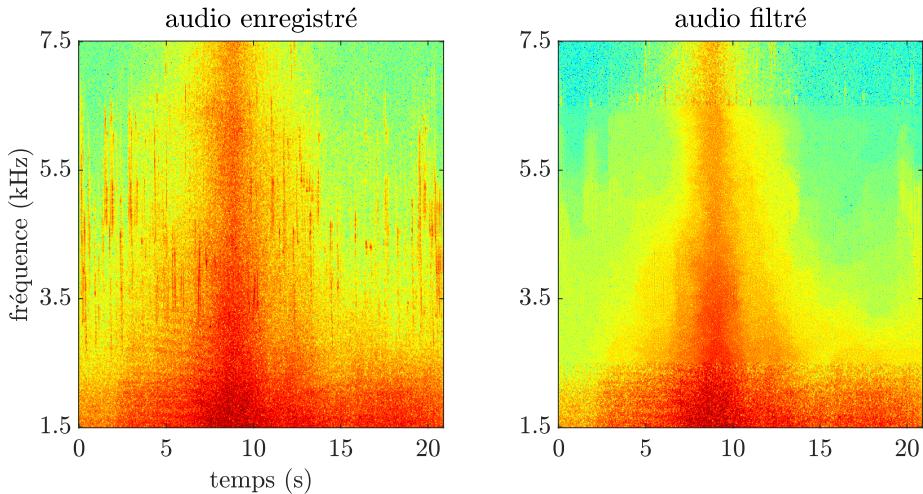


FIGURE 1.2 – Zoom du spectrogramme (nombre de point  $w = 2^{12}$  avec 50 % de recouvrement) dans la bande de fréquence [1500 – 7500] Hz d'un enregistrement de passage de véhicule (véhicule Renault, rapport 3, 40 km/h). À gauche, l'enregistrement original, à droite l'enregistrement filtré par le filtre médian.

de cette fenêtre est présente en Figure 1.2

Même si elle reste persistante sur certains enregistrements, la présence des oiseaux est fortement atténue sans toutefois dégrader la qualité du signal global du véhicule.

### 1.3.3 Composition du corpus élémentaire complet

La base de données est alors divisée en deux catégories. Une première comprend les évènements sonores courts allant de 1 seconde (klaxon, aboiement de chien) à plusieurs dizaines de secondes (passages de voitures, sirènes d'ambulances). Ces éléments permettent de générer les évènements sonores émergeant dans une scène. Une seconde catégorie est composée des sons de durées plus longues (1 min à 2 min) qui vont permettre de construire le bruit de fond utile à la création de l'ambiance sonore générale de la scène (chants d'oiseaux continu, voix d'enfants dans une cours de récréation, trafic routier continu ...).

Les enregistrements des passages de voitures sont, quant à eux, séparés en deux parties : les enregistrements issus des deux premiers véhicules (Renault Mégane, Renault Clio) sont inclus dans le corpus élémentaire, les autres échantillons des deux autres voitures (Renault Sénic, Dacia Sandero) serviront dans la partie ???. Puis, les échantillons sont séparés en deux classes de sons : *voiture Ville* (si la vitesse stabilisée ou finale est inférieure ou égale à 50 km/h) et *Voiture Route* (si la vitesse stabilisée ou finale est supérieure à 50 km/h). L'ensemble des fichiers audio est en format .wav échantillonnés à 44,1 kHz. La base de données finales est résumée dans le Tableau 1.2 pour les évènements sonores et dans le Tableau 1.3 pour les bruits de fond sonores.

La classe de son *bruit rue* résume les nombreux bruits, le plus souvent très bref, dont la

## 1.4. CORPUS DE SCÈNES AMBIANCE

---

Classe de son	Nombre	Classe de son	Nombre
Aboiement de chien	34	Porte de voiture	5
Balais	6	Roulement de valise	5
Bruit de chantier (marteau, perceuse ...)	12	Sirène	9
Bruit de rue	24	Sonnette	5
Camion	4	Toussotement	7
Cloches d'églises	8	Train	7
Klaxon	24	Tram	7
Oiseaux	30	Voiture à l'arrêt	7
Orage	3	Voiture ville	28
Pas dans la ville	11	Voiture route	16
Pas dans un parc	16	Voix (rire, 1 ou 2 mots)	24
Porte de maison	5	<b>Total</b>	<b>321</b>

TABLEAU 1.2 – Composition de la base de données pour les évènements sonores

Classe de son	Nombre	Classe de son	Nombre
Brouhaha de foule	15	Pluie	14
Brouhaha parc	25	Trafic routier	9
Chantier	28	Vent dans les arbres	15
Cours de récréation	12	Ventilation	10
Oiseaux	25	<b>Total</b>	<b>153</b>

TABLEAU 1.3 – Composition de la base de données pour les bruits de fond

source sonore n'a pas pu être déterminée. De la même façon, les sons relatifs à un chantier en construction (marteau-piqueur, marteau, perceuse) sont regroupés en une seule classe par soucis de simplification.

À partir de ce corpus constitué, disponible en ligne<sup>5</sup>, il est possible de réaliser des corpus de scènes sonores urbaines. En vue de tester à la foi les limites des estimateurs et leur performances lors d'une utilisation proche de mesures faites en ville, deux corpus de scènes sonores urbaines sont construits.

## 1.4 Corpus de scènes *Ambiance*

Dans un premier temps le choix est fait de générer un corpus où la présence de chaque source est défini selon sa classe de son et où les niveaux sonores du trafic sont calibrés. Ce corpus a vocation à estimer le comportement de la NMF selon certaines sources sonores isolées et selon la prédominance du trafic routier dans les scènes.

Nommé *Ambiance*, ce premier corpus consiste en un ensemble de 6 sous-corpus de 25 scènes ayant chacune une durée de 30 secondes. Chaque sous-corpus mélange une composante *trafic*

---

5. <https://zenodo.org/record/1213793>

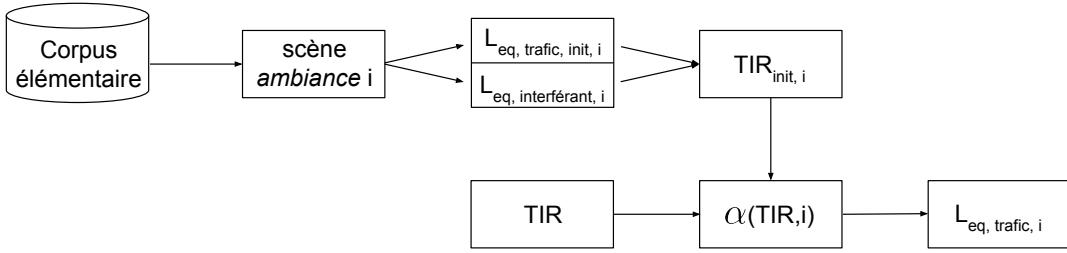


FIGURE 1.3 – Diagramme bloc de la pondération du signal trafic selon la scène  $i$  et le TIR

avec une classe de son spécifique (appelée classe *interférante*). Ces 6 classes de sons sont : *alerte* (qui inclut des sons de klaxons et de sirènes), *animaux* (siflement d’oiseaux et aboiement de chien), *climat* (pluie, vent et orage), *humains* (voix humaine), *mécanique* (bruit métallique et ventilation), *transport* (trains, avion et tramway). Chaque scène générée possède alors un niveau sonore trafic initial  $L_{eq,trafic,init}$  et un niveau sonore *interférant*,  $L_{eq,interferant}$  répliquée ensuite 5 fois afin d’y calibrer le niveau sonore du trafic telle que  $L_{eq,trafic} - L_{eq,interferant} = TIR$  avec  $TIR = \{-12 -6 0 6 12\}$ . Pour cela, les fichiers audio relatifs au trafic sont pondérés par un coefficient  $\alpha$  afin d’obtenir le niveau sonore souhaité selon le  $TIR$  avec

$$\alpha(TIR, i) = 10^{TIR - TIR_{init,i}/20} \quad (1.1)$$

où  $TIR_{init,i} = L_{eq,trafic,init,i} - L_{eq,interferant,i}$ . Lorsque  $TIR < 0$  dB, le signal trafic est plus faible que le signal interférant, à l’inverse lorsque  $TIR > 0$ , le trafic devient la classe sonore prépondérante. En tout 750 scènes sont ainsi disponibles (6 sous-corpus  $\times$  25 scènes  $\times$  5 TIR). Ce corpus est donc totalement artificiel dans son aspect et ne peut pas être assimilable à des enregistrements sonores réalisés en villes.

## 1.5 Corpus de scènes grafic

Pour palier à l’artificialité de corpus *Ambiance*, un second corpus est généré, basé sur des enregistrements sonores réalisées en villes. Ce corpus a pour vocation à tester les performances de la NMF sur des scènes similaire à des enregistrements sonore faits en ville.

### 1.5.1 Présentation des scènes *GRAFIC*

Les enregistrements audio de références sont issus du projet GRAFIC [Aumond *et al.*, 2017] et ont été enregistrés à pied dans le 13<sup>e</sup> arrondissement de la ville de Paris sur un parcours comprenant 19 points d’arrêts (Figure 1.4). Le parcours définit présente l’avantage de couvrir plusieurs ambiances sonores représentatifs d’un environnement sonore urbain (Tableau 1.4).

Ce trajet a été parcouru sur deux jours (le 23/05/2015, jour 1, et le 30/05/2015, jour 2),

## 1.5. CORPUS DE SCÈNES GRAFIC

---



FIGURE 1.4 – Parcours réalisé par l'étude avec les 19 points de mesures avec le niveau sonore mesuré équivalent

deux fois par jour (le matin puis l'après-midi) dans un sens (d'est en ouest, EW) et dans l'autre (d'ouest en est, WE). L'enregistrement est réalisé par un système d'acquisition équipé d'un microphone ASA Sense omnidirectionnel situé sur un sac à dos porté par l'opérateur. En tout, 76 enregistrements audio (19 points  $\times$  4 trajets) de 1 à 4 minutes sont disponibles.

Point	Description	Point	Description
1	Large rue à deux voies	10	Rue sans trafic près d'une école
2	Large rue à deux voies	11	Rue silencieuse sans trafic
3	Parc calme	12	Rue avec un faible débit de trafic
4	Rue animée avec restaurant/bar	13	Rue avec un faible débit de trafic
5	Rue très calme	14	Rue avec un faible débit de trafic
6	Rue animée avec restaurant/bar	15	Rue avec un fort débit de trafic
7	Rue animée avec restaurant/bar	16	Rue avec un fort débit de trafic
8	Parc situé le long d'une rue	17	Rue piétonne calme située entre deux rues bruyantes
9	Rue avec un trafic modéré	18	Grand carrefour avec un trafic constant
		19	Grand parc

TABLEAU 1.4 – Résumé des 19 points de mesures avec l'ambiance générale.

### 1.5.2 Écoutes des scènes sonores

La première étape établit un classement selon quatre ambiances sonores (*parc, rue calme, rue animée, rue très animée*), comme défini par [Can et Gauvreau, 2015], des enregistrements sonores à partir des indications fournies dans [Aumond *et al.*, 2017] (résumé dans le Tableau 1.4) et des écoutes faites (Tableau 1.5).

Jour	trajet	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																			

Parc		Rue calme		Rue animée		Rue très animée		Non renseigné	
------	--	-----------	--	------------	--	-----------------	--	---------------	--

TABLEAU 1.5 – Classification des scènes par ambiances sonores.

événements	t <sub>init</sub> (s)	t <sub>fin</sub> (s)
bruit rue	0,00	8,50
voix	0,00	44,00
camion	1,00	56,10
voix	36,50	42,30
voiture Ville	52,00	63,00
voix	59,00	66,50

TABLEAU 1.6 – Exemple d'un fichier d'annotation pour la scène 1-EW-07.

Une majorité de scènes appartiennent à l'ambiance sonore *rue calme* (35 scènes), 23 scènes appartiennent à l'ambiance *rue animée* et 8 scènes à l'ambiance *parc* et *rue très animée*. Plus de la moitié des points de mesures possèdent la même ambiance sur les 4 trajets. À l'exception du point 10, les autres points de mesures possèdent deux ambiances sonores voisines. Ces variations proviennent des variations des activités dans la journée (matin ou l'après-midi). Enfin, les points 3 et 19 du parcours 1-WE ne sont pas exploitables : le point 3 est pollué par un camion balayeur et le point 19 n'a pas été correctement enregistré. Au final, c'est 74 fichiers audio qui sont disponibles et utilisés pour créer des scènes sonores. Ces 74 enregistrements forment le *corpus de référence*.

### 1.5.3 Annotation des enregistrements sonores

L'annotation des 74 enregistrements est ensuite réalisée. Il consiste à écouter chaque fichier audio et à estimer les sources sonores présentes ainsi que leur temps de présence. Pour chaque enregistrement, l'ensemble des annotations est résumé dans un fichier .txt. Un exemple d'annotation est présenté dans le Tableau 1.6.

De ces annotations, il est alors possible d'estimer par ambiance sonore, un niveau sonore moyen, les classes de sons qui caractérisent leur bruit de fond et également les classes de sons catégorisées en événements sonores et leur densité (nombre d'événement par minute). Ces informations sont alors suffisantes pour pouvoir recréer ces scènes par le mode *abstract* de *SimScene* (Tableau 1.7).

Sur l'ensemble des scènes sonores, 11 classes de sons sont identifiés en tant qu'évènement

## 1.5. CORPUS DE SCÈNES GRAFIC

---

Environnement sonore	Niveau sonore (dB)	Bruit de fond	Évènement	Nombre évènement/min	Rapport Évènement-Bruit de fond (dB)
Parc	69,0	voix siflements d'oiseaux	voiture ville	1,6	3,0 ( $\pm$ 6,0)
			voix	0,5	6,5 ( $\pm$ 5,0)
			siflements d'oiseaux	0,5	0,0 ( $\pm$ 9,5)
			bruit de rue	0,5	6,7 ( $\pm$ 4,5)
			bruit de pas	0,3	4,0 ( $\pm$ 7,0)
Rue calme	70,2	trafic routier siflements d'oiseaux	voiture ville	1,7	7,6 ( $\pm$ 4,6)
			voix	0,7	8,2 ( $\pm$ 4,0)
			bruit de rue	0,7	7,6 ( $\pm$ 4,2)
			bruit de pas	0,5	8,0 ( $\pm$ 5,0)
			siflements d'oiseaux	0,2	3,0 ( $\pm$ 5,8)
			porte de maison	0,2	9,0 ( $\pm$ 3,3)
			porte de voiture	0,2	7,7 ( $\pm$ 4,2)
			chantier	0,1	3,7 ( $\pm$ 5,1)
			voiture ville	9,4	3,3 ( $\pm$ 2,5)
Rue animée	73,5	trafic routier	voix	0,6	1,3 ( $\pm$ 2,6)
			bruit de pas	0,5	-3,6 ( $\pm$ 6,4)
			bruit de rue	0,4	5,2 ( $\pm$ 4,6)
			klaxon	0,3	3,5 ( $\pm$ 3,9)
			siflements d'oiseaux	0,2	1,6 ( $\pm$ 5,0)
			porte de voiture	0,2	4,4 ( $\pm$ 5,4)
			sirène	0,1	2,0 ( $\pm$ 6,2)
			sonnette	0,1	1,7 ( $\pm$ 3,5)
			voiture ville	40,9	2,3 ( $\pm$ 1,3)
Rue très animée	76,0	trafic routier	voix	0,3	1,3 ( $\pm$ 1,1)
			klaxon	0,3	2,7 ( $\pm$ 4,1)
			porte de voiture	0,3	3,6 ( $\pm$ 5,4)
			sirène	0,2	-3,0 ( $\pm$ 4,2)
			bruit de pas	0,2	-3,6 ( $\pm$ 5,8)
			bruit de rue	0,2	5,1 ( $\pm$ 4,7)

TABLEAU 1.7 – Niveau sonore et description des classes de sons les plus récurrentes dans l'environnement urbain (nombre d'évènements sonore par minute  $> 0.1/\text{min}$ ).

sonore (trafic routier, voix, sifflements d'oiseaux, bruit de rue, bruit de pas, porte de maison, porte de voiture, chantier, klaxon, sonnette, sirène) et 3 classes de sons sont présentes en tant que bruit de fond sonore (brouhaha de foule, sifflements d'oiseaux, trafic routier continu). Les sources sonores les plus communes sont *voiture*, *voix* et *bruit rue*. En outre, en plus des classes de sons résumées dans le Tableau 1.7, de nombreuses autres classes de sons (*abolement de chien*, *bruit de balais*, *toussotement*, *passage d'avion*, *roulement de valise*) entendus intervient plus sporadiquement (nombre d'évènement/min < 0,1) et sont susceptibles d'intervenir dans les quatre ambiances sonores.

La composition des environnements sonore diffère entre eux : dans *parc* la voix et les oiseaux sont les bruits de fond sonores principales permettant d'établir l'ambiance sonore adéquat, puis, plus la rue est animée plus la part de la classe *trafic* et celles de l'activité humaine (*voix*, *bruit de pas*) sont prédominantes. À l'inverse, les classes de sons « naturel » (*oiseaux*) disparaissent progressivement.

Notons que dans *rue calme*, *animée* et dans *parc*, le décompte des voitures est assez aisé. Il l'est beaucoup moins dans *rue très animée* où c'est un flot de véhicules peut être présent, le comptage y est alors très délicat car les véhicules peuvent être considérés à la fois comme bruit de fond et évènements sonore. Ainsi, lorsque le flux de véhicule est trop important, on considère en moyenne 1 véhicule par seconde. Ce nombre est donc soumis à une forte incertitude mais reste cependant cohérent avec les indications du débit moyen fournis dans [?].

#### 1.5.4 Reproduction des enregistrements audio

Afin d'obtenir des scènes les plus réalistes possibles, le choix a été fait de reproduire les 74 enregistrements à l'aide de leur annotation et du mode *replicate* de *SimScene*. Ce choix permet ainsi de s'assurer que la disposition des évènements sonores dans les mixtures sonore est la plus proche possible d'une structure temporelle ayant déjà été réalisée. La difficulté réside surtout dans l'estimation du *ebr* pour les évènements sonores qui doit être cohérent par rapport à l'ambiance souhaitée. Son estimation et sa variance s'est donc faite empiriquement et a été ajusté progressivement afin d'obtenir un rendu satisfaisant. Le niveau sonore global de la scène simulée est enfin modifié pour être similaire à celui de la scène réelle.

Dans la suite du document, les scènes issues du mode *replicate* de *SimScene* seront appelées « scènes répliquées » en raison du processus de duplication. Les scènes originelles sont quant à elle nommées « scènes enregistrées ». L'ensemble des ces scènes répliquées forment le *corpus d'évaluation* (voir Figure 1.5).

Afin de vérifier que le rendu global des scènes répliquées est suffisamment réaliste pour être assimilables aux enregistrements faits en ville, celles-ci sont soumises à un test perceptif.

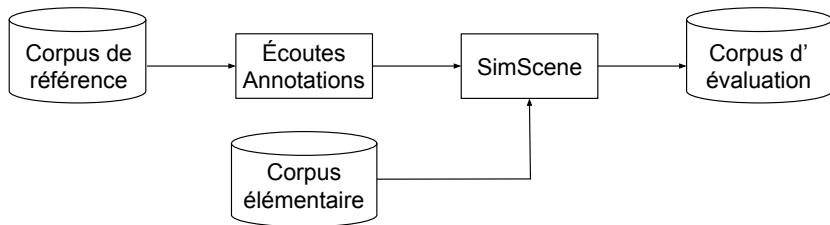


FIGURE 1.5 – Diagramma bloc résumant la création du corpus d'évaluation

### 1.5.5 Mise en place du test perceptif

Ce test consiste à faire écouter, à un panel d'auditeurs, un ensemble de scènes sonores comprenant autant d'enregistrements sonores que de scènes reconstitués. À chaque scène, l'auditeur doit alors évaluer, sur une échelle de Likert à 7 points allant de « très peu réaliste » à « extrêmement réaliste », le réalisme de la scène qu'il vient d'entendre. L'objectif est que l'ensemble des scènes répliquées soient perçues de façon similaire aux scènes réalistes.

Sur l'ensemble des 148 scènes (74 enregistrées, 74 répliquées), un ensemble de 40 scènes sont testés. Cet ensemble est composé dans une première moitié de scènes enregistrées choisis aléatoirement parmi les 74 enregistrements tout en prenant soin d'avoir une répartition équitable entre les ambiances sonores afin d'avoir suffisamment de diversité sonore. On extrait alors 5 scènes issues d'une ambiance *Parc*, 6 issues de *Rue calme*, 4 de *Rue animée* et 5 de *Rue très animée*. Pour chaque audio, 30 secondes sont ensuite sélectionnés aléatoirement. Dans la seconde moitié, les mêmes scènes répliquées sont sélectionnés (même audio mais en version répliquée, même 30 secondes). L'hypothèse faite est que si le réalisme de ces 20 scènes répliquées est perçu de la même manière que les 20 scènes enregistrées, celui-ci pourra être étendu aux 54 autres scènes répliquées. Un récapitulatif des fichiers audio sélectionnés et de la position des 30 secondes extraites sont résumés dans le Tableau 1.8.

Un seul auditeur n'écoute toutefois pas les 40 scènes disponibles car le test serait trop long ( $\approx 20$  minutes) et la capacité de concentration de l'auditeur ne pourrait pas être constante tout le long du test. Ainsi, chaque auditeur écoute un sous-corpus de 20 audio ; la durée du test n'excède alors pas 10 minutes.

Comme les auditeurs n'évaluent plus l'ensemble des scènes mais seulement une partie, il faut définir un plan d'écoute qui répartit équitablement l'ordre de succession des écoutes. Pour cela, on réalise un « Bloc Équilibré Incomplet » (BEI) [Pagès et Périnel, 2007].

En analyse sensorielle, un BEI permet d'élaborer l'ordre d'évaluation des produits testés pour chaque panéliste en évitant que des biais statistiques apparaissent (effet de rang, du juge, de succession ...). Il se construit à partir de plusieurs variables :

- le nombre de juges  $J$  (appelé aussi *blocs*),

ambiance	t <sub>deb</sub>	t <sub>fin</sub>		id	scènes enregistrées		id	scènes répliquées
Parc	41,7	71,7		1	1_EW_03		21	replicate_1_EW_03
	20,5	50,5		2	1_EW_08		22	replicate_1_EW_08
	38,2	68,2		3	1_EW_10		23	replicate_1_EW_10
	56,2	86,2		4	2_EW_03		24	replicate_2_EW_03
	38,5	68,5		5	2_WE_19		25	replicate_2_WE_19
Rue Calme	20,0	50,0		6	1_EW_05		26	replicate_1_EW_05
	135,5	165,5		7	1_WE_06		27	replicate_1_WE_06
	28,6	58,6		8	1_WE_14		28	replicate_1_WE_14
	38,6	68,6		9	2_EW_13		29	replicate_2_EW_13
	110,7	140,7		10	2_WE_10		30	replicate_2_WE_10
	109,3	139,3		11	2_WE_05		31	replicate_2_WE_05
Rue animée	19,8	49,8		12	1_EW_01		32	replicate_1_EW_01
	211,6	241,6		13	1_EW_18		33	replicate_1_EW_18
	8,8	38,8		14	2_EW_02		34	replicate_2_EW_02
	57,5	87,5		15	1_WE_02		35	replicate_1_WE_02
Rue très animée	69,9	99,9		16	1_EW_16		36	replicate_1_EW_16
	75,6	105,6		17	1_WE_16		37	replicate_1_WE_16
	34,6	64,6		18	2_EW_16		38	replicate_2_EW_16
	87,3	117,3		19	2_WE_15		39	replicate_2_WE_15
	87,1	117,1		20	2_WE_18		40	replicate_2_WE_18

TABLEAU 1.8 – Résumé des 40 audio composant l’ensemble des scènes testés avec les temps d’extraction des 30 secondes d’audio, l’identifiant et le nom des fichiers audio originaux.

- le nombre de produits à tester,  $B$  (appelé aussi *variétés* ou *traitements*),
- le nombre de produits testé par juge,  $K$
- le nombre de réplications d’un produit,  $R$
- le nombre de répétabilités d’une paire de produit,  $\lambda$ .

Plusieurs conditions sont à remplir entre ces variables pour réaliser un BEI correct :

$$B \geq K, \quad (1.2a)$$

$$JK = BR, \quad (1.2b)$$

$$\lambda = R \frac{K - 1}{B - 1}. \quad (1.2c)$$

avec  $[J, B, K, R, \lambda] \in \mathbb{N}$ .

La dénomination « incomplète » provient de l’évaluation des juges que d’une partie de l’ensemble des produits à tester (condition 1.2a). La dénomination « équilibré », quant à elle, provient de la constance de  $\lambda$  pour les différents couples de  $B$ .

Plusieurs paramètres ont été choisis et justifiés au début de la partie : le nombre de produit

## 1.5. CORPUS DE SCÈNES GRAFIC

---

testé a été établi à 40 ( $B = 40$ ) pour un nombre de produit testé par juge fixé à 20, ( $K = 20$ ).

La principale difficulté reste à obtenir la participation de  $J$  personnes pour ce test. Ce nombre est alors fixé à  $J = 50$  en cela que ce nombre est suffisant et facilement atteignable en peu de temps.

À partir des variables  $J$ ,  $B$  et  $K$ , le nombre  $R$  de réPLICATION est défini à 25. Toutefois, ces valeurs impliquent que la condition 1.2c n'est pas validée ( $\lambda = 9,69 \notin \mathbb{N}$ ) et donc que les contraintes que l'on s'impose ne permettent pas d'obtenir un plan équilibré. Deux solutions sont alors possibles : la première serait de modifier certains paramètres pour trouver l'équilibre. Or le nombre de juges,  $J = 50$ , paraît un nombre limite raisonnable à atteindre tout comme le nombre de fichiers audio à tester  $K$ . Avec ces 2 contraintes fixées, il n'est pas possible d'obtenir un plan d'écoute adéquat. La deuxième solution, qui semble alors la plus adaptée, est de réaliser un plan optimal [Pagès et Périnel, 2007]. Dans ce cas, pour une configuration  $[J, K, R]$  donnée, un algorithme d'échange détermine un « plan optimal » qui satisfait au mieux son équilibre (sans toutefois l'atteindre parfaitement).

Le plan optimal  $X_{opt}$  en fonction des conditions  $J$ ,  $K$  et  $R$  est réalisé sous le logiciel  $R$  à l'aide la fonction *optimaldesign* fourni par le package *SensomineR* [Lê et Husson, 2008]. Cette méthode établit dans un premier temps, le nombre de combinaison total possible ( $J \times B$ ) puis un premier plan des combinaisons possibles (appelé  $X$  de dimension  $J \times K$ ) est élaboré de façon aléatoire. Celui-ci est ensuite mis à jour itérativement en remplaçant chaque combinaison possible  $\tau_{j,k}$  par une autre combinaison  $\tau_{j,b}^*$  extrait de la matrice de combinaison totale, de telle façon à minimiser le produit matriciel 1.3. Ce procédé est le principe de l'algorithme d'échange.

$$\min_{\tau_{j,b}} \det(X'X)^{-1}. \quad (1.3)$$

Cet algorithme est dit *D-optimal* car il fait intervenir l'opérateur *Déterminant* mais il peut être *A-optimal* en faisant appel à l'opérateur *Trace* à la place. Le résultat est alors un plan  $X_{opt}$  de dimensions  $J \times K$  résument l'ordre d'écoutes des fichiers audio pour chaque juge. La Figure 1.6 résume le nombre de réPLICATION de chaque scène dans le plan obtenu.

L'optimisation du plan ne permet alors pas d'avoir un nombre de réPLICATION  $R$  constant mais variable évoluant dans l'intervalle [20 – 30]. On s'assure ensuite que la répartition entre les scènes enregistrées et simulées pour chaque juge est la plus équilibrée que possible (Figure 1.7).

Le plan généré permet bien d'avoir en moyenne une répartition équilibrée entre les scènes réelles et simulées par juge, même si certain ont jusqu'à 12 scènes d'un même type.

Une page web<sup>6</sup> est mis en ligne le 8 février 2017 permettant l'accès au test à une large public et s'est clôturé 12 jours plus tard. Chaque juge écoute donc une succession de 20 audio de 30 se-

---

6. <http://soundthings.org/research/xpRealism>

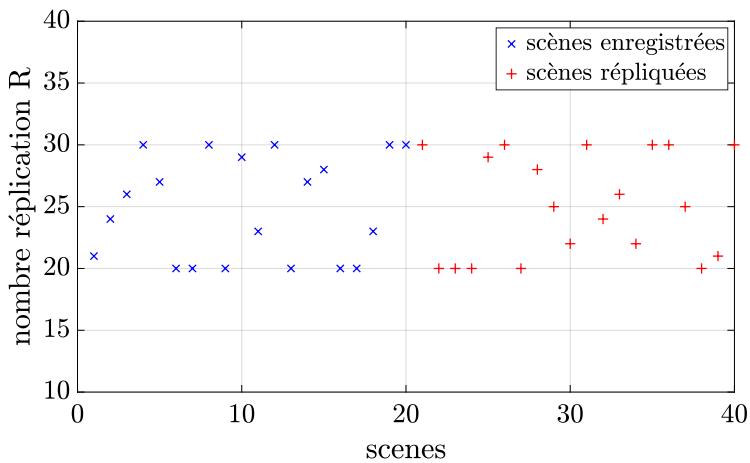


FIGURE 1.6 – Nombre de réPLICATION,  $R$ , pour chaque scène obtenu dans  $X_{opt}$  avec comme combinaison  $J = 50$ ,  $B = 40$ ,  $K = 20$ . Les 20 premières scènes sont les scènes issues des enregistrements du projet GRAFIC, les 20 suivantes sont les scènes répliquées sous *SimScene*.

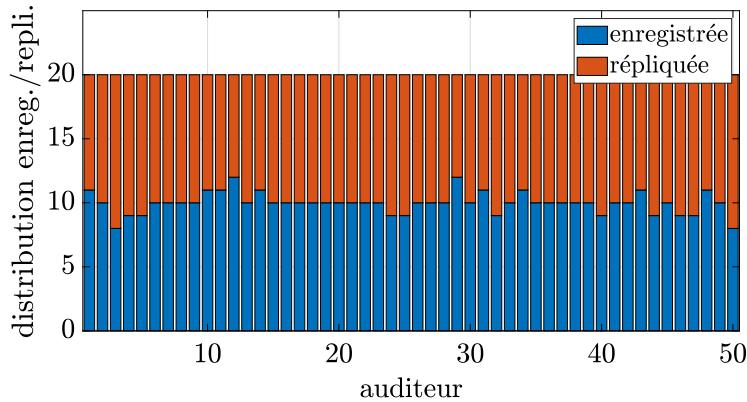


FIGURE 1.7 – Répartition entre les scènes enregistrées et répliquées par juge. La somme cumulée des deux ensembles pour chaque juge correspond au nombre d’écoute  $K$ .

condes dans un ordre établit par le plan optimal. Chaque audio peut être réécouté autant de fois que possible avant d’être évalué sans qu’il soit toutefois possible de revenir sur son évaluation. L’auditeur a également la possibilité de laisser un commentaire sur chaque audio pour pouvoir justifier son choix. En fin de test, afin de connaître le panel d’évaluateur, il est demandé aux juges de renseigner leur âge, leur sexe (H/F) et leur expérience quant à l’écoute de mixtures sonores urbaines.

Les fichiers résultats sont stockés également sous une page web<sup>7</sup> et téléchargeable sous le format .json pour ensuite être traités sous le logiciel Matlab.

7. <http://soundthings.org/research/xpRealism/responses/>

### 1.5.6 Résultats

#### 1.5.6.1 Constitution du panel

La figure 1.8 résume, sous forme d'histogrammes, l'âge, le sexe et l'expérience des auditeurs. 2 personnes ont renseigné aucun de ces champs et une troisième personne a seulement omis de préciser son sexe.

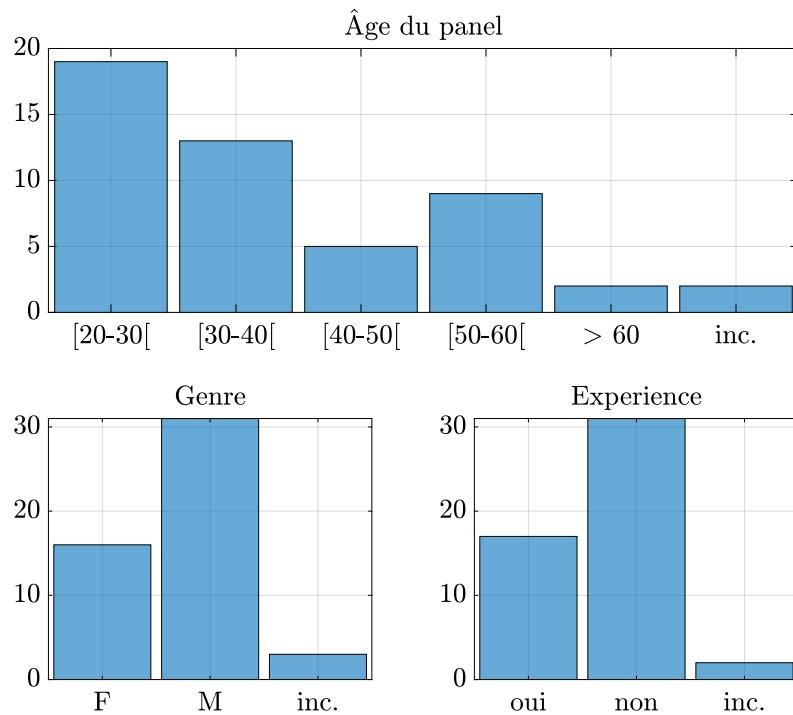


FIGURE 1.8 – Résumé des informations relatifs aux auditeurs

Le panel est composé à 62 % d'hommes et à 32 % de femmes. La classe d'âge [20 – 30[ est la plus représentée suivie de la classe [30 – 40[ (26 %), [50 – 60[ (18 %), [40 – 50[ (10%) et enfin de la classe > 60 (4 %) . 62 % du panel a déclaré n'avoir pas d'expérience dans l'écoute d'ambiances sonores urbaines. Cette dernière caractéristique indique que la majorité des jugements provient d'auditeurs inexpérimentés dans ce domaine et se sont donc plus attardés sur une évaluation générale de la scène là où les auditeurs plus expérimentés se sont attardés en plus sur des aspects plus particulier comme la différence de réverbération entre les sources sonores.

On présente différents test statistiques afin de voir si les scènes répliquées ont été perçues de façon similaires aux scènes enregistrées. Cette vérification permet ensuite de valider l'utilisation du corpus d'évaluation pour tester les performances de la NMF.

### 1.5.6.2 Représentation et test de Student

Dans un premier temps, la distribution de toutes les notes des scènes enregistrées et répliquées est exprimée au travers d'une boîte à moustache (Figure 1.9). Cette représentation graphique permet de comparer plusieurs distributions en résumant pour chaque boîte la médiane (trait plein rouge), les valeurs du premier quartile au troisième quartile (boîte en bleue), la valeur maximale et minimale de la distribution (respectivement trait supérieur et inférieur en noir). À cela est également ajoutée la moyenne.

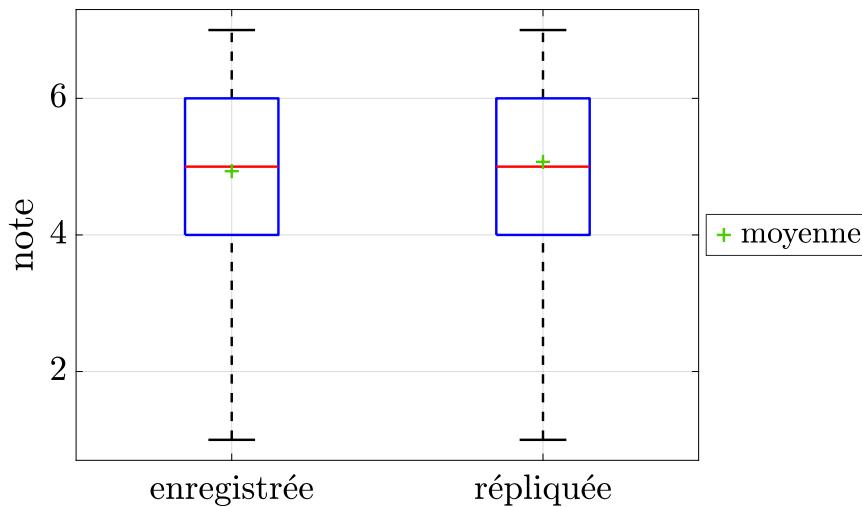


FIGURE 1.9 – Représentation en diagramme en boîte à moustache entre les scènes réelles et simulées

La répartition des notes pour les deux types de scènes est fortement similaire. Chaque type présente des valeurs identiques (médiane, valeurs extrêmes, quantiles). Seule la note moyenne permet de différencier les deux ensembles :  $m_{En} = 4.93(\pm 1.64)/7$  et  $m_{Re} = 5.06(\pm 1.56)/7$ .

À cette première observation, un test  $t$  de Student est considéré pour chaque scène entre les notes de la catégorie *enregistrée* et la catégorie *répliquée*. Un test de Student consiste à comparer les moyennes de 2 groupes d'échantillons pour déterminer si elles sont significativement différentes d'un point de vue statistique. Toutefois, puisque pour chaque scène, les évaluations entre le pendant *enregistré* et *répliqué* sont réalisées par des individus différents, que le nombre d'évaluation par catégorie n'est pas identique et que les variances entre les deux catégories ne sont pas égales, c'est une variante du test- $t$  de Student qui est réalisée : le test- $t$  de Welch [Ruxton, 2006]. Dans ce test, pour chaque scène, deux hypothèses sont émises sur les distributions :

- les distributions des échantillons des deux catégories sont semblables (hypothèse *nulle*  $H_0$ ),
- les deux distributions sont différentes, (hypothèse *alternative*  $H_1$ ).

Plusieurs statistiques sont alors définis :

- la valeur  $t$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \quad (1.4)$$

où  $\bar{X}_i$ ,  $s_i$  et  $N_i$  sont, respectivement, la moyenne de l'échantillon, la variance et le nombre d'échantillon de la catégorie  $i$ ,

- les degrés de liberté  $DDL$

$$DDL = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}} \quad (1.5)$$

Ces statistiques sont alors utilisées avec une loi de Student pour déterminer une valeur  $p$  qui permet de rejeter (ou non) l'hypothèse  $H_0$  selon une valeur seuil de référence  $\alpha$  (défini à 5 %) :

- si  $\alpha > p$ , il existe alors au moins deux distributions différentes, l'hypothèse  $H_0$  est rejetée et  $H_1$  est acceptée,
- si  $\alpha < p$ , l'hypothèse  $H_0$  n'est pas considérée comme *vraie* mais on considère qu'il n'y a pas de raison à rejeter  $H_0$ . Cette nuance provient du fait que cette décision se base sur un nombre limité d'informations (le nombre total d'observations) qui ne permet pas de rejeter totalement l'hypothèse  $H_1$ .

L'ensemble des 20 valeurs  $p$  et les boîtes à moustaches de chaque scène sont résumées dans les Figures 1.10 et 1.11.

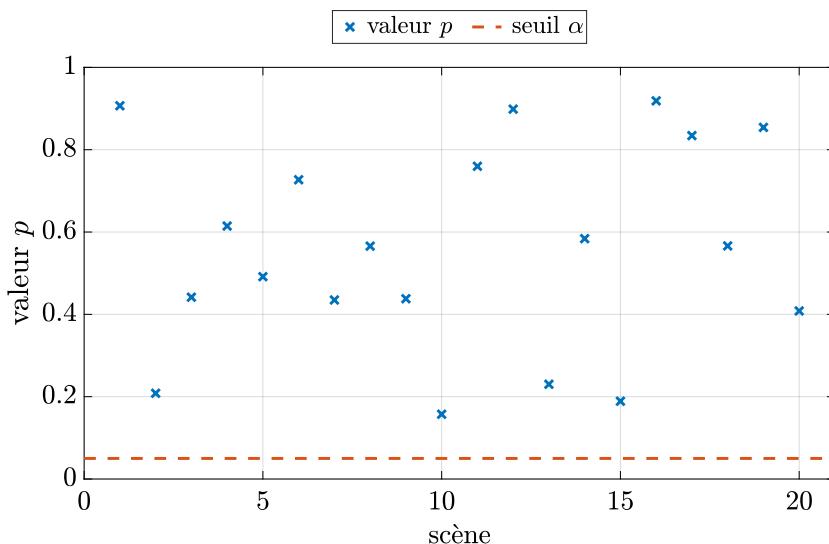


FIGURE 1.10 – Résumé des valeurs  $p$  calculé pour chaque scène entre les notes issues du type *enregistré* et du type *répliqué*.

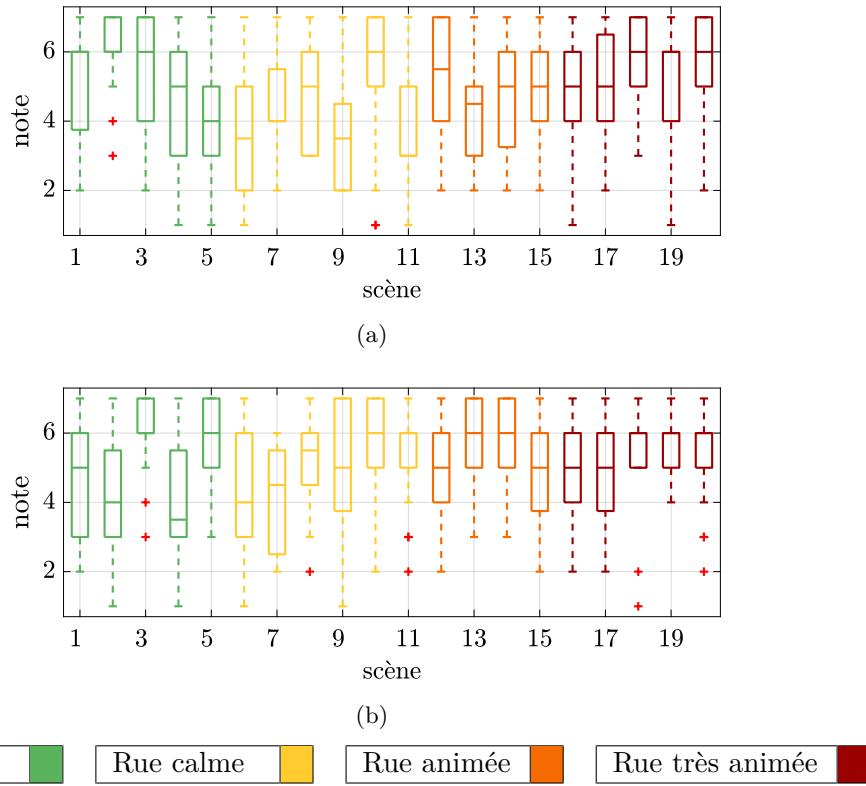


FIGURE 1.11 – Boites à moustaches pour les scènes enregistrées (a) et pour les scènes répliquées (b) classé selon leur ambiance sonore.

L'ensemble des tests de Student mené sur les 20 couples de scènes révèlent des valeurs  $p$  supérieures au seuil de signification  $\alpha$  de 5 %. L'hypothèse  $H_0$  n'est donc pas rejeté sur l'ensemble des scènes : le réalisme des scènes du type *répliquée* est alors considéré comme similaire à celui du type *enregistrée*.

#### 1.5.6.3 Effets des auditeurs sur l'évaluation des scènes

Pour aller plus loin, une analyse de variance (abrégée ANOVA pour *ANalyse Of VAriance* en anglais) est réalisée afin de déterminer l'influence de chaque auditeur sur l'évaluation des scènes. En effet, selon l'auditeur, l'échelle des notes émises peut varier (un peut noter sur l'ensemble de l'échelle, d'autres peuvent noter une échelle réduite), influençant l'interprétation des résultats.

C'est ainsi une ANOVA à deux facteurs avec interaction qui est considérée (le type de scènes (*enregistré*, *répliqué*) et auditeur (50 auditeurs en tout)). De la même manière que le test de Student, l'ANOVA est un outil statistique qui permet de comparer des moyennes d'échantillons et d'étudier l'effet des variables qualitatives (ou facteurs), pouvant prendre plusieurs valeurs (ou niveaux), sur une variable quantitative. Pour cela, plusieurs statistiques sont calculées (somme des carrés des écarts, le degrés de liberté du facteurs, la variance, la statistique de Fischer  $F$ ). De ces indices, la valeur  $p$  établie, là encore, la probabilité d'obtenir une valeur limite du test si  $H_0$  est vraie par rapport à la valeur seuil  $\alpha$ . Les définitions de ces indices sont résumées en

## 1.5. CORPUS DE SCÈNES GRAFIC

---

annexe ???. Les résultats sont résumés dans le Tableau 1.9.

Source	SCE	DDL	variance	F	p-valeur
<b>auditeur</b>	687,93	49	14,03	7,89	<1e-4
<b>type</b>	3,61	1	3,61	1,82	0,18
<b>auditeur/type</b>	93,29	49	0,90	0,55	0,55
<b>erreur</b>	1780,42	899	1,98		
<b>total</b>	2572	998			

TABLEAU 1.9 – Résultat de l'ANOVA avec interaction avec les facteurs *type* et *auditeur*

Le facteur *auditeur* a une influence significative (valeur  $p < \alpha$ ) révélant que les auditeurs n'ont pas les même échelle d'évaluation. L'influence du facteur *type* reste toujours non significative. Son interaction avec le facteur *auditeur* est non-significatif également. Le phénomène d'interaction traduit l'influence des différents niveaux d'un facteur sur l'autre facteur. Ici, puisque l'interaction entre le facteur *auditeur* et *type* est non-significative, le choix du juge n'influe pas sur la similarité perçu des scènes enregistrées et répliquées, même si entre chaque juge des dissimilarités existent.

### 1.5.6.4 Effets de l'ambiance sonore

Une seconde ANOVA à deux facteurs avec intéraction est générée avec pour facteur le *type* (*enregistrée, répliquée*) et l'*ambiance sonore* (*parc, rue calme, rue bruyante, rue très bruyante*). afin de déterminer si la perception du réalisme est différent selon l'*ambiance sonore*. Les résultats de l'ANOVA sont résumés dans le Tableau 1.10

Source	SCE	DDL	variance	F	p-valeur
<b>type</b>	5,72	1	5,72	2,28	0,13
<b>ambiance</b>	42,65	3	14,21	5,66	8,00e-4
<b>type/ambiance</b>	36,83	3	12,27	4,89	2,20e-3
<b>erreur</b>	2488,49	991	2,55		
<b>total</b>	2572	998			

TABLEAU 1.10 – Résultat de l'ANOVA avec interaction avec les facteurs *type* et *ambiance*

L'impact du facteur *type* est toujours non significatif. Celui du facteur *ambiance* et l'interaction entre les deux facteurs sont toutefois significatifs ( $p < \alpha$ ). L'influence principale de l'*ambiance* signifie qu'il y a une distinction entre les distributions des notes entre les différents ambiances sonores. Le phénomène d'interaction traduit alors que selon l'*ambiance sonore* la perception du type de la scène varie.

Pour visualiser ce phénomène d'interaction entre le type de scènes et l'*ambiance sonore*, on trace l'évolution de la note moyenne dans chaque cas (Figure 1.12). On observe que selon l'*ambiance sonore*, la note de réalisme des scènes répliquées peut être inférieure ou supérieure par rapport aux scènes enregistrée. Cette évolution traduit une interaction croisée. Leur origine est toutefois difficile à estimer. Il est possible, pour mieux comprendre ces différences, de s'intéresser aux commentaires laissés par les auditeurs sur certaines scènes. Ces derniers relèvent notamment

des sons trop forts ou qui s'inscrivent mal dans les scènes (bruit de rue dans la scène, oiseaux dans la scène). Enfin certains extraits de voix ne sont pas suffisamment réalistes. En effet, lors de la phase d'écoutes des enregistrement, on remarque que les voix perçues, en dehors d'un brouhaha de foule, sont le plus souvent des bribes de conversations de personnes entre elles ou au téléphone. Malheureusement, il n'a pas été possible de trouver des bases de données de conversation suffisamment réalistes pour être inclus dans les scènes. De nombreuses bases de données se concentrent, par exemple, sur la lecture de textes récités ce qui ne permet pas d'atteindre le réalisme souhaité.

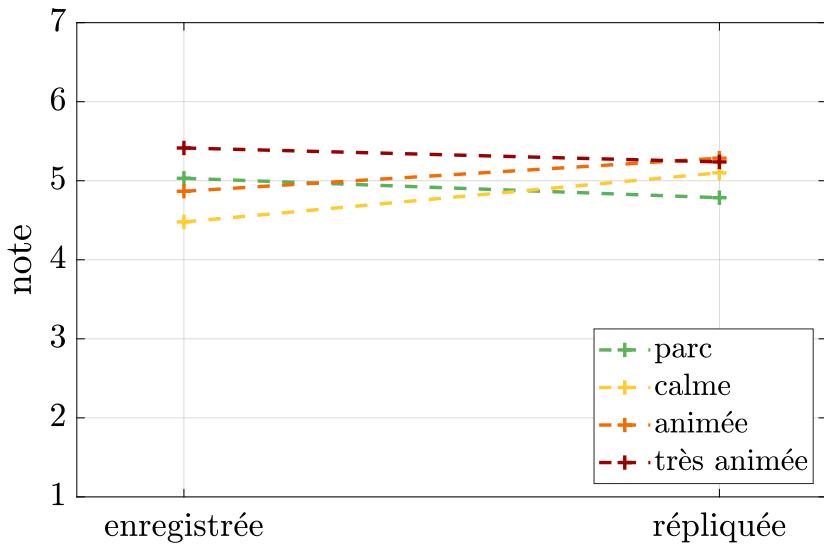


FIGURE 1.12 – Évolution de la note moyenne du réalisme par ambiance et selon le type.

En conclusion, même si les moyennes globales et les distributions entre les scènes enregistrées et repliquées sont similaires, des disparités existent selon les auditeurs ou les ambiances sonores sans que celle-ci remettent en cause les similarités entre les deux types.

## 1.5. CORPUS DE SCÈNES GRAFIC

# Bibliographie

- [Adams *et al.*, 2008] ADAMS, M. D., BRUCE, N. S., DAVIES, W. J., CAIN, R., JENNINGS, P., CARLYLE, A., CUSACK, P., HUME, K. et PLACK, C. (2008). Soundwalking as a methodology for understanding soundscapes. In *Institute of Acoustics*, volume 30, Reading, U.K.
- [Aumond *et al.*, 2017] AUMOND, P., CAN, A., DE COENSEL, B., BOTTELOOREN, D., RIBEIRO, C. et LAVANDIER, C. (2017). Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(3):430–443.
- [Bruce *et al.*, 2009] BRUCE, N., DAVIES, W. et ADAMS, M. (2009). Development of a soundscape simulator tool. In *Proceedings of the INTERNOISE Congress*, Ottawa, Canada.
- [Can et Gauvreau, 2015] CAN, A. et GAUVREAU, B. (2015). 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.
- [Embleton *et al.*, 1976] EMBLETON, T. F. W., PIERCY, J. E. et OLSON, N. (1976). Outdoor sound propagation over ground of finite impedance. *The Journal of the Acoustical Society of America*, 59(2):267–277.
- [Fitzgerald, 2010] FITZGERALD, D. (2010). Harmonic/Percussive Separation Using Median Filtering. *Conference papers*.
- [Forssén *et al.*, 2009] FORSSÉN, J., KACZMAREK, T., ALVARSSON, J., LUNDÉN, P. et NILSSON, M. E. (2009). Auralization of traffic noise within the listen project—preliminary results for passenger car pass-by. *Euronoise 2009*.
- [Guastavino *et al.*, 2005] GUASTAVINO, C., KATZ, B. F. G., POLACK, J., LEVITIN, D. J. et DUBOIS, D. (2005). Ecological validity of soundscape reproduction. *Acta Acustica united with Acustica*, 91(2):333–341.
- [Guillaume *et al.*, 2014] GUILLAUME, G., GAUVREAU, B. et L’HERMITE, P. (2014). Estimation expérimentale des propriétés acoustiques des surfaces végétalisées : influences de la variabilité spatiale et de la configuration de mesure. In *Congrès Français d’Acoustique 2014*, page 6p, Poitiers, France.
- [Guillaume *et al.*, 2015] GUILLAUME, G., GAUVREAU, B. et L’HERMITE, P. (2015). Numerical study of the impact of vegetation coverings on sound levels and time decays in a canyon street model. *Science of The Total Environment*, 502:22–30.

## BIBLIOGRAPHIE

---

- [Lê et Husson, 2008] Lê, S. et HUSSON, F. (2008). SensoMineR : A package for sensory data analysis (PDF Download Available). *Journal of Sensory Studies*, pages 14 – 25.
- [Lafay *et al.*, 2014] LAFAY, G., ROSSIGNOL, M., MISDARIIS, N., LAGRANGE, M. et PETIOT, J.-F. (2014). A New Experimental Approach for Urban Soundscape Characterization Based on Sound Manipulation : A Pilot Study. In *International Symposium on Musical Acoustics*, Le Mans, France.
- [Leroy *et al.*, 2010] LEROY, O., GAUVREAU, B., JUNKER, F., DE ROCQUIGNY, E. et BERENGIER, M. (2010). Uncertainty assessment for outdoor sound propagation. In *20th International Congress on Acoustics, ICA 2010*, page 7p, France.
- [Lihoreau *et al.*, 2006] LIHOREAU, B., GAUVREAU, B., BÉRENGIER, M., BLANC-BENON, P. et CALMET, I. (2006). Outdoor sound propagation modeling in realistic environments : Application of coupled parabolic and atmospheric models. *The Journal of the Acoustical Society of America*, 120(1):110–119.
- [Misra *et al.*, 2007] MISRA, A., WANG, G. et COOK, P. (2007). Musical Tapestry : Re-composing Natural Sounds†. *Journal of New Music Research*, 36(4):241–250.
- [Pagès et Périnel, 2007] PAGÈS, J. et PÉRINEL, E. (2007). Blocs incomplets équilibrés versus plans optimaux. *Journal de la Société Française de Statistique*, 148(2):99–112.
- [Picaut *et al.*, 2005] PICAUT, J., LE POLLÈS, T., L'HERMITE, P. et GARY, V. (2005). Experimental study of sound propagation in a street. *Applied Acoustics*, 66(2):149–173.
- [Raimbault *et al.*, 2003] RAIMBAULT, M., LAVANDIER, C. et BÉRENGIER, M. (2003). Ambient sound assessment of urban environments : field studies in two French cities. *Applied Acoustics*, 64(12):1241–1256.
- [Rossignol *et al.*, 2015] ROSSIGNOL, M., LAFAY, G., LAGRANGE, M. et MISDARIIS, N. (2015). SimScene : a web-based acoustic scenes simulator. In *1st Web Audio Conference (WAC)*.
- [Ruxton, 2006] RUXTON, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann-whitney u test. *Behavioral Ecology*, 17(4):688–690.
- [Salamon *et al.*, 2014] SALAMON, J., JACOBY, C. et BELLO, J. P. (2014). A dataset and taxonomy for urban sound research. In *22st ACM International Conference on Multimedia (ACM-MM 14)*, Orlando, FL, USA.
- [Schafer, 1993] SCHAFER, R. M. (1993). *The soundscape : Our sonic environment and the tuning of the world*. Simon and Schuster.
- [Stienen et Vorländer, 2015] STIENEN, J. et VORLÄNDER, M. (2015). Auralization of urban environments—concepts towards new applications. In *Proc. EuroNoise*, Maastricht, Pays-Bas.
- [Vorländer, 2007] VORLÄNDER, M. (2007). *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media.