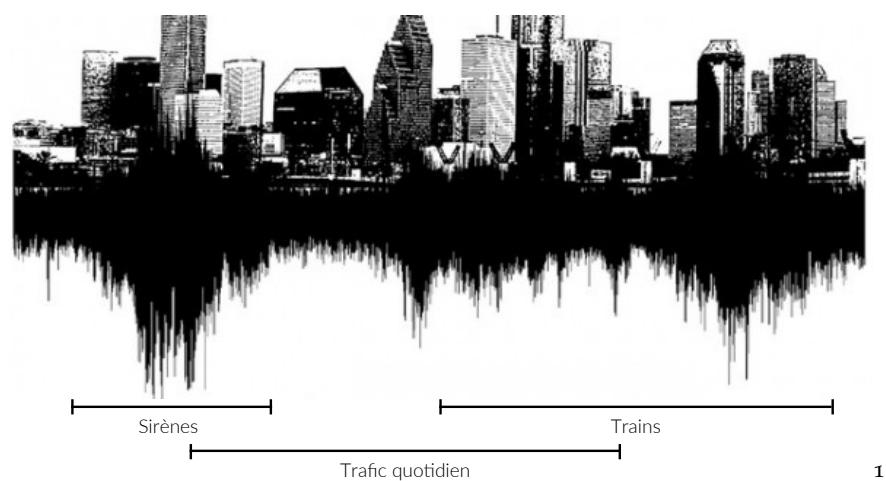


# SIMULATION DE SCÈNES SONORES ENVIRONNEMENTALES

Application à l'analyse sensorielle et à l'analyse automatique



GRÉGOIRE LAFAY

Doctorant

Équipe Analyse et Décision en Traitement du Signal et des Images (ADTSI)  
Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN)  
École Centrale de Nantes, France

8 Décembre 2016

<sup>1</sup> d'après <http://www.joesdaily.com/art-design/your-words-turned-into-art/>



*À ma famille.*



## RÉSUMÉ

---

La présente thèse traite de l'analyse de scènes extraites d'environnements sonores, résultat auditif du mélange de sources émettrices distinctes et concomitantes. Ouvrant le champ des sources et des recherches possibles au-delà des domaines plus spécifiques que sont la parole ou la musique, l'environnement sonore est un objet complexe. Son analyse, le processus par lequel le sujet lui donne sens, porte à la fois sur les données perçues et sur le contexte de perception de ces données.

Tant dans le domaine de la perception que de l'apprentissage machine, toute expérience suppose un contrôle fin de l'expérimentateur sur les stimuli proposés. Néanmoins, la nature de l'environnement sonore nécessite de se placer dans un cadre écologique, c'est à dire de recourir à des données réelles, enregistrées, plutôt qu'à des stimuli de synthèse.

Conscient de cette problématique, nous proposons un modèle permettant de simuler, à partir d'enregistrements de sons isolés, des scènes sonores dont nous maîtrisons les propriétés structurelles – intensité, densité et diversité des sources. Appuyé sur les connaissances disponibles sur le système auditif humain, le modèle envisage la scène sonore comme un objet composite, une somme de sons sources.

Nous investissons à l'aide de cet outil deux champs d'application. Le premier concerne la perception, et la notion d'agrément perçu dans des environnements urbains. L'usage de données simulées nous permet d'apprécier finement l'impact de chaque source sonore sur celui-ci. Le deuxième concerne la détection automatique d'événements sonores et propose une méthodologie d'évaluation des algorithmes mettant à l'épreuve leurs capacités de généralisation.

## ABSTRACT

---

This thesis deals with environmental scene analysis, the auditory result of mixing separate but concurrent emitting sources. The sound environment is a complex object, which opens the field of possible research beyond the specific areas that are speech or music. For a person to make sense of its sonic environment, the involved process relies on both the perceived data and its context.

For each experiment, one must be, as much as possible, in control of the evaluated stimuli, whether the field of investigation is perception or machine learning. Nevertheless, the sound environment needs to be studied in an ecological framework, using real recordings of sounds as stimuli rather than synthetic pure tones.

We therefore propose a model of sound scenes allowing us to simulate complex sound environments from isolated sound recordings. The high level structural properties of the simulated scenes – such as the type of sources, their sound levels or the event density – are set by the experimenter. Based on knowledge of the human auditory system, the model abstracts the sound environment as a composite object, a sum of sound sources.

The usefulness of the proposed model is assessed on two areas of investigation. The first is related to the soundscape perception issue, where the model is used to propose an innovative experimental protocol to study pleasantness perception of urban soundscape. The second tackles the major issue of evaluation in machine listening, for which we consider simulated data in order to powerfully assess the generalization capacities of automatic sound event detection systems.

## PUBLICATIONS

---

### Revues

- Benetos, Emmanouil, Grégoire Lafay, and Mathieu Lagrange (2016). "Polyphonic Sound Event Tracking using Linear Dynamical Systems." In: *IEEE/ACM Transactions on audio, speech and language processing, Special issue on Sound Scene and Event Analysis*, (accepted).
- Lafay, Grégoire, Mathieu Lagrange, Emmanouil Benetos, Mathias Rossignol, and Axel Roebel (2016a). "A Morphological Model for Simulating Acoustic Scenes and Its Application to Sound Event Detection." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10, pp. 1854–1864.
- Lafay, Grégoire, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol (2016b). "Semantic browsing of sound databases without keywords." In: *Journal of the Audio Engineering Society* 64.9, pp. 628–635.
- Lagrange, Mathieu, Grégoire Lafay, Boris Defreville, and Jean-Julien Aucouturier (2015). "The bag-of-frames approach : a not so sufficient model for urban soundscapes." In: *The Journal of the Acoustical Society of America, express letter* 138.5, pp. 487–492.

### Conférences

- Benetos, Emmanouil, Grégoire Lafay, Mathieu Lagrange, and Mark D. Plumbley (2016). "Detection of overlapping acoustic events using a temporally-constrained probabilistic model." In: *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE. Shanghai, China, pp. 6450–6454.
- Lafay, Grégoire, Mathias Rossignol, Nicolas Misdariis, Mathieu Lagrange, and Jean-François Petiot (2014). "A new experimental approach for urban soundscape characterization based on sound manipulation : A pilot study." In: *Proceedings of the International Symposium on Musical Acoustics (ISMA)*. SFA. Le Mans, France.
- Rossignol, Mathias, Mathieu Lagrange, Grégoire Lafay, and Emmanouil Benetos (2015a). "Alternate level clustering for drum transcription." In: *Proceedings of the European Signal Processing Conference (EUSIPCO)*. IEEE. Nice, France, pp. 2023–2027.
- Rossignol, Mathias, Grégoire Lafay, Mathieu Lagrange, and Nicolas Misdariis (2015b). "SimScene : a web-based acoustic scenes simulator." In: *Proceedings of the Web Audio Conference (WAC)*. IRCAM. Paris, France.

**Revue en lien avec la thèse, mais dont les résultats ne sont pas présentés dans ce document**

Lostanlen, Vincent, Grégoire Lafay, Joakim Anden, and Mathieu Lagrange (2016). "Auditory Scene Similarity Retrieval and Classification with Relevance-based Quantization of Scattering Features." In: *IEEE/ACM Transactions on audio, speech and language processing, Special issue on Sound Scene and Event Analysis, (AQ)*.

## REMERCIEMENTS

---

Je remercie tout d'abord très chaleureusement les membres du jury, Bertrand David et Catherine Lavandier, rapporteurs, Jean-julien Autourier, examinateur, et Alain de Cheveigné, examinateur et président du jury, pour l'intérêt qu'ils ont porté à mes travaux, et leurs retours précieux pour la finalisation du présent document.

Pour la grande qualité de leur encadrement, je remercie non moins chaleureusement mes encadrants : Jean-François Lafay, sans qui cette thèse n'aurait pas pu démarrer, Jérôme Idier, qui a accepté de diriger cette thèse, et qui, en outre, a pris le temps d'expliquer au jeune étudiant que j'étais les réalités du monde de la recherche, Jean-François Petiot, qui m'a apporté toute son expertise dans le domaine de l'analyse sensorielle. Je lui suis, par ailleurs, infiniment reconnaissant de la confiance qu'il m'a accordée en m'offrant de monter, puis de donner, à Central Nantes, des cours ayant trait à mes sujets de recherche. L'expérience, qui a débuté il y a trois ans, et qui se poursuit encore, aura été enrichissante, et plus que cela sans doute.

Il est enfin des rencontres qui changent le cours d'une vie. Celle avec Mathieu Lagrange a été décisive, tant du point de vue scientifique que du point de vue personnel. Pour la liberté qu'il m'a laissée dans la façon d'aborder mon sujet, non sans me canaliser quand, par enthousiasme, je m'égarais, pour son investissement et sa motivation sans faille, pour la richesse de nos débats (tant de visions et d'intuitions partagées), pour m'avoir initié au *Machine-Listening* et, d'une manière générale, pour tout ce qu'il m'a appris, je lui témoigne ici ma profonde gratitude, et lui dit, je le peux bien aujourd'hui, toute mon affection.

Pour leur accueil et nos discussions autour de thèmes scientifiques connexes, je remercie les membres de l'équipe Analyse et Décision en Traitement du Signal du laboratoire IRCCyN, Marie-Françoise Lucas, Saïd Moussaoui, Sébastien Bourguignon et Éric Le Carpentier.

Pour tous nos échanges durant ces trois années, je remercie mes collègues doctorants, en particulier Robin Tournemenne, Corentin Friedrich, Mohamed-Anis Dhuieb, Maxime Sirbu et Jean-Remy Gloaguen.

Pour notre collaboration autour de la transformée en *scattering*, je remercie tout spécialement Vincent Lostanlen et Joakim Andén.

Pour avoir permis d'inscrire mes travaux sur l'évaluation des algorithmes en *Machine-Listening* dans le cadre du challenge international DCASE, et pour nos multiples collaborations scientifiques, je remercie vivement Emmanouil Benetos.

Pour m'avoir enseigné les bases de l'analyse des signaux audio, et également pour m'avoir donner le goût de la pluridisciplinarité, je remercie l'équipe pédagogique du master ATIAM. Notamment, les membres de l'équipe Perception et Design Sonore de l'IRCAM qui m'ont accueilli lors de mon stage de Master. Un salut tout particulier à Nicolas Misdariis, pour son excellent encadrement, et pour avoir été celui qui m'a fait découvrir la psychologie cognitive.

Pour être le premier à avoir élaboré l'outil de simulation de scènes sonores, pour son expertise scientifique et son expertise en programmation, pour ses conseils avisés et son aide de tous les instants, mais également, pour son amitié, je remercie chaleureusement Mathias Rossignol. Cela fait maintenant quatre ans que nous collaborons, et je suis heureux que l'expérience se poursuive.

Pour leur motivation, leur bonne humeur et leur intérêt, je tiens à remercier les élèves de l'École Centrale de Nantes que j'ai eu le plaisir d'encadrer au cours de nombreux projets étudiants, Simon Dubois, Adrien Urso, Nicolas Dany, Jean-Baptiste Kaiser, Thomas Forgue, Lucas Sanchez, César Lacroix, Théo Cordoliani, Hugo Pagnier, Baptiste Parquier, Oriane Cosson et Pablo Bonachela-Guhmann. Ces projets ont permis d'explorer plusieurs pistes de recherches intéressantes, et certains ont par ailleurs directement participé à la mise en place des protocoles expérimentaux présentés dans cette thèse.

Pour avoir accepté d'être mes « cobayes », je remercie grandement les très nombreux étudiants de l'École Centrale de Nantes qui ont participé aux non moins nombreuses expériences perceptives réalisées dans la cadre de cette thèse. Inutile de dire que ce travail n'aurait pas été possible sans leur participation volontaire et motivée.

Pour leur soutien inconditionnel, j'adresse encore mes remerciements à ma famille, mes grands parents, ma sœur et bien sûr mes parents, à qui je dois beaucoup.

Enfin, j'embrasse Camille qui fut si loin et à la fois si proche durant ces trois années.

## TABLE DES MATIÈRES

---

<b>I PRÉAMBULE</b>	<b>1</b>
<b>1 PRÉAMBULE</b>	<b>3</b>
<b>1.1 Introduction Générale</b>	<b>3</b>
<b>1.1.1 L'environnement sonore</b>	<b>3</b>
<b>1.1.2 Pourquoi modéliser une scène sonore ?</b>	<b>5</b>
<b>1.2 Motivations des cas d'études</b>	<b>5</b>
<b>1.2.1 Un cadre applicatif pluridisciplinaire</b>	<b>5</b>
<b>1.2.2 La perception des paysages sonores urbains</b>	<b>6</b>
<b>1.2.3 La détection automatique d'événements sonores</b>	<b>9</b>
<b>1.3 Plan</b>	<b>10</b>
<b>II UN MODÈLE MORPHOLOGIQUE</b>	<b>11</b>
<b>2 ÉTAT DE L'ART</b>	<b>13</b>
<b>2.1 Introduction</b>	<b>13</b>
<b>2.2 Perception et Cognition</b>	<b>14</b>
<b>2.2.1 Définitions</b>	<b>14</b>
<b>2.2.2 Théorie classique de la cognition</b>	<b>14</b>
<b>2.2.3 Une approche ancrée de la cognition</b>	<b>15</b>
<b>2.2.4 Une approche écologique de la cognition</b>	<b>17</b>
<b>2.2.5 Discussion</b>	<b>18</b>
<b>2.3 Structure catégorielle des représentations mentales</b>	<b>20</b>
<b>2.3.1 La notion de catégorie</b>	<b>20</b>
<b>2.3.2 Le processus de catégorisation</b>	<b>22</b>
<b>2.3.3 Organisation de la structure catégorielle</b>	<b>23</b>
<b>2.3.4 Théories de la catégorisation</b>	<b>27</b>
<b>2.3.5 Catégorisation et contexte sensoriel</b>	<b>31</b>
<b>2.3.6 Similarité et catégorisation</b>	<b>32</b>
<b>2.3.7 Discussion</b>	<b>32</b>
<b>2.4 L'étude psychologique du système auditif</b>	<b>34</b>
<b>2.4.1 La psychoacoustique</b>	<b>34</b>
<b>2.4.2 La psychologie cognitive</b>	<b>35</b>
<b>2.4.3 Paradigme de la psychologie cognitive</b>	<b>36</b>
<b>2.4.4 Reproduire l'environnement sonore</b>	<b>37</b>
<b>2.4.5 Le Soundwalk</b>	<b>38</b>
<b>2.4.6 Discussion</b>	<b>39</b>
<b>2.5 Une vue générale du système auditif</b>	<b>40</b>
<b>2.5.1 La chaîne de traitement</b>	<b>40</b>
<b>2.5.2 Processus <i>Bottom-up</i> et processus <i>Top-down</i></b>	<b>42</b>
<b>2.5.3 Discussion</b>	<b>43</b>
<b>2.6 Analyse de scènes acoustiques</b>	<b>44</b>
<b>2.6.1 Définition</b>	<b>44</b>
<b>2.6.2 Une approche psychoacoustique</b>	<b>45</b>

2.6.3	Régularités et processus primitifs	45
2.6.4	Perception de la forme	46
2.6.5	Flux auditif et stratégie de groupement	49
2.6.6	L'approche par les neurosciences	51
2.6.7	Attention et saillance	52
2.6.8	Discussion	53
2.7	L'étude des paysages sonores	54
2.7.1	La notion de paysage sonore	54
2.7.2	Approches catégorielle et dimensionnelle	56
2.7.3	Descripteurs perceptifs des paysages sonores	64
2.7.4	Catégoriser les sources et paysages sonores	70
2.7.5	Classifier les sources et environnements sonores	74
2.7.6	Contributions des différentes sources sonores	76
2.7.7	Discussion	78
2.8	Événements et textures sonores	79
2.8.1	Définition	79
2.8.2	Percevoir les textures	80
2.8.3	Discussion	82
3	MODÈLE ET SIMULATION	85
3.1	Introduction	85
3.2	Fondement perceptif du modèle morphologique	85
3.2.1	L'unité : la source sonore	85
3.2.2	L'objet : la séquence sonore	86
3.2.3	Une typologie source-action	86
3.2.4	Événements et textures	88
3.3	Description du modèle morphologique	90
3.3.1	Classe et collection de samples	90
3.3.2	Séquences de samples	91
3.3.3	Paramètres	91
3.3.4	Formalisation du modèle	93
3.4	Un modèle pour la simulation	94
3.4.1	Choix de conception	94
3.4.2	Simulation et perception des paysages sonores	94
3.4.3	Simulation et détection automatique d'événe- ments sonores	99
3.5	Conclusion	100
III	UTILISATION PRATIQUE DE LA SIMULATION	101
4	DONNÉES SIMULÉES EN ANALYSE SENSORIELLE	103
4.1	Introduction	103
4.1.1	Protocole expérimental basé sur la simulation	104
4.2	Agrément perçu et composition sémantique	110
4.2.1	Objectif	110
4.2.2	Banque de données de sons isolés	111
4.2.3	Typologie des sources sonores	111
4.2.4	Acquisition des sons isolés	112

4.2.5	Planification expérimentale	114
4.2.6	Données et méthodes d'analyses	117
4.2.7	Validité écologique de l'expérience	121
4.2.8	Vérification de l'agrément des scènes simulées	122
4.2.9	Descripteurs structurels	123
4.2.10	Descripteurs structurels et agrément perçu	125
4.2.11	Descripteurs sémantiques	128
4.2.12	Espaces de représentation induits par les descripteurs sémantiques	131
4.2.13	Marqueurs sonores et agrément perçu	133
4.2.14	Discussions	135
4.3	Modification de la composition sémantique	140
4.3.1	Objectif	140
4.3.2	Planification expérimentale	140
4.3.3	Données et méthodes d'analyses	142
4.3.4	Détection de valeurs extrêmes	143
4.3.5	Influence des descripteurs structurels des scènes avec marqueurs sur l'agrément perçu : une analyse comparative	143
4.3.6	Influence de la présence des marqueurs sur l'agrément perçu	146
4.3.7	Influence des descripteurs structurels des scènes sans marqueurs sur l'agrément perçu	147
4.3.8	Discussions	148
4.4	Composition sémantique et catégorisation	150
4.4.1	Objectif de l'expérience	150
4.4.2	Planification expérimentale	151
4.4.3	Données et méthodes d'analyses	152
4.4.4	Stratégie de catégorisation	154
4.4.5	Analyse lexicale des descriptions	155
4.4.6	Descripteurs sémantiques subjectifs	158
4.4.7	Descripteurs sémantiques objectifs	161
4.4.8	Descripteurs perceptifs et structurels	163
4.4.9	Discussions	165
4.5	Conclusion	166
5	DONNÉES SIMULÉES EN ANALYSE AUTOMATIQUE	169
5.1	Introduction	169
5.2	Le challenge DCASE	169
5.2.1	Présentation	169
5.2.2	Évaluation	170
5.2.3	Métrique	171
5.3	Application au challenge DCASE 2013	173
5.3.1	Objectif	173
5.3.2	Génération des corpus	174
5.3.3	Métrique	180
5.3.4	Données et analyses	181

5.3.5	Système de détection	183
5.3.6	Résultats	184
5.3.7	Discussion	190
5.4	Application au challenge DCASE 2016	191
5.4.1	Objectifs	191
5.4.2	Génération des corpus	191
5.4.3	Métrique	194
5.4.4	Données et analyses	194
5.4.5	Systèmes de détection	195
5.4.6	Résultats	196
5.4.7	Discussion	203
5.5	Conclusion	203
<b>IV CONCLUSIONS ET PERSPECTIVES</b>		<b>205</b>
6	<b>CONCLUSIONS ET PERSPECTIVES</b>	<b>207</b>
6.1	Analyse sensorielle	207
6.1.1	Agrément des paysages sonores urbains	207
6.1.2	Simulation et cognition	207
6.1.3	Perspectives	208
6.2	Analyse automatique	209
6.2.1	Détection automatique d'événements sonores	209
6.2.2	Perspectives	209
6.3	Contributions	210
6.3.1	Valorisation scientifique	210
6.3.2	Programmes et banques de données	211
<b>V APPENDICES</b>		<b>213</b>
A	<b>OUTILS D'ANALYSE STATISTIQUE UNI-VARIÉE</b>	<b>215</b>
A.1	Tests paramétriques à deux échantillons	215
A.2	Tests paramétriques à plus de deux échantillons	216
A.3	Comparaisons multiples	219
A.4	Mesures de corrélation paramétriques	220
B	<b>EXPÉRIENCE ANNEXE : PÉRIODE D'ATTENTION</b>	<b>223</b>
B.1	Objectif de l'expérience	223
B.2	Banque de données	223
B.3	Planification expérimentale	224
B.4	Méthodologie et outils statistiques	225
B.5	Résultats	225
C	<b>EXPÉRIENCE ANNEXE : CONGRUENCE</b>	<b>227</b>
C.1	Objectif de l'expérience	227
C.2	Banque de données	227
C.3	Planification expérimentale	228
C.4	Méthodologie et outils statistiques	229
C.5	Résultats	229
<b>BIBLIOGRAPHIE</b>		<b>231</b>

## TABLE DES FIGURES

---

FIGURE 1	Théories de la cognition.	16
FIGURE 2	Processus cognitifs et perceptifs.	18
FIGURE 3	Les trois niveaux d'abstraction de l'axe vertical de la structure catégorielle.	26
FIGURE 4	Prototype et caricature.	30
FIGURE 5	Paradigme de la psychologie cognitive	38
FIGURE 6	Principaux processus de traitement de l'information auditive et leurs interactions.	40
FIGURE 7	Le phénomène de bistabilité : l'illusion du canard-lapin.	44
FIGURE 8	Groupement séquentiel : proximité temporelle.	46
FIGURE 9	Groupement séquentiel : proximité fréquentielle.	47
FIGURE 10	Groupement simultané : régularité harmonique.	48
FIGURE 11	Groupement ancien-plus-nouveau.	49
FIGURE 12	Compétition entre groupement séquentiel et groupement simultané.	50
FIGURE 13	Tâche de description et tâche de tri ou de catégorisation.	60
FIGURE 14	Les dimensions de calme et de dynamisme permettant de caractériser l'environnement sonore urbain.	69
FIGURE 15	Catégorisation des paysages sonores urbains.	74
FIGURE 16	Taxonomie des sources sonores urbaines.	75
FIGURE 17	Information potentielle contenue dans les séquences d'événements, les textures, et le bruit.	80
FIGURE 18	Planification expérimentale de l'expérience de discrimination de textures sonores et d'exemplaires de textures sonores	82
FIGURE 19	Taxonomie des sources sonores urbaines suivant la nomenclature source-action.	84
FIGURE 20	Organisation hiérarchique de la banque de sons isolés utilisée pour la simulation.	91
FIGURE 21	Représentation schématisée des pistes du modèle de scènes sonores.	92
FIGURE 22	Relation entre l'analyse psycholinguistique et la simulation.	98
FIGURE 23	Etape de processus de simulation pour l'analyse sensorielle	105
FIGURE 24	Paradigme du protocole expérimental basé sur la simulation.	107

FIGURE 25	L'interface de sélection aveugle de l'outil de simulation <i>Simscape</i> . <a href="#">108</a>
FIGURE 26	L'outil de simulation <i>Simscape</i> . <a href="#">109</a>
FIGURE 27	Planification expérimentale des expériences de simulation et d'évaluation de l'agrément <a href="#">111</a>
FIGURE 28	Taxonomies des classes de sons utilisées pour la simulation des environnements sonores urbains. <a href="#">113</a>
FIGURE 29	Dispersions des notes données par les sujets lors de l'expérience 1.b moyennées suivant les sujets ( $\mathcal{A}_{\text{scene}}$ : a), et suivant les scènes ( $\mathcal{A}_{\text{subject}}$ : b), en fonction du type de scènes (i ou ni). <a href="#">123</a>
FIGURE 30	Dispersions des descripteurs structurels de niveaux sonores L (a, d), L(E) (b, e) et L(T) (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu $\mathcal{A}_{\text{scene}}$ de l'expérience 1.b (d, e, f). <a href="#">125</a>
FIGURE 31	Dispersions des descripteurs structurels de densité D (a, c) et D(E) (b, d), en fonction du type de scènes (a, b) et de l'agrément perçu $\mathcal{A}_{\text{scene}}$ de l'expérience 1.b (c, d). <a href="#">126</a>
FIGURE 32	Moyenne et écart type de la diversité des classes utilisées en considérant l'ensemble des classes (DIV), les classes d'événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i- et ni-scènes ainsi que les différents niveaux d'abstraction. <a href="#">127</a>
FIGURE 33	Pourcentage de scènes simulées comportant une classe de son particulière. <a href="#">130</a>
FIGURE 34	P@5 obtenues en considérant la matrice de dissimilarité résultant des distances par paires de Hamming calculées entre les vecteurs des descripteurs sémantiques des scènes. <a href="#">133</a>
FIGURE 35	Dispersions des descripteurs structurels de densité relatifs à la présence des marqueurs $D_m$ (a, c) et $D(E)_m$ (b, d), en fonction du type de scènes (a, b) et de l'agrément perçu $\mathcal{A}_{\text{scene}}$ de l'expérience 1.b (c, d). <a href="#">136</a>
FIGURE 36	Dispersions des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs $L_m$ (a, d), $L(E)_m$ (b, e) et $L(T)_m$ (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu $\mathcal{A}_{\text{scene}}$ de l'expérience 1.b (d, e, f). <a href="#">137</a>

- FIGURE 37** Dispersions des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs  $L_b$  (a, d),  $L(E)_b$  (b, e) et  $L(T)_b$  (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{scène}$  de l'expérience 1.b (d, e, f). [138](#)
- FIGURE 38** Dispersions des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs  $L_m - L_b$  (a, d),  $L(E)_m - L(E)_b$  (b, e) et  $L(T)_m - L(T)_b$  (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{scène}$  de l'expérience 1.b (d, e, f). [139](#)
- FIGURE 39** Dispersions des notes données par les sujets lors de l'expérience 2 aux i/am-scènes (vert) et ni/am-scènes (rouge). [144](#)
- FIGURE 40** Dispersion des notes données par les sujets lors de l'expérience 2 moyennées suivant les sujets ( $\mathcal{A}_{sujet}$  : a), suivant les scènes ( $\mathcal{A}_{scène}$  : b et c), en fonction du type de scènes (a et b) et des  $\mathcal{A}_{scène}$  relevés à l'expérience 1.b. [144](#)
- FIGURE 41** Moyennes et écarts types de la diversité des classes utilisées en considérant l'ensemble des classes (DIV), les classes d'événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i/sm- et ni/sm-scènes ainsi que les différents niveaux d'abstraction. [148](#)
- FIGURE 42** Dispersions des descripteurs structurels de densité D (a, c) et  $D(E)$  (b, d), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b) et de l'agrément perçu  $\mathcal{A}_{scène}$  de l'expérience 2 (c, d). [149](#)
- FIGURE 43** Dispersions des descripteurs structurels de niveaux sonores  $L$  (a, d),  $L(E)$  (b, e) et  $L(T)$  (c, f), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{scène}$  de l'expérience 2 (d, e, f). [150](#)
- FIGURE 44** Partitions P établies suivant la classification ascendante hiérarchique. [153](#)
- FIGURE 45** Pourcentage de scènes étant décrites par un label sémantique subjectif donné (a, c), un label de qualité subjectif donné (b, d), en considérant l'ensemble des scènes (a, b) ou les i- et ni-scènes séparément (c, d). [157](#)
- FIGURE 46** Processus de génération des corpus de scènes simulées utilisés dans l'évaluation du challenge DCASE 2013. [175](#)

FIGURE 47	Distributions des notes de réalisme $\mathcal{R}_{\text{ sujet }}$ pour les scènes enregistrées <i>test-QMUL</i> et les scènes simulées <i>instance-IRCCYN</i> . <a href="#">180</a>
FIGURE 48	Vision schématisée des systèmes de détection d'événements du challenge DCASE 2013. <a href="#">183</a>
FIGURE 49	Performances des systèmes évalués dans le cadre du challenge DCASE 2013 sur les corpus QMUL et IRCCYN en considérant $\text{Fcw}_{\text{eb}}$ . <a href="#">185</a>
FIGURE 50	Performances des systèmes évalués dans le cadre du challenge DCASE 2013 sur les corpus <i>instance-QMUL</i> simulés avec différents EBR (6, 0, -6 et -12dB). <a href="#">188</a>
FIGURE 51	Performances globales des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique $\text{Fcw}_{\text{eb}}$ . <a href="#">197</a>
FIGURE 52	Influence de la polyphonie sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique $\text{Fcw}_{\text{eb}}$ . <a href="#">198</a>
FIGURE 53	Influence du niveau de bruit (EBR) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique $\text{Fcw}_{\text{eb}}$ . <a href="#">199</a>
FIGURE 54	Influence du nombre d'événements (nec) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique $\text{Fcw}_{\text{eb}}$ . <a href="#">202</a>
FIGURE 55	Événement ou texture sonore : influence de la période d'attention. <a href="#">226</a>
FIGURE 56	Pourcentage de réponses correctes en fonction de la position du son cible et de la congruence. <a href="#">230</a>

## LISTE DES TABLEAUX

---

TABLE 1	Indicateurs acoustiques.	58
TABLE 2	Indicateurs psychoacoustiques : modèles mathématiques illustrant des qualités affectives perçues.	59
TABLE 3	Les catégories sonores les plus citées.	71
TABLE 4	Résumé des étapes de l'expérience de simulation.	115
TABLE 5	Acronyme des variables utilisées dans le cadre des expériences sensorielles.	119
TABLE 6	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 1.b et les descripteurs structurels.	128
TABLE 7	Classes d'événements et de textures identifiées comme étant des marqueurs sonores.	131
TABLE 8	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 1.b et les descripteurs structurels relatifs à la présence des marqueurs sonores.	134
TABLE 9	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 2, et les descripteurs structurels globaux relatifs à la présence des marqueurs sonores pour les i/am-scènes et ni/am-scènes.	145
TABLE 10	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 2 et les descripteurs structurels pour les i/sm-scènes et ni/sm-scènes.	146
TABLE 11	Stratégies de catégorisation et nombre de groupements effectués.	155
TABLE 12	Labels relevés sur les descriptions verbales des groupements effectués par les sujets, en considérant séparément ceux relatifs aux descripteurs de qualité subjectifs, et ceux relatifs aux descripteurs sémantiques subjectifs.	156
TABLE 13	Répartition des labels relatifs aux sources sonores relevées par les sujets en fonction des partitions établies par la classification ascendante hiérarchique.	158
TABLE 14	Répartition des labels relatifs aux qualités affectives perçues en fonction des partitions établies par la classification ascendante hiérarchique.	159

TABLE 15	Résultats des ANOVA à mesures répétées pratiquées sur les différents descripteurs structurales en tenant compte du partitionnement des scènes. <a href="#">164</a>
TABLE 16	Répartition des classes de sons en fonction des partitions établies par la classification ascendante hiérarchique. <a href="#">168</a>
TABLE 17	Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2013. <a href="#">174</a>
TABLE 18	Description synthétique des systèmes soumis dans le cadre de la tâche 2 de challenge DCASE 2013. <a href="#">183</a>
TABLE 19	Résultats mesurés par Fcweb pour chacun des systèmes évalués dans le cadre du challenge DCASE 2013 en considérant les corpus <i>test-QMUL</i> , <i>insQ-EBR(0)</i> et <i>abstract-QMUL</i> . <a href="#">186</a>
TABLE 20	Nombre maximum de faux positifs pour chaque système évalué et pour chaque corpus <a href="#">187</a>
TABLE 21	Résultats mesurés par Fcweb pour chacun des systèmes évalués dans le cadre du challenge DCASE 2013 en considérant les corpus <i>test-QMUL</i> , <i>instance-IRCCYN</i> et <i>abstract-IRCCYN</i> . <a href="#">189</a>
TABLE 22	Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2016. <a href="#">192</a>
TABLE 23	Description synthétique des systèmes soumis dans le cadre de la tâche 2 du challenge DCASE 2016. <a href="#">196</a>
TABLE 24	Interprétation du coefficient de corrélation de Pearson. <a href="#">220</a>
TABLE 25	Théorie de la détection du signal. <a href="#">224</a>

Première partie

## PRÉAMBULE



## PRÉAMBULE

---

### 1.1 INTRODUCTION GÉNÉRALE

#### 1.1.1 *L'environnement sonore*

Émuler la capacité humaine à percevoir et analyser les objets de l'environnement à partir de données sensibles brutes est aujourd'hui un défi central de l'informatique appliquée. Dans ces recherches, le domaine de l'image s'est taillé la part du lion, stimulé par les nombreuses applications possibles, tant dans le domaine de la recherche d'information que dans celui de la robotique – avec aujourd'hui l'apparition des premiers véhicules intelligents, capables d'analyser leur environnement. La sphère sonore, pour sa part, a d'abord vu se développer les recherches portant sur les objets spécifiques que sont la voix, puis la musique : chacun bénéficie maintenant de champs de recherche dédiés, que ce soit en perception ou en intelligence artificielle (SP : *Speech Processing*, traitement automatique de la parole, et MIR : *Music Information Retrieval*, recouvrement de l'information musicale). Plus récemment, suivant un intérêt croissant pour les problématiques de nuisance sonore, d'une part, et de prise en compte du contexte (*context awareness*), de l'autre, la question de l'analyse d'environnements sonores, c'est-à-dire de scènes sonores « ordinaires » ne relevant ni de la parole, ni de la musique a gagné en importance. C'est dans ce domaine que se positionne le travail présenté dans cette thèse.

On peut tout d'abord remarquer que cette définition des sons environnementaux<sup>1</sup> par exclusion de la parole, et de la musique, n'est pas satisfaisante. D'un côté, elle réduit les sons environnementaux à des entités secondaires. D'un autre, l'opposition suggérée entre sons environnementaux et parole ou musique, deux domaines où le sens donné aux sons est de primordial, peut mener à penser que l'influence de la valeur sémantique des sons environnementaux est anecdotique, induisant, *de facto*, la prédominance de leurs caractéristiques physiques. Postulat largement réfuté dans la littérature (Ballas and Howard, 1987).

Nous préférerons ici la définition donnée par Vanderveer, 1980 (cité par Ballas and Howard, 1987) qui se pose en quatre points. Un son environnemental :

---

<sup>1</sup> Dans ce document, par souci rédactionnel, nous parlerons indifféremment de son(s) environnemental(aux), d'environnement(s) sonore(s), de scène(s) sonore(s), et de scène(s) sonore(s) environnementale(s), pour désigner les sons environnementaux.

1. est produit par une source réelle ;
2. a un sens, en vertu de l'action qui en est la cause ;
3. est par essence plus complexe qu'un stimulus de synthèse produit en laboratoire, comme un son pur ;
4. ne fait pas partie d'un système de communication.

Les deux premiers points caractérisent directement les sources émettrices, précisant qu'il s'agit de sources réelles, et insistant sur l'importance du sens qu'elles portent. Nous remarquons cependant que la définition pose la valeur sémantique des sources uniquement par rapport à l'action à l'origine du son. Or, le contexte d'émission/réception, contexte relatif au sujet ou à son environnement, est déterminant. Une même scène sonore peut être perçue différemment par deux individus, et il nous paraît nécessaire de renforcer le point 2 de la définition comme suit :

- a un sens, en vertu de l'action qui en est la cause, ainsi que du contexte d'écoute.

Les deux derniers points positionnent les sons environnementaux par rapport aux autres stimuli sonores couramment étudiés, les opposant spécifiquement aux sons de synthèse produits en laboratoire, ainsi qu'aux sons ayant une portée communicationnelle comme la parole ou la musique.

La définition insiste sur le fait que la perception d'un environnement sonore relève avant tout de l'interprétation sémantique des événements qui le peuplent, *i.e.* de l'identification de la nature des sources sonores émettrices. Cette importance de la composition sémantique sur les qualités sensibles des scènes nous permet d'envisager la scène comme un objet composite, le résultat de l'association des sources sonores qui la constituent.

Partant de cette vision composite des scènes, l'objectif de nos travaux est triple :

- proposer un modèle morphologique des scènes sonores environnementales, fondé sur une étude approfondie de la littérature ayant trait aux mécanismes régissant la perception des sons environnementaux ;
- montrer l'utilité d'un tel modèle :
  - dans le cadre de l'analyse sensorielle ;
  - dans le cadre de l'analyse automatique.

### 1.1.2 Pourquoi modéliser une scène sonore ?

Que ce soit dans le domaine de la perception ou de l'apprentissage machine, tout protocole expérimental suppose un niveau de contrôle maximal de l'expérimentateur sur les caractéristiques des stimuli proposés. En ce qui concerne les environnements sonores, peu de travaux ont porté sur le développement d'outils pouvant permettre aux chercheurs d'agir sur ces stimuli.

Conscients de cette problématique, nous proposons ici un modèle génératif permettant de simuler, à partir d'enregistrements de sons isolés, des scènes sonores dont nous maîtrisons les propriétés structurelles, à savoir, l'intensité, la densité et la diversité des sources sonores en présence. Le modèle envisage la scène sonore comme un objet composite, une somme de sons sources. Le niveau d'abstraction choisi est motivé par les connaissances disponibles sur le système auditif humain.

Fort des banques de données ainsi constituées, nous investissons deux champs d'application. Le premier concerne la perception des paysages sonores, et questionne plus spécifiquement la notion d'agrément perçu dans des lieux urbains. L'utilisation de données simulées nous permet d'apprécier finement les contributions de chacune des sources sonores dans l'agrément perçu. Elle nous permet encore de retravailler les scènes en modifiant les paramètres afin de mesurer leurs effets.

Le deuxième concerne la détection automatique des événements sonores, et propose une méthodologie novatrice afin d'évaluer les performances des algorithmes dédiés à cette tâche. Les données simulées se révèlent un outil précieux afin d'évaluer notamment la capacité de généralisation des algorithmes.

## 1.2 MOTIVATIONS DES CAS D'ÉTUDES

### 1.2.1 Un cadre applicatif pluridisciplinaire

Comme précédemment évoqué, l'application des données simulées issues de notre modèle d'environnements sonores porte à la fois sur l'analyse sensorielle, et sur l'analyse automatique des environnements.

Par analyse sensorielle, on entend l'ensemble des processus qui constituent le système perceptif de l'homme, système par lequel il comprend son environnement, lui donne sens. Ces processus comprennent, d'une part, les mécanismes d'acquisition de l'information, d'autre part, les mécanismes de traitement de l'information.

Par analyse automatique on entend l'apprentissage machine. Dans ce domaine, l'objectif des recherches est d'élaborer des algorithmes propres à la simulation de la perception humaine. Ici encore on distingue les étapes d'acquisition de l'information, et de traitement de

l'information. Les études portant sur l'acquisition se focalisent sur les descripteurs mathématiques permettant d'extraire une information utile des données brutes du signal. Les études portant sur le traitement se penchent sur les techniques permettant de trier l'information ainsi collectée, c'est-à-dire regrouper les parties de l'information qui se « ressemblent ». Ce tri peut s'opérer soit dans un cadre non-supervisé, *i.e.* en effectuant les groupements sur la seule base de l'information collectée, soit dans un cadre supervisé, *i.e.* en effectuant les groupements sur la base de classes d'objets pré-considérées.

L'analyse sensorielle a donc trait à la perception humaine, et l'analyse automatique à l'intelligence artificielle. Les deux domaines peuvent, à première vue, paraître éloignés. Ils portent cependant sur l'acquisition, la structuration et l'utilisation des connaissances, et constituent les deux disciplines d'une même quête, *i.e.* la pensée humaine, l'une visant à comprendre son fonctionnement, l'autre cherchant à le simuler. À ce titre, perception humaine et intelligence artificielle font toutes deux partie d'un même champ de recherche : les sciences cognitives.

### *1.2.2 La perception des paysages sonores urbains*

#### *1.2.2.1 Historique et application*

La ville est un environnement bruyant. Elle l'a été de tous temps. Déjà dans les rues de la Rome antique, le bruit des chariots pose problème<sup>2</sup>, à tel point que le consul Jules César interdit à ceux-ci de circuler la nuit. Ce qui a changé, par contre, c'est la perception du bruit. Dans les années 70-80, le bruit « devient » pollution, facteur de dégradation de la qualité de vie. Cette pollution est d'autant plus critique que d'ici 2050, 68% de la population mondiale sera urbaine (Park et al., 2014).

Les chercheurs se concentrent alors sur l'identification des sources du bruit, et sur les moyens d'abaisser les niveaux sonores. Les premières législations anti-bruit apparaissent, qui proposent/imposent une réduction du niveau des bruits produits, essentiellement, par les transports et l'industrie.

La notion de bruit cependant est subjective. Le fait est que bien des lieux urbains sont aussi appréciés pour leur « atmosphère vivante ». Ville agréable ne rime pas nécessairement avec ville silencieuse. La problématique du bruit se complexifie.

Il est aujourd'hui communément admis que des mesures objectives de niveaux sonores (*e.g.*  $L_{Aeq}$ ) ne peuvent, seules, rendre compte de la qualité d'un environnement, et que vouloir réhabiliter cet environnement en s'attaquant uniquement aux paramètres acoustiques, par définition objectifs, est illusoire (Aletta et al., 2016; Kang and Zhang,

---

<sup>2</sup> Juvenal, Satire 3.232–238

2010; Schulte-Fortkamp and Fiebig, 2006; Yang and Kang, 2005). Il faut désormais envisager le bruit non plus seulement comme un objet physique, mais encore comme un objet cognitif (Guastavino, 2003). Le problème n'est plus de savoir à partir de quand un son est gênant, mais pourquoi il est perçu comme tel, et par tel individu.

C'est l'objet de la recherche sur les paysages sonores : envisager l'environnement sonore du point de vue de celui qui le perçoit.

Centrées sur le sujet, ces recherches sont par essence interdisciplinaires (Aletta et al., 2016; Davies et al., 2013), faisant appel à des outils et des méthodes provenant de champs de recherche variés comme l'acoustique, la psychologie cognitive, la psycho-linguistique, la sociologie, et plus récemment, l'intelligence artificielle.

Depuis vingt ans, l'approche par les paysages sonores a permis de développer une base de descripteurs qualitatifs et acoustiques grâce auxquels nous jugeons mieux, et sommes mieux à même d'améliorer l'environnement sonore urbain (Kang, 2006; Schulte-Fortkamp et al., 2007). L'enjeu est aujourd'hui de relier ces données perceptives à des mesures acoustiques, afin de pouvoir établir une politique de réduction du bruit efficace, adaptée à chaque situation (Schulte-Fortkamp, 2013).

Cependant, le caractère pluridisciplinaire de ces recherches, et l'utilisation de protocoles expérimentaux variés pour évaluer l'environnement sonore, rendent l'intégration des résultats difficile (Davies et al., 2013). De plus, il n'y a toujours pas de consensus sur les descripteurs (acoustiques ou perceptifs) à utiliser pour caractériser un paysage sonore (Aletta et al., 2016; Brocolini et al., 2012), ce qui empêche la communauté, d'une part, de présenter aux décideurs des indicateurs génériques d'évaluation des paysages sonores, et d'autre part, d'élaborer/proposer des modèles crédibles sur la base de ces expertises.

Récemment, plusieurs projets internationaux ont été lancés afin de standardiser les pratiques expérimentales des recherches portant sur les paysages sonores, notamment *the European Cooperation in Science and Technology Action*<sup>3</sup> (Schulte-Fortkamp and Kang, 2010) et *the Positive Soundscape project* (Davies et al., 2009, 2013). Mais les difficultés persistent (Ribeiro et al., 2013; Schulte-Fortkamp, 2013). Afin d'acquérir la masse de données nécessaire pour évaluer la qualité de l'environnement sonore sur un temps long, les caractéristiques de l'environnement variant au cours de la journée, comme au cours des saisons, d'autres projets (Park et al., 2014) ont pour objet le déploiement d'un réseau de senseurs capable de capturer, en continu et en temps réel, toutes les informations relatives à la qualité de l'environnement évalué.

<sup>3</sup> TD0804, *soundscape of European Cities and Landscapes* : [http://www.cost.eu/COST\\_Actions/tud/TD0804](http://www.cost.eu/COST_Actions/tud/TD0804)

### 1.2.2.2 Évaluation perceptive des paysages sonores

#### **Une description holistique insatisfaisante**

Un des objectifs premiers des études sur les paysages sonores est d'identifier les informations contenues dans l'environnement qui influent sur les sensations perçues. La majorité des études considèrent des descriptions holistiques, calculant des indicateurs (physiques ou perceptifs) sur l'ensemble de l'environnement.

Cependant, de plus en plus de travaux tendent à montrer que toutes les sources ne participent pas de manière égale à la perception de la scène sonore. En conséquence, les recherches se portent maintenant sur l'étude des contributions spécifiques des différents éléments qui composent l'environnement.

Ces recherches peuvent bénéficier de l'utilisation de données simulées, données dont les caractéristiques structurelles, en particulier la nature des sources présentes, leurs positions, et leurs caractéristiques physiques, sont connues.

#### **Un accès partiel à la représentation mentale du paysage sonore**

Questionner le lien entre sensation (*e.g.* calme) et environnement (*e.g.* scène sonore urbaine), revient à objectiver la représentation mentale que se fait un sujet d'une scène sonore en particulier (*e.g.* une scène sonore urbaine calme).

Si on demande au sujet d'évaluer un environnement donné (évaluez le « calme » de cette scène sonore), le gain d'information de l'expérimentateur est conditionné par la nature des stimuli disponibles, que ces derniers soient des scènes sonores enregistrées (expérience en laboratoire), ou réelles (expérience *in situ*). Cependant, le fait que ces stimuli existent lui permet d'en établir une description physique précise.

Si, à l'inverse, on demande au sujet de décrire un environnement donné (décrivez une scène sonore urbaine « calme »), l'expérimentateur obtient bien une information riche, symbolique et sémantique, de la représentation que se fait le sujet de cet environnement. Mais, en l'absence de données sonores, il ne peut la caractériser physiquement.

Nous pensons que la simulation permet de faire un lien élégant entre ces deux approches auparavant distinctes. Elle peut être vue comme une description modale, *i.e.* définie sur des dimensions physiques (le signal de la scène simulée), de la représentation mentale du sujet, description modale dont il est possible d'extraire les caractéristiques physiques. Elle ouvre la voie à de vastes champs d'analyses et d'investigations.

### 1.2.3 La détection automatique d'événements sonores

#### 1.2.3.1 Historique et application

La quantité de données audio enregistrées à partir de notre environnement sonore a considérablement augmenté au cours des dernières décennies. Afin de mesurer l'effet de l'activité humaine et du changement climatique sur la biodiversité du règne animal, les chercheurs travaillant dans le domaine de l'eco-acoustique (Pijanowski et al., 2011; Sueur et al., 2014) ont récemment entrepris le déploiement massif de capteurs acoustiques à travers le monde (Ness et al., 2013; Stowell and Plumbley, 2013a,b; Warren et al., 2006).

Par ailleurs, de nouveaux travaux démontrent l'intérêt de moniter l'environnement acoustique des zones urbaines afin d'en caractériser l'agrément (Guyot et al., 2005; Ricciardi et al., 2015), ainsi que la gêne due aux bruits de circulation (Gloaguen et al., 2016). De par leur impact sociétal, et du fait des nombreux défis scientifiques qu'elles soulèvent, l'intérêt de ces études est majeur.

Afin de répondre aux problématiques soulevées par les recherches ci-dessus exposées, l'Analyse Automatique de Scènes Sonores ( $A^2S^2$ ) (Stowell et al., 2015) vise à développer des approches et des systèmes permettant d'extraire automatiquement des environnements sonores une information utile.

On distingue alors deux tâches dans l' $A^2S^2$  :

- la classification des scènes acoustiques (ASC : *acoustic scene classification*), dont l'objectif est de reconnaître automatiquement le type d'environnement (parc, rue calme, marché) d'un enregistrement donné ;
- la détection d'événements sonores (SED : *sound event detection*), dont l'objectif est d'identifier les différents événements qui composent l'enregistrement d'un environnement donné (parc ⇒ chant d'oiseau, ballon, bruit de pas, voix).

C'est de cette dernière tâche que nous allons traiter.

#### 1.2.3.2 Évaluation des algorithmes de détection d'événements

Si la reconnaissance automatique de la parole (Rabiner and Juang, 1993), ou encore le recouvrement automatique de l'information musicale (Müller, 2007) sont des domaines d'investigation aujourd'hui bien établis, l' $A^2S^2$  est un domaine jeune, et ne profite pas encore des banques de données suffisantes, en volume et en qualité, ce qui contraint l'effort de recherche.

Ces banques de données sont cependant essentielles pour obtenir une évaluation saine et informative des systèmes proposés. Une mauvaise conception de ces dernières peut mener à des erreurs d'appréciation des systèmes.

Ce fait a déjà été constaté dans les domaines liés à l'analyse automatique en image<sup>4</sup>, en musique (Sturm, 2014), mais également dans le cadre de nos travaux, en ce qui concerne le recouvrement de similitudes entre scènes sonores environnementales (Lafay et al., 2016b).

Nous montrons dans ce document comment l'utilisation de données simulées permet de répondre à ces problématiques de fond en fournissant à l'expérimentateur un degré de contrôle plus élevé sur les données.

### 1.3 PLAN

Ce document comprend 4 parties.

La partie **i** est constituée du chapitre 1, la présente introduction.

La partie **ii** introduit le modèle morphologique de scènes sonores. Elle regroupe les chapitres 2, et 3. Le chapitre 2 présente un état de l'art des connaissances sur les processus mis en œuvre par le système auditif dans la perception des environnements sonores. Sur la base de ces considérations perceptives, le chapitre 3 présente un modèle morphologique de scènes sonores environnementales. Il introduit également les outils permettant de simuler, à partir du modèle proposé, des environnements sonores.

La partie **iii** présente des cas d'applications concrets pouvant profiter de l'utilisation de données simulées. Elle est composée des chapitres 4, et 5. Le chapitre 4 présente une série d'expériences montrant comment le modèle introduit permet d'étendre les possibilités des méthodologies traditionnellement utilisées en analyse sensorielle. Le cadre applicatif choisi par ces expériences est l'évaluation de l'agrément dans les environnements sonores urbains. Le chapitre 5 précise comment le modèle de scènes sonores proposé peut être appliqué à l'évaluation des algorithmes de détection automatique d'événements sonores. Il montre notamment comment il permet de gagner en connaissance quant à la capacité de généralisation des systèmes de détection proposés.

La partie **iv**, partie conclusive, comprend le chapitre 6, qui résume les différentes contributions de cette thèse, conclut quant au travail effectué, et propose de nouvelles pistes à explorer.

---

<sup>4</sup> Neural Network Follies : <https://neil.fraser.name/writing/tank/>

Deuxième partie  
UN MODÈLE MORPHOLOGIQUE



# 2

## ÉTAT DE L'ART

---

### 2.1 INTRODUCTION

Avant d'aller plus loin dans l'exposé de ces travaux, il semble nécessaire de dresser un état des lieux des connaissances liées au système auditif. Les savoirs rapportés ici sont, dans leur grande majorité, le fruit de recherches adoptant une approche méthodologique propre à la psychologie. Ils ne rendent compte que très partiellement des connaissances issues d'autres disciplines comme par exemple les neurosciences.

Par ailleurs, plusieurs sujets traités dans ce chapitre, comme l'analyse de scènes acoustiques, la représentation mentale de l'environnement sonore, l'étude des paysages sonores, ou encore les événements et textures sonores, sont le plus souvent étudiés isolément. De fait, le vocabulaire employé pour décrire ces phénomènes, ou le regard posé sur les problèmes liés à l'audition varient suivant les approches, et le lien entre ces domaines est rarement fait. Cet état de l'art a pour ambition de connecter ces sujets.

Le chapitre comprend sept sections. La première aborde de manière théorique certains mécanismes mis en œuvre par l'homme quand il interagit avec son environnement, notamment les processus perceptifs, liés à l'acquisition de l'information issue du monde physique, et les processus cognitifs, liés au traitement de cette information. La seconde interroge plus spécifiquement la manière dont nous nous représentons le monde perçu. Elle montre par ailleurs comment ces représentations influent sur notre interaction avec l'environnement, et notamment sur notre capacité à catégoriser, *i.e.* à identifier deux objets comme étant semblables. La troisième présente les différentes approches méthodologiques adoptées par les études psychologiques pour étudier le fonctionnement des processus perceptifs et cognitifs dans le cadre de l'audition. La quatrième aborde le système auditif, et propose une vue d'ensemble des différents systèmes fonctionnels intervenant dans le traitement de l'information sonore. La cinquième s'intéresse aux processus perceptifs dits d'Analyse de Scènes Acoustiques (ASA), processus par lesquels le système auditif ségrégue les informations contenues dans l'environnement sonore afin d'en dégager des objets cohérents. La sixième introduit la notion de paysage sonore, et examine l'impact que cette notion a sur les recherches en perception des environnements. Elle dresse un état de l'art des connaissances dans le domaine, et tente de dégager les grands axes méthodologiques suivis par ces études. La septième, enfin, précise

les concepts d'événement et de texture sonore, notions clefs qui interviennent, par la suite, dans le modèle de scènes sonores proposé.

A la fin de chaque partie, les points abordés sont discutés. Lors de ces discussions, nous précisons les notions ou théories que nous privilégions dans le cadre de nos travaux, et celles dont nous nous détachons. Ces notions interviennent notamment dans la construction du modèle de scènes sonores proposé. Elles interviennent, par après, dans l'interprétation des résultats du cas d'étude sur la perception de l'agrément dans un environnement sonore urbain (cf. Chapitre 4).

## 2.2 PERCEPTION ET COGNITION

### 2.2.1 *Définitions*

**Perception.** Le mot désigne l'ensemble des processus de traitement de l'information sensorielle. La perception est à l'origine de l'interaction que nous entretenons avec notre environnement, l'interface par laquelle nous passons d'une information externe, physique, à une information interne, mentale.

**Cognition.** Le mot désigne l'ensemble des processus de traitement de l'information interne, processus comprenant des fonctionnalités aussi essentielles que le langage, le raisonnement ou la mémoire.

L'étude de la cognition a pris son essor avec le mouvement cognitiviste, dans les années 50, mouvement né en réaction au *Béhaviorisme*, courant fondé sur l'étude spécifique des comportements objectivement observables (externes) de l'être humain, et négligeant, de fait, le rôle des processus cognitifs (internes).

Une définition de la cognition est donnée par U. Neisser (Neisser, 1976, p. 1)<sup>1</sup>

« Cognition is the activity of knowing : the acquisition, organisation and the use of knowledge. »<sup>2</sup>

### 2.2.2 *Théorie classique de la cognition*

Selon la théorie classique, perception et cognition dépendent de deux groupes de systèmes fonctionnels du cerveau distincts.

La perception met en œuvre des systèmes de traitement dits modaux. Ces systèmes, supportés par les organes sensoriels (oreilles, yeux etc ...), transposent l'information physique d'un objet perçu,

<sup>1</sup> Ulric Neisser est considéré comme un des pères du cognitivisme, notamment grâce à son livre (Neisser, 1967). Il a par la suite critiqué la direction prise par le mouvement, dénonçant une « approche laboratoire » trop éloignée de la réalité terrain.

<sup>2</sup> La cognition est l'activité liée au « savoir » : l'acquisition, l'organisation et l'utilisation des connaissances.

en une « image » mentale. Le cerveau forme ces images mentales (Barsalou, 2003b; Martin, 2001). Nous sommes ainsi capables de nous représenter une chaise, sa taille, l'écartement des pieds, la courbe du dossier *etc..* Le fait de pouvoir associer à l'image les dimensions physiques de l'objet dit qu'il s'agit là d'une information modale.

La cognition, elle, s'appuie sur des « représentations » mentales.

Une définition de ces représentations est donnée par (Houdé et al., 1998):

« La représentation mentale peut être vue comme une entité interne, le correspondant cognitif individuel des réalisations externes expérimentées par un sujet. »

Les représentations font office de sauvegardes de l'information. Elles rendent compte à la fois de notre compréhension du monde, et de la manière dont nous l'abordons.

Selon la théorie classique de la cognition, elles sont conservées en mémoire sous une forme amodale (symbolique et sémantique) (McAdams and Bigand, 1994, p. 357), c'est à dire, « nettoyées » des informations modales relatives aux dimensions physiques perçues. Par exemple, lorsque nous percevons l'objet physique chaise, ce n'est pas la forme de la chaise qui est traitée par les processus cognitifs, mais plutôt le fait qu'il s'agisse d'une chaise, avec quatre pieds, un dossier, le tout en bois, et de couleur marron (cf. Figure 1a).

On parle de concept concret lorsqu'on se réfère à la représentation mentale d'objets concrets (chaise), et de concept abstrait lorsqu'on se réfère à des notions qui n'ont pas d'existence physique comme une émotion (concept de bonheur, de vérité), ou un raisonnement (concept de preuve)<sup>3</sup>.

Dans l'approche classique, la perception est vue comme la porte d'entrée de la cognition. Les processus perceptifs transforment l'information sensorielle en une image mentale (modale), qui, suivant sa nature, va solliciter différentes représentations conceptualisées, symboliques et sémantiques (amodale), dont le traitement est supporté, lui, par les processus cognitifs. Perception et cognition fonctionnent ainsi de manière indépendante, agissant chacune sur un type d'information donné (modale ou amodale), les représentations amodales étant supposément stockées dans des systèmes mémoriels séparés de ceux supportant la perception (Barsalou, 2008).

### 2.2.3 Une approche ancrée de la cognition

Cette dichotomie entre perception et cognition est aujourd'hui remise en question.

---

<sup>3</sup> On parlera, d'image, dans ce document, lorsqu'on se réfère à l'image mentale d'un objet (une chaise particulière), et de représentation, lorsqu'on se réfère aux connaissances stockées en mémoire liées au concept de cet objet (le concept de chaise)

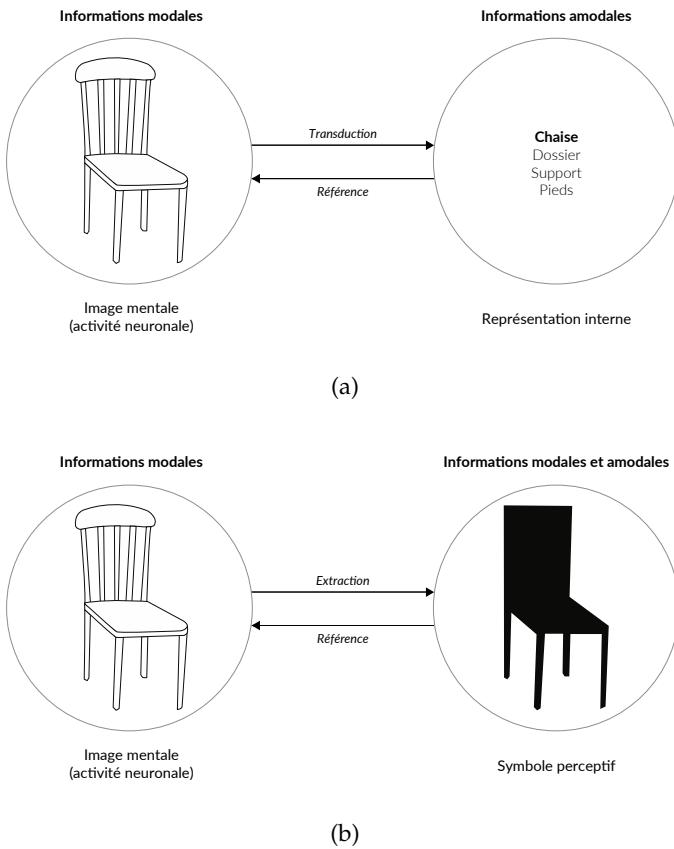


FIGURE 1 : Théories de la cognition : (a) la théorie classique ; (b) la théorie ancrée, d'après (Barsalou, 1999).

Dans une approche « ancrée » de la cognition (*Grounded Cognition*), L.W. Barsalou réfute la thèse selon laquelle les processus cognitifs s'appuient uniquement sur des ressources amodales, et défend l'idée qu'ils profitent également d'une information modale.

Ces représentations modales sont appelées des « symboles perceptifs » (Barsalou, 1999). On peut voir le symbole perceptif comme une image mentale schématisée d'un objet, *i.e.* une image dont seule une partie de l'information est conservée. Pour reprendre l'exemple de la chaise, le symbole perceptif ne conservera de l'image mentale que la forme globale, la couleur ou les dimensions de celle-ci (cf. Figure 1b), les autres propriétés étant dégradées.

Le choix de l'information à conserver dépend du contexte, qu'il s'agisse de la situation dans laquelle est perçu l'objet, ou bien encore de l'état de l'individu (heureux, triste, fatigué *etc.*) (Barsalou, 2003b).

#### 2.2.3.1 *La simulation cognitive*

Les symboles perceptifs ne sont pas des entités indépendantes. Différents symboles perceptifs représentant différentes formes de chaises sont connectés de par leur origine commune (l'objet chaise).

Le concept « chaise » est alors lié à l'ensemble des symboles perceptifs de chaise acquis des expériences passées. Ces symboles perceptifs ne capturant qu'un aspect du concept (forme, couleur, *etc.*), ils peuvent être combinés. Il existe, *a priori*, un nombre infini de réalisations possibles du concept chaise. À ce titre, le concept est un simulateur.

Cette notion de simulation est fondamentale dans la théorie ancrée de la cognition. Les processus perceptifs extraient de l'environnement des images mentales. De ces images, seule une partie de l'information est conservée, la sélection étant fonction du contexte. Cette information est traduite en symboles perceptifs. Ces symboles activent des simulateurs (concepts) qui vont composer, à partir des connaissances de l'individu, une image simulée de l'environnement perçu (Barsalou, 2003b).

On le voit, l'approche « ancrée » tente de réunir les processus perceptifs et cognitifs (Goldstone and Barsalou, 1998). Les deux processus agissent de concert, et profitent l'un et l'autre d'une information modale, bâtie, entre autre, à partir de l'information sensorielle (externe), et des représentations mentales (interne).

Ces processus sont, de fait, dépendants l'un de l'autre. Les possibilités de simulation sont restreintes par l'information sensorielle provenant des processus perceptifs. Ces derniers profitent alors de l'image simulée par les processus cognitifs, afin d'optimiser l'extraction de l'information sensorielle.

L'information externe intervient dans la cognition comme un complément contextualisé des représentations internes, représentations qui sont elles mêmes « situées » suivant la nature des informations sensorielles.

A noter cependant que cette approche n'a pas été encore formalisée, et qu'aucune théorie n'explique encore concrètement son fonctionnement (Barsalou, 2010). Elle reste par ailleurs encore très discutée (Barsalou, 2016; Leshinskaya and Caramazza, 2016), bien que de nombreuses études semblent montrer son bien-fondé (Barsalou, 1999, 2003a, 2010), y compris dans le cadre de l'audition (Kiefer et al., 2008). Approche classique et approche ancrée de la cognition sont illustrées sur la Figure 2.

#### 2.2.4 Une approche écologique de la cognition

L'approche ancrée n'est pas la première à postuler que l'environnement joue un rôle dans le traitement perceptif de l'information.

Dans le domaine de la vision, Gibson (Gibson, 1966), dès 1966, se demande si les « lois structurant les objets sont porteuses d'informations, ou si cette information est tirée de comparaisons » (Gibson, 1978). Il introduit « l'approche écologique ».

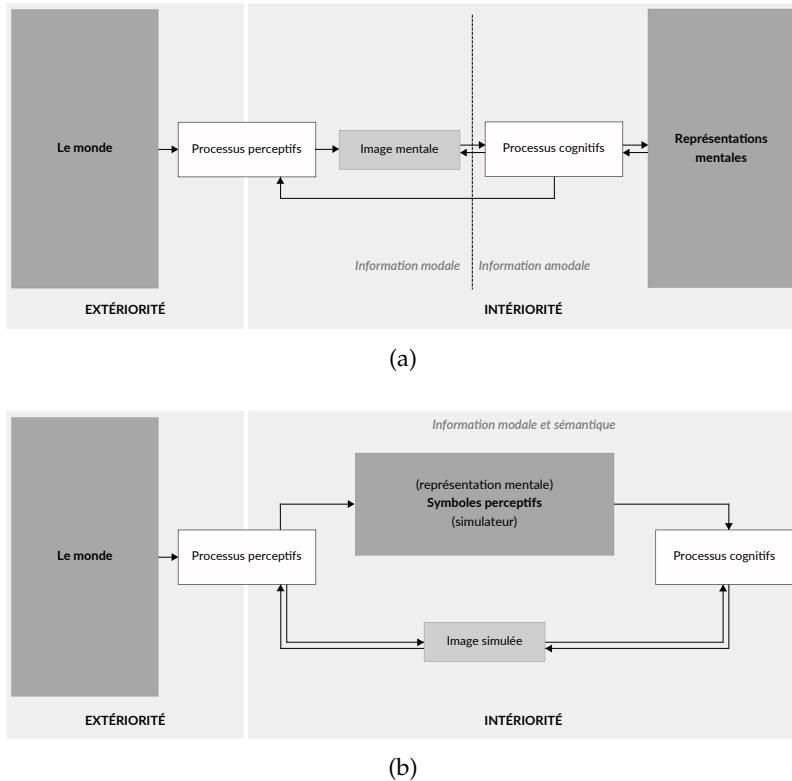


FIGURE 2 : Processus cognitifs et perceptifs : (a) la théorie classique ; (b) la théorie ancrée.

La théorie écologique envisage la perception en mettant au centre le couple homme-environnement. Elle postule que les données nécessaires à la perception de l'environnement sont contenues entièrement dans l'information sensorielle, et donne ainsi la primauté aux processus perceptifs dans le traitement de l'information de l'environnement.

C'est l'accoutumance à son environnement qui permet au sujet de sélectionner, dans la profusion des informations disponibles, les invariants structurels dont il a besoin pour produire du sens. Le système perceptif est entièrement « accordé » par l'environnement.

L'approche écologique néglige l'importance des connaissances propres du sujet. On la considère comme une approche directe de la perception, par opposition aux approches indirectes, qui valorisent, elles, l'action des représentations internes. A ce titre, la théorie écologique a été vivement critiquée (Ullman, 1980).

### 2.2.5 Discussion

Dans le débat Perception/Cognition, nous nous inscrivons dans une vision « ancrée » de la cognition. Cela a trois conséquences :

- nous considérons que le traitement de l'information interne dépend des connaissances de l'individu, mais également du statut de l'information externe ;
- nous considérons que l'information interne, *i.e.* les représentations mentales, ne sont pas exclusivement amodales, mais intègrent une dimension sémantique, et une dimension physique (Barsalou, 2008) ;
- nous considérons que le système auditif est capable, en interne, de simuler la réalisation d'un concept sonore perçu. Cette réalisation revêt, par définition, une information modale. Ses attributs dépendent des symboles perceptifs conservés en mémoire par l'individu, mais également de l'information sensorielle externe, autrement dit, du contexte. Nous envisageons ici le concept à la fois dans une perspective concrète (chants d'oiseau), et dans une perspective abstraite (environnement sonore calme). Bien que cela soit encore peu étudié, il existe en effet des indices montrant qu'il est possible de créer une image (une réalisation) d'un concept abstrait (Barsalou, 2003a; Barsalou and Wiemer-Hastings, 2005). Une telle image est plus complexe que les images tirées des concepts concrets, et peut même les incorporer (*e.g.* chant d'oiseau ⊂ environnement sonore calme) (Barsalou and Wiemer-Hastings, 2005).

Ce choix de la cognition ancrée est en particulier motivé par l'explication élégante que cette dernière donne au phénomène d'introspection. Le cerveau entretient un dialogue constant avec l'environnement, et ce même si aucun objet n'est présent. L'introspection est la faculté de se représenter mentalement un objet absent de l'environnement<sup>4</sup>.

Ce phénomène est particulièrement prégnant en musique. A la seule vue d'une partition, le musicien est capable d'entendre l'œuvre intérieurement. Nombre de compositeurs soutiennent d'ailleurs que le processus créatif doit se passer de supports sonores (instrument, ordinateur), leur habilité à utiliser ces outils pouvant potentiellement restreindre leur expressivité.

Ainsi, le processus de simulation interne se conçoit très bien en musique. Pour le lecteur qui « écoute » la musique à partir de la représentation qu'il se fait des instruments, la simulation est contrainte par une information sensorielle, la partition. Pour le compositeur, cette simulation, à la fois plus libre et plus complexe, n'est contrainte que par un contexte fonctionnel, *i.e.* son intention de compositeur.

---

<sup>4</sup> Formellement, l'introspection englobe d'autres types de processus cognitifs. Ces derniers n'étant pas abordés dans ce document, nous conservons ici le sens donné ci-devant

Nous ne souscrivons pas à l'approche écologique, mais retenons de celle-ci l'importance de considérer le stimulus dans son contexte d'écoute.

## 2.3 STRUCTURE CATÉGORIELLE DES REPRÉSENTATIONS MENTALES

Nous venons de le voir, (cf. Section 2.2), l'interaction environnement/individu dépend des représentations internes que le sujet se fait du monde.

Nonobstant la nature de ces représentations (modale, amodale), elles forment une vision discrète du monde réel continu (Houdé et al., 1998), une structure modulaire groupant (entre autre) une information sémantique. C'est par le passage du continu au discret que nous sommes à même d'organiser nos connaissances, afin de les réutiliser de manière efficace. Les objets discrets, issus de cette organisation, sont appelés des catégories. L'action consistant à juger si un événement perçu appartient à une catégorie est appelée la catégorisation.

### 2.3.1 *La notion de catégorie*

#### 2.3.1.1 *Définition*

Le propre de tout être humain est de segmenter son environnement, *i.e.* d'en regrouper des objets n'étant pas identiques suivant un système de classification par lui-même élaboré (Rosch and Lloyd, 1978, p. 1). On appelle catégorisation, l'action consistant à regrouper des objets du monde physique considérés comme équivalents, et catégorie, l'entité mentale contenant le groupe d'objets ainsi rassemblés.

D'un point de vue écologique, la catégorisation est un processus essentiel. Nous sommes constamment en train de catégoriser l'environnement, et devons être à même, à tout moment, de prendre une décision sur l'appartenance catégorielle d'un objet. Ce processus est adaptatif. La prise de décision est toujours fonction du sujet, d'une situation et d'un contexte. Ainsi, un même objet, perçu à deux moments distincts, pourra être affecté à des catégories différentes.

(Anderson, 1991) propose trois exemples de manifestations quotidiennes des catégories :

- le langage : Le langage est le lieu, par excellence, des catégories. Catégoriser, c'est considérer un objet comme un élément distinct du monde. Cette distinction s'accompagne généralement (pas toujours) d'une désignation. C'est l'essence même des processus d'identification que de chercher à nommer les objets, une fois qu'ils ont été isolés. Il est raisonnable de penser que, de la même manière, une catégorie possède un label associé. On parlera alors de catégorie sémantique. Cette relation entre langage et catégorie nourrit le débat sur l'universalité de la caté-

gorisation. L'opération permanente consistant à isoler un objet, et à lui donner un nom, est à l'opposé du « geste adamique » (d'Adam, le premier homme révélé dans la bible), *i.e.* l'attribution libre d'un nom, hors toute influence et/ou contexte. La langue est un code partagé par une communauté. Dans une certaine mesure, ce code, *i.e.* le sens donné aux mots, peut varier suivant les groupes de cette communauté. Ainsi, catégoriser ne dépend pas seulement d'une réalité physique du monde, mais également d'un contexte socioculturel. La catégorisation peut être vue comme une action intermédiaire entre, d'une part, l'organisation d'une connaissance individuelle résultant d'une expérience sensorielle personnelle, d'autre part, la constitution d'une représentation collective pouvant être partagée par le biais d'un langage commun (Dubois et al., 2006).

- le regroupement par similarité des caractéristiques : Nous sommes capables de regrouper des objets possédant des caractéristiques similaires, et ce, même si ces objets nous sont inconnus (Fried and Holyoak, 1984).
- le regroupement par similarité fonctionnelle : Nous sommes capables de regrouper des objets possédant des fonctions similaires, et ce, même si ces objets ont des caractéristiques distinctes. Par exemple : deux hommes descendant d'un camion de pompier pour éteindre un feu de forêt seront catégorisés comme pompiers, ce indépendamment des tenues qu'ils portent. Quand nous parlons de la catégorisation comme du processus de discréétisation du monde réel, ce « monde réel » englobe et la réalité des faits physiques, et la réalité des faits sociologiques.

### 2.3.1.2 *La nature des catégories*

Toute opération permettant de « voir un objet comme étant ... » plutôt que de simplement « voir un objet » relève de la catégorisation. Tous les objets peuvent être catégorisés, quelle que soit leur nature (Goldstone and Kersten, 2003). Reconnaître un animal comme étant un éléphant est un acte catégoriel. Identifier qu'un morceau de musique est le premier mouvement d'une sonate, et qu'il est issu de la période classique, relève également d'un processus de catégorisation. La catégorisation intervient donc sur des objets de différentes natures. On distingue généralement trois types de catégories :

- catégories de concepts naturels : elles regroupent des objets concrets existant à l'état naturel (animaux, fleurs, *etc.*) ;
- catégories de concepts artificiels : elles regroupent des objets concrets fabriqués par l'homme (voitures, outils) ;

- catégories de concepts abstraits : elles regroupent des objets abstraits qui ne sont pas ancrés dans une réalité physique (art, stratégie, sentiments).

Ces trois types de catégories groupent des objets sur la base de leurs ressemblances. Pour les catégories de concepts naturels et artificiels, ces ressemblances s'établissent, entre autre, à partir de leurs caractéristiques physiques. Pour les catégories de concepts abstraits, ces similarités relèvent d'attributs cognitifs de plus haut niveau.

Qu'ils soient abstraits ou concrets, ces objets ont une existence avérée, *i.e.* indépendante du contexte. Cependant, certaines situations particulières poussent à grouper des objets parfaitement dissimilaires, en fonction d'un contexte situationnel, *e.g.* la liste de courses. Les catégories inhérentes à de tels groupements sont nommées *ad hoc* (Barsalou, 1983) :

- catégories de concepts *ad hoc* : elles regroupent des objets afin de répondre à un besoin spécifique.

### 2.3.2 *Le processus de catégorisation*

#### 2.3.2.1 *Catégorisation et prédiction*

La structure catégorielle forme la base des ressources cognitives sur lesquelles nous nous appuyons afin d'isoler des objets du monde. Ce processus procède de deux mécanismes :

- mécanisme inductif : associer un objet à une catégorie sur la base des propriétés perçues de ce dernier ;
- mécanisme déductif : associer à un objet les propriétés de la catégorie à laquelle il appartient.

Le mécanisme déductif nous permet de généraliser nos connaissances, *i.e.* d'inférer les propriétés d'un objet sans pour autant les avoir perçues. Ces propriétés transmises peuvent être physiques ou conceptuelles. Exemple : la vue seule de la croupe de l'animal nous fait « déduire » cheval. Autre exemple : le bourdonnement d'un insecte laissant supposer la présence d'une guêpe nous fait « déduire » danger. On le voit, le mécanisme déductif nous permet d'aller au-delà de l'information perçue, mais peut également mener à des erreurs d'interprétation. Notre capacité d'adaptation est très liée à ce mécanisme.

#### 2.3.2.2 *Catégorie et langage*

Comme nous l'avons vu (cf. Section 2.3.1.1), catégorie et langage relèvent d'un même principe. En attribuant le même nom à des objets distincts nous les regroupons *de facto* dans la même catégorie.

Les catégories n'étant pas accessibles directement par l'expérimentateur, l'analyse linguistique des descriptions verbales est un moyen d'objectiver les représentations mentales d'un individu (cf. Section 2.7.2.1).

L'analyse linguistique, comme outil d'approche des processus de catégorisation des sons, a été particulièrement étudiée par Dubois et ses collaborateurs (Dubois, 2000; Dubois et al., 2006; Guastavino, 2006; Rimbault and Dubois, 2005). Leurs travaux montrent entre autre que, contrairement à ce qui se constate dans le domaine de la vision, il n'existe pas en audition (ainsi qu'en olfaction (Dubois, 2000)) de consensus entre les sujets sur le vocabulaire à utiliser pour décrire les phénomènes sonores. L'analyse est ainsi pratiquée sur des descriptions verbales ayant la forme de phrases longues et complexes, plutôt que de mots ou mots+adjectifs isolés. Étudier la construction de ces phrases (nom+adjectif+verbe) permet de se renseigner sur les indicateurs et les processus inhérents au groupement catégoriel (Dubois, 2000; Guastavino, 2006).

Par ailleurs, le langage étant un élément partagé par un groupe de personnes semblables, l'analyse linguistique permet de faire un lien entre la description de l'expérience sensible d'un individu, et les représentations mentales collectives, partagées par sa communauté (Dubois, 2000).

### 2.3.2.3 Catégorisation et identification

On distingue généralement les processus de catégorisation, et les processus d'identification. La catégorisation, *i. e.* regrouper des objets en classes d'équivalences, est un processus pouvant s'opérer dans un cadre non-supervisé, *i. e.* sans avoir besoin de nommer les classes. L'identification, elle, est nécessairement supervisée, *i. e.* nous ne pouvons identifier des objets qu'à partir des catégories que nous connaissons. Ainsi, un enfant qui voit pour la première fois des hyènes dans un zoo comprend que ces animaux appartiennent à la même espèce. Mais il est très probable qu'il les identifie comme étant « une sorte de chien ».

Les deux processus sont pourtant très liés (Goldstone and Kersten, 2003), l'identification pouvant être vue comme un cas particulier de la catégorisation (Schyns, 1998). Nos travaux ne requérant pas de distinguer ces deux mécanismes, nous considérons, dans ce document, la catégorisation au sens large, incluant l'identification.

### 2.3.3 Organisation de la structure catégorielle

Le cerveau doit en permanence faire sens d'une information riche et variée, et ce, de manière productive. Afin de satisfaire à cette exigence d'efficacité, l'organisation de la structure catégorielle doit répondre à deux grands principes (Rosch and Lloyd, 1978, p. 29):

- l'économie cognitive ;
- la redondance structurelle.

#### 2.3.3.1 *L'économie cognitive*

La catégorisation doit fournir un maximum d'informations pour un minimum d'efforts. C'est pourquoi la logique catégorielle prend en compte le contexte sensoriel. En résumé, le traitement de l'objet s'opère et par rapport à lui, et par rapport au traitement des objets perçus et catégorisés simultanément. Comme énoncé par D. Dubois (Dubois, 1991, p. 33):

« Catégoriser un stimulus signifie le considérer dans la finalité de cette catégorisation, non seulement comme équivalent des autres stimuli de la même catégorie, mais également différent des stimuli qui n'appartiennent pas à cette catégorie. »

Du principe d'économie cognitive, il découle que la catégorisation de l'objet n'est pas une catégorisation dans l'absolu. Elle ne dépend pas uniquement de l'observation des propriétés particulières de l'objet, mais également du contexte dans lequel il est appréhendé.

#### 2.3.3.2 *La redondance structurelle*

L'ensemble des objets physiques ne vit pas dans un espace fini, identifié, et dont les valeurs seraient équiprobables. Le monde ne se résout pas à des paramètres dimensionnés, indépendants et manipulables, comme dans le cadre d'études en laboratoire. Au contraire, il peut exister des discontinuités saillantes entre objets, de même que ces objets peuvent être liés entre eux par des patterns de co-occurrence de propriétés (exemple : un chien possède "quatre pattes et un museau" plus souvent que "deux pattes et un museau"). Ces discontinuités et corrélations présentes dans les propriétés perçues étayent la structure catégorielle de notre représentation mentale, et gouvernent ainsi le processus de catégorisation.

#### 2.3.3.3 *Catégorie et abstraction*

Pour des catégories de concepts concrets (naturels ou artificiels), Rosch propose de voir la structure catégorielle suivant deux axes (Rosch and Lloyd, 1978, p. 30-41):

- *axe vertical* : L'axe vertical fixe l'organisation hiérarchique des catégories, et permet d'appréhender l'imbrication de ces catégories les unes par rapport aux autres. Ce faisant, il en dresse la taxonomie, les catégories de haut niveau concernant des concepts ayant un haut niveau d'abstraction (Trafic), et incluant un grand

nombre de sous catégories, et les catégories de bas niveau concernant des concepts liés à des objets plus spécifiques (Voiture), incluant peu de sous catégories. Ainsi, plus le niveau d'abstraction est grand, plus les similitudes entre instances<sup>5</sup> de mêmes catégories (intra-catégorielles), ou de catégories distinctes (inter-catégorielles) sont faibles. Inversement, plus le niveau d'abstraction est faible, plus les similitudes intra- et inter-catégorielles sont élevées. Rosch décompose cette dimension verticale en trois niveaux d'abstraction (cf. Figure 3) : superordonné, base, subordonné. Le niveau superordonné regroupe les catégories à haut niveau d'abstraction. Les périmètres de ces catégories sont larges (Mobilier, Véhicule, etc.) *i.e.* les objets qu'elles regroupent peuvent être très distincts. Le niveau subordonné regroupe les catégories à bas niveau d'abstraction. Les périmètres de ces catégories sont plus précis (Chaise longue, Cabriolet, etc.), et les instances qu'elles contiennent sont nécessairement très similaires. On notera ici que les instances de la classe Cabriolet présentent plus de propriétés communes avec les instances de la classe Berline, qu'il n'en existe entre les instances des classes Mobilier et Véhicule. Au niveau de base, les instances d'une même catégorie partagent encore beaucoup de propriétés, tout en maintenant une dissimilarité inter-catégorielle élevée.

- *axe horizontal* : L'axe horizontal fixe, lui, l'organisation « géographique » des catégories, et permet d'appréhender, d'une part, les périmètres de ces catégories au sein d'un même niveau d'abstraction, d'autre part, la typicalité des instances contenues dans une même catégorie (cf. Section 2.3.3.4). Les catégories ne sont pas des entités strictement discrètes, et les propriétés des objets qu'elles regroupent peuvent se retrouver sur d'autres objets, correspondant à d'autres catégories. Ainsi, les frontières entre les différentes catégories ne sont pas figées, et peuvent même se recouvrir.

On remarque que l'organisation de nos connaissances, ainsi représentée par la structure catégorielle, forme un miroir de la redondance structurelle inhérente au monde physique. Selon (Rosch and Lloyd, 1978, p. 28), c'est le niveau de base qui rend compte au mieux de cette structure. Il s'agit d'un niveau privilégié, proposant le meilleur compromis entre le nombre de catégories, et l'information qu'elles véhiculent. Il permet d'obtenir le maximum d'informations au prix d'un moindre effort cognitif.

---

<sup>5</sup> Nous employons le terme objet pour désigner l'objet physique, et le terme instance pour désigner une représentation mentale de ce dernier, représentation conservée en mémoire dans la catégorie du concept correspondant. A noter que l'instance d'une catégorie d'un niveau d'abstraction élevé (Trafic), peut elle-même être une catégorie (Voiture).

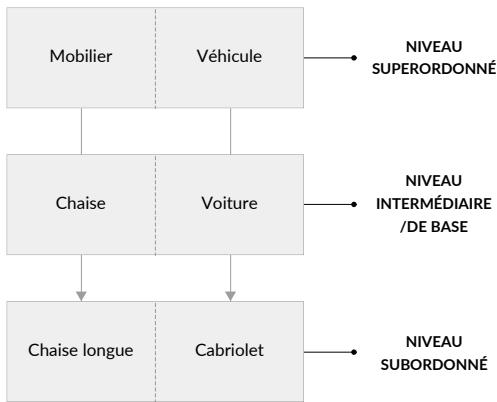


FIGURE 3 : Les trois niveaux d'abstraction de l'axe vertical de la structure catégorielle.

Cependant, si cette vision bidimensionnelle de la structure catégorielle est adaptée aux catégories de concepts concrets, elle est moins pertinente dans le cas de catégories de concepts abstraits, comme les catégories sociales (Dubois, 1991, p. 72-88). Considérer l'organisation catégorielle comme le reflet du monde perçu vaut surtout en ce qui concerne le monde physique.

#### 2.3.3.4 *La notion de typicalité*

Une notion clef, dans les processus catégoriels, est la typicalité. Toutes les instances d'une catégorie ne sont pas égales. Il y a une gradation dans l'appartenance catégorielle. Certaines instances, partageant les propriétés dominantes d'une catégorie, sont considérées comme très représentatives de celle-ci. D'autres, n'en possédant que peu d'attributs caractéristiques, ont une appartenance moins marquée.

Le fait qu'il existe des objets plus typiques que d'autres est un constat empirique (Mervis and Rosch, 1981; Rosch and Lloyd, 1978, p. 37), établi à partir d'échelles de jugements (quel est l'objet le plus typique ?), ou au moyen d'épreuves de vérification chronométrées (un chien est un animal : vrai ou faux ?) (Dubois, 1991, p. 41).

La typicalité agit sur différents processus de traitement (Houix, 2003; Mervis and Rosch, 1981, p. 51):

- *temps de traitement* : un objet typique est catégorisé plus rapidement qu'un objet moins typique ;
- *apprentissage* : un enfant bâtit sa structure catégorielle en commençant d'abord par des objets typiques ;
- *ordre mémoriel* : lorsqu'un sujet énumère les membres d'une catégorie, il commence par les membres typiques ;

- *langage* : certains termes du langage courant sont directement connectés à la typicalité, ainsi, un « moineau est un *vrai* oiseau », alors qu'un « pingouin est une *sorte* d'oiseau » (Mervis and Rosch, 1981);
- *asymétrie des jugements de ressemblance* : il existe un phénomène d'attraction autour des objets typiques d'une catégorie. Si nous considérons la catégorie couleur, l'orange ressemble plus au rouge que le rouge ne ressemble à l'orange. L'asymétrie dans les processus perceptifs a été extensivement étudiée (Krumhansl, 1978; Tversky, 1977).

#### 2.3.4 Théories de la catégorisation

##### 2.3.4.1 Théorie classique

Suivant les principes de la théorie classique, l'appartenance d'un objet à une catégorie se fait sur la base de règles. L'objet doit posséder un certain nombre de propriétés afin d'être assimilé à une catégorie, propriétés qui sont inhérentes à ladite catégorie.

Cette approche logique (dite aussi approche par règles), qui soutient que toutes les instances d'une catégorie doivent partager des propriétés communes, est aujourd'hui critiquée. Selon (Goldstone and Kersten, 2003) :

- l'appartenance catégorielle n'est pas figée : deux personnes peuvent catégoriser un même objet de deux manières. Qui plus est, une même personne peut modifier sa stratégie de catégorisation (McCloskey and Glucksberg, 1978);
- les objets ne partagent pas le même degré d'appartenance : comme vu précédemment (cf. Section 2.3.3.4), tous les objets à l'intérieur d'une catégorie ne sont pas égaux, certains étant plus typiques que d'autres.
- il est difficile de définir des règles d'appartenance : définir des catégories comme « célibataire » nécessite d'élaborer des stratégies afin d'isoler les cas « enfant », « veuf » ou encore « pape » qui, intuitivement, n'ont rien à voir avec « célibataire ».

De plus, (Houix, 2003, p. 49) souligne que, dans l'approche logique, les classes subordonnées héritent des règles d'appartenance des classes superordonnées, niant ainsi le fait qu'il existe des niveaux d'abstractions privilégiés.

##### 2.3.4.2 Théorie prototypique

Un autre parti consiste à envisager la catégorie non plus comme relevant de règles, mais comme découlant des ressemblances (ou « air de

famille ») liant ses instances (Ludwig, 1953). Partant de cette idée, E. Rosch et B. B. Lloyd (Rosch and Lloyd, 1978) ont formalisé la théorie prototypique. Suivant celle-ci, la catégorie est définie par rapport aux objets qu'elle englobe, et non dans le but d'englober ces objets.

Pour discriminer les catégories, Rosch propose de ne pas raisonner en terme de frontières, mais plutôt de décrire chaque catégorie par un nombre de cas non ambigus (*clear case*) (Rosch and Lloyd, 1978, p. 36). Toutes les instances d'une catégorie ne sont pas également représentatives de cette dernière. Il a été montré que des sujets peuvent très bien s'accorder sur la typicalité d'un objet par rapport à une catégorie, tout en n'étant pas d'accord sur les frontières de celle-ci (Rosch, 1975; Rosch and Lloyd, 1974). Les cas non ambigus peuvent être vus comme les instances les plus typiques de la catégorie. Le terme prototypique, caractérisant la théorie, vient de l'assertion que, parmi ces cas non ambigus, il en existe un, le prototype, plus représentatif que les autres, et qui forme le noyau de la catégorie.

Ainsi, les catégories sont structurées en interne, en référence à un prototype, *i.e.* l'instance possédant les attributs typiques de celle-ci. L'appartenance d'un objet à une catégorie dépend alors de la ressemblance qu'entretient ce dernier avec le prototype. Plusieurs propositions ont été faites afin de définir le prototype d'une catégorie : Pour Tversky (Tversky, 1977), l'élément prototype est celui dont la somme des similarités avec les autres éléments de la catégorie est la plus élevée. Pour Rosch (Rosch and Mervis, 1975), il s'agit de l'instance possédant le plus de propriétés en commun avec les instances de la catégorie, et le moins de propriétés en commun avec les instances des catégories externes. Autrement dit, la typicalité d'un élément d'une catégorie s'évalue à la fois en fonction de son degré d'appartenance à celle-ci, et de son degré de différenciation vis à vis des autres catégories.

Toutes ces approches supposent que le prototype est la représentation mentale d'un objet réel. Cependant, le prototype peut être aussi vu comme un objet stéréotypé, un assemblage des attributs les plus représentatifs de la catégorie. Ainsi, en se limitant à l'observation d'attributs vivant dans un espace métrique, (Reed, 1972; Rosch et al., 1976) ont montré que le prototype est un centroïde, une instance définie comme étant la moyenne des attributs des instances de la catégorie. Cette distinction entre le meilleur représentant et le représentant moyen est analogue à celle faite dans les algorithmes de *clustering* entre le médoïde (ici prototype) et le centroïde.

Cette théorie prototypique de la catégorisation, bien que se basant sur des faits expérimentaux, est avant tout une vision pratique, qui n'a pas été clairement définie et dont l'implication dans les processus de catégorisation reste floue (Rosch and Lloyd, 1978, p. 36-40) (Dubois, 1991, p. 49-54).

### 2.3.4.3 Théorie des exemplaires

La théorie des exemplaires nie l'existence d'un prototype. Au contraire, elle propose que la catégorie soit représentée par l'ensemble des instances (exemplaires) la constituant, en tenant compte de leurs degrés de typicalité respectifs (Medin and Schaffer, 1978; Nosofsky, 1986, 1992). Ainsi, les mécanismes déductifs (cf. Section 2.3.2.1) peuvent profiter de tous les exemplaires de la catégorie afin d'inférer les propriétés des objets perçus. En analyse automatique, la philosophie de l'approche par les exemplaires est proche de celle des cartes auto-organisées (*Self organized map*) (Kohonen, 1995), l'organisation du réseau et de ses noeuds étant réactualisée en fonction des propriétés de tous les items.

Plusieurs versions de cette théorie existent, entre autres, le modèle de contexte (Medin and Schaffer, 1978), le modèle MINERVA (Hintzman, 1986) et le modèle de contexte généralisé (Nosofsky, 1986, 1992). Dans tous les cas, l'appartenance catégorielle se fait sur la base de la somme pondérée des similarités entretenues entre l'objet à catégoriser, et les exemplaires de la catégorie, la pondération visant à favoriser les exemplaires les plus proches de l'objet perçu.

L'approche par les exemplaires lève cependant deux questions (Goldstone and Kersten, 2003):

- Comment justifier que nombre d'études montrent que l'appartenance catégorielle s'effectue via une comparaison à un prototype ?

Dans certains cas, calculer la somme pondérée des similarités entretenues entre un objet à catégoriser, et les exemplaires d'une catégorie, équivaut à calculer la similarité entre l'objet et le représentant moyen des exemplaires. L'existence d'un prototype n'est alors qu'un artefact. Notons par ailleurs que la théorie des exemplaires ne nie pas l'existence d'un gradient de typicalité entre les exemplaires d'une catégorie.

- Comment justifier que le principe d'économie cognitive reste valide, si le cerveau utilise l'ensemble des items d'une catégorie pour la représenter ?

Bien qu'on puisse montrer, dans le cas d'un bruit blanc, que le cerveau peut stocker en mémoire la totalité d'un signal sur une période de plusieurs semaines (Agus et al., 2010), il est évidemment peu probable que, pour chaque catégorie, le cerveau sauvegarde la totalité des exemplaires. Deux phénomènes peuvent alors intervenir : soit il existe un processus de sélection des exemplaires (Palmeri and Nosofsky, 1995), soit les exemplaires émanant d'une même entité physique (deux exemples de pigeon) sont résumés par un même représentant (Barsalou et al., 1998).

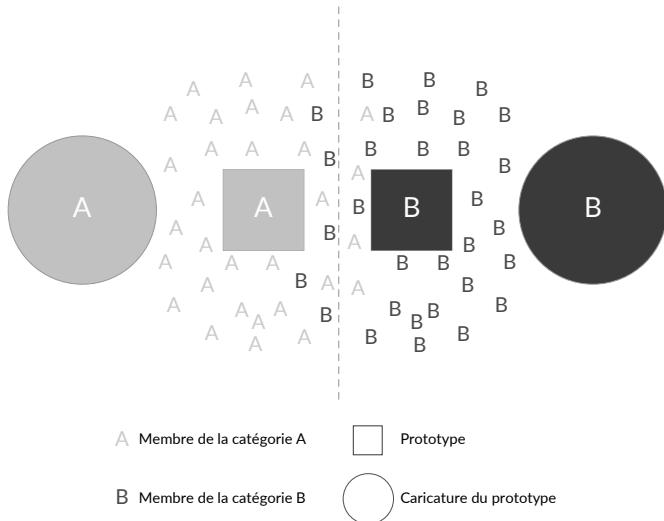


FIGURE 4 : Prototype et caricature, d'après (Davis and Love, 2010).

#### 2.3.4.4 Théorie des frontières

En opposition directe avec le modèle prototypique, la théorie des frontières représente les catégories par leur périphérie. L'importance des frontières est un fait expérimental mis en avant par plusieurs études. Notamment (Davis and Love, 2010) qui montre que, dans le cas de deux catégories proches, l'instance représentative putative n'est pas le prototype, mais une caricature de ce dernier. La caricature du prototype est une entité dont les propriétés ont été distordues afin qu'elle s'éloigne de la frontière séparant les deux catégories, distorsion qui peut substantiellement éloigner la caricature des instances de sa catégorie (cf. Figure 4).

(Goldstone and Kersten, 2003) souligne que les théories du prototype et des frontières peuvent se compléter :

- pour des catégories très éloignées, la distance au prototype (représentant moyen des instances d'une catégorie) est une information suffisante pour associer l'objet à une catégorie ;
- pour des catégories très similaires, l'appartenance catégorielle, afin d'être efficace, doit s'appuyer sur une information plus spécifique, à savoir la frontière.

Par ailleurs, la théorie des frontières n'envisage pas la périphérie d'une catégorie comme quelque chose de statique. Il existe une frontière a priori, certes, mais cette dernière peut bouger en fonction du contexte.

### 2.3.5 Catégorisation et contexte sensoriel

La catégorisation d'un objet est dépendante de l'environnement dans lequel il est perçu, c'est à dire des sons co-occurants. La manière dont ce contexte sensoriel influe sur l'identification est cependant méconnue.

(Ballas and Howard, 1987) propose de voir l'environnement sonore comme une forme de langage à part entière. Comme pour la parole, les propriétés sémantiques des sons, à savoir la nature perçue de la source émettrice (*trafic, humain, etc.*), mais aussi la manière avec laquelle cette source s'accorde avec les autres sons présents, influent significativement sur leur perception. Pour la parole, il est connu que le contexte grammatical ainsi que le contexte sémantique (ici le sens de la phrase) ont une incidence sur la reconnaissance d'un mot (Bilger et al., 1984). Au terme de leur comparaison, les auteurs concluent que la source d'un son est plus facilement identifiable, si ce dernier est en adéquation avec son contexte environnemental (*e.g.* identifier un *oiseau* dans un environnement de *parc*).

Dans (Ballas and Mullins, 1991), ils montrent même que l'effet du contexte est suppressif, à savoir qu'un contexte inadéquat a tendance à diminuer la capacité d'un individu à identifier un son. Aucun effet bénéfique n'est cependant observé pour un contexte adéquat.

En appliquant ces idées à la détection automatique d'événements sonores, (Niessen et al., 2008) montrent que la modélisation du contexte environnemental, *i.e.* la probabilité d'entendre un événement dans un type d'environnement donné et *a fortiori* la probabilité de co-occurrence des événements, permet d'améliorer les performances de détection.

Le fait est, cependant, que l'inverse a aussi été prouvé, *i.e.* que le caractère incongru d'un son par rapport à son environnement peut faciliter son identification. Dans une étude approfondie, (Gygi and Shafiro, 2011) mettent en évidence que la source d'un événement est plus facilement identifiable, si ce dernier apparaît dans un environnement où il n'est pas « censé » apparaître. Il est ainsi plus facile d'isoler et d'identifier un son de canard dans un aéroport, qu'un son d'avion. Les auteurs appellent ce phénomène : « l'Avantage de l'Incongrue » (AI). Ils montrent, par ailleurs, que l'AI dépend du rapport entre le niveau de l'événement, et celui du fond sonore (*background*). Il n'opère pas pour des rapports inférieurs à -7.5dB, et son effet est constant pour des rapports supérieurs à ce seuil.

Une étude menée dans le cadre de cette thèse tend à montrer que l'AI dépend également de la structure temporelle de l'environnement. Elle est présentée en annexe C.

### 2.3.6 Similarité et catégorisation

Comme vu précédemment, au moment notamment d'évoquer les théories prototypique, des exemplaires ainsi que des frontières, similarité et catégorisation apparaissent comme des concepts très proches, la catégorie étant un groupement d'objets similaires pour l'individu réalisant le groupement.

Les ressemblances entretenues par les objets d'une même catégorie sont globales, liées à la fois aux propriétés physiques desdits objets, mais également à leurs propriétés sémantiques et affectives, ces dernières relevant de la connaissance de l'individu. De plus, le processus de catégorisation est lui-même contextuel, relevant à la fois de la diversité des stimuli en présence, mais également d'un objectif à atteindre, comme c'est notamment le cas pour les catégories *ad hoc*.

La similarité, quant à elle, est habituellement comprise comme une notion beaucoup plus stable, n'impliquant que la comparaison des propriétés intrinsèques des objets (forme, taille, fréquence, durée *etc.*). Cependant, il faut noter que la similarité, comme la catégorisation (cf. Section 2.3.5), dépendent du contexte sensoriel d'exposition, *i.e.* du nombre et de la nature des objets présents lors de la mesure. (Tversky, 1977; Tversky and Gati, 1978) ont montré que les propriétés physiques rentrant en compte dans les mesures de similarité diffèrent suivant la diversité du corpus d'objets étudiés.

On fait donc la distinction entre ces deux mécanismes, la similarité apparaissant comme un processus dirigé par les données (stimuli), la catégorisation étant, elle, dirigée par les connaissances de l'individu et un contexte (Houix, 2003, p. 59). Il est néanmoins possible de réunir ces deux notions en considérant la similarité comme un des mécanismes (essentiel) du processus de catégorisation (Houix, 2003, p. 61-65).

### 2.3.7 Discussion

De cette section nous retenons que les représentations mentales s'organisent autour de catégories. Pour les catégories de concepts concrets, la structure catégorielle semble être taxonomique.

Nous constatons également la centralité des processus de catégorisation dans l'interaction que nous avons avec le monde. Toute réponse qui suppose l'identification d'un objet (concret ou abstrait) relève d'un processus catégoriel. Tout processus catégoriel permet d'inferer des propriétés non-observées.

Enfin, nous notons l'importance du contexte dans la catégorisation, celui-ci pouvant l'influencer (cf. Section 2.3.5), voire la déterminer (catégorie *ad hoc*; cf. Section 2.3.1.1).

Concernant les connaissances liées à une catégorie, nous nous rapprochons des vues proposées par les théories du prototype et des

exemplaires. Nous admettons que l'information relative au concept d'une catégorie est une forme résumée des connaissances relatives à ses membres.

Là où nous divergeons des théories ci-devant évoquées (cf. Section 2.3.4), c'est sur la nature des représentations. En effet, toutes ces théories s'appuient sur une vision classique de la cognition, et assument que l'information catégorielle est amodale. Pour notre part, nous nous plaçons dans une vision ancrée de la cognition, qui reconnaît aux représentations mentales une dimension modale. Il convient ici de détailler les liens entre catégorisation et cognition ancrée.

Une catégorie est la représentation mentale d'un concept. Dans une approche ancrée de la cognition, ce concept est un simulateur. La catégorie regroupe ainsi les ressources mentales sur lesquelles s'appuie le simulateur. Autrement dit, elle est formée par l'ensemble des symboles perceptifs liés au concept (Barsalou, 2003b). Cette approche ne change en rien les propriétés des processus catégoriels évoqués précédemment (cf. Section 2.3.2), à savoir :

- mécanisme inductif : lorsqu'un objet est perçu, une image mentale se forme. L'appartenance catégorielle de cet objet dépend ainsi des symboles perceptifs extraits de l'image ;
- mécanisme déductif : lorsqu'une catégorie est sollicitée par un symbole perceptif, son simulateur s'enclenche afin de générer une image simulée d'une instance de la catégorie. Cette image infère certaines des propriétés non-perceptibles de l'objet, en fonction des connaissances mémorisées ;
- influence du contexte : la simulation cognitive est guidée par les processus perceptifs. L'éventail des instances pouvant potentiellement être générées est ainsi contraint par l'information sensorielle montante. L'image simulée est alors, par définition, contextualisée par rapport à l'environnement. On parlera d'une représentation située (Barsalou, 2003b; Barsalou and Wiemer-Hastings, 2005).

La simulation est l'étape centrale de la catégorisation, celle par laquelle se fait le lien entre les processus perceptifs et cognitifs. Enfin, il est à noter que la notion de représentation située atténue celle de typicalité (cf. Section 2.3.3.4). En effet, dans la vision ancrée de la cognition, la typicalité ne peut se comprendre dans l'absolu, mais toujours par rapport à un contexte.

## 2.4 L'ÉTUDE PSYCHOLOGIQUE DU SYSTÈME AUDITIF

### 2.4.1 La psychoacoustique

Dans le domaine de la psychologie expérimentale, la psychophysique est une discipline dont l'objet est de trouver des corrélats entre des mesures de quantités physiques, objectives, et décrites sur des dimensions établies par les sciences de la nature, et les sensations qu'elles produisent chez l'homme.

La psychoacoustique est une branche de la psychophysique qui applique au domaine de l'acoustique les notions et méthodes de la psychophysique. Formellement, il s'agit d'obtenir du sujet une mesure subjective quantitative d'une variable physique donnée, comme par exemple le niveau sonore. Les sensations du sujet sont obtenues à partir de protocoles fermés, *i.e.* où les réponses des sujets sont contraintes. On considère deux types de protocoles expérimentaux (Guastavino, 2003, p. 29-30).

- seuil de détection : considérant un stimulus dont on fait varier une dimension physique particulière (fréquence, niveau, *etc.*), il s'agit d'établir le seuil à partir duquel le sujet ne fait plus la distinction entre les stimuli. Ces derniers sont habituellement présentés soit séquentiellement, soit par deux ou par trois (2 sons identiques, 1 son différent).
- estimation de grandeur : il s'agit de demander au sujet d'évaluer la grandeur d'une dimension physique particulière, soit dans l'absolu, en utilisant une échelle graduée ou continue, soit par comparaison à une référence (*e.g.* dans le cas du niveau sonore : est ce que le stimulus A est plus fort ou moins fort que le stimulus B).

Dans tous les cas, la psychoacoustique cherche à établir le lien entre une dimension physique, et sa perception par le sujet. L'objectif ultime étant de proposer une fonction mathématique permettant, pour une dimension physique donnée, de rendre compte de l'intensité perçue en fonction de l'intensité mesurée. L'approche psychoacoustique traitant de dimensions « bas niveau » (fréquence, niveau, *etc.*), définissables analytiquement, et finement contrôlées, elle privilégie les études en laboratoire, et utilise comme stimuli des sons de synthèse, bien souvent purs ou complexes<sup>6</sup>.

Le point fondamental de cette approche est l'hypothèse d'une relation directe entre le stimulus et la réponse du sujet, passant outre les processus cognitifs internes de ce dernier, lesquels, dans ce cas, sont vus comme des « boîtes noires ».

---

<sup>6</sup> Un son pur est un son composé d'une seule sinusoïde, *i.e.* possédant une seule fréquence. Un son complexe est, lui, composé de plusieurs composantes fréquentielles

#### 2.4.2 La psychologie cognitive

La psychologie cognitive s'inscrit dans le mouvement *cognitiviste*. Elle est l'une des méthodes scientifiques, avec la psycholinguistique, les neurosciences ou encore l'intelligence artificielle (pour ne citer qu'elles), permettant d'étudier expérimentalement le fonctionnement du système cognitif.

En psychologie cognitive, le sujet n'est pas traité comme un système entrée/sortie, dont l'entrée serait les organes sensoriels, et la sortie une réponse directement observable et quantifiable. La réponse du sujet est supposée tenir compte non seulement des traitements perceptifs, mais aussi des représentations internes, celles-ci résultant, d'une part, de la mémoire individuelle de l'individu, sa relation sensible au monde, d'autre part, de sa mémoire collective, construite à travers le développement des connaissances partagées. L'individu est ici considéré comme un tout, défini entre autre par une expérience passée, une culture, une sphère sociale et une activité.

La psychologie cognitive envisage l'ensemble des étapes du traitement auditif de manière globale, et tente, entre autre, de faire le lien entre l'information sensorielle (modale) et l'information symbolique (sémantique) (McAdams and Bigand, 1994). A ce titre, elle interroge à la fois les processus perceptifs et les processus cognitifs. La distinction est faite d'ailleurs entre les approches cognitivistes, qui s'intéressent plus particulièrement aux processus montants (*bottom-up*, cf. Section 2.5.2) relatifs au traitement de l'information perçue, et les approches cognitives, qui interrogent, avant tout, les processus descendants (*top-down*, cf. Section 2.5.2) liés aux représentations internes, ainsi qu'au contexte (Guastavino, 2003, p. 34).

Trois remarques concernant la psychologie cognitive :

- les processus cognitifs ne sont pas considérés comme des « boîtes noires », mais intégrés comme faisant partie intégrante du système auditif ;
- il n'existe pas de relation directe entre le stimulus et la réponse, cette dernière étant déterminée non seulement par la nature du stimulus, mais également par les connaissances du sujet, le contexte d'écoute, et les interactions multi-sensorielles putatives ;
- les connaissances du sujet (*i. e.* ses représentations internes), non directement observables, restent néanmoins accessibles au chercheur par le biais d'expériences d'objectivation (voir section 2.7.2.1).

D'un point de vue expérimental, la psychologie cognitive emprunte une méthodologie moins fermée que celle de la psychoacoustique, considérant les aspects quantitatifs, mais également qualitatifs, des réponses du sujet (Maffiolo, 1997, 1999).

Enfin, le fait que la psychologie cognitive interroge le fonctionnement des processus cognitifs, relatifs à une information sémantique, implique qu'elle utilise des stimuli réels (enregistrés, dans le cas des sons), des stimuli de synthèse simples n'ayant aucune valeur sémantique.

#### 2.4.3 *Paradigme de la psychologie cognitive*

La psychologie cognitive conçoit l'individu comme un système de traitement de l'information. Elle admet que ce dernier adopte une stratégie afin d'optimiser le traitement des stimuli. Cette stratégie est déterminée par la nature du stimulus, mais aussi par son contexte, et par les connaissances *a priori* du sujet.

Maffiolo (Maffiolo, 1999) propose une présentation des présupposés sur lesquels repose la psychologie cognitive (cf. Figure 5). Ces présupposés sont résumés ci-après :

- le monde est discrétisé en dimensions ou propriétés issues de la physique, et considérées comme vraies ;
- ces dimensions ou propriétés peuvent être mesurées objectivement par des instruments. Elles rendent compte ainsi de la réalité ;
- le sujet intègre de manière séquentielle ces dimensions ou propriétés en fonction du contexte ;
- l'évaluation subjective du sujet est interprétée comme un décalage par rapport à la mesure objective considérée comme vraie.

Le paradigme de la psychologie cognitive repose ainsi sur la dualité communément acceptée entre objectivité et subjectivité, autrement dit, la différenciation opérée entre, d'un côté, le monde « réel » (et *a fortiori* les objets qui le composent) considéré comme « vrai », et que l'on peut décrire suivant des dimensions et propriétés physiques objectives, et, de l'autre côté, la perception du sujet, *i.e.* la représentation « biaisée » qu'il se fait du monde.

Au regard de ce paradigme, Maffiolo met en évidence quatre points discutables :

1. la pertinence des dimensions et propriétés physiques utilisées pour le découpage du monde ;
2. un traitement par les sujets tenant spécifiquement compte de ces dimensions ;
3. une séparation nette entre stimulus et contexte ;
4. le caractère subjectif du jugement humain comparé à l'objectivité d'un appareil de mesure.

Les points 1 et 2 se rejoignent. Les dimensions physiques utilisées pour décrire le monde sont établies par les sciences naturelles, *via* des procédés qui ne prennent pas en compte la perception du sujet. Dans le cas de l'analyse sensorielle, questionner la pertinence de ces dimensions, *i.e.* leur capacité à diriger notre relation sensible au monde, est naturel. Il est connu aujourd'hui que ces dimensions ne peuvent expliquer seules les processus perceptifs mis en œuvre, notamment pour ce qui est de l'audition et de l'olfaction (Dubois, 2000). On peut citer ici la capacité limitée des descripteurs psychoacoustiques comme la *loudness* (descripteurs établis à partir de stimuli synthétiques ; cf. Section 2.7.2) à caractériser l'intensité perçue, et encore plus, la gêne occasionnée par cette intensité, pour des stimuli réels enregistrés (cf. Section 1.2.2.1).

Concernant le point 3, bien que le paradigme suppose l'influence d'un contexte, ce contexte est envisagé comme une entité séparée du stimulus, un élément modificateur qui agirait sur la perception de ce dernier. Ce fait implique qu'il existe une représentation mentale de l'objet perçu (*i.e.* le stimulus), déconnectée de tout contexte d'exposition. Un objet n'étant jamais perçu de manière isolé, ce point peut, à juste titre, être considéré comme une hypothèse restrictive.

Enfin, le fait de considérer les réponses « subjectives » du sujet comme une déviation par rapport à une réalité physique externe implique que le monde est indépendant des représentations mentales que nous nous en faisons, et qu'il est possible d'en « trier » l'information sans tenir compte de la manière dont nous l'interprétons (Dubois, 2000). Cela suppose que l'existence des objets qui composent le monde est avérée, et indépendante de notre perception. Or, comme nous le verrons, la perception d'un objet varie d'un sujet à l'autre, que l'on considère son identification (quel est l'objet perçu ?), sa description (comment, et à partir de quels descripteurs l'objet est-il perçu ?), ou son interprétation (quel est l'effet de l'objet sur le sujet ?).

Questionner ce dernier point revient à interroger : « Est-ce que l'objet existe parce que je le perçois, ou est-ce que je le perçois parce qu'il existe ? »

#### 2.4.4 Reproduire l'environnement sonore

Reproduire un environnement en laboratoire peut avoir un impact non négligeable sur les qualités intrinsèques du son. Une reproduction stéréophonique ou monophonique n'est pas sans effet sur l'information spatiale, en particulier la position des sources et/ou les caractéristiques structurelles du lieu d'écoute.

Le problème d'une reproduction écologique des environnements sonores en laboratoire a été particulièrement étudié par Guastavino. (Guastavino and Cheminée, 2003; Guastavino and Katz, 2004; Guastavino et al., 2005). En comparant les descriptions verbales produites à

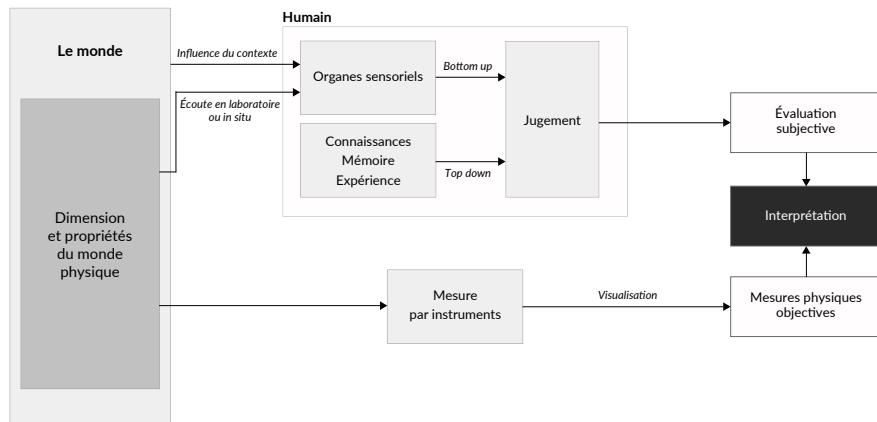


FIGURE 5 : Paradigme de la psychologie cognitive, d'après (Maffiolo, 1999)

la suite d'écoutes *in situ*, et d'écoutes en laboratoire, via des systèmes de reproduction stéréophoniques et multi-phoniques, (Guastavino et al., 2005) montre que les événements sonores peuplant les scènes sont décrits de la même manière, quel que soit le contexte d'écoute.

Cependant des différences apparaissent au niveau de la description des fonds sonores (*backgrounds*), entre les écoutes stéréophoniques, et les écoutes multi-phoniques *in situ*, suggérant de fait que le système de reproduction influe sur les processus cognitifs mis en œuvre. Des conclusions similaires sont faites dans (Guastavino and Katz, 2004), s'agissant cette fois de systèmes de reproduction mono, stéréo, et multi-phoniques.

#### 2.4.5 *Le Soundwalk*

Appliquée à la perception sonore, l'approche écologique introduite par Gibson (cf. Section 2.2.4) suggère de prendre en compte l'environnement et du sujet, et du stimulus auquel il est exposé. La démarche s'oppose aux méthodes expérimentales traditionnelles, celle de la psychophysique en particulier, sur l'aspect décontextualisant de l'écoute en laboratoire, qui affecte la perception du sujet contraint à un effort d'abstraction supplémentaire pour obtenir l'illusion de la réalité.

Nombre d'études désormais sont réalisées dans un cadre *in situ*. On parle d'ailleurs de *soundwalk*<sup>7</sup> pour désigner les expériences où le sujet est immergé dans l'environnement qu'il doit évaluer (Adams et al., 2008; Jeon et al., 2013).

La méthode des *soundwalk* permet entre autre :

- de contextualiser le sujet, à savoir, l'évaluer dans un environnement qu'il connaît (lieu de vie, de travail...);

<sup>7</sup> *Soundwalk* est un terme anglais introduit par R. Murray Schafer (Schafer, 1969) signifiant littéralement « marche sonore ». Ce terme étant couramment utilisé en français, il ne sera pas traduit dans ce document.

- d'évaluer l'environnement sonore, tout en maintenant actif les autres sens (vision, olfaction) ;
- d'éluder le problème de la reproduction des environnements sonores en laboratoire.

Malgré tout, les études *in situ*, bien que valides écologiquement, présentent elles aussi des inconvénients. Dans l'hypothèse où tous les sujets ne passent pas l'expérience en même temps, il est impossible de garantir à chacun les mêmes stimuli et/ou le même environnement. Inversement, dans l'hypothèse où tous les sujets passent l'expérience en même temps, il peut se poser des problèmes d'organisation, de nature à compromettre une égale réceptivité, disponibilité chez tous les sujets.

#### 2.4.6 Discussion

Nous reconnaissons le caractère restrictif de la psychoacoustique, comparé à la psychologie cognitive. Nous prenons acte, par ailleurs, des réserves émises par Maffiolo, 1999 quant aux présupposés sur lesquels repose le paradigme *cognitiviste*, en particulier en ce qui concerne la *séparation nette entre stimulus et contexte*. Cette vision ne peut s'envisager que si l'on considère des représentations mentales amodales. Les concepts étant alors « déconnectés » de la réalité du monde, on peut supposer que le contexte agit ici comme un élément perturbateur. Dans l'approche ancrée de la cognition, les représentations mentales (les symboles perceptifs) encodent une information modale, qui, par définition, va varier en fonction du contexte d'écoute. Ainsi, les caractéristiques spatiales d'un lieu influent sur le son. De même, la luminosité influe sur la couleur d'un objet. Et de fait, le contexte se retrouve complètement intégré dans la représentation des concepts.

Concernant les approches *in situ*, nos recherches portant sur un modèle permettant de simuler des environnements sonores, les expériences d'analyse sensorielle présentées ici (cf. Chapitre 4) sont toutes pratiquées en « laboratoire ». Nous notons cependant que les résultats de ces expériences sont conditionnés à une perception faiblement contextualisée, et de surcroît mono-modale, *i. e.* n'impliquant qu'une modalité sensorielle.

Par ailleurs, nous notons que l'absence d'enregistrements *i. e.* de stimuli figés, pose le problème de la reproductibilité des expériences *in situ*, les stimuli ne pouvant, par définition, être recréés à l'identique. Compte tenu du débat actuel sur le faible taux de réPLICATION DES RÉSULTATS obtenus dans le domaine de la psychologie (Spellman, 2015), ce point n'est pas négligeable. Nous pensons qu'il motive d'ailleurs l'intérêt d'explorer des méthodes génératives permettant d'obtenir des stimuli réels et contrôlés, comme cela est proposé dans ce document. Enfin, soulignons que plusieurs études tentent de reproduire

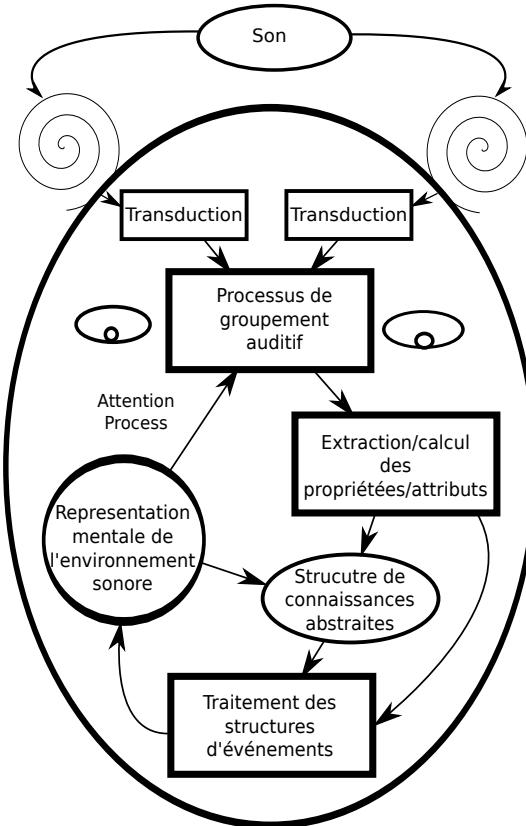


FIGURE 6 : Principaux processus de traitement de l'information auditive et leurs interactions, d'après (McAdams and Bigand, 1994).

le caractère multi-modal des conditions *in situ* en laboratoire, ces approches étant cependant assez lourdes à mettre en œuvre (Morel et al., 2016; Woloszyn, 1997).

Concernant la reproduction écologique des environnements, nous relevons le caractère destructif des écoutes monophoniques et stéréophoniques. Nous remarquons cependant que cette dégradation ne semble intervenir que pour les sons d'arrière plan (*background*), la restitution des sons de premier plan (événements) paraissant produire des réponses similaires à celles obtenues dans un cadre *in situ*.

## 2.5 UNE VUE GÉNÉRALE DU SYSTÈME AUDITIF

### 2.5.1 *La chaîne de traitement*

Le son est une vibration émise par une source d'excitation, et transmise à l'air. Cette vibration se propage ensuite jusqu'à atteindre un récepteur, le tympan, qui va capter le différentiel de pression résultant de cette vibration. C'est le point de départ du processus de traitement de l'information auditive.

Si on adopte une approche *traitement de l'information*, on peut décomposer ce processus en plusieurs systèmes interconnectés. Ces sys-

tèmes forment une chaîne qui, au fur et à mesure des traitements, interprète le signal acoustique afin d'en extraire l'information sémantique. Plus on se place loin dans la chaîne de traitement, plus on accède à une information symbolique et sémantique, potentiellement utilisable par des processus de plus haut niveau. La figure 6, extraite de (McAdams and Bigand, 1994), donne un aperçu des principales fonctionnalités du système de traitement auditif :

- *transduction* : lors de cette étape, les vibrations sonores parvenant au tympan, sont analysées puis traduites en impulsions nerveuses transmises au cerveau. Ces impulsions rendent compte des attributs spectraux et temporels de l'onde. L'extraction des composantes fréquentielles intervient dans la cochlée. C'est à l'intérieur de celle-ci que les différentes parties de la membrane basilaire vont être excitées, en fonction des fréquences composant le signal, suivant un axe tonotopique. Les vibrations, captées à chaque point d'excitation de cette membrane basilaire, sont transmises au cerveau via les nerfs auditifs, chaque point codant une information correspondant à une bande fréquentielle limitée ;
- *groupement auditif* : c'est une étape d'intégration temporelle. Lors de cette étape, l'information est analysée en images auditives cohérentes. Contrairement à ce que pensaient les Grecs anciens, nous ne possédons pas de « canaux » séparés pour chaque objet sonore présent dans l'environnement (Yost, 1994). C'est notre cerveau qui se charge de fusionner et de discréteriser les éléments sonores simultanés, afin de créer un flux auditif structuré. En d'autres termes, il détermine le nombre d'objets présents, identifie leur provenance, et en définit le sens. Afin d'illustrer notre propos, mettons nous dans la peau du mélomane écoutant un Choral de Bach : c'est le processus de *groupement auditif* qui, sur la base des paramètres spectro-temporels du signal, nous permet de distinguer les voix de basse, ténor, alto et soprano ;
- *extraction/calcul des propriétés/attributs* : lors de cette étape sont extraites les qualités perceptives des objets groupés, ces qualités pouvant être vues comme des propriétés cognitives de haut niveau. Pour revenir à notre exemple, c'est à partir d'une analyse des attributs perceptifs que le mélomane est capable de percevoir les mélodies du Choral comme des objets unitaires, même si celles-ci sont développées entre les différentes voix ;
- *structuration des connaissances abstraites* : les phases de groupement et d'extraction concernent l'élaboration et l'analyse d'entités mentales. Lors de l'étape de *structuration des connaissances abstraites*, ces entités sont identifiées, et un sens interprétatif leur

est donné. En pratique, notre mélomane détermine ici si le Choral est plaisant ou non ;

- *traitement des structures d'événements* : lors de cette étape sont intégrés dans le processus cognitif différents contextes, comme par exemple le contexte fonctionnel (le cadre à l'intérieur duquel le son est entendu), ou encore le contexte sensoriel (l'information visuelle, ou la mémoire des événements sonores précédemment entendus). Le *traitement des structures d'événements* permet au mélomane, toujours lui, d'envisager un morceau dans son ensemble, et, dans le cas d'une fugue, d'entendre que la strette finale est un résumé condensé des sujets précédemment exposés ;
- *élaboration des représentations mentales* : lors de cette dernière étape sont organisées et conservées les informations extraites des environnements perçus sous forme de représentations mentales. Ces représentations, d'une part, vont structurer les connaissances acquises, d'autre part, et par effet de rétroaction, vont influer sur les processus de traitement de l'information auditive ci-devant exposés.

### 2.5.2 Processus Bottom-up et processus Top-down

L'interaction entre l'homme et son environnement est fonction, d'une part, de l'information sensorielle captée par l'individu, d'autre part, de la rétroaction exercée par lui sur ces données.

Typiquement, si nous reprenons la figure 6, les étapes de *transduction* et de *groupement auditif*, dépendent, entre autre, de la nature du signal perçu. Elles mobilisent des mécanismes innés, qui opèrent sur l'information sensorielle à partir de la présence de régularités physiques. Ces régularités sont universelles. Elles apparaissent quelle que soit la nature de l'environnement, et sont expérimentées par l'ensemble des individus. Pour exemple, si la fréquence fondamentale d'un son harmonique change au cours du temps, toutes ses harmoniques changent également afin de maintenir la structure harmonique du son [p. 38](Bregman, 1994) (cf. Section 2.6.3).

La rétroaction, quant à elle, dépend de la mémoire de l'individu, *i.e.* sa représentation mentale interne des réalités externes du monde (cf. Section 2.3 et Figure 2.3). Cette mémoire est à la fois :

- individuelle : déterminée par son expérience sensible, *i.e.* la mémoire des interactions sensorielles passées ;
- collective : déterminée par des connaissances transmises, connaissances qui dépendent de sa sphère d'appartenance socio-culturelle, ainsi que de sa langue maternelle.

C'est par la rétroaction que se manifeste l'expérience du sujet, autrement dit, qu'il optimise l'analyse des stimuli, et intègre les effets de contexte dus à l'environnement. Cette rétroaction est également l'expression de son individualité. C'est cette individualité qui explique que deux personnes ayant des capacités sensorielles semblables peuvent percevoir différemment un même environnement.

Deux exemples pourraient venir illustrer les effets de l'individualité en matière d'analyse de l'information sensorielle. Dans le domaine de la vision d'abord, le phénomène dit de bi-stabilité, *i.e.* la faculté, chez un sujet, de tirer d'un même stimulus deux analyses différentes, mais jamais simultanément (Schwartz et al., 2012) (cf. Figure 7). Dans le domaine de l'audition ensuite, le cas cité par McAdams et Bigand (McAdams and Bigand, 1994, p. 2):

« ...Imaginez vous un instant en pleine forêt amazonienne : vous entendriez exactement les mêmes bruits que le guide qui vous accompagne, mais, étant donné votre manque de connaissance du milieu, vous seriez incapable d'extraire du fond sonore les sons correspondant aux cris de l'iguane, aux singes macaques, aux chants des ouistitis ou aux bruissements des arbres tropicaux. De ce fait vous seriez dans l'incapacité d'attribuer une signification à l'ensemble de la structure sonore, ce qui pourrait être important pour votre survie dans l'environnement. »

Suivant la ressource, externe ou interne, le système auditif mobilise deux formes de traitements (cf. Figure 2) :

- les traitements dits ascendants (*bottom-up*), processus dirigés par les données ;
- les traitements dits descendants (*top-down*), processus dirigés par les concepts et les représentations.

Réduire l'interaction entre l'homme et son environnement à une simple association de sensations établies à partir du signal capté, ne permet pas de rendre compte de l'éventail des processus entrant dans le décodage de l'environnement. Étudier le système auditif requiert de prendre en compte aussi bien l'information issue des processus ascendants (information externe), que celle issue des processus descendants (information interne).

La distinction opérée entre ces deux formes de traitement fait directement écho aux notions de perception et de cognition (cf. Section 2.2).

### 2.5.3 Discussion

Les systèmes de traitement décrits dans cette section ne relèvent pas d'une théorie rendant compte du fonctionnement du système auditif,

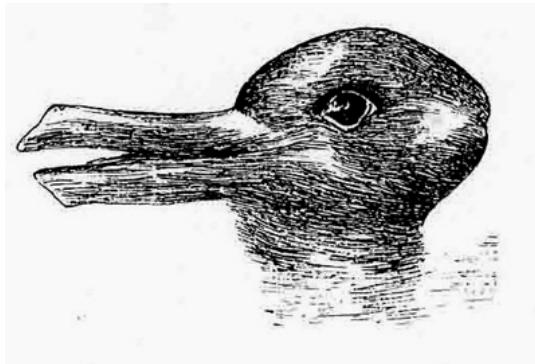


FIGURE 7 : Le phénomène de bistabilité : l'illusion du canard-lapin. Première publication dans *Fliegende Blätter*, 23 octobre 1892, p. 147.

à l'instar de celles évoquées, dans un cadre plus général, à la section 2.2. Leur existence est avant tout avérée empiriquement. De fait, les connexions qu'entretiennent ces systèmes entre eux (cf. Figure 2.3), et notamment leur dépendance vis à vis des représentations internes, restent floues.

Nous notons cependant l'existence de deux types de traitements, et faisons le lien entre les processus montants et perceptifs, tout deux dirigés par l'information sensorielle, et les processus descendants et cognitifs, tout deux dirigés pas les représentations mentales.

## 2.6 ANALYSE DE SCÈNES ACOUSTIQUES

### 2.6.1 Définition

L'analyse de scène est un procédé issu de la recherche dans le domaine de la vision, où les études portent notamment sur les stratégies suivies par l'ordinateur pour parvenir à isoler un(des) objet(s), ou une structure, d'une image (McAdams and Bigand, 1994, p. 12). Dans le domaine de l'audition, un procédé analogue est appelé Analyse de Scène Acoustique. L'ASA a été introduite par A. S. Bregman dans son ouvrage de référence (Bregman, 1994).

L'ASA désigne l'ensemble des traitements perceptifs permettant d'isoler, dans une mixture sonore, les informations émanant de sources distinctes, et de les « organiser » en un tout cohérent. Ces regroupements sont nécessaires au cerveau, et donc au sujet, pour comprendre, pour donner sens à l'environnement. On parle de processus de ségrégation ou de processus de groupement (Winkler et al., 2009). Comme vu à la section 2.5.1, ces processus mobilisent à la fois les traitements ascendants ou *bottom-up*, intervenant au niveau de l'information auditive transduite, et les traitements descendants ou *top-down*, intervenant au niveau du bagage mémoriel. Les processus *bottom-up* sont appelés « processus primitifs », et les processus *top-down*, « processus basés sur des schémas ».

Les processus primitifs sont innés, et opèrent à partir des régularités du signal, afin d'en regrouper les composantes fréquentielles produites par une même source. Le mot régularité désigne ici les propriétés constantes de l'environnement, perçues par tous les individus, et en tous lieux.

Les processus basés sur des schémas sont, eux, conditionnés, et opèrent sur la base des connaissances (schémas) issues de notre représentation mentale du monde, représentation construite à partir des écoutes antérieures.

### 2.6.2 *Une approche psychoacoustique*

Si la plupart des recherches sur l'ASA adoptent une approche cognitiviste, se concentrant sur l'étude des *processus primitifs*, elles suivent généralement une méthodologie expérimentale très inspirée de la psychoacoustique. De fait, les sujets sont soumis à des stimuli décrits analytiquement dans un espace multidimensionnel de dimensions physiques (fréquence, intensité, etc.) (Dubois et al., 2006). Dans la majorité des cas, ces stimuli, ou sons, qu'ils soient purs ou qu'ils soient complexes, sont synthétisés en laboratoire.

Ces sons sont émis de manière séquentielle. Au cours de l'expérience, les paramètres d'écoute sont modifiés (intensité, fréquence, espace entre les séquences, etc.) afin d'évaluer le seuil à partir duquel la capacité du sujet à distinguer les sources sonores est altérée.

L'ASA se concentre donc sur l'analyse de l'effet de descripteurs « bas niveau » dans le processus d'intégration, sans tenir compte d'attributs perceptifs « haut niveau », comme la valeur sémantique attribuée aux sons, sans tenir compte non plus de considérations écologiques (cf. Section 2.4.2).

### 2.6.3 *Régularités et processus primitifs*

L'existence des *processus primitifs* est une conséquence de l'efficience, dans le monde sonore, de régularités universelles affectant l'ensemble des stimuli auditifs. Bregman distingue 4 types de régularités [p. 19, 21, 31, 33] (McAdams and Bigand, 1994):

1. *synchronicité* : il est rare que des sons n'ayant aucun rapport entre eux démarrent et s'arrêtent au même moment ;
2. *continuité* :
  - les propriétés d'un son isolé tendent à se modifier lentement et de façon continue ;
  - les propriétés d'une séquence de sons émis par la même source tendent à se modifier lentement.

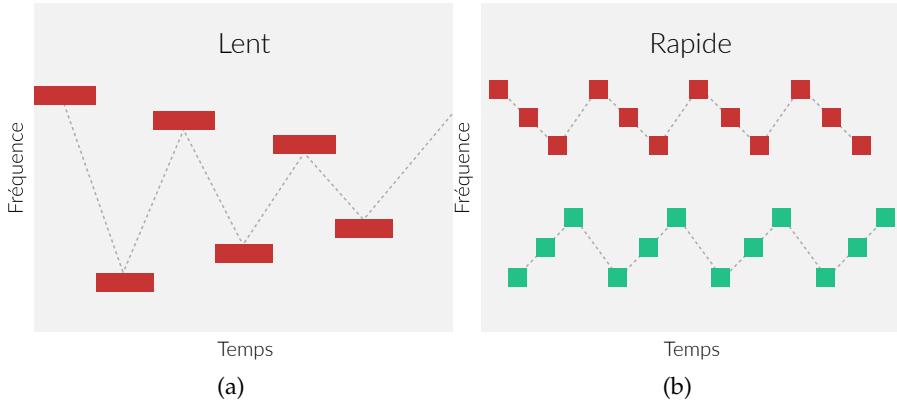


FIGURE 8 : Groupement séquentiel : proximité temporelle. Dans l'exemple (a), la durée entre les événements est importante, aucun regroupement n'est effectué, les sons sont perçus comme des événements distincts. Dans l'exemple (b), la durée entre les événements est réduite, un regroupement s'opère suivant la proximité fréquentielle. Deux flux sont ainsi créés, le premier (rouge) regroupe les sons haute fréquence, le deuxième (vert), les sons basse fréquence.

3. *harmonicité* : lorsqu'un corps sonore vibre à une période répétée, ses vibrations donnent naissance à un motif acoustique dont les fréquences des composants sont des multiples d'une même fréquence fondamentale ;
4. *uniformité* : la plupart des modifications qui surviennent dans un signal acoustique affectent tous les composants du son résultant, de manière identique, et simultanée.

Notre perception du monde est assujettie à ces régularités. Les *processus primitifs*, sensibles aux stimuli exclusivement, nous permettent d'isoler du monde sonore des objets cohérents, perçus à travers elles (Ballas and Howard, 1987). Le fait est qu'un principe similaire de perception des formes s'applique également au domaine de la vision.

#### 2.6.4 Perception de la forme

Que ce soit en vision ou en audition, notre cerveau est en permanence stimulé par une multitude de sources distinctes. Percevoir un objet dans cet agglomérat, c'est être capable d'isoler tous les signaux émis par une même source, et de les réunir en une unité perceptive cohérente.

Parmi les premiers travaux qui se sont intéressés à ces processus de groupement, on trouve la psychologie de la forme, en allemand *Gestalttheorie*. Cette théorie, introduite par Ernst Mach et Christian von Ehrenfels à la fin du XIXème siècle, explicite les principes selon lesquels des stimuli sensoriels sont combinés afin de former un pattern

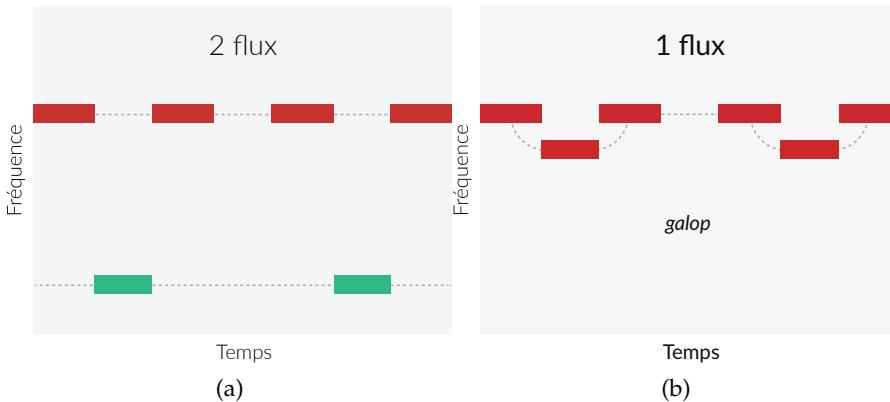


FIGURE 9 : Groupement séquentiel : proximité fréquentielle. Deux groupes de sons sont joués à deux fréquences. Dans l'exemple (a), la distance entre les fréquences des sons est importante, deux flux sont créés, le premier (rouge) regroupe les sons haute fréquence, le deuxième (vert), les sons basse fréquence. Dans l'exemple (b), la distance entre les fréquences est réduite, un groupement s'opère suivant la proximité fréquentielle. Un seul flux est créé, et l'on perçoit un motif temporel, ici, le galop d'un cheval.

mental rendant compte de la présence d'un objet dans un environnement donné.

Mis en évidence en perception visuelle, ces principes restent vrais en perception auditive (Bregman, 1994, ch. 1). Parmi ces principes, nous en détaillons ici cinq :

1. *proximité* : des éléments proches les uns des autres ont tendance à être groupés ensemble. En audition, ce principe de proximité opère suivant les différentes caractéristiques du son, à savoir, la fréquence, l'*onset*<sup>8</sup> et l'intensité.
2. *similarité* : des éléments qui se ressemblent ont tendance à être groupés ensemble. Dans le domaine de la vision, la proximité est une notion de spatialité. La similarité, elle, s'applique aux caractéristiques physiques de l'objet (forme, couleur, *etc.*), caractéristiques qui ne peuvent se décrire dans une dimension unique. Dans le domaine de l'audition, la proximité est généralement admise comme notion de temporalité (*onsets* proches). Pour le reste des descripteurs, il est cependant difficile de distinguer les principes de proximité et de similarité. Bregman propose de parler de proximité lorsqu'on traite d'une dimension physique particulière, et de parler de similarité dès lors qu'on considère un ensemble de descripteurs, ou lorsque l'on traite d'attributs

<sup>8</sup> En traitement du signal audio, on désigne par les mots anglais *onset* et *offset* le début et la fin du signal. Ces termes étant couramment utilisés en français, nous ne les traduirons pas dans ce document.

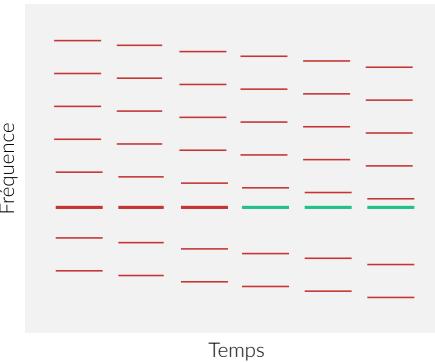
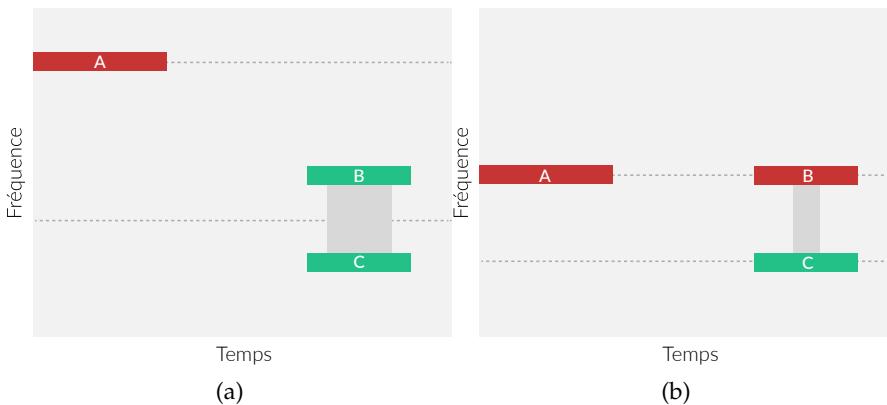


FIGURE 10 : Groupement simultané : régularité harmonique. Un son complexe est joué plusieurs fois. À chaque occurrence, on abaisse les hauteurs de la fréquence fondamentale et des harmoniques, de manière uniforme. Un harmonique est conservé à fréquence constante (trait gras). Au début, un flux est créé, *i.e.* les harmoniques et la fréquence fondamentale sont perçus comme étant un seul objet. Au fur et à mesure que la régularité harmonique est brisée, le cerveau tend à percevoir l'harmonique à fréquence constante dans un flux séparé, *i.e.* comme étant un second objet.

qui ne peuvent être clairement décomposés suivant des dimensions distinctes. Exemple : le timbre.

3. *continuité* : des éléments qui varient de manière non abrupte ont tendance à être groupés ensemble. Ainsi, des objets distincts mais proches temporellement (en vision, proches spatialement) ont tendance à être "vus" comme le prolongement des uns par les autres. C'est ce principe qui nous permet de percevoir comme une seule entité un objet dont les caractéristiques varient dans le temps, *e.g.* le son d'une sirène. *A contrario*, un changement abrupt indique généralement l'apparition d'une nouvelle source. De récentes études en neurosciences ont montré l'importance de ce principe dans les processus de groupement. (Winkler et al., 2009) propose de voir l'ASA comme un processus prédictif, le cerveau cherchant à anticiper la nature des stimuli qui lui parviennent, sur la base de régularités extraites des objets détectés dans l'instant précédent.
4. *clôture* : des éléments discontinus, qui suggèrent la forme d'un objet continu, ont tendance à être groupés ensemble. De manière automatique, le cerveau tend à percevoir un ensemble d'objets distincts comme un tout. En audition, ce principe est très lié à la notion de masquage. En effet, les sons que nous percevons sont régulièrement masqués par d'autres sons concurrents, éventuellement plus forts. Le principe de clôture nous permet de compenser ce phénomène de masquage, et de percevoir le signal sans discontinuité. Ainsi lorsqu'un son pur est régulièrement entrecoupé de silences, nous percevons une série



**FIGURE 11 : Groupement ancien-plus-nouveau.** Dans l'exemple (a), les sons A et B ont des fréquences éloignées. Le cerveau génère deux flux, le premier relatif au son A (rouge), et le deuxième comprenant les sons B et C (vert). Dans l'exemple (b), les sons A et B ont la même fréquence. Le cerveau interprète le son B comme étant une continuité du son A. L'attraction entre B et C en est réduite. Le cerveau génère toujours deux flux, le premier regroupant cette fois les sons A et B (rouge), et le deuxième le son C (vert).

de sons purs, mais, si ces silences sont comblés par un bruit blanc, nous percevons un son pur continu. Ce phénomène est parfois appelé « l'illusion de continuité » (Dannenbring, 1976), et s'applique particulièrement dans le contexte de la perception de la parole (Carlyon et al., 2002).

5. *destin commun* : des éléments qui varient de manière synchrone et uniforme ont tendance à être groupés ensemble. C'est ce principe qui permet de percevoir comme un tout les différents harmoniques qui composent un son complexe. C'est également ce principe qui nous incite à percevoir de manière unie des stimuli ayant le même *onset temporel*.

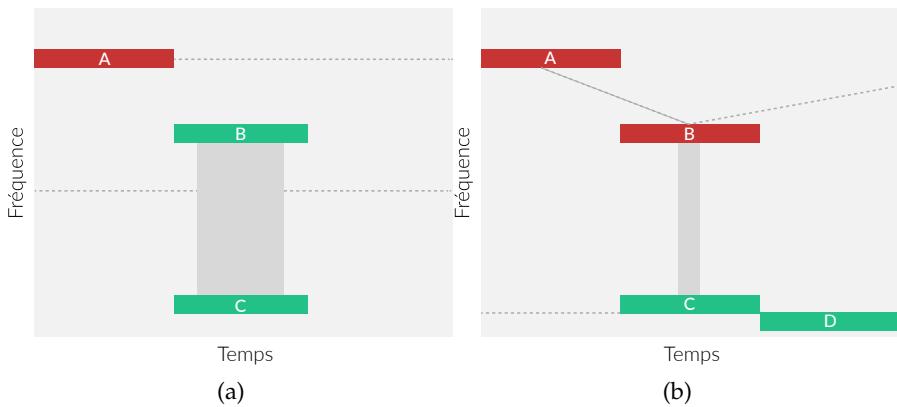
Ces principes agissent de concert afin de grouper les composantes du son en flux auditifs.

#### 2.6.5 Flux auditif et stratégie de groupement

L'une des notions les plus importantes en ASA est le concept de flux auditif (*auditory stream*), que Bregman définit comme « le groupement perceptif que nous faisons des parties du spectre qui vont ensemble ».

Contrairement au terme « son », qui peut faire référence aux réalités physiques des phénomènes acoustiques, aussi bien qu'aux représentations mentales que nous nous en faisons, le flux auditif désigne spécifiquement une entité perceptive.

Le terme flux se veut volontairement généraliste. Ce dernier peut désigner aussi bien un son isolé (un coup de marteau), que plusieurs



**FIGURE 12 :** Compétition entre groupement séquentiel et groupement simultané. Dans l'exemple (a), le cerveau perçoit deux flux, le premier regroupant le son A (rouge), et le deuxième regroupant les sons B et C (vert). Dans l'exemple (b), un quatrième son (D) est ajouté, ce dernier possédant une fréquence très proche de celle de (C). Le groupement séquentiel par proximité fréquentielle entre les couples A-B et C-D est favorisé, au détriment du groupement simultané entre les sons B-C.

sons, à condition que ces derniers soient perçus comme formant une seule entité (une série de coups de marteau rapprochés).

On désigne par « formation de flux auditif » (*auditory streaming*), le processus à l'origine de la création de flux. (Winkler et al., 2009) proposent la définition suivante en ce qui concerne la formation de flux auditifs :

« Un phénomène perceptif dans lequel une séquence de sons est perçue comme étant composée de deux ou plusieurs flux auditifs »

Comme nous le voyons, le flux désigne une représentation perceptive d'un son. Il est l'équivalent auditif de l'objet pour la vision (Bregman, 1994, p. 11). Cependant, nous notons que la notion de flux, et plus particulièrement celle de formation de flux, est souvent utilisée dans un contexte où une dimension temporelle est sollicitée. On parle d'ailleurs de construction de flux (*build-up of streaming*) pour désigner la période durant laquelle le cerveau accumule des indices afin de générer et de stabiliser les flux auditifs (Cusack et al., 2004; Snyder and Alain, 2007). Ainsi, dans ce document, nous réservons les termes « flux » pour désigner les représentations mentales des stimuli en train d'être traités (intégrés temporellement), et « formation de flux auditifs » pour désigner les processus perceptifs de groupement des sons. Nous conservons le mot « objet » pour désigner, de manière générale, les représentations mentales, stockées en mémoire, des phénomènes acoustiques.

En ce qui concerne les processus primitifs, on distingue trois stratégies de groupement :

- *groupement séquentiel* : désigne le groupement de sons ne partageant pas un même *onset*. Dans le cas de sons purs, le groupement séquentiel s'appuie beaucoup sur le principe de *proximité*, et notamment de *proximité fréquentielle*, principe qui veut que, plus deux sons sont proches en fréquence, plus ils ont tendance à être regroupés dans le même flux (cf. Figure 9). Pour des sons complexes, c'est plutôt le principe de similarité qui entre en jeu. La *proximité temporelle*, *i.e.* la durée séparant chaque son, rentre cependant en ligne de compte. Plus cette dernière est faible, plus deux sons ont des chances d'être regroupés (cf. Figure 8) ;
- *groupement simultané* : aussi appelé groupement spectral, désigne le groupement des composantes fréquentielles qui partagent un même *onset*. Le groupement simultané s'appuie principalement sur la régularité harmonique. Une composante fréquentielle venant briser la suite harmonique tend à être isolée dans un autre flux auditif (cf. Figure 10) ;
- *groupement ancien-plus-nouveau* : ce groupement est l'application directe du principe de continuité. Lorsque le spectre de l'environnement sonore s'enrichit subitement, tout en conservant ses composantes fréquentielles de départ, le cerveau tend à interpréter l'ajout comme une continuation de l'ancien, et à l'intégrer dans le même flux sonore (cf. Figure 11).

Dans le cas d'une compétition entre un groupement séquentiel et un groupement simultané, c'est l'organisation issue du groupement séquentiel qui prime (cf. Figure 12). Ce phénomène fait sens d'un point de vue écologique. En effet la plupart des sons, et en particulier ceux utilisés pour la communication, n'existent que dans une certaine durée, et sont intermittents. Il est alors nécessaire de faire des associations entre des sons parfois séparés par un intervalle de temps long, afin de percevoir le sens du message (Winkler et al., 2009).

Les exemples cités plus haut se placent tous dans un contexte *bottom-up*, en évacuant d'éventuels processus attentionnels. Bien que cela ait été encore très peu étudié, il paraît cependant évident que ces stratégies de groupement s'opèrent également dans un contexte *top-down*, en s'appuyant sur une mémoire à plus long terme.

#### 2.6.6 *L'approche par les neurosciences*

L'étude de l'ASA a connu un regain d'intérêt avec les neurosciences. Cette section présente certains résultats des travaux menés dans le domaine, sans toutefois détailler les méthodes expérimentales utilisées.

Plusieurs études ont montré que les mécanismes de l'ASA, notamment ceux relatifs aux processus perceptifs (primitifs) liés au groupement et à la ségrégation de la mixture sonore, se retrouvent lorsque que l'on observe l'activité neuronale du premier cortex auditif (A1) (Carlyon, 2004; Nelken, 2004; Snyder and Alain, 2007). (Dyson and Alain, 2004) montrent une variation de l'activité neuronale lorsque la régularité harmonique est brisée. Utilisant une séquence alternée de deux stimuli A et B, dont l'un (B) souffre d'une condition d'écoute défavorable (fréquence extrême), (Fishman et al., 2001) montrent que, dans le cas d'une répétition lente, on observe des réponses neuronales distinctes pour chacun des stimuli (A et B). Dans le cas d'une répétition rapide, la réponse du stimulus défavorisé (B) tend à disparaître (A et B fusionnent en un seul flux), illustrant ainsi le phénomène de formation de flux auditif.

Une notion capitale abordée par ces études est celle de « l'objet auditif » (Nelken, 2004). L'objet auditif se définit comme une image mentale (information modale) d'une source sonore. Il est représenté dans A1, et caractérise les propriétés « bas niveau » du son (propriété spectrale, temporelle), tout en possédant également une information de plus haut niveau (qualité abstraite, identité, intelligibilité) (Nelken and Bar-Yosef, 2008). La formation de cette image est ainsi le résultat direct d'un processus de formation de flux, mais également de processus cognitifs. Plusieurs études tendent à montrer que, lorsqu'il est soumis à un environnement complexe, A1 se représente celui-ci comme un ensemble « d'objets auditifs » (Kocsis et al., 2016).

Enfin, il semble que les objets auditifs aient une fonction prédictive (Winkler and Schröger, 2015). (Winkler et al., 2009) proposent notamment un modèle dans lequel les régularités cherchées par les processus primitifs à un instant  $t + 1$  sont celles des « objets auditifs » représentés à un instant  $t$ . Dans le cas où les régularités d'un « objet auditif » ne sont plus détectées, ce dernier disparaît. A l'inverse, si les régularités perçues ne correspondent à aucune image mentale, elles sont candidates à la formation d'un nouvel « objet auditif ».

#### 2.6.7 *Attention et saillance*

L'attention est la capacité de notre système auditif à focaliser sur des composantes spécifiques de notre environnement sonore en ignorant le reste. En fonction du contexte de la scène, certains flux ont tendance à attirer plus facilement notre « attention ». Un des paramètres pouvant susciter l'attention est la saillance.

La saillance d'un flux audio peut se voir comme l'impact potentiel d'un stimulus sur notre perception, et notre comportement. Cette saillance est fonction du contexte d'écoute de la scène sonore. L'attention et la saillance ont une influence dans l'identification des sources. Cette identification, et l'attribution de « sens » qui en découle, est

une étape primordiale dans le processus de création de l'image mentale d'un environnement, à partir de la perception de son empreinte sonore. Ainsi, un élément saillant est facilement identifiable. A l'inverse, les sources d'un fond sonore (*background*), par définition peu saillant (*i.e.* attirant potentiellement moins l'attention), sont moins discernables (Elhilali et al., 2009).

De Coensel et Botteldooren proposent plusieurs modèles permettant de simuler l'attention (Botteldooren and De Coensel, 2009; De Coensel and Botteldooren, 2010; De Coensel et al., 2010). Les modèles calculent une « carte de saillance » décrivant l'évolution de la saillance d'une scène en fonction du temps. Les deux chercheurs partent du principe que le cerveau ne peut pas traiter toutes les informations en même temps, et sélectionne donc l'information utile. Ces modèles ne prennent cependant pas en compte les traitements de type *top-down* et se concentrent sur les processus *bottom-up* relatifs à l'analyse des caractéristiques propres aux stimuli. Ce modèle d'attention a été inclus dans un modèle plus général permettant de détecter les sources sonores actives dans un environnement (Oldoni et al., 2012, 2013).

#### 2.6.8 Discussion

Concernant l'ASA, nous retenons qu'une des premières fonctions du système auditif est de séparer, dans la mixture sonore, les informations relatives aux différentes sources qui la composent. Cette ségrégation résulte d'une part de processus primitifs (innés), dépendant uniquement du signal, d'autre part de « processus basés sur des schémas », dépendant des connaissances du sujet.

Au vu de leurs définitions respectives, nous considérons les processus perceptifs montants et primitifs comme relevant d'un seul mécanisme. Nous considérons par ailleurs les processus descendants et basés sur des schémas comme faisant partie d'un même type de processus cognitifs.

Le système auditif ne perçoit pas le monde sonore comme une succession d'objets distincts, mais intègre ces derniers dans des flux. Le processus de formation des flux dépend à la fois des caractéristiques physiques, les sons successifs d'une même source étant groupés dans un même flux, mais également du contexte et de l'attention portée. A ce titre, un ensemble de sons ne partageant pas *a priori* de caractéristiques communes peut être groupé dans un même flux, si présenté simultanément avec une séquence sonore plus saillante.

Par ailleurs, nous observons des connexions entre la notion « d'objet auditif », et celle « d'image simulée » introduite dans l'approche ancrée de la cognition (*i.e.* la réalisation d'un concept sonore activé par la perception de l'objet physique correspondant ; cf. Section 2.2.3). Les deux comportent une information relative aux représentations internes, et, de fait, constituent plus qu'une simple image mentale d'un

objet perçu. De plus, la qualité prédictive attribuée aux « objets auditifs » est aussi présente dans les « images simulées », ces dernières étant supposées rétroagir sur les processus montants. Rien *a priori* n'empêche d'envisager que la génération des « objets auditifs » relève du même processus de simulation que celui mis en œuvre à partir des symboles perceptifs. La simulation étant considérée comme dynamique (Barsalou, 1999), elle peut s'intégrer dans le modèle prédictif (dynamique de par sa nature adaptative).

Précisions encore que « objets cognitifs » et « images simulées » ne sont pas des entités isolées, mais participent toutes deux à la création d'une image mentale (simulée) plus large, représentant l'ensemble des éléments perçus de l'environnement.

Un point cependant peut poser problème : l'existence d'un « objet auditif » semble être conditionnée à la perception d'un objet physique, ce qui n'est pas forcément le cas pour les « images simulées » (*e.g.* phénomène d'introspection). On peut néanmoins poser que « l'objet auditif » est le résultat d'une simulation conditionnée à la perception d'un objet physique.

Pour finir, nous notons que les études de l'ASA présentées ci-devant (en psychologie et en neurosciences), utilisent, dans leur immense majorité, des sons de synthèse. Il est difficile de faire un parallèle entre la notion de source sonore prévalant en ASA, et celle admise en psychologie cognitive (cf. les études portant sur les paysages sonores abordées à la section 2.7). Dans l'une et l'autre approche, le terme source sonore s'applique *a priori* à l'objet source (*e.g.* voiture). C'est évident en psychologie cognitive, où le stimulus est le plus souvent un enregistrement de ladite source. Ça l'est moins en ASA, où le stimulus, synthétisé, est un objet abstrait (agglomérat de sinusoïdes), éloigné de la réalité des phénomènes acoustiques, et dont l'existence n'est avérée que dans la mesure où il est interprété par le sujet comme un tout, une entité. Les résultats obtenus à partir de ces stimuli restent difficiles à généraliser à des applications plus incarnées.

## 2.7 L'ÉTUDE DES PAYSAGES SONORES

### 2.7.1 *La notion de paysage sonore*

La notion de paysage sonore (*soundscape*) a été introduite par R. M. Schafer dans les années soixante-dix dans son livre (Schafer, 1969), et détaillée dans l'ouvrage de référence (Schafer, 1977). La question que pose Schafer est :

« Quelle est la relation entre l'homme et les sons de l'environnement qui est le sien, et que se produit-il lorsque ces sons viennent à changer ? »

Une première définition du paysage sonore a été donnée par (Truax, 1978):

« [a]n environment of sound (or sonic environment) with emphasis on the way it is perceived and understood by the individual, or by a society. »<sup>9</sup>

Aujourd’hui, cependant, on s’accorde sur la définition suivante (Aletta et al., 2016):

« Un environnement sonore tel qu'il est perçu, expérimenté et/ou compris par un individu ou une communauté, dans son contexte. »<sup>10</sup>

L'une et l'autre définitions sont larges. Tout environnement peut être considéré comme un paysage sonore, dès lors qu'on lui associe un ensemble de sons entendus par un sujet donné. Le problème est d'envisager l'environnement sonore par rapport à l'évaluation subjective de l'auditeur, et non uniquement par rapport à ses paramètres acoustiques. Schafer, déjà, explique la nécessité de ne plus considérer le bruit seul, mais aussi la perception de ce bruit par les individus, et le contexte dans lequel il est perçu, ceci afin d'améliorer la qualité de leur environnement. On parle d'environnement sonore lorsqu'on se réfère au phénomène acoustique physique, et de paysage sonore lorsqu'on se réfère à la représentation que l'on se fait de l'environnement.

Ainsi, les études sur les paysages sonores suivent le paradigme de la psychologie cognitive (Dubois et al., 2006; Maffiolo, 1999) (cf. Section 2.4.3). L'environnement sonore est décrit en utilisant à la fois des descripteurs acoustiques (mesures), et des descripteurs perceptifs, l'analyse de l'interaction entre ces descripteurs permettant de comprendre les processus cognitifs mis en œuvre dans l'évaluation perceptive des paysages sonores.

### 2.7.1.1 *L'approche positive*

Si beaucoup d'efforts sont faits afin de réguler les niveaux de bruit des sons non-désirés, l'approche inverse, *i.e.* ajouter des sons positivement connotés, reste très peu considérée. Cette approche, consistant à identifier et agir sur les sons acceptés, ou plaisants, afin d'améliorer la qualité d'un environnement, est nommée l'approche positive par Schafer. De récentes études ont donné des résultats prometteurs, notamment (Hong and Jeon, 2013) qui montre que l'ajout de sons

<sup>9</sup> Un environnement sonore étudié en mettant l’accent sur la manière dont il est perçu et compris par un individu ou une communauté.

<sup>10</sup> Cette définition a été publiée dans le cadre de la norme ISO-12913 (*ISO 12913-1 :2014 acoustics-soundscape-part 1 : definition and conceptual framework* 2013)

d'oiseaux, ou d'eau, à des sons de trafic urbain, permet de significativement améliorer l'appréciation de ces derniers. (Galbrun and Ali, 2012) montre qu'un son d'eau ayant un niveau sonore similaire ou inférieur de -3dB à celui du trafic permet de correctement masquer ce dernier. L'étude indique également que des sons de cours d'eau possédant un contenu fréquentiel basse-fréquence sont préférés aux sons de fontaines et de chutes d'eaux.

### 2.7.2 Approches catégorielle et dimensionnelle

Deux grandes problématiques intéressent la recherche sur les paysages sonores :

- la première concerne la *représentation mentale des paysages sonores*. Comment nous représentons nous, en mémoire, un paysage sonore perçu ? La question en amène deux autres :
  1. Quelles sont les différentes catégories de paysages sonores ?
  2. Comment caractériser ces catégories ?<sup>11</sup>
- la deuxième concerne les *dimensions perceptives*. Quelles dimensions perceptives entrent en jeu dans l'évaluation subjective des paysages sonores ? Là encore la question en amène deux autres :
  1. Quels descripteurs perceptifs permettent de caractériser les dimensions perceptives à partir desquelles nous appréhendons l'environnement sonore ?
  2. Quels indicateurs influent sur ces descripteurs perceptifs ?

Les descripteurs perceptifs caractérisent les dimensions selon lesquelles nous interprétons l'environnement. Pour exemple, un des descripteurs perceptifs communément utilisé est l'agrément (cf. Section 2.7.3 pour plus de détails sur les descripteurs couramment utilisés). L'enjeu est ici de trouver les indicateurs qui influent sur ces descripteurs perceptifs. On distingue quatre types d'indicateurs :

- *indicateurs acoustiques/physiques* : il s'agit d'indicateurs objectifs, obtenus via des mesures. Parmi ces indicateurs, certains caractérisent le niveau sonore par une approche holistique ( $L_{Aeq}$ ), d'autres par une approche statistique ( $L_{A10-90}$ ), d'autres encore en considérant séparément les différents canaux fréquentiels. On inclut, d'autre part, dans les indicateurs acoustiques/physiques, des indicateurs permet-

---

<sup>11</sup> Répondre à cette dernière question revient à comprendre quels sont les éléments qui constituent un paysage sonore, et comment la nature de ces éléments influe sur le processus de catégorisation de l'environnement. Intuitivement, les éléments constitutifs d'un paysage sonore sont les sources sonores. Il s'agit alors, également, d'étudier la manière dont nous nous représentons ces sources.

tant de décrire les caractéristiques spectrales du son (cf. tableau 1) ;

- *indicateurs perceptifs* : les dimensions affectives suivant lesquelles nous percevons l'environnement sonore ne sont pas indépendantes. Ainsi certains descripteurs, comme l'agrement ou le confort, peuvent eux-mêmes servir d'indicateurs pour d'autres descripteurs plus généraux comme la qualité sonore. On inclut, d'autre part, dans les indicateurs perceptifs, des indicateurs procédant d'une évaluation subjective d'un attribut physique (*e.g.* niveau sonore perçu) ;
- *indicateurs psychoacoustiques* : ces indicateurs sont à mi-chemin entre les indicateurs acoustiques et les indicateurs perceptifs. Comme les premiers, ils sont objectifs, calculés sur le signal sonore. Comme les seconds, ils sont perceptivement inspirés, *i.e.* construits afin de rendre compte d'une réalité perceptive. Pour exemple, la *loudness* de Zwicker (Zwicker and Fastl, 1990) qui rend compte du niveau sonore perçu. Le tableau 2 présente quelques-uns des indicateurs les plus utilisés ;
- *indicateurs extra-sonores* : on regroupe ici tous les indicateurs qui ne sont pas liés au son. Certains sont liés au sujet (âge, genre, humeur), d'autres aux stimuli visuels, d'autres encore au moment de la journée. Contrairement aux indicateurs acoustiques, psychoacoustiques ou perceptifs, qui sont tous évalués/mesurés suivant des échelles ordonnées, discrètes ou continues, certains indicateurs extra-sonores sont eux évalués sur des échelles de catégories<sup>12</sup> (*e.g.* genre : homme/femme). On parlera alors plutôt de contexte extra-sonore.

Ces problématiques, pour rappel, la représentation mentale des paysages sonores, et les dimensions perceptives, sont à la base des deux grandes approches méthodologiques adoptées par la communauté scientifique : l'approche catégorielle, et l'approche dimensionnelle.

#### 2.7.2.1 Méthodologie de l'approche catégorielle

Les objectifs de l'approche catégorielle sont triples. Il s'agit :

- d'appréhender les principes psychologiques qui sous-tendent la formation des représentations mentales ;
- d'objectiver la nature de ces représentations ;

---

<sup>12</sup> Le terme catégorie, employé ici pour décrire les différents niveaux d'une échelle, est sans rapport avec le terme catégorie relatif aux représentations mentales

Nom	Description
$L_A$	Niveau sonore instantané pondéré A
$L_{Aeq,T}$	Niveau sonore équivalent pondéré A calculé sur une période T
$L_{A10-90}$	10-90ème quantiles des $L_A$
$L_{Amin}, L_{Amax}$	minimum maximum des $L_A$
Facteur crête	Ratio entre la valeur de pression maximale et la valeur RMS

TABLE 1 : Indicateurs acoustiques.

- de comprendre l'influence de ces représentations sur le traitement de l'information sonore.

A ce titre, l'approche catégorielle peut être vue comme une approche cognitive.

Afin d'objectiver la nature des catégories mentales de paysages sonores, ou de sources sonores, l'approche catégorielle peut avoir recours à trois types d'expériences (cf. Figure 13) :

- *tâche de description* : on demande au sujet de décrire l'environnement sonore auquel il a été exposé (Axelsson et al., 2005; Guastavino, 2006; Rimbault, 2006; Rimbault and Dubois, 2005), soit de la manière la plus libre possible, soit en contrignant la description par le biais d'un questionnaire. Là encore, les réponses peuvent être libres (questionnaire semi-dirigé) ou à choix forcés (questionnaire dirigé). Plus la description est libre, plus on accède à des représentations mentales spécifiques au sujet. *A contrario*, plus le questionnaire est contraint, plus on accède à des représentations stéréotypées.

L'analyse linguistique et lexicale des données ainsi collectées permet d'en faire émerger les catégories sémantiques. La richesse des descriptions, résultant de la liberté de réponse laissée au sujet, rend cependant ce travail d'analyse complexe. Ces expériences de description peuvent être réalisées en laboratoire, ou dans un cadre *in situ* ;

- *tâche de tri ou catégorisation* : on demande au sujet d'organiser les stimuli auxquels il vient d'être soumis, via une interface graphique le plus souvent, en groupes ou paquets, suivant une consigne fixée en fonction des objectifs mêmes de l'expérience

Psychoacoustiques	
Nom	Description
<i>loudness</i> de Zwicker	Niveau sonore perçu
Acuité ( <i>sharpness</i> )	Contenu fréquentiel
Rugosité ( <i>roughness</i> )	Modulation enveloppe temporelle (15-70Hz)
Fluctuation ( <i>Fluctuation strength</i> )	Modulation enveloppe temporelle (4Hz)
Brillance	Centre de gravité spectral

TABLE 2 : Indicateurs psychoacoustiques : modèles mathématiques illustrant des qualités affectives perçues.

(Guastavino, 2007; Maffiolo, 1999). L'analyse de ces groupes permet d'en faire émerger les catégories, et de comprendre quels sont les attributs perceptifs à l'origine de l'organisation catégorielle proposée par le sujet. Il est par ailleurs possible de demander au sujet de nommer, voire de décrire ces groupes, afin d'acquérir encore plus de connaissances sur la nature des groupements effectués. On parle de catégorisation forcée lorsque que le nombre de groupes est contraint, et de catégorisation libre lorsque le nombre de groupes est laissé à l'appréciation du sujet. Ces expériences de tri sont pratiquées en laboratoire, et utilisent habituellement des enregistrements sonores comme stimuli ;

- *comparaison par paires* : on demande au sujet de noter la similarité entre des paires de stimuli (Gygi et al., 2007). L'association des mesures par paires permet d'obtenir une matrice de similarités illustrant les ressemblances entre tous les stimuli. Via un positionnement multidimensionnel (*Multidimensional scaling*), il est alors possible de retrouver l'espace rendant compte au mieux de ces similarités. De la position des stimuli dans l'espace on peut déduire des groupements catégoriels. Des outils de clustering (*e.g.* clustering hiérarchique ascendant) peuvent être également appliqués directement sur la matrice afin de faire émerger des groupes d'objets similaires. Ces méthodes ont notamment été utilisées en perception musicale afin d'étudier les indicateurs acoustiques/psychoacoustiques qui déterminent la notion de timbre (Caclin et al., 2005)

L'avantage de ces pratiques expérimentales est double :

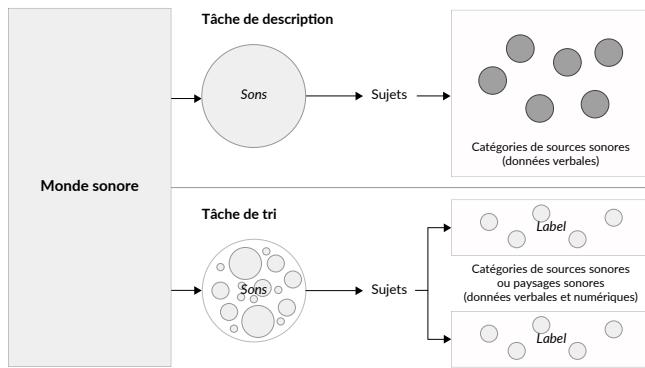


FIGURE 13 : Tâche de description et tâche de tri ou de catégorisation.

1. elles laissent une grande liberté au sujet dans ses réponses. En particulier, les tâches de comparaisons et de catégorisation peuvent permettre de caractériser des stimuli sans imposer au sujet des dimensions ou attributs particuliers à partir desquels évaluer les sons, comme c'est notamment le cas pour l'analyse sémantique différentielle (cf. Section 2.7.2.2). Présupposer des dimensions intervenant dans la comparaison des stimuli, c'est en effet prendre le risque que ces dimensions ne fassent pas sens du point de vue du sujet, mais également des stimuli. Ces tâches permettent ainsi d'apprécier les ressemblances globales pouvant exister entre des stimuli sonores, ressemblances qui découlent à la fois de similarités physiques et sémantiques ;
2. elles profitent d'une information riche via l'utilisation de descriptions verbales (tâche de description ou de catégorisation avec verbalisation). L'analyse de ces descriptions, qu'elles soient libres ou rattachées à des groupes, permet à l'expérimentateur d'approfondir ses connaissances sur les processus cognitifs sous-jacents à la perception des stimuli, le renseignant notamment sur l'influence putative d'attributs sémantiques ne relevant pas (ou peu) des caractéristiques physiques des sons (cf. Section 2.3.2.2). Le langage agit ici comme senseur qualitatif.

#### 2.7.2.2 Méthodologie de l'approche dimensionnelle

L'approche dimensionnelle tente, elle, de caractériser les environnements sur la base de dimensions perceptives pré-établies. Comme nous l'avons vu, ces dimensions sont décrites par des descripteurs perceptifs.

Pour élaborer ces descripteurs, l'approche dimensionnelle a communément recours à l'analyse sémantique différentielle. Au cours de ces expériences, le sujet, à partir de ses ressentis, doit évaluer des descripteurs imposés en s'aidant d'échelles sémantiques bipolaires, ou échelles de Likert. Ces échelles répertorient l'ensemble des valeurs

pouvant être prises par les différents descripteurs d'une scène sonore en cours d'évaluation. Elles forment un questionnaire à réponses fermées. En fonction des besoins de l'étude, elles sont discrètes ou continues, paires ou impaires. Cependant, dans le cadre de l'évaluation des environnements sonores, on utilise généralement des échelles impaires et graduées en 7 (Raimbault, 2006), 9 (Hall et al., 2013) ou 11 points (Ricciardi et al., 2015).

La valeur sémantique des échelles tient au fait que les extrémités en sont bornées par des mots. Pour exemple, (Ricciardi et al., 2015) évalue la qualité d'un environnement, ainsi que la présence de voitures, à partir de deux échelles de 11 points chacune, et délimitées, l'une, par les termes désagréable (1) / plaisant (11) (*unpleasant/pleasant*), l'autre, par les termes rare (1) / fréquent (11) (*rarely/frequently*). Ces termes cadrent les réponses des sujets afin de s'assurer que tous interprètent l'échelle de la même manière, *i.e.* en attribuant peu ou prou la même valeur à chacune des graduations. Ce fait est néanmoins difficilement vérifiable en pratique, les termes extrêmes pouvant revêtir un sens différent en fonction des sujets.

L'évaluation à partir d'échelles sémantiques peut être réalisée en laboratoire, via une interface machine. Elle peut également être réalisée dans un cadre *in situ*, au moyen de questionnaires papiers (Jeon et al., 2013; Torija et al., 2013), ou, comme c'est de plus en plus le cas, au moyen d'une application sur téléphone portable (Ricciardi et al., 2015). L'outil présente plusieurs avantages : il permet une collecte des données sur serveur directement, et offre la possibilité d'enregistrer les environnements en train d'être évalués, ce qui répond en partie aux problèmes inhérents aux études *in situ*, notamment en ce qui concerne la reproductibilité des stimuli (cf. Section 2.4.4).

L'intérêt que suscite l'approche dimensionnelle réside dans le fait que les résultats obtenus sont facilement analysables et interprétables. Évaluer un environnement sonore au moyen d'échelles sémantiques et d'indicateurs objectifs permet d'obtenir une description sous la forme de descripteurs quantitatifs. Or il existe nombre de tests statistiques (cf. Annexe A), ou d'outils d'analyse dimensionnelle, directement applicables à ces données.

En fonction des objectifs de l'étude, on peut distinguer trois approches méthodologiques :

- *identification des descripteurs perceptifs* : dans cette approche, on distingue déjà deux méthodes :
  - dans la première, l'expérimentateur identifie des descripteurs perceptifs pertinents, sans avoir d'idée pré-établie sur leur nature. Ces descripteurs sont habituellement détectés à partir de l'analyse lexicale des descriptions verbales fournies par les sujets ;

- dans la seconde, l'expérimentateur sélectionne, au sein d'un groupe de descripteurs perceptifs et/ou objectifs donné, ceux qui rendent compte au mieux de l'évaluation des paysages sonores. Les descripteurs objectifs sont calculés à partir du signal sonore, les descripteurs perceptifs sont évalués sur des échelles sémantiques. Différentes techniques d'analyse statistique multidimensionnelle, comme l'analyse en composantes principales ou le positionnement multidimensionnel (*Multidimensional scaling*), permettent de faire émerger des dimensions linéairement non-correlées, qui rendent compte au mieux de la variabilité des données (Cain et al., 2013; Torija et al., 2013). Ces nouvelles dimensions n'ayant pas de valeurs physiques ou perceptives a priori, une inspection qualitative des scènes sonores est alors nécessaire, afin de les caractériser. Il est par ailleurs possible de tester d'éventuelles corrélations entre les descripteurs perceptifs et/ou les descripteurs objectifs. Par exemple, (Torija et al., 2013) évalue les corrélations entre 15 descripteurs perceptifs et 49 indicateurs acoustiques, via l'utilisation du coefficient de Pearson ;
- *étude de l'influence des indicateurs sur le descripteur* : l'expérimentateur, à partir d'un descripteur perceptif et d'une série d'indicateurs objectifs ou perceptifs donnés, évalue dans quelle mesure l'évolution du descripteur est contrainte par les indicateurs, l'objectif à terme étant de d'obtenir un modèle prédictif de la variation du descripteur. Dans ce but, (Lavandier and Defréville, 2006; Ricciardi et al., 2015) se servent tous les deux de la régression linéaire multiple afin de modéliser la variation de la qualité sonore en fonction d'indicateurs perceptifs (niveau sonore perçu, familiarité avec l'environnement, présence des sources sonores de voix) et objectifs ( $L_{Aeq}$ ,  $L_{A10}-L_{A90}$ ) ;
- *classification non-supervisée des scènes sonores* : là encore, dans cette approche, on distingue deux méthodes :
  - dans la première, l'expérimentateur, considérant des classes de paysages sonores données (e.g. parc, rue, marché), vérifie si les descripteurs varient d'un type d'environnement à l'autre. Il dispose d'outils statistiques (cf. Annexe A) permettant de tester l'existence de différences significatives entre les types d'environnements sonores (Hong and Jeon, 2013);
  - dans la seconde, l'expérimentateur, considérant un ensemble de paysages sonores, analyse directement l'espace décrit par les descripteurs afin de faire émerger des groupes de scènes sonores similaires, au sens des descripteurs. Il peut avoir recours à des techniques de clustering, comme par

exemple le clustering hiérarchique ascendant (Torija et al., 2013), ou encore à d'autres méthodes non-supervisées, inspirées des réseaux de neurones, comme les cartes auto-organisées (*Self Organizing Map, SOM*) (Ricciardi et al., 2015). Les différents descripteurs pouvant être corrélés entre eux, il lui est également possible d'utiliser des outils d'analyse dimensionnelle, comme l'analyse en composante principale, permettant de générer de nouvelles dimensions décorrélées, et de sélectionner celles qui expliquent le mieux la variance des données.

Contrairement à l'approche catégorielle, l'approche dimensionnelle laisse peu de liberté au sujet, ce dernier étant contraint d'utiliser les échelles qui lui sont présentées pour décrire l'environnement. L'utilisation de ces échelles suppose que les attributs qu'elles décrivent puissent être évalués de manière linéaire et uni-dimensionnelle, ce que le sujet n'est pas toujours en mesure de réaliser. (Rimbault, 2006) montre notamment que des échelles sensées évaluer la structure temporelle d'une scène sonore (stable/instable ou figé/évolutif) ne conviennent pas, cette notion n'étant pas comprise par les sujets comme étant bipolaire.

L'utilisation d'échelles comprend par ailleurs plusieurs risques :

- les échelles peuvent être mal interprétées par le sujet, ou même ne pas faire sens. Une description détaillée des échelles, ainsi que l'utilisation de plusieurs mots pour en définir les extrémités, permettent de réduire ce biais potentiel. (Hall et al., 2013) évalue ainsi l'agrément en utilisant une échelle de 9 points dont les extrémités sont décrites ainsi : désagréable-mécontent-insatisfait / agréable-content-satisfait. Il est, en outre, possible de s'assurer de la pertinence des termes associés aux extrémités en soumettant le sujet à un questionnaire libre (Guastavino and Katz, 2004), réalisable en condition *in situ* (Hong and Jeon, 2013; Kang and Zhang, 2010), ou en lui demandant de commenter verbalement l'annotation desdites extrémités (Rimbault, 2006);
- tous les sujets peuvent ne pas utiliser les échelles de la même manière. Certains sont portés à en utiliser toutes les valeurs. D'autres à n'en privilégier que certaines, écartant les extrêmes en particulier. Une normalisation des données, avant analyse, est possible, pour réduire l'impact de ce biais (Defréville et al., 2004; Hong and Jeon, 2013; Lavandier and Defréville, 2006; Nielbo et al., 2013). Cette normalisation est obligatoire, s'agissant d'échelles de notation (*e.g.* attribution d'une note entre 0-10, 0-100 *etc.*), échelles aux extrémités, *a priori*, non décrites/expliquées. Rien ne garantit, en effet, que la valeur subjective donnée à une note (*e.g.* 5/20) soit la même pour tous les sujets. Elle s'impose dans une moindre mesure, s'agissant d'échelles

sémantiques. Les valeurs aux extrémités de ces échelles étant fixées (*e.g.* agréable *vs.* désagréable), il est difficile d'apprécier si des écarts de résultats trahissent, chez les sujets, un problème d'interprétation des valeurs, ou rendent compte de la subjectivité des sensations. Le fait est, par ailleurs, que les données provenant d'analyses sensorielles comprennent souvent des réponses extrêmes (*outliers*). La normalisation, dans ce cas, peut fausser sensiblement les données ;

- en général, pour un environnement donné, la valeur finale d'une échelle est calculée en moyennant les réponses de plusieurs sujets. Pour être valide, cette approche suppose que la distribution des réponses sur l'échelle soit unimodale. Or il a déjà été montré que ces distributions peuvent être multi-modales, du fait, entre autre, des variations d'interprétations de l'échelle entre les sujets, ou des différences d'appréciation relatives à d'autres facteurs (Raimbault, 2006). Il peut être utile d'inspecter les distributions des réponses avant de considérer des résultats moyen-nés.

Ainsi, dans le cadre de l'approche dimensionnelle, il est important de s'assurer que :

1. les échelles soient aptes à décrire les attributs qu'elles évaluent ;
2. les échelles soient correctement interprétées par les sujets.

### 2.7.3 Descripteurs perceptifs des paysages sonores

Nous détaillons, dans la suite de cette section, les descripteurs perceptifs ayant fait l'objet d'une attention particulière dans les approches dimensionnelles. Il est à noter qu'il n'existe pas de consensus dans la communauté, ni concernant la définition de ces descripteurs, ni concernant les pratiques expérimentales, permettant d'étudier ces descripteurs.

Par pratique expérimentale, nous comprenons, entre autre, la nature des échelles à utiliser (nombre de points, termes aux extrémités), leur analyse, ainsi que l'application d'éventuelles étapes de normalisation (Aletta et al., 2016).

#### 2.7.3.1 Gêne et bruit

La gêne provoquée par un paysage sonore, et en particulier par un paysage sonore urbain, est un des descripteurs perceptifs les plus étudiés. Ce fait est notamment lié au besoin pressant de trouver une solution à la pollution sonore en ville. Une récente étude indique que 86% des français se disent gênés par le bruit extérieur lorsqu'ils se trouvent chez eux (Bendavid and Chasles-Parot, 2014). La question

posée est :

« Quels sont les sons responsables de la gêne, et comment est-il possible de prévoir leurs effets en considérant, d'une part, leurs caractéristiques physiques, d'autre part, les facteurs extra-sonores ? ».

La problématique des bruits urbains s'est imposée avant celle des paysages sonores. Elle a déjà été étudiée en profondeur, (Marquis-Favre et al., 2005a,b). C'est notamment dans ce cadre qu'ont été introduits, dans les années 1990, la grande majorité des indicateurs psychoacoustiques (Zwicker and Fastl, 1990)(cf. Tableau 2), indicateurs encore utilisés aujourd'hui (Fiebig et al., 2009; Hall et al., 2013; Yang and Kang, 2013).

On évalue principalement la gêne en considérant l'influence des bruits issus des transports (routiers, aériens, ferroviaires), ou de l'industrie (Gille and Marquis-Favre, 2016; Gille et al., 2016a; Klein et al., 2015; Trollé et al., 2015). Plusieurs modèles permettant de prédire la gêne générée par ces sources ont déjà été proposés (Miedema, 2004; Miedema and Oudshoorn, 2001), et ces derniers continuent d'être revisités/améliorés (Gille et al., 2016b). Aujourd'hui, une attention particulière est portée à l'influence des facteurs extra-sonores comme par exemple l'activité du sujet, sa sensibilité au bruit, mais également le sentiment de peur potentiellement suscité par les sources sonores auxquelles il est exposé (trafic, industrie) (Marquis-Favre and Morel, 2015; Morel et al., 2016). Afin d'être valides écologiquement, certaines de ces études ont recours à des dispositifs expérimentaux assez lourds, allant par exemple jusqu'à recréer en laboratoire l'environnement d'un salon, et demander aux sujets de pratiquer des activités du quotidien durant l'exposition aux stimuli (Marquis-Favre and Morel, 2015).

Ces études sur la gêne mettent cependant l'accent sur les sons non-souhaités, responsables du bruit, et n'intègrent pas ou peu l'effet compensatoire d'autres sources mieux acceptées (Aletta et al., 2016).

#### 2.7.3.2 Qualité sonore

La qualité sonore se veut être un descripteur général, prenant en compte de manière globale les qualités affectives perçues. La question posée est alors :

« Est-ce que l'environnement est bon ou mauvais ? ».

Plusieurs études ont tenté de proposer des modèles ou indicateurs permettant de prédire cette notion de qualité.

En comparant des indicateurs objectifs relatifs au niveau sonore, au contenu spectral, ainsi qu'à la fluctuation temporelle, (Nilsson and

Berglund, 2006; Nilsson et al., 2007) montrent que c'est le niveau sonore qui permet d'expliquer l'essentiel de la variance des qualités perçues. Le contenu spectral et la fluctuation temporelle n'ont, eux, qu'un intérêt limité.

(García Pérez et al., 2012) propose un indicateur acoustique de la qualité, nommé *ESEI*, qui prend en compte, à la fois, un indicateur objectif de niveau global, un indicateur objectif de la présence de différentes sources, et un indicateur subjectif fixe de la qualité hédonique de chacune des sources. La qualité des sources est établie sur la base de questionnaires. Elle dépend notamment du lieu dans lequel est entendue la source. Par exemple, les auteurs indiquent que les voix d'enfants sont majoritairement bien acceptées, sauf sur les places publiques.

La régression linéaire multiple est un outil souvent utilisé afin de modéliser la qualité d'un environnement (Ricciardi et al., 2015). (Brocolini et al., 2012) montrent notamment que cet outil permet d'obtenir des prédictions comparables à celles obtenues via l'utilisation de méthodes non-linéaires comme les réseaux de neurones artificiels. Notons néanmoins ici que le faible nombre de données disponibles pour entraîner le réseau peut limiter sa capacité de généralisation, et donc ses performances. Très souvent, ces modèles sont construits à partir de descripteurs globaux, relatifs aux sons, mais également au contexte visuel. Ils intègrent par ailleurs des descripteurs caractérisant de manière séparée les contributions spécifiques des différentes sources sonores (cf. Section 2.7.6) (Brocolini et al., 2012; Ricciardi et al., 2015). La forte influence du contexte visuel sur la qualité de l'environnement a aussi été montrée dans (Hong and Jeon, 2013).

On utilise également la notion de préférence afin d'évaluer la qualité globale d'un environnement (Yu and Kang, 2010). (Hong and Jeon, 2013) montre, par ailleurs, que la préférence est influencée par le confort acoustique ressenti (cf. Section 2.7.3.4).

### 2.7.3.3 Agrément

La notion d'agrément interroge la qualité hédonique de l'environnement. La question posée est :

« Est-ce que l'environnement est agréable ou désagréable ? »

Contrairement aux recherches sur la gêne et le bruit, les études sur l'agrément adoptent une approche positive, et s'intéressent aux sons bénéfiques pour la qualité des environnements. Cette démarche implique généralement de considérer séparément la contribution des différentes sources (cf. Section 2.7.6) (García Pérez et al., 2012; Lavandier and Defréville, 2006).

Le contexte (physique, visuel, social, personnel, cf. Section 2.7.3.8) semble être d'une grande importance dans l'évaluation de l'agrément (Guillén and López Barrio, 2007).

#### 2.7.3.4 *Confort acoustique*

Une autre notion très proche de l'agrément est celle du confort acoustique (Jeon et al., 2011, 2013; Tse et al., 2012). Comme pour l'agrément, le confort semble dépendre plus du type de source perçu (Yang and Kang, 2005), et du contexte d'exposition (Meng et al., 2013), que des caractéristiques physiques globales de l'environnement. En s'appuyant sur l'utilisation d'un réseau de neurones, (Yu and Kang, 2009) montre notamment qu'il est peu probable qu'un même modèle puisse rendre compte efficacement du confort sonore d'un environnement urbain, sans faire de distinction entre les différents types de lieux d'écoute.

#### 2.7.3.5 *Calm et tranquillité*

(Delaitre et al., 2012) a effectué une analyse lexicale du vocable français utilisé depuis le XVI<sup>e</sup> siècle pour décrire la notion d'environnement calme. Il propose la définition suivante :

« An area in spatial or temporal break from the outside activities, whose acoustic environment is favorable to physical or psychological rest. »<sup>13</sup>

Concernant les environnements tranquilles, (Pheasant et al., 2008) propose la définition suivante :

« A quiet, peaceful and attractive place to be in, *i.e.*, a place to get away from everyday life. »<sup>14</sup>

Bien qu'il puisse exister des différences, les notions de tranquillité et de calme sont très proches, et la distinction entre les deux est rarement faite dans la littérature (Delaitre et al., 2012).

Les études sur le calme sont complémentaires des études sur la gêne. Les qualités d'un environnement peuvent impacter les réponses physiologiques d'un sujet (Hume and Ahtamad, 2013). Si l'on admet que le bruit peut être la cause d'une dégradation de la santé (Stansfeld et al., 2005), on reconnaît au calme des vertus régénératrices (De Coensel and Botteldooren, 2006; Payne, 2013).

Le calme semble être lié à la régularité temporelle de l'environnement (Delaitre et al., 2012). Une scène stable et amorphe (cf. Section 2.7.4.2 pour la définition de amorphe), composée de peu d'événements saillants, peut être vue comme un environnement très calme.

<sup>13</sup> Une zone séparée spatialement ou temporellement des activités extérieures, et dont l'environnement acoustique favorise le repos physique ou psychologique.

<sup>14</sup> Un endroit calme, paisible et attrayant, permettant de s'échapper du quotidien.

Suivant cette idée, un indicateur du calme perçu (nommé *slope*) a été proposé par (Memoli et al., 2008). Cet indicateur prend en compte l'évolution temporelle du niveau sonore, le nombre d'événements occurrents dans l'environnement, et la manière dont ces éléments émergent du fond sonore.

En utilisant la régression linéaire multiple, (Pheasant et al., 2009; Pheasant et al., 2008) ont proposé un modèle de la tranquillité perçue dans un environnement urbain (nommé *Tranquillity Rating*) tenant compte du niveau sonore, ainsi que du pourcentage d'éléments naturels contenus dans l'environnement visuel. L'effet bénéfique sur le calme ressenti des sons d'origine naturelle, comme d'origine humaine, a été aussi observé par (De Coensel et al., 2013).

Considérant un milieu rural, (De Coensel and Botteldooren, 2006) ont montré que le calme perçu est en partie dû à des facteurs extra-sonores relatifs aux caractères congrus de l'environnement. Partant de l'hypothèse qu'un paysage sonore rural est par essence calme et « revigorant », les auteurs proposent de considérer des indicateurs centrés sur les sons venant briser la tranquillité inhérente à cet environnement (voiture, tracteur).

#### 2.7.3.6 Propriétés combinées

Au lieu de ne considérer qu'un descripteur, il est également possible d'évaluer l'environnement sur la base d'une combinaison de descripteurs.

Dans une première étude, (Kang, 2006) montre que les dimensions liées à la relaxation et au dynamisme, entre autres, sont pertinentes dans l'évaluation des paysages sonores. En demandant à des sujets de noter 116 descripteurs perceptifs sur des échelles sémantiques unidirectionnelles, et en appliquant une analyse en composantes principales, (Axelsson et al., 2010) montrent que 3 d'entre ces descripteurs permettent d'expliquer 74% de la variance des données, en particulier l'agrément (50%), la présence d'événements (18%, *eventfulness*) et la familiarité (6%). Enfin, Cain et al. (Cain et al., 2013) proposent de caractériser l'environnement urbain suivant deux dimensions orthogonales (cf. Figure 14), l'une caractérisant le calme, et l'autre le dynamisme (*vibrancy*).

Si l'on admet que les descripteurs d'agrément, de relaxation et de calme sont proches, ces trois études présentent des résultats constants (Davies et al., 2013).

Un paysage sonore est essentiellement perçu à travers les curseurs de qualité que sont l'agrément (agréable/calme) et le dynamisme (nombre d'événements).

Les études précédemment citées considèrent comme stimuli des enregistrements de sources sonores isolées. (Hall et al., 2013) montrent que dans le cas d'enregistrements de mixtures sonores, ces deux mêmes dimensions (agrément et dynamisme) permettent d'expliquer 71% de

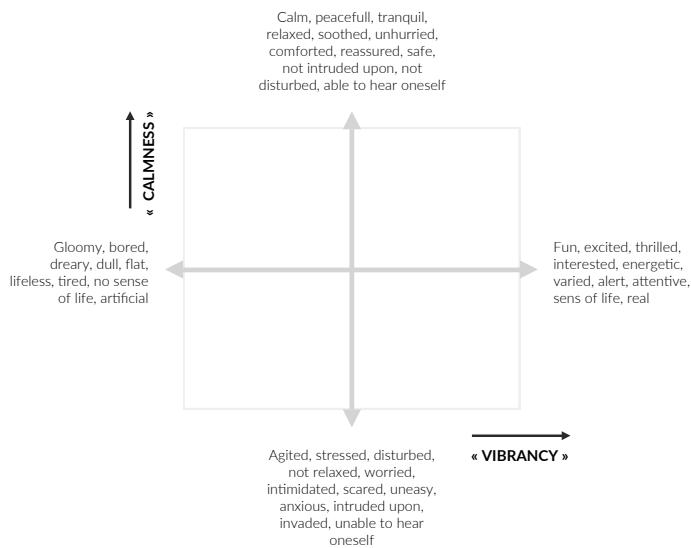


FIGURE 14 : Les dimensions de calme et de dynamisme permettant de caractériser l'environnement sonore urbain, d'après (Cain et al., 2013).

la variance. Cependant les auteurs indiquent, d'une part, qu'il n'y a pas de relation évidente entre ces deux dimensions, d'autre part, que des descripteurs objectifs acoustiques seuls ne permettent pas de prédire avec précision les valeurs perceptives de ces dimensions.

#### 2.7.3.7 Autres descripteurs

Plusieurs autres descripteurs sont utilisés pour rendre compte de la qualité d'un paysage sonore, qualifiant les propriétés communicationnelle, spatiale et dynamique de l'environnement (Kang and Zhang, 2010), mais également son caractère « musical » (Botteldooren et al., 2006).

#### 2.7.3.8 Influence d'attributs extra-sonores

Comme les exemples précédents le suggèrent, la perception d'un paysage sonore, et *a fortiori* les processus cognitifs activés, sont très liés à un contexte.

Les recherches sur les paysages sonores ont permis de montrer que les qualités sonores d'un environnement dépendent, entre autre, du contexte environnemental (température, chaleur, humidité), (Jeon et al., 2011; Meng et al., 2013), de la configuration spatiale du lieu d'exposition (Hall et al., 2013), des éléments visuels (De Coensel and Botteldooren, 2006; Guillén and López Barrio, 2007), et enfin d'un contexte de « justesse » (*appropriateness*) (De Coensel and Botteldooren, 2006; Nielbo et al., 2013), *i.e.* la manière dont l'environnement sonore s'accorde avec l'activité du lieu. Ces recherches ont également

conduit au constat que ces qualités perçues sont spécifiques au sujet, et relatives notamment à son âge, son sexe, et sa sphère socio-culturelle (Guillén and López Barrio, 2007; Hall et al., 2013; Yu and Kang, 2010).

#### *2.7.4 Catégoriser les sources et paysages sonores*

Alors que les études précédemment citées envisagent l'environnement sonore de manière holistique, celles que nous présentons dans la suite comprennent la scène sonore comme un object composite.

##### *2.7.4.1 Catégories de sources sonores*

Parmi les travaux les plus influents sur la catégorisation des sources sonores, on trouve ceux de W. W. Gaver (Gaver, 1993a,b). Il propose d'envisager le problème sous un angle phénoménologique, en considérant l'action physique à l'origine, plutôt que l'objet. L'action étant très dépendante de la nature physique de l'objet, il trie les causes suivant trois types d'objets :

- solide : heurter, gratter, rouler, déformer ;
- liquide : goutter, éclabousser, clapoter ;
- gaz : exploser, souffler.

D'autres études ont montré l'importance des phénomènes physiques originels dans les processus de catégorisation et d'interprétation des sons (Lemaitre et al., 2010; Marcell et al., 2000). (Houix et al., 2012) notamment demande à des sujets de catégoriser librement 60 sons environnementaux en se concentrant sur l'action, et de commenter leurs groupes. Une analyse des groupements révèle que la catégorisation s'opère suivant deux étages hiérarchisés, le premier, le plus général, comprenant des catégories d'objets très proches de celles proposées par Gaver (Solide, Liquide, Gaz et Machine), le deuxième, plus spécifique, comprenant des catégories d'actions. Une seconde expérience similaire, réalisée uniquement sur des objets solides, montre que la nature du pattern temporel (continu ou discret) résultant de l'action à l'origine du son influe de manière significative sur la catégorisation. Ce dernier point est également observé par (Gygi et al., 2007). Sur la base d'une matrice de similarité obtenue à partir de comparaisons par paires, et via un positionnement multidimensionnel en trois dimensions, Gygi et al. montrent que les sons s'organisent en trois clusters incluant les sons harmoniques, les sons d'impacts (discrets), et les sons continus. Une épreuve de catégorisation semi-libre (les sujets devant réaliser au minimum 5 clusters), avec verbalisation, pratiquée dans la même étude, montre par ailleurs que les sujets catégorisent les sons principalement en fonction du type de sources (animaux,

Nom des catégories les plus citées

Sons naturels	Voix	Voiture
Sons humains	Enfant	Machine
Sons technologiques/artificielles	Cloche	Vent
Trafic	Background	Aboiement chien
Oiseaux	Événement	Bruits de pas
Musique	Avion	
Travaux	Eaux	

TABLE 3 : Les catégories sonores les plus citées, d'après (Niessen et al., 2010).

homme, véhicule, mécanique, musique, eau), moins fréquemment en fonction du contexte et du lieu (extérieur, sport, bar), et rarement sur la base de caractéristiques physiques isolées (hauteur, fréquence), ou d'émotions ressenties (ennuyeux, alarmant).

L'influence de la nature de la source, et de la sémantique associée, sur les processus de catégorisation a particulièrement été étudiée. A partir d'une étude de 35 articles traitant de catégories sonores, (Niessen et al., 2010) établissent une liste des 20 catégories de sons les plus citées. La liste est présentée dans le Tableau 3.

La grande majorité de ces catégories sont des catégories de sources sonores. Seules deux font référence à des catégories d'objets sonores abstraits (événement, *Background*). Ces catégories de sources ne s'expriment pas toutes au même niveau d'abstraction. Certaines sont précises (*aboïement chien*), d'autres sont très larges (*sons naturels*). Par ailleurs, certaines sont incluses dans d'autres (*bruits de pas*<*sons humains*), les trois catégories ayant le périmètre le plus large, et englobant toutes les autres, sont *sons naturels*, *sons humains* et *sons technologiques/artificiels*. Comme nous allons le voir (cf. Section 2.7.4.2 et 2.7.6), c'est en partie suivant ce découpage catégoriel que s'opère la perception de l'environnement.

Plusieurs études se basent sur une analyse linguistique de descriptions spontanées et libres d'environnements sonores, afin d'établir des catégories de sources sonores. Dans ces études, il est d'usage de demander explicitement au sujet de distinguer les aspects plaisants et désagréables du paysage étudié.

En réalisant une étude *in situ* d'environnements de parcs, (Szermeta and Zannin, 2009) mettent en évidence 9 catégories de sources sonores. Certaines sont systématiquement connotées positives (*oiseaux*, *nature*), d'autres sont systématiquement connotées négatives (*machine*, *alarme/signaux*, *train*). Quelques unes cependant peuvent être jugées alternativement positives ou négatives, comme *personne* (majoritairement positive), *trafic routier* (majoritairement négative), *musique*, *trafic aérien*.

L'étude de (Guastavino, 2006) utilise une méthode d'analyse similaire, mais en demandant aux sujets de décrire un environnement urbain idéal (plaisant), sur la base de leur mémoire uniquement. Des résultats similaires sont observés, *i.e.* les catégories *oiseaux* et *nature* sont systématiquement positivement perçues, les catégories *klaxon* et *travaux* sont systématiquement négativement perçues, les catégories *personne* et *musique* ayant une connotation variable.

L'auteure fait remarquer que les sujets décrivent les sons en s'appuyant sur la source émettrice de ces derniers. Il y a donc une assimilation entre l'objet et le phénomène acoustique. En conséquence, la sémantique (le sens) liée à l'objet intervient dans le processus perceptif (dans ce cas le jugement hédonique), au même titre que les propriétés acoustiques. L'observation des appréciations relatives aux catégories de véhicules vont dans ce sens : les catégories *trafic* (*voiture, moto/scooter, camion*) sont toujours mal perçues, à la différence des catégories *transports publics* (*bus* et *train*), toujours bien perçues. La représentation positive que nous avons des *bus* fait que ces sons, bien que proches de ceux de voiture, sont le plus souvent bien acceptés.

Des travaux de (Guastavino, 2006) nous retenons que, dans le cas d'une expérimentation sans support audio, les sujets travaillant de mémoire, les attributs sémantiques entrent pour une large part dans la représentation des environnements sonores urbains. Nous approfondissons ce point à la section [4.2.11.1](#).

#### 2.7.4.2 Catégories de paysages sonores

Outre les catégories de sources sonores, plusieurs études s'intéressent à la formation de catégories d'objets plus complexes, les paysages sonores.

V. Maffiolo (Maffiolo, 1999) met en évidence deux processus distincts engagés, en fonction de la capacité de l'auditeur à identifier des événements sonores. Dans cette étude, les sujets doivent 1) catégoriser des enregistrements d'environnements sonores urbains, 2) décrire les groupements effectués. A partir d'une analyse linguistique des descriptions verbales, Maffiolo montre l'existence de deux catégories cognitives abstraites d'environnements sonores respectivement appelées : « les séquences événementielles » et « les séquences amorphes ». Les séquences événementielles sont des environnements composés d'événements saillants et identifiables (*démarrage de voiture, voix d'homme*). Les séquences amorphes sont des environnements dont il est difficile d'isoler des éléments distincts.

Chacune de ces catégories a été sous catégorisée suivant différentes stratégies :

- les scènes événementielles ont été sous-catégorisées en fonction :

- 1. des types de sources présentes ;
- 2. de la qualité affective de l'environnement (agréable, désagréable, ennuyant, agressif, insupportable, calme).
- les scènes amorphes ont été sous-catégorisées en fonction :
  - 1. de l'agrément perçu (agréable/désagréable) ;
  - 2. de l'évaluation des propriétés acoustiques, à savoir l'intensité sonore, le contenu spectral (haute/basse fréquence), et la structure temporelle (continu, discontinu).

On remarque ainsi que les scènes événementielles profitent d'une analyse descriptive basée sur l'identification des sources sonores, alors que les scènes amorphes bénéficient d'une analyse holistique, à partir d'indicateurs acoustiques (subjectifs) globaux. On note que les deux catégories suscitent un jugement hédonique (plaisant/non-plaisant).

Cette distinction (événementiel/amorphe) s'opère aussi au niveau de la source sonore. Analysant des descriptions libres des sources sonores peuplant l'environnement urbain, Guastavino (Guastavino, 2006) montre que les descriptions des sons à basse fréquence peuvent se diviser en deux catégories appelées « événements sonores » et « bruit de fonds ». S'agissant des bruits de fond, aucune des sources les composant ne peut être identifiée.

Rimbault et Dubois (Rimbault and Dubois, 2005), combinant les résultats obtenus par trois thèses (Guastavino, 2003; Maffiolo, 1999; Rimbault, 2002), montrent que la catégorisation des paysages sonores s'opère suivant leurs compositions en termes de sources sonores (cf. Figure 15).

Une première distinction intervient entre, d'un coté, les paysages sonores comportant des sons de *transports motorisés* et/ou de *travaux*, de l'autre, des paysages sonores comportant des sons suggérant une présence humaine. Ces derniers se subdivisent encore entre, d'une part, les paysages « vivant », d'autre part, les paysages « relaxant », composés également de sons de *nature*. Le rôle prédominant joué par l'activité humaine dans la catégorisation des environnements était déjà pressenti par Schaefer (Schaefer, 1977).

Des résultats très similaires sont obtenus par (Guastavino, 2007). Passant par une tâche de catégorisation libre avec verbalisation, Guastavino montre que cette catégorisation tient compte de l'occurrence de sons d'origine humaine, de même qu'elle s'appuie sur un jugement hédonique des sources. En présence de sons d'origine humaine, une distinction est faite entre les environnements dominés par les sons humains, et ceux dominés par des sons mécaniques. Enfin, les « environnements humains » sont scindés en fonction de l'activité et du lieu (parc calme/marché animé), les « environnements mécaniques » en fonction de la présence ou non de sons humains.

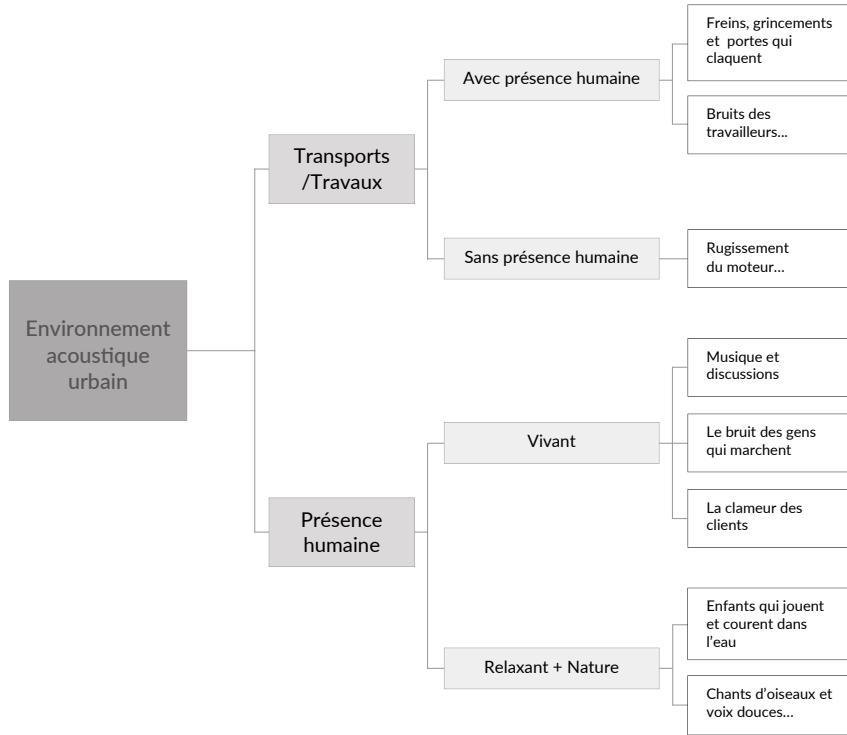


FIGURE 15 : Catégorisation des paysages sonores urbains, d'après (Rimbault and Dubois, 2005)

### 2.7.5 Classifier les sources et environnements sonores

Contrairement à la section précédente, où il est question de catégories, *i.e.* représentation mentale, nous traitons ici de classes. Par classe on entend un groupe d'objets qui ne fait pas référence à une entité mentale particulière, mais dont l'association vient d'une volonté de classer, d'organiser des environnements, ou des sources, suivant leurs caractéristiques physiques, morphologiques, ou encore suivant leurs fonctions. Le but est alors, sur la base de descripteurs objectifs, d'étudier les similarités existant entre ces groupes, (cf. Section 2.7.2.2).

#### 2.7.5.1 Classes de sources sonores

Un des buts premiers des études sur les classes sonores est d'établir la typologie complète de tous les types de sources peuplant un environnement donné.

Sur la base de l'étude de (Rimbault and Dubois, 2005), et dans l'idée de proposer une nomenclature générique pour décrire les sources sonores présentes en milieu urbain, (Brown et al., 2011) propose une taxonomie reprise à la figure 16. Cette classification est centrée sur l'objet. Partant de là, (Salamon et al., 2014) propose, lui, une nouvelle taxonomie, plus détaillée, centrée à la fois sur l'objet et sur l'action

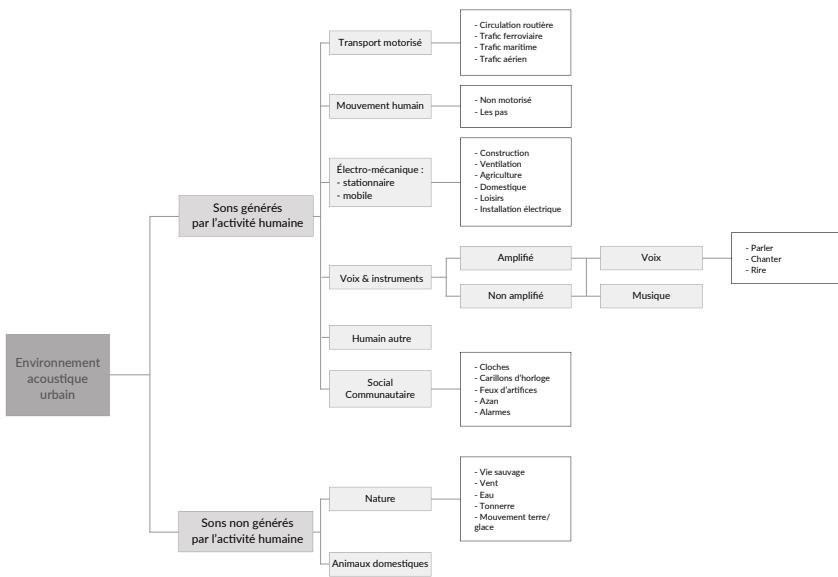


FIGURE 16 : Taxonomie des sources sonores urbaines, d'après (Brown et al., 2011).

(cf. Figure 19). Pour les auteurs, la réalité sonore d'un objet diffère en fonction de son utilisation (*passage de voiture vs. freinage de voiture vs. klaxon de voiture*). Pour rendre compte de ce fait, certaines classes d'objets de plus bas niveau sont subdivisées en classes d'actions, labellisées par des verbes.

Outre organiser les sources, il est aussi utile de comprendre quelles sont les différences acoustiques qui peuvent se manifester entre plusieurs classes de sons. (Yang and Kang, 2013), sur la base d'indicateurs acoustiques et psychoacoustiques, compare des classes de sons d'environnements urbains, (*musique, mécaniques, trafic*), et des classes de sons d'environnements naturels, (*eau, vent, oiseaux*). Chaque indicateur est calculé sur le signal, à l'aide d'une fenêtre glissante, et moyenné. En réalisant une analyse en composante principale de ces indicateurs, les auteurs montrent que l'intensité (*loudness de Zwicker*), le contenu spectral *sharpness* et la structure temporelle *fluctuation* sont les trois principaux indicateurs permettant d'expliquer la variance entre ces différents types de sons. Ce fait avait déjà été observé dans d'autres études mêlant différents stimuli (Botteldooren et al., 2006; De Coensel and Botteldooren, 2006).

#### 2.7.5.2 Classes de paysages sonores

Nombre d'études analysent l'existence de similarités entre environnements sonores à partir d'indicateurs quantitatifs objectifs (Rychtáriková and Vermeir, 2013), subjectifs (Jeon et al., 2013), ou même les

deux à la fois (Ricciardi et al., 2015; Torija et al., 2013). La méthodologie est presque toujours la même :

1. pour chaque environnement, calculer des indicateurs acoustiques/psychoacoustiques, ou évaluer des indicateurs perceptifs à l'aide d'échelles sémantiques ;
2. utiliser des outils de clustering afin d'établir des classes d'environnements similaires.

Sur la base de descripteurs subjectifs uniquement, (Jeon et al., 2013) identifient 4 classes comprenant respectivement, des environnements dominés par le bruit urbain, des environnements dominés par des sons de nature, des environnements urbains ouverts (place), ou encore des environnements équilibrés (sons urbains et naturels). La distinction se fait en grande partie à partir d'indicateurs de préférence liés au confort acoustique, mais également à l'impression visuelle, et à la configuration spatiale du lieu.

Sur la base de descripteurs subjectifs et objectifs, (Torija et al., 2013) établit 15 classes de paysages sonores, les distinctions s'opérant au niveau des sources sonores présentes/absentes (trafic, oiseaux, fontaine, moto, sirène, parc, humain). Parmi les indicateurs acoustiques, ceux tenant compte de la dynamique du niveau sonore (*crest factor*), ainsi que du niveau des basses fréquences (L à 125Hz), suffisent à expliquer 84% de la variance des données. Les auteurs concluent que l'utilisation de descripteurs acoustiques peut permettre, seule, d'isoler les paysages sonores similaires. Conclusion reprise par (Rychtáriková and Vermeir, 2013).

#### *2.7.6 Contributions des différentes sources sonores*

Comme nous l'avons vu, plusieurs études adoptant l'approche catégorielle ont permis de montrer que l'identification de certaines sources sonores, ainsi que leur sémantique associée, jouent un rôle important dans l'évaluation perceptive des paysages, en particulier au niveau de l'agrément perçu (Defréville et al., 2004; Gozalo et al., 2015; Guastavino, 2006; Nilsson, 2007; Szeremeta and Zannin, 2009).

Dans la continuité, d'autres études, adoptant, elles, l'approche dimensionnelle, tendent de plus en plus à compléter les indicateurs globaux avec des indicateurs caractérisant les contributions spécifiques de différentes sources sonores. Pour ce faire, elles partent toutes d'une liste de catégories de sources pré-établie. A partir de cette liste, elles calculent des indicateurs acoustiques spécifiques à ces sources, et/ou demandent à des sujets d'en évaluer les caractéristiques perceptives.

En menant des travaux *in situ* sur la qualité de différents environnements, (Nilsson, 2007; Nilsson et al., 2007) mettent en évidence que l'identification des sources sonores permet une meilleure approche

de la qualité globale de l'environnement que la simple mesure du niveau sonore. Ils montrent ainsi que les sons *technologique/mécanique* ont un impact négatif sur l'environnement alors que les sons *naturels* ont un impact positif. Les sons *humains* restent, eux, neutres. En outre, dans les conditions d'une exposition modérée aux bruits de trafic, l'ajout de sons positivement perçus (*naturels* dans leur cas) peut potentiellement améliorer la qualité de l'environnement. Observation déjà faite (Galbrun and Ali, 2012; Hong and Jeon, 2013). Cependant, dans les conditions d'une exposition forte à ces mêmes bruits, une politique de réduction des niveaux est obligatoire.

(Defréville et al., 2004; Lavandier and Defréville, 2006) évaluent l'impact séparé de différentes sources de trafic, (*voiture, moto, scooter, bus*), de sons humains, (*voix adultes, voix enfants*), et de sons naturels, (*oiseaux*), sur l'agrément perçu. Pour chacune de ces sources ils calculent des indicateurs objectifs de niveaux ( $L_{Aeq}$ ,  $L_{A10}$ ) et de présence (nombre d'occurrences, pourcentage de temps présent), ainsi que des indicateurs perceptifs (présence, proéminence, proximité). Des indicateurs globaux de niveau (objectif : *loudness de Zwicker*, et subjectif : niveau perçu) sont également pris en compte. La régression linéaire multiple est utilisée afin de mesurer l'influence de ces indicateurs sur l'agrément.

Que l'on considère les indicateurs subjectifs ou objectifs, l'utilisation combinée de l'indicateur de niveau global avec des indicateurs spécifiques aux différentes sources permet d'augmenter la capacité de prédiction de la qualité sonore, comparé à l'utilisation seule de l'indicateur de niveau global. Là encore, les auteurs montrent que, sur les environnements peu exposés au trafic, les sons d'*oiseaux* et d'*humain* ont un effet bénéfique. Ils notent, par ailleurs, que l'appréciation de la présence des *voitures* diffère en fonction du type d'environnement. Dans un parc, elles sont vues négativement, alors que dans la rue, elles font partie de l'environnement, et n'influencent pas (de manière isolée) la qualité perçue.

Dans une étude d'envergure, comprenant 3400 réponses collectées sur deux villes (Paris et Milan), et utilisant une méthodologie proche de celle de (Lavandier and Defréville, 2006), Ricciardi *et al.* (Ricciardi et al., 2015) testent plusieurs modèles permettant de prédire la qualité sonore. Ces modèles sont bâtis à partir d'indicateurs perceptifs globaux, sonores et visuels, ainsi que d'indicateurs perceptifs sonores spécifiques à différentes sources. Les modèles tenant compte des indicateurs visuels produisent des sorties corrélées à 72% avec la qualité mesurée. Ce chiffre décroît pour tomber à 58%, dès lors qu'on supprime les indicateurs visuels, et à 19%, en ne considérant plus que le niveau sonore global (sans les indicateurs spécifiques aux sources). Les auteurs clusterisent les différents environnements sur la base de ces indicateurs. 6 classes émergent, les regroupements s'opérant, cette

fois encore, en fonction de la présence/absence de diverses sources sonores. Plus spécifiquement, certains groupements sont liés :

1. à la possibilité de distinguer, ou non, des sources sonores dans les scènes (scènes événementielles *vs.* amorphes) ;
2. à la présence majoritaire d'une classe de sons en particulier (*traffic, humain, nature*) ;
3. à la présence simultanée de plusieurs sources.

En recalculant des modèles pour chacune des classes, les auteurs montrent que les indicateurs relatifs à des sources sont plus représentés dans les modèles par classes, mais varient significativement d'une classe à l'autre. Par exemple, l'indicateur correspondant à *oiseaux* n'apparaît, dans le modèle, que pour la classe dominée par des sons *naturels*. Ces résultats questionnent l'utilité et l'efficacité d'une modélisation de la qualité sonore qui se voudrait générale, *i.e.* applicable pour tous types de situations et d'environnements.

#### 2.7.7 *Discussion*

De cette section nous retenons qu'il existe deux manières d'aborder la problématique des paysages sonores : l'approche dimensionnelle qui étudie les indicateurs permettant de les caractériser, et l'approche catégorielle, qui détermine leurs éléments d'intérêts.

En ce qui concerne l'approche catégorielle, nous observons que catégoriser ou classer les sources s'effectue principalement sur la base de leurs caractéristiques physiques (niveau sonore, contenu fréquentiel, pattern temporel). *A contrario*, lorsqu'il s'agit de scènes sonores, c'est avant tout à partir de la présence des sources et des qualités affectives perçues que s'opèrent les groupements. On a ainsi une gradation de l'importance des caractéristiques physiques en fonction de l'objet. Pour des objets simples et décontextualisés (sources), la primauté va aux caractéristiques physiques. Pour des objets composites (scènes), les caractéristiques physiques perdent de leur importance au bénéfice d'indicateurs perceptifs ou sémantiques, *i.e.* relatifs à la présence des différentes sources.

En ce qui concerne l'approche dimensionnelle, nous observons que les indicateurs acoustiques seuls ne permettent pas de rendre compte de la perception d'un environnement sonore. Les indicateurs perceptifs jouent un rôle primordial. Enfin, nous notons l'intérêt d'étudier les contributions des différentes sources sonores de manière séparée, pour un descripteur donné, cette approche étant, entre autre, dictée par les résultats des études catégorielles sur la prédominance des indicateurs sémantiques dans les jugements de ressemblance inter-scènes.

Nous pensons que la théorie ancrée de la cognition motive naturellement l'approche dimensionnelle des paysages sonores. En effet,

dans le cas d'une représentation amodale, les qualités affectives, qui découlent des processus cognitifs, n'ont *a priori* aucune raison de dépendre des propriétés physiques de l'environnement, mais plutôt de la combinaison des concepts activés. Dans le cas d'une approche ancrée, il est tout à fait envisageable de chercher des corrélats entre une mesure physique, et une qualité, cette dernière étant codée, via les symboles perceptifs, par une information modale, entre autre.

Cependant, le penchant naturel de la perception à être catégorielle, (à isoler des objets/flux, et à les associer à des catégories), et le fait que les représentations des concepts liés à ces catégories soient situées par rapport à un contexte, peuvent nous amener à penser que la relation entre une caractéristique physique et une qualité varie 1) en fonction de l'objet qui la porte, 2) en fonction du contexte d'écoute de l'objet. Ce fait va ainsi dans le sens des études ci-devant évoquées (cf. Section 2.7.6) sur les contributions spécifiques des différentes sources sonores sur les qualités perçues.

## 2.8 ÉVÉNEMENTS ET TEXTURES SONORES

### 2.8.1 Définition

S'éloignant de l'approche des paysages sonores, plusieurs études se sont concentrées sur l'analyse perceptive d'un certain type de sons, appelés textures sonores.

Pour définir la texture sonore, nous nous appuyons sur la définition donnée par (Saint-Arnaud, 1995, p. 25):

- « les textures sonores sont des objets composites, formés d'éléments de base appelés atomes ; »
- « les atomes apparaissent suivant un pattern haut-niveau pouvant être soit périodique (galop), soit aléatoire (pluie) ; »
- « les caractéristiques haut-niveau des textures restent constantes sur de longues périodes de temps, ce qui implique qu'elles ne peuvent comporter aucun message complexe ; »
- « le pattern haut-niveau doit être présenté au moins une fois dans sa totalité pendant une période de temps n'excédant pas quelques secondes. Cette période est nommée période d'attention (*attention span*). »

Cette définition est avant tout morphologique, la texture étant définie en fonction de ses caractéristiques physiques. Cela vient, entre autre, du fait que la texture a d'abord été étudiée dans le cadre du traitement du signal, beaucoup d'applications multimédia ayant besoin de modèles permettant de synthétiser de tels sons (Schwarz, 2011). La notion de texture s'oppose intuitivement à la notion d'événement

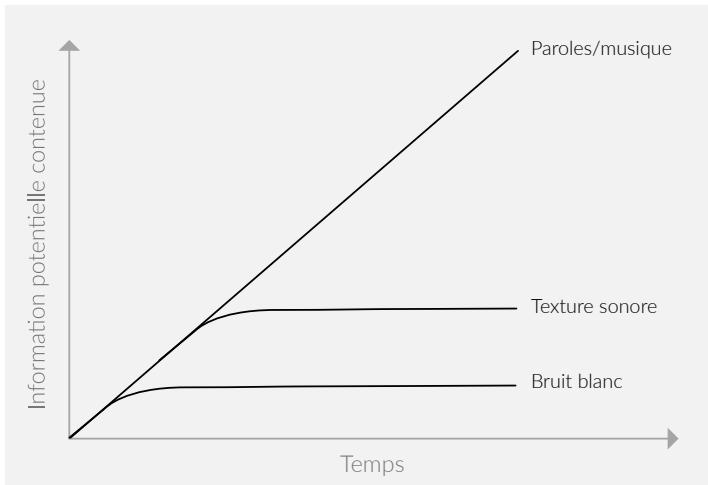


FIGURE 17 : Information potentielle contenue dans les séquences d'événements, les textures, et le bruit. D'après (Saint-Arnaud, 1995).

sonore, ou encore à celle de séquence d'événements. Par opposition, l'événement est vu comme un élément discret, un son court et non homogène.

C'est par la notion d'information transmise que semble se faire la distinction entre texture et séquence d'événements. Les caractéristiques des textures restant stables au cours du temps, l'information transmise finit par atteindre une asymptote. En revanche, une succession d'événements distincts, comme c'est le cas pour une séquence musicale ou de parole, transmet de plus en plus d'informations dans le temps (cf. Figure 17). En poussant le raisonnement à l'extrême, le bruit blanc peut être vu comme la représentation la plus « pure » d'une texture, ce dernier étant porteur d'une information très limitée.

Cette dimension événement/texture est orthogonale à celles de « bruit de fond » / « événements de premier plan » (*background/foreground*), utilisées dans le langage courant pour discriminer l'environnement urbain. Concernant les notions de *background* et de *foreground*, nous considérons que l'une et l'autre peuvent être vues comme des flux auditifs, ces derniers pouvant être composés de textures et d'événements regroupés dans le but de faciliter le traitement auditif de la scène.

### 2.8.2 Percevoir les textures

Contrairement aux événements sonores, la texture est un objet simple, dont le traitement cognitif ne requiert pas une analyse poussée.

Cela a été mis en évidence par Josh H. McDermott et ses co auteurs (McDermott and Simoncelli, 2011; McDermott et al., 2013). S'inspirant du fonctionnement de l'oreille humaine, et notamment des processus auditifs intervenant depuis la cochlée jusqu'au thalamus, ils ont pu

établir un modèle permettant de re-synthétiser des textures sonores en ne se servant que de statistiques simples, calculées à partir de représentations temps-fréquence de signaux de textures enregistrés.

Dans une première expérience (McDermott and Simoncelli, 2011), la capacité des sujets à identifier les textures synthétisées a été testée. Les résultats ont montré que les sons de synthèse étaient aussi bien identifiés que les sons enregistrés. McDermott démontre ainsi qu'une information résumée sous la forme de statistiques est suffisante, d'un point de vue cognitif, à la reconnaissance. Dans le cas des textures, ces statistiques constituent même l'unique information disponible, le système auditif ayant fait fi de toute autre représentation plus détaillée (Nelken and Cheveigné, 2013).

Dans une seconde expérience (McDermott et al., 2013), les sujets ont dû reconnaître, parmi une triade de sons synthétisés, celui produit par une source différente (*i.e.* un type de texture différent, cf. Figure 18). Les résultats ont montré que la capacité de discrimination est fonction de la durée des textures. Plus cette dernière est élevée, plus la capacité à discriminer est importante. Ce constat valide les hypothèses formulées par (Saint-Arnaud, 1995) sur l'existence d'une période d'attention nécessaire au cerveau afin de percevoir le stimulus comme une texture. Ces résultats ont aussi montré que le processus de traitement de l'information sonore comprend une prise de décision quant à la nature des stimuli, laquelle va ensuite influer sur la manière d'analyser l'information montante. L'expérience prouve que cette prise de décision n'a rien d'anodin, car, dans le cas où le cerveau perçoit une texture, il décide sciemment de dégrader l'information, en la résumant de manière statistique.

Le fait qu'un jugement perceptif s'améliore avec la durée des stimuli est un principe bien connu en perception des sons (Moore, 1973). Une troisième expérience de (McDermott et al., 2013) a montré, cependant, que cette vérité n'était pas toujours vérifiée. Au cours de cette expérience, les sujets, soumis à trois exemplaires d'un même type de textures (*e.g.* trois sons synthétisés de pluie) dont deux étaient produits à partir des mêmes statistiques extraites, ont dû identifier le troisième, issu de statistiques différentes (Figure 18). Les résultats ont montré que la capacité des sujets à discriminer le bon stimulus décroît avec la durée des stimuli. Ce fait, qui peut sembler paradoxal, est une conséquence directe du choix du cerveau de ne traiter les textures que sur la base de statistiques. Le signal sonore étant analysé suivant des fenêtres d'intégrations successives (Poeppel, 2003; Yabe et al., 1998), plus les stimuli sont longs, plus le système auditif est confiant dans le fait qu'il a à faire à des textures, et plus il tend à conserver une information réduite. La réduction de cette information finit éventuellement par gommer les différences fines qui existent entre les stimuli, ce qui ne permet plus de faire la distinction entre eux.

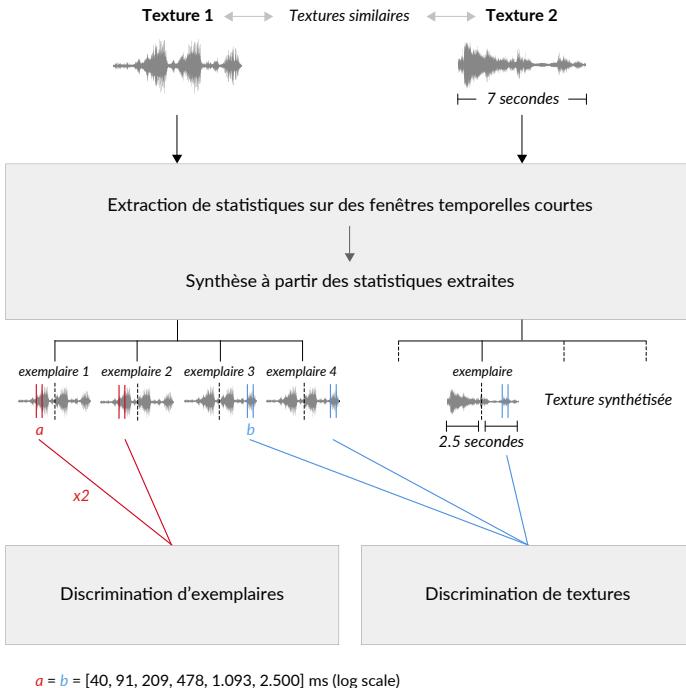


FIGURE 18 : Plannification expérimentale des expériences de discrimination de textures sonores et d'exemplaires de textures sonores menées par (McDermott et al., 2013)

Une des avancées majeures de ces études est qu'elles apportent de nouvelles réponses sur la nature des images mentales représentées dans le système auditif. Dans le cas des textures, il s'agirait ainsi de descripteurs bas-niveau, résumés sous la forme de statistiques simples. Cette découverte fait sens d'un point de vue écologique, car elle respecte le principe d'économie de moyens. Le cerveau, reconnaissant que les caractéristiques des textures n'évoluent pas au cours du temps, ne conserve qu'une information condensée, qui lui permet pourtant de traiter des sons potentiellement longs.

### 2.8.3 Discussion

Nous commençons pas remarquer que la distinction opérée entre une texture et une séquence d'événements est d'ordre structurel. Il est possible de voir une texture comme une séquence composée d'événements qui ont perdu leur signification individuelle à cause de l'organisation temporelle de la séquence. Une étude préliminaire, menée dans le cadre de cette thèse, tend à montrer qu'il existe en effet un seuil minimal d'espacement entre les événements, seuil en deçà duquel il n'est plus possible de les discriminer. Elle est présentée en annexe B. A l'inverse, il semble qu'il soit aussi possible « d'individualiser » un extrait de texture. (Agus et al., 2010) montre qu'une

séquence d'une seconde, composée de deux segments de bruit blanc (par définition une texture) identiques de 0.5 seconde, peut être identifiée parmi d'autres séquences de bruits blancs composées d'un seul segment d'une seconde. De tels extraits peuvent être reconnus même plusieurs semaines après l'écoute. Dans ce cas le cerveau emmagasine bien la totalité du signal acoustique. Les segments contenant une répétition ne sont donc pas interprétés comme des textures.

La perception « statistique » des textures peut s'inscrire dans une vision ancrée de la cognition. En effet, le fait que le cerveau s'appuie sur un résumé statistique des propriétés du signal pour reconnaître les textures suggère que la représentation mentale de tels objets comprend cette information statistique. La nature des informations contenues par ces représentations n'est ainsi pas seulement amodale, mais également modale (L'expérience de (Agus et al., 2010) sur la reconnaissance des bruits blancs va aussi dans ce sens). Par ailleurs, on est tenté de conjecturer que la décision prise par le système auditif de ne conserver en mémoire qu'une information condensée du signal est conditionnée par l'identification de la nature globale de l'objet perçu (S'agit il d'un événement ou d'une texture). Si tel est le cas, alors on est bien en présence de processus perceptifs (résumer l'information) qui fonctionnent de concert avec des processus cognitifs (identification la nature de l'objet).

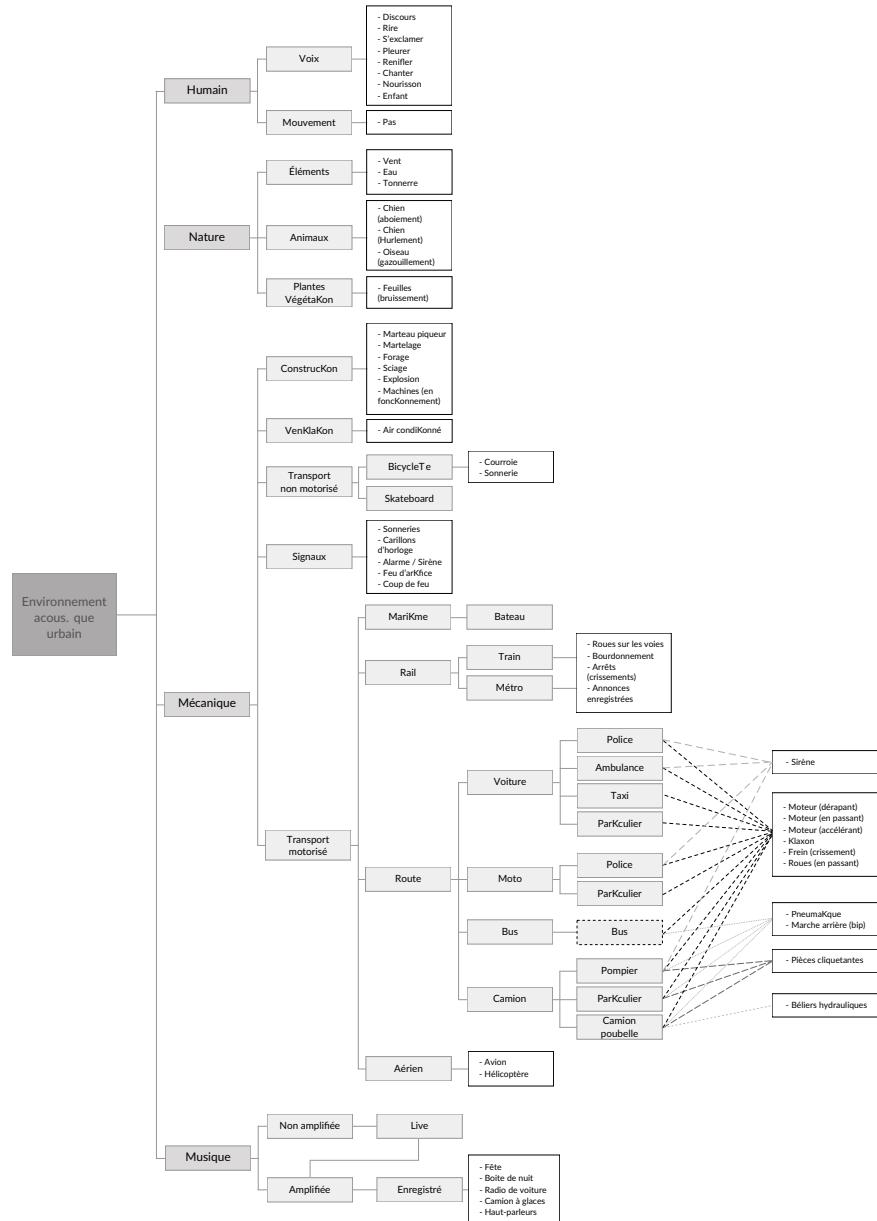


FIGURE 19 : Taxonomie des sources sonores urbaines suivant la nomenclature source-action, d'après (Salamon et al., 2014).

# 3

## MODÈLE MORPHOLOGIQUE ET SIMULATION DE SCÈNES SONORES ENVIRONNEMENTALES

---

### 3.1 INTRODUCTION

Nous appuyant sur les connaissances relatives au fonctionnement du système auditif présentées au chapitre précédent (cf. Chapitre 2), nous développons, en phase préliminaire à nos travaux, un modèle morphologique de scènes sonores environnementales. Il ne s'agit pas ici de proposer un modèle analytique de scènes sonores au sens physique ou psychophysique, mais plutôt de satisfaire à un intérêt pratique, *i. e.* proposer un modèle permettant de facilement simuler des scènes sonores dont nous contrôlons les caractéristiques structurelles. Ces caractéristiques sont supposées être des facteurs importants entrant en compte dans l'analyse des scènes, qu'il s'agisse d'analyse sensorielle, ou d'analyse automatique.

Ce chapitre comprend trois sections. La première explicite les considérations perceptives sur lesquelles s'ancre le modèle. La seconde présente une formalisation du modèle morphologique. La troisième motive l'utilisation du modèle dans le cadre des deux cas d'étude choisis : la perception de l'agrément dans un milieu urbain, et l'évaluation des algorithmes de détection d'événements sonores.

### 3.2 FONDEMENT PERCEPTIF DU MODÈLE MORPHOLOGIQUE

#### 3.2.1 *L'unité : la source sonore*

Les études portant sur l'ASA, et plus spécifiquement sur les processus de ségrégation, montrent, d'une part, que l'homme fait sens de son environnement en isolant les informations relatives aux différentes sources qui le composent, d'autre part, que ce groupement intervient très tôt dans la chaîne de traitement, et se base sur des règles génériques innées (cf. Section 2.6).

Dans le même temps, l'approche de l'ASA par les neurosciences montre que le système auditif tend à générer des images mentales (« objets auditifs ») des sources ainsi isolées, et que c'est à partir de ces images qu'il adapte son traitement de l'information montante (cf. Section 2.6.6).

Enfin, la recherche sur les paysages sonores, adoptant l'approche catégorielle, met en évidence que les processus de catégorisation s'appuient également sur la composition sémantique des scènes, *i. e.* les sources sonores identifiées (cf. Section 2.7.4.2).

Pour modéliser un environnement sonore, il semble pertinent de considérer comme élément de base la source sonore. Comme vu précédemment (cf. Section 2.3.3), la notion de source sonore est variable, un même objet pouvant être reconnu suivant plusieurs degrés d'abstraction.

### 3.2.2 *L'objet : la séquence sonore*

Les études portant sur l'ASA (psychophysiques et neurosciences), montrent que des événements émis par une même source ont tendance à être groupés dans un même flux, et traités comme une seule entité, par le système auditif (cf. Section 2.6.5).

Par ailleurs, la nature catégorielle des représentations mentales suggère encore cette tendance *a priori* naturelle de l'homme à considérer comme équivalents des objets partageant un certain nombre de propriétés en commun (cf. Section 2.3).

De fait, il n'apparaît pas nécessaire, pour le modèle, de gérer séparément deux éléments sonores proches et relevant d'une même source. Ces éléments sont regroupés dans une séquence sonore. Cette séquence forme l'objet de base du modèle, celui dont nous contrôlons les caractéristiques.

### 3.2.3 *Une typologie source-action*

La constitution de banques de sons peuplant l'environnement urbain est incontournable. Avant d'acquérir ces sources, *i.e.* de les enregistrer, il est nécessaire de les identifier. Une démarche naïve consisterait alors à établir la liste exhaustive de toutes les classes de sources sonores composant l'environnement. Une telle approche soulèverait deux problèmes :

- une source sonore peut se décrire en fonction de plusieurs niveaux d'abstraction. Identifier et nommer sont des actions déterminées par notre représentation mentale du monde (cf. Section 2.3). Cette représentation s'organise, entre autre, suivant l'axe vertical des niveaux d'abstraction sur lequel s'appuie notre travail de catégorisation. Ainsi, si deux individus entendent un même son de voiture, il est possible que le premier le nomme « voiture » et le deuxième « moteur ». Le dénombrement de l'ensemble des sources pouvant être utilisées par le modèle doit prendre en compte ce fait. Ces sources doivent être regroupées en classes hiérarchisées, afin de bâtir une structure taxonomique ;
- il n'existe pas de taxonomie standardisée des sources sonores. Pourtant, c'est une tradition des sciences modernes de classer et nommer les éléments avant de les étudier. Dans les domaines

de la faune ou de la flore, une observation longue et minutieuse des sujets d'étude a permis d'élaborer un système de classification (taxonomie), et ainsi d'organiser et trier les objets en fonction de leurs propriétés partagées. Elle a permis encore d'élaborer une terminologie précise des classes d'objets. Grâce à quoi, la Biologie est devenue une science, laquelle science devait donner naissance à la théorie de l'évolution (Lecointre and Le Guyader, 2006). Dans le domaine du son, en revanche, point de système de classification (Dubois, 2000; Niessen et al., 2010). Nous trouvons à cela deux explications :

- *champ lexical limité* : l'identification et la description d'un son sont des processus subjectifs étroitement liés au langage (cf. Section 2.3.2.2). Deux sujets appartenant à deux groupes sociaux différents n'utiliseront pas les mêmes mots pour décrire un même objet. Pour établir un système de classification, il faut prendre une décision quant à la définition des termes utilisés. Or, contrairement au domaine de la vision, où une terminologie de base pour décrire les objets (couleur, forme etc.) est globalement partagée, le champ lexical applicable aux phénomènes acoustiques est, d'une part, limité (durée, fréquence...) (Dubois, 2000), d'autre part, emprunté, dans une large mesure, à d'autres domaines perceptifs. On parle ainsi de brillance, ou de rugosité des sons. La diversité des termes descriptifs, et l'absence de consensus sur ce qu'ils désignent, rend difficile l'élaboration d'une classification standardisée ;
- *influence du contexte* : L'identification et la description d'un son sont dépendantes du contexte (cf. Section 2.3.5), *i. e.* de la nature des sources co-occurentes dans la scène (Ballas and Howard, 1987; Gygi and Shafiro, 2011; Niessen et al., 2008).

Il apparaît clairement que les classes de sons peuplant notre environnement doivent être organisées autour d'une taxonomie : un système de classes hiérarchisé. Cependant il y a un choix à faire quant à la manière de regrouper les sons à l'intérieur de cette taxonomie.

Comme vu à la section 2.7.4, plusieurs études ont montré que la catégorisation des sources sonores s'opère suivant des attributs sémantiques. Parmi ceux-ci, deux reviennent souvent :

- la source (agent, objet, fonction), *i. e.* l'objet émettant le son ;
- l'action, *i. e.* le mouvement physique à l'origine du son.

Ces deux attributs fonctionnent de pair. S'inspirant de l'organisation catégorielle verticale à trois niveaux de Rosch (cf. Section 2.3.3.3),

Guyot et al. (Guyot et al., 1997) proposent un système de catégorisation où les auditeurs identifient des groupements de sources abstraites au niveau superordonné (« Bruit généré par une excitation mécanique »), des actions au niveau de base (« gratter », « frotter ») et des sources concrètes au niveau subordonné (« vaisselle », « stylo »). Reprenant à son tour ce système, (Houix et al., 2012) montre que les sons sont catégorisés, en premier lieu, à partir du type de sources concrètes, et ensuite, seulement, à partir d'actions.

L'association source-action semble être une base sensée sur laquelle bâtir une taxonomie où les classes haut-niveau sont des classes abstraites de sources sonores (« véhicule »), les classes intermédiaires, des classes de sources sonores (« voiture »), et les classes basses, des actions sonores (« passage »). Pour les classes de bas niveau, la perméabilité intra-classe est minimale.

Cette association source-action n'est cependant pas suffisante. Le choix des labels utilisés doit faire l'objet d'une sélection particulière. Ces labels doivent être génériques, compréhensibles, et décrire de manière non ambiguë les objets de la classe considérée. Afin de les déterminer, il est possible de se référer aux travaux de Gaver (Gaver, 1993b), qui propose une taxonomie phénoménologique des sons, à ceux de Niessen *et al.* (Niessen et al., 2010) qui, sur la base d'une étude bibliographique de près de 35 publications, établit la liste des catégories sonores les plus utilisées, également à ceux de Salamon *et al.* (Salamon et al., 2014), qui, partant des travaux de (Brown et al., 2011), et reprenant l'association source-action, élaborent une taxonomie de sons urbains.

### 3.2.4 Événements et textures

Nous avons montré que l'utilisation de la nomenclature basée sur l'association source-action nous permet de dénombrer et de trier l'ensemble des sons présents dans l'environnement.

Afin de permettre le développement d'un outil de création de scènes sonores simulées à partir d'enregistrements de sons réels, il est nécessaire de constituer un corpus d'enregistrements, corpus structuré suivant la taxonomie définie.

Le problème est alors, sur la base de cette taxonomie, d'enregistrer, pour chacune des classes, un nombre de sons suffisant. Considérant des environnements denses comme la ville ou la forêt, cette approche pose une question pratique de faisabilité.

Afin de contourner le problème, on peut s'appuyer sur des considérations perceptives pour définir, dans un contexte expérimental donné, quels sons requièrent d'être enregistrés séparément, quels autres peuvent l'être simultanément.

En effet, tous les sons n'ont pas le même intérêt. Une voix humaine peut facilement être isolée du reste des sons concurrents (Carlyon,

2004). Inversement, un fond sonore de trafic urbain est moins informatif que d'autres sons ponctuels et proches (Southworth, 1969). Maffiolo montre à ce sujet (cf. Section 2.7.4.2) l'existence de deux processus cognitifs distincts dont l'activation dépend de la nature des environnements : l'analyse holistique, s'agissant de scènes amorphes, *i. e.* sans événements apparents, et l'analyse descriptive (sur la base d'une information sémantique extraite à partir des événements connus), s'agissant de scènes événementielles, *i. e.* comprenant des événements identifiables.

Par ailleurs, le cerveau a tendance à résumer l'information extraite, lorsqu'il détecte qu'une séquence n'est composée que d'un mélange de sons similaires, et que ces sons n'enrichissent pas l'information. Cette particularité a déjà été évoquée (cf. Section 2.8).

Ces éléments nous amènent à penser que les processus de ségrégation dépendent de la nature structurelle de l'environnement. Lorsque des événements émergent d'un environnement sonore, le cerveau traite l'information des différentes sources de manière séparée. Plusieurs flux auditifs sont ainsi générés *i. e.* un pour chaque séquence d'événements émis par la même source. Inversement, quand le cerveau ne parvient pas à isoler d'événement, la scène est traitée globalement, tous ces constituants étant agglomérés dans un même flux.

Ainsi, quatre types de sons semblent pouvoir être isolés :

- événement sonore : un son isolé, ponctuel, dont les caractéristiques physiques varient au cours du temps ;
- texture sonore : un son, isolé, long, dont les caractéristiques physiques restent stables au cours du temps, et analysé à partir de statistiques extraites d'une représentation temps-fréquence ;
- *scène événementielle* : une séquence contenant une information sémantique élevée ;
- *scène amorphe* : une séquence contenant une information sémantique faible.

Une scène événementielle est une séquence composée soit uniquement d'événements, soit d'événements et de textures, les événements porteurs d'une information plus riche primant quant au choix du processus de traitement à mettre en œuvre.

Les textures et les scènes amorphes, elles, sont traitées de manière holistique, à partir de propriétés acoustiques globales pour les scènes amorphes (Dubois et al., 2006; Maffiolo, 1999), et sur la base d'une information résumée statistiquement pour les textures (McDermott et al., 2013). Toutes deux portent une information limitée (Nelken and Cheveigné, 2013; Saint-Arnaud, 1995). Cependant, les séquences amorphes sont spontanément décrites par les sujets comme des « fonds sonores » (Guastavino, 2006; Maffiolo, 1999), induisant qu'elles n'existent

que grâce à un processus de construction de flux auditifs, alors que les textures sont des objets définis seulement sur la base de leur nature physique. Il est cependant possible d'assimiler une scène amorphe à une texture, ses caractéristiques physiques demeurant stables au cours du temps. De fait, nombre de scènes amorphes (« brouhaha de rue », « brouhaha de trafic ») sont citées comme textures. Cependant, l'inverse, considérer une texture comme une scène amorphe, n'est pas forcément vrai. Un exemple de texture souvent cité est le son du « galop », son qui peut très bien apparaître au premier plan de la scène.

Afin de limiter le nombre d'enregistrements nécessaire, il est donc possible d'enregistrer directement des mixtures de sons, à la condition qu'elles puissent être considérées comme des textures, la définition de cette dernière notion englobant les scènes amorphes.

### 3.3 DESCRIPTION DU MODÈLE MORPHOLOGIQUE

#### 3.3.1 Classe et collection de samples

La scène sonore est vue comme une somme de sources sonores, ou, si l'on admet la distinction opérée entre événements et textures sonores, « un squelette d'événements sur un lit de textures » (Nelken and Cheveigné, 2013).

D'un point de vue pratique, ces éléments sonores sont enregistrés. Ils sont nommés samples.

**Définition 1** *Un sample est un enregistrement d'un son isolé, qu'il s'agisse d'un événement ou d'une texture.*

Les samples, regroupés en classes de sons hiérarchisées, sont organisés selon la taxonomie préétablie. Un exemple est donné figure 20. Les niveaux hiérarchiques de la taxonomie sont appelés niveaux d'abstraction. Les classes ayant un niveau d'abstraction élevé constituent un regroupement conceptuel de samples ayant potentiellement des caractéristiques variées (*e.g.* Humain). Plus le niveau de la classe est bas, plus le regroupement est précis, englobant des samples similaires (*e.g.* voix-adulte-cri).

**Définition 2** *Une classe est une collection de samples jugés perceptivement équivalents. Si le niveau d'abstraction d'une classe est tel que cette dernière possède des sous-classes, alors sa collection de samples est la somme des collections respectives de chacune des sous-classes.*

Les classes de niveau d'abstraction élevé sont nommées uniquement à l'aide de termes abstraits désignant, de manière globale, les samples qu'elles regroupent (*e.g.* transport). Les classes de niveau

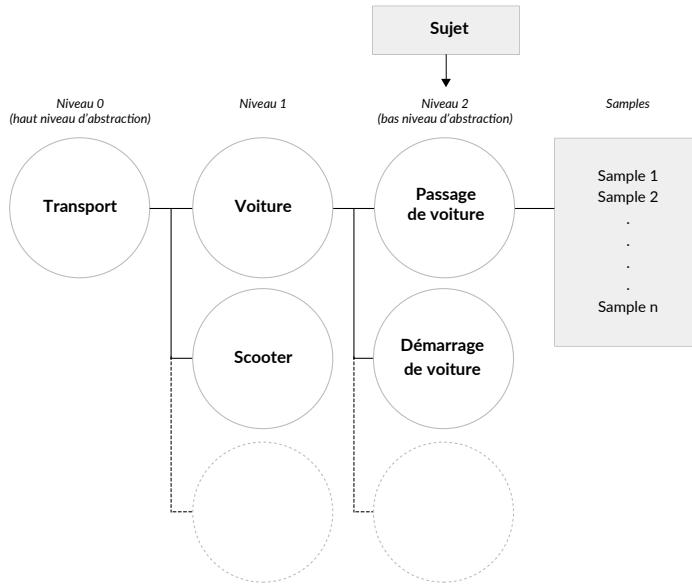


FIGURE 20 : Organisation hiérarchique de la banque de sons isolés utilisée pour la simulation.

moindre utilisent la nomenclature source-action (*e.g. passage de voiture*). Les classes des niveaux les plus bas correspondent à des collections de samples, par définition, équivalents les uns aux autres.

### 3.3.2 Séquences de samples

Chaque classe de sons faisant partie de la scène est liée à une piste. Cette piste est une séquence temporelle où sont positionnés les différents samples. Elle est le correspondant simulé du flux auditif.

**Définition 3** *Une piste est une séquence temporelle composée de samples appartenant à une même classe de sons.*

La construction de la taxonomie (nombre de classes, nombre de niveaux d'abstraction), dépend, évidemment, de la tâche considérée.

L'ensemble des pistes, ainsi que leurs paramètres, forment ce que nous appelons un scénario sonore.

**Définition 4** *Le scénario sonore désigne l'ensemble des propriétés des pistes composant une scène, à savoir, les classes de sons liées aux pistes, et leurs paramètres structurels (niveau, espacement, début et fin, cf. Section 3.3.3).*

### 3.3.3 Paramètres

En suivant la terminologie ci-devant introduite, une scène sonore est vue comme une somme de pistes. Chaque piste est une séquence

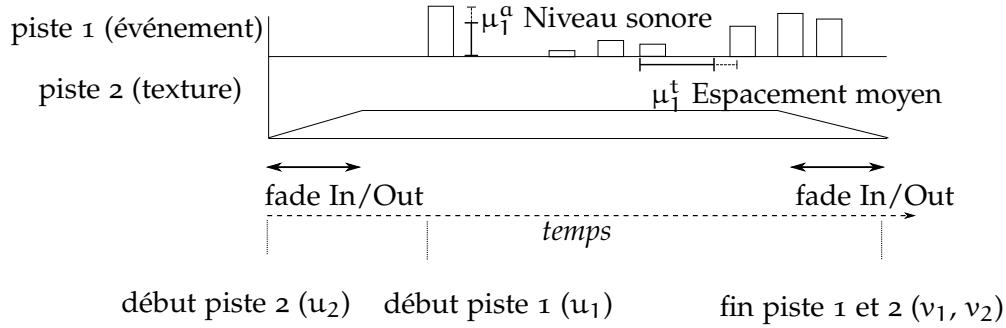


FIGURE 21 : Représentation schématisée des pistes du modèle de scènes sonores.

temporelle, dont la structure dépend d'une série de paramètres (cf. Figure 21). Une description exhaustive de la scène supposerait de connaître au moins les positions et les niveaux exacts de chaque occurrence de chaque sample. Ces informations seraient trop nombreuses pour être correctement maîtrisées et interprétées. Nous proposons donc de réaliser une approximation en posant l'hypothèse d'une distribution normale des valeurs. Le modèle ne propose pas d'interagir avec un sample en particulier, mais toujours avec une séquence de samples.

Nous isolons trois attributs globaux permettant de contrôler une piste :

- *niveau* : la moyenne/variance des niveaux des samples ;
- *espacement* : la moyenne/variance des espacements inter-onsets entre les samples ;
- *durée* : le début et la fin de la piste.

Le modèle fait une distinction explicite entre la gestion des pistes d'événements, et la gestion des pistes de textures. En effet, la notion de texture ne peut se comprendre que pour un son continu. Une piste de texture est donc composée de samples concaténés les uns aux autres, sans espacement (cf. figure 21). Pour qu'une piste de texture soit « plausible », *i.e.* qu'on ne détecte pas de discontinuité flagrante, elle doit être une séquence composée de samples provenant de la même source, et obtenus avec un matériel (et des réglages) identique(s).

Le scénario sonore est défini à partir de paramètres stochastiques. Il existe une infinité de réalisations pour un scénario donné. Nous nommons partition, une réalisation particulière d'un scénario.

**Définition 5** *La partition désigne l'ensemble des propriétés des samples composant une réalisation particulière d'un scénario sonore donné (classe, niveau, onsets et offsets de chacun des samples).*

### 3.3.4 Formalisation du modèle

La formalisation présentée ici vaut uniquement pour les classes d'événements sonores. Nous décrivons par la suite les diverses contraintes qui s'appliquent pour une classe de textures.

En considérant  $s$ , une scène composée de  $z$  classes de sons, le modèle de  $s$  se définit comme suit :

$$s(n) = \sum_{i=1}^z p_i(n) \quad (1)$$

avec  $n$  un indice temporel discret, et  $p_i$  la piste correspondant à la classe  $c_i$ . La classe  $c_i$  est composée de  $|c_i|$  samples  $c_{i,m}$ ,  $1 < m < |c_i|$ .

Soit  $\mathcal{U}(x, y)$ , une distribution uniforme d'entiers allant de  $x$  à  $y$ , avec  $x < y$ . On définit  $E_j^i$  ( $j = \{1, 2, \dots, k_i\}$ ) une suite de variables aléatoires indépendantes et identiquement distribuées (iid) suivant la loi  $\mathcal{U}(1, |c_i|)$ .

$$E_j^i \text{ iid : } \mathcal{U}(1, |c_i|) \quad \forall j \quad (2)$$

Une piste  $p_i$  est vue comme une séquence de  $k_i$  samples d'événements. On définit  $e_j^i(n)$ , un événement choisi aléatoirement parmi les  $|c_i|$  samples de la classe  $c_i$  :

$$e_j^i = c_{i, E_j^i} \quad (3)$$

On définit  $A_j^i$  et  $T_j^i$  ( $j = \{1, 2, \dots, k_i\}$ ), deux suites de variables aléatoires iid. Pour chaque piste  $p_i$ , les  $A_j^i$  sont les facteurs d'amplitudes appliqués aux événements  $e_j^i$ , dont nous modélisons la distribution, par souci de simplicité, par une loi normale de moyenne  $\mu_a^i$  et de variance  $\sigma_a^i$ . De même, pour chaque piste  $p_i$ , les  $T_j^i$  sont les espacements inter-onsets, lesquels suivent une loi normale de moyenne  $\mu_t^i$  et de variance  $\sigma_t^i$ .

Soit  $\mathcal{N}(\mu, \sigma)$ , une distribution normale de moyenne  $\mu$  et de variance  $\sigma$ , on a alors :

$$A_j^i \text{ iid : } \mathcal{N}(\mu_a^i, \sigma_a^i) \quad \forall j \quad \text{et} \quad T_j^i \text{ iid : } \mathcal{N}(\mu_t^i, \sigma_t^i) \quad \forall j \quad (4)$$

Lors de la génération d'une scène, les valeurs des  $A_j^i$  et  $T_j^i$  sont instantanées par tirage au sort, suivant les distributions correspondantes.

On définit  $u^i$  et  $v^i$  les indices temporels de début et de fin de chaque piste  $p_i$  respectivement.

Formellement, une piste  $p_i$  se définit alors comme suit :

$$p_i(n) = \sum_{j=1}^{k_i} A_j^i e_j^i(n - n_j^i) \quad \text{avec} \quad n_j^i = n_{j-1}^i + T_j^i \quad (5)$$

où, par convention,  $n_0^i = u^i$  et  $p_i(n) = 0$  si  $n > v^i$ .

Les paramètres du modèle sont,  $\mu_a^i$ ,  $\sigma_a^i$ ,  $\mu_t^i$ ,  $\sigma_t^i$ ,  $u^i$  et  $v^i$ , et doivent être fixés pour chaque piste  $p_i$ . La figure 21 offre une illustration de l'action des paramètres introduits.

Pour les textures, deux distinctions sont à observer avec le modèle défini précédemment :

1. afin d'éviter toute sensation de discontinuité, deux samples de texture sont concaténés en considérant un recouvrement fixé, sur lequel est appliqué un fondu enchaîné (*cross-fade*) à valeur d'énergie constante entre les samples, afin de donner l'illusion de continuité ;
2. il n'y a qu'un facteur d'amplitude par piste ( $A^i : \mathcal{N}(\mu_a^i, \sigma_a^i)$ ), sa valeur s'appliquant à tous les samples.

### 3.4 UN MODÈLE POUR LA SIMULATION

#### 3.4.1 Choix de conception

Comme évoqué dans l'introduction, les choix de conception du modèle sont gouvernés par un intérêt pratique. Celui-ci, afin d'être facilement utilisable, doit comprendre un minimum de paramètres réglables, tout en offrant une grande expressivité. Ces considérations ont largement guidé le design des paramètres de contrôles (cf. Section 3.3.4).

Ainsi, les niveaux et les positions ne sont pas réglés individuellement pour chaque sample, mais de manière stochastique, par piste. Le fait de considérer des paramètres structurels qui s'appliquent à tous les événements d'une même classe, *i.e.* provenant d'une même source, fait sens pour les deux cas d'étude considérés :

- la détection automatique d'événements, dont le principe est de regrouper les événements sonores en fonction de leur classe d'appartenance ;
- la perception des paysages sonores, le système auditif traitant comme une seule entité les éléments successifs émis par une même source.

#### 3.4.2 Simulation et perception des paysages sonores

##### 3.4.2.1 Simulation et objectivation

Nous l'avons vu, les représentations mentales agissent sur la manière dont nous percevons les sons (cf. Section 2.3). Dans l'approche ancrée de la cognition (cf. Section 2.2), cette rétroaction s'effectue via la simulation cognitive, étape au cours de laquelle l'individu, sur la

base et de l'information sensorielle montante, et des représentations mentales, génère une image de l'environnement qui l'entoure. Cette image revêt alors une dimension modale, *i.e.* relative à des caractéristiques physiques.

Par ailleurs, nous avons vu que cette simulation cognitive peut également s'opérer sans stimuli, de manière introspective (cf. Section 2.2.5). Elle est alors conditionnée par des facteurs autre que perceptifs, dépendant de la tâche que l'individu cherche à accomplir.

Nous pensons que demander à un sujet de simuler un environnement sonore donné, permet d'objectiver l'image qu'il se fait de cet environnement, image dont la génération procède elle-même d'un processus de simulation (cognitif). La nature de l'image mentale générée est alors conditionnée par la consigne de l'expérience (*e.g.* simuler un environnement sonore urbain et calme).

Le modèle permettant au sujet de sélectionner les sources, et d'en régler les caractéristiques structurelles (niveau sonore, espace-ment, *etc.*), il est possible d'atteindre l'ensemble des informations véhiculées par l'image mentale simulée, *i.e.* les informations modales et sémantiques.

Cependant, nous soulignons ici qu'il serait inexact de penser que l'expérience de simulation ne nécessite, de la part du sujet, qu'un effort introspectif, et ce pour deux raisons :

- les capacités expressives du sujet sont conditionnées à sa maîtrise du simulateur. Il est donc primordial de considérer une interface de simulation simple et intuitive, afin de fluidifier le passage de l'image mentale à la scène sonore ;
- si les ressources cognitives du sujet sont potentiellement infinies, les ressources matérielles, *i.e.* les enregistrements de sons isolés, ne le sont pas. L'interprétation des scènes simulées est donc fonction de la diversité de la banque de sons disponible.

#### 3.4.2.2 *Un lien entre les approches catégorielles et dimensionnelles*

Comme vu à la section 2.7.2, l'étude expérimentale des paysages sonores adopte deux approches : l'approche catégorielle et l'approche dimensionnelle.

L'approche catégorielle vise à mettre en évidence des catégories d'environnements ou sources sonores. Pour ce faire, elle identifie les objets d'intérêt de l'environnement. L'approche dimensionnelle vise à mettre en évidence les dimensions perceptives engagées dans les processus cognitifs, et les indicateurs dont ces dimensions perceptives dépendent. Pour ce faire, elle identifie les éléments caractérisant les qualités affectives perçues d'une scène.

Dans une certaine mesure, si l'on interroge les influences qu'ont les différents éléments constituant une scène sur les qualités affectives perçues, les deux approches se complètent. L'approche catégorielle

permet d'établir la liste des éléments d'intérêt, cette liste servant de base à une annotation des stimuli utilisés par l'approche dimensionnelle afin d'étudier les contributions spécifiques de leurs éléments respectifs.

La simulation nous paraît offrir un cadre élégant permettant de faire le lien entre les deux approches.

En recomposant les environnements sonores à partir de banques de sons isolés (cf. Figure 22), la simulation rejoint la démarche catégorielle, même si, à l'inverse, celle-ci discrétise ces environnements sur la base de tris et/ou de descriptions verbales émanant des sujets (cf. Section 2.7.2.1). Ainsi, les catégories sonores, point de sortie des épreuves catégorielles, constituent-elles la banque de sons, point d'entrée de la simulation.

En produisant des environnements dont les caractéristiques structurelles et compositionnelles sont connues, la simulation propose des stimuli à partir desquels la démarche dimensionnelle évalue les contributions des différents éléments.

La simulation présente en outre :

- *un intérêt pratique* : afin d'étudier l'importance relative des différentes sources, il est indispensable de disposer de stimuli dont nous pouvons extraire les informations inhérentes auxdites sources. Une première approche, adoptée par (Lavandier and Defréville, 2006), a été d'annoter les stimuli. La méthode cependant est limitée. D'une part, l'opération est difficile à réaliser sur de grandes banques de données. D'autre part, connaître la position des différentes sources dans une mixture sonore ne permet pas d'isoler leurs caractéristiques physiques respectives, et donc de calculer des indicateurs acoustiques dédiés. En traitement du signal, la séparation des sources reste un problème ouvert(Vincent et al., 2014).

Par la simulation, nous obtenons directement le stimulus et son annotation. Qui plus est, celle-ci est produite par le sujet lui-même, et non par un tiers. Enfin, le fait de posséder des sons isolés permet de facilement calculer des indicateurs acoustiques spécifiques à chaque source sonore ;

- *un intérêt écologique* : la validité écologique des stimuli est un problème fondamental en analyse sensorielle. Dans le cas de l'analyse des qualités affectives perçues, où l'on demande au sujet « que pensez vous de la qualité Q de cet environnement ? », il s'agit de garantir que les stimuli proposés fassent sens par rapport à la représentation mentale que le sujet se fait du monde sonore, d'une part, de la qualité Q, d'autre part.

Il est possible, dans les approches classiques, de résoudre ces problèmes en étudiant, au préalable, les stimuli à enregistrer (cf. Section 2.7.2.1).

La simulation, en renversant la question posée (« générer un environnement qui corresponde à une certaine valeur de Q »), garantit la validité écologique des stimuli, par définition connectés à la représentation sonore du sujet ;

- *un intérêt en terme de représentativité des stimuli* : toute étude sensorielle, qu'elle soit *in situ* ou en laboratoire, doit sélectionner un nombre restreint d'environnements sonores à évaluer. Il s'agit, tant que faire se peut, de garantir que le substrat de stimuli proposé soit représentatif de l'ensemble des environnements étudiés, un déséquilibre dans l'élaboration dudit substrat pouvant affecter, *in fine*, l'évaluation des stimuli.

Dans le cas des études sur la perception des environnements urbains, il est d'usage d'isoler des zones d'intérêts (parc, rue, place, cf. Section 2.7.2.1), et de répartir équitablement les stimuli parmi ces zones. Cependant, l'environnement d'une même zone est changeant, aussi bien s'agissant du type de sources présent, que s'agissant de la structure des patterns temporels émis par ces sources (*e.g.* pour une même rue passante, un son de trafic sera plus dense, composé de plus d'événements de voitures à certaines heures du jour). Il est donc nécessaire de contrôler la diversité des sources qui y occurrent, ainsi que la diversité structurelle de leurs séquences d'émission, *a fortiori* si l'on cherche à étudier l'influence spécifique des différentes sources. Cette étape est complexe.

Si la structure interne des paysages sonores est variable, la diversité des sources sonores qui les composent est plus maîtrisable. Des environnements sonores de parcs et de rues peuvent comprendre des voix humaines, des bruits de pas, des sons de voitures *etc..* Seules les caractéristiques physiques, ainsi que les patterns d'occurrences de ces sources, vont varier. Évaluer des scènes simulées, à partir d'une banque de sons isolés (sources sonores), peut constituer une solution au problème de la diversité des stimuli. Considérons l'étude de l'agrément sonore dans l'environnement urbain. Dans un premier temps, les stimuli sont obtenus via une épreuve de simulation. Dans cette simulation, seule la qualité affective des stimuli est fixée (agréable/désagréable). Les sujets construisent alors les scènes directement en fonction de l'image qu'ils se font d'un environnement urbain agréable/désagréable, adaptant ainsi la structure de la scène à la qualité de l'environnement. Dans un deuxième temps, les scènes ainsi élaborées peuvent constituer des stimuli pour une analyse sémantique différentielle de l'agrément. Cette approche est celle utilisée dans nos travaux (cf. Chapitre 4).

Enfin, la plupart des environnements que nous percevons sont relativement neutres, et ne provoquent pas en nous de réactions

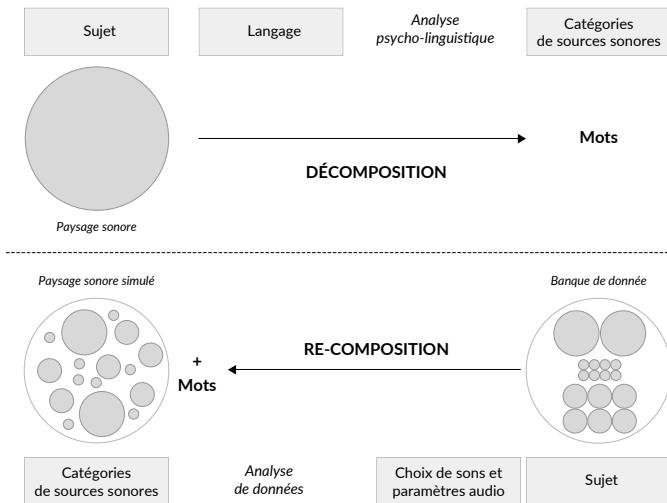


FIGURE 22 : Relation entre l'analyse psycholinguistique et la simulation.

particulières. Il peut n'être pas évident d'évaluer des dimensions perceptives comme l'agrément, la gêne ou le confort, de ces environnements. Des scènes simulées, sur la base d'une qualité affective imposée (*e.g.* agréable), proposent, quant à elles, des versions stéréotypées des environnements ainsi qualifiés. On peut voir dans ces scènes un « résumé cognitif », riche et condensé, des environnements étudiés. Isoler les éléments d'intérêt de ces scènes peut s'avérer plus facile.

### 3.4.2.3 Utilisations antérieures de la simulation

Plusieurs outils de simulation de scènes sonores ont déjà été proposés (Finney and Janer, 2010; Misra et al., 2006, 2007; Schirosa et al., 2010; Valle et al., 2009). Ils ont souvent pour but de générer automatiquement l'ambiance sonore d'un environnement virtuel (Finney and Janer, 2010; Valle et al., 2009). Ils peuvent être vus comme des systèmes semi-autonomes : la simulation étant contrôlée par un utilisateur, mais dépendant aussi, soit d'un environnement visuel à illustrer, soit d'un environnement sonore à reproduire. D'autres outils, entièrement contrôlés par un utilisateur, servent, eux, d'aide à la composition (Misra et al., 2006, 2007). Ces systèmes s'éloignent tous sensiblement du cadre expérimental de l'analyse sensorielle.

Bruce *et al.* (Bruce and Davies, 2014; Bruce et al., 2009) se sont servis de la simulation afin d'étudier la perception des paysages sonores. Ils proposent un système de simulation permettant au sujet d'agir sur un environnement en ajoutant ou supprimant des sources sonores spécifiques. Ce système permet par ailleurs de modifier le niveau sonore des sources, et leurs positions spatiales.

A l'aide de cet outil, les auteurs demandent à leurs sujets de manipuler des sources, afin de recréer un environnement urbain. Ce que faisant, ils montrent que l'inclusion ou l'exclusion des sources dépendent plus de considérations sociales/sémantiques, que des caractéristiques physiques des sources. Ils soulignent néanmoins que le manque d'enregistrements disponibles limite l'analyse. Ils suggèrent de regrouper les enregistrements similaires en « groupes sémantiques » afin de faciliter l'analyse, ce qui est fait dans le cas de notre modèle.

En utilisant le simulateur proposé par (Bruce et al., 2009), (Davies et al., 2014) montrent que, lorsqu'on demande à des participants de simuler un environnement sonore, les simulations font référence à ce que ces derniers s'imaginent être un environnement typique, sans tenir compte de leurs propres préférences pour des sources sonores particulières. Les auteurs soulignent là encore que le faible nombre de sources disponibles (16), ainsi que le nombre réduit de paramètres de contrôle (niveaux des sources, positions spatiales), limitent l'expressivité du sujet, et par conséquent, les capacités de l'analyse.

Dans notre démarche, nous faisons le choix d'une restitution simplifiée (monophonique) des scènes, au profit d'un nombre plus important de paramètres et de sources disponibles, ce afin de garantir, à la sortie, des données viables et expressives.

### 3.4.3 *Simulation et détection automatique d'événements sonores*

En apprentissage machine, la valeur des systèmes proposés par la communauté est bien souvent indexée sur les performances obtenues par ces systèmes sur des corpus d'évaluation.

Comme nous l'avons évoqué à la section 1.2.3, un corpus mal construit, *i.e.* présentant une caractéristique cachée, sur/sous représentée, peut entraîner des erreurs d'interprétations importantes. Pour limiter ces biais potentiels, deux approches peuvent être envisagées :

- augmenter le nombre des données, afin que les biais potentiels se retrouvent « dilués » face à la diversité des données proposées ;
- considérer des données très contrôlées.

La première approche est celle suivie habituellement dans l'évaluation des algorithmes d'analyse automatique. La deuxième, consistant à travailler avec des données peu nombreuses, mais maîtrisées, tient plutôt des pratiques expérimentales ayant cours en analyse sensorielle, ou, par définition, il n'est pas envisageable de soumettre le sujet à un trop grand nombre de stimuli.

C'est cette deuxième approche que nous suivons afin d'évaluer les algorithmes de détection automatique d'événements sonores. Nous nous appuyons sur le modèle proposé pour générer des scènes dont nous maîtrisons les caractéristiques structurelles, à savoir :

- le nombre de classes ;
- le nombre d'événements par classe ;
- le rapport entre le niveau sonore de l'événement et celui du *background*.

Le nombre et la nature des classes sont déterminés par la typologie de la banque de données. Dans le cadre de l'analyse automatique, la typologie n'a cependant pas besoin d'être organisée en une structure taxonomique, le nombre de classes d'événements étant souvent assez faible.

Le nombre d'événements par classe est une résultante directe de l'espacement inter-onset. Comme nous le verrons par la suite, nous considérons également une variante du modèle où le nombre d'événements est directement contrôlé (cf. Section 5.4.2.2).

Enfin, une propriété essentielle de tout algorithme de détection est d'être robuste à différents niveaux de bruit de fond (*background*). Le bruit ici est représenté par l'ensemble des sons n'appartenant pas aux classes à détecter. Dans le modèle présenté, les classes d'événements sont modélisées par des pistes d'événements, tandis que le bruit de fond est modélisé par une unique piste de *texture*. Dans le cas de l'analyse sensorielle, les facteurs  $\mu_i^a$  et  $\sigma_i^a$  déterminant l'amplitude des samples d'événements de la classe  $i$  ne déterminent pas le niveau sonore absolu d'un sample, mais le rapport entre le niveau de ce sample, et celui du *background*.

Nous invitons le lecteur à se référer aux sections 5.3.2.2, 5.3.2.3 et 5.4.2.2 pour une description détaillée des différents processus de simulation utilisés pour générer les corpus de scènes dans le cadre de l'évaluation des algorithmes de détection automatique d'événements.

### 3.5 CONCLUSION

Dans ce chapitre, nous introduisons un modèle fondé sur des considérations perceptives. Le modèle est pensé afin de faciliter la génération de données simulées. Concernant l'analyse sensorielle, nous montrons en quoi la simulation permet de capturer la représentation mentale que se fait un individu d'un environnement en particulier, et motivons son utilisation dans le cadre des études sur les paysages sonores urbains. Concernant l'analyse automatique, nous montrons l'importance de considérer des scènes dont la structure des séquences d'événements est contrôlée par l'expérimentateur, afin de finement apprécier les performances des algorithmes de détection d'événements sonores.

Les deux chapitres suivant présentent les deux cas d'étude.

### Troisième partie

## UTILISATION PRATIQUE DE LA SIMULATION



# 4

## DONNÉES SIMULÉES EN ANALYSE SENSORIELLE

---

### 4.1 INTRODUCTION

Comme nous l'avons vu (cf. Section 2.7.6), la recherche sur les paysages sonores a besoin d'outils permettant d'analyser séparément les influences des différentes sources sur les qualités affectives de l'environnement. La simulation offre des possibilités intéressantes (cf. Section 3.4), car elle nous permet d'obtenir des scènes sonores dont nous connaissons tous les paramètres structuraux, en particulier les caractéristiques distinctes des différentes sources.

Pour démontrer l'intérêt de la simulation en analyse sensorielle, nous choisissons, comme cadre applicatif, le problème de l'agrément perçu dans les environnements sonores urbains.

Ce chapitre présente les résultats d'une série d'expériences s'appuyant sur la simulation, et visant, chacune, à comprendre comment les différentes sources sonores qui composent une scène influent sur la perception de l'agrément :

1. *expérience de simulation* : au cours de cette expérience, les sujets simulent les environnements qui serviront de stimuli pour les étapes suivantes ;
2. *évaluation de l'agrément* : les sujets doivent évaluer l'agrément des scènes simulées à partir d'une échelle sémantique ;
3. *évaluation de l'agrément après modification des scènes* : comme pour l'expérience précédente, les sujets doivent évaluer l'agrément des scènes simulées à partir d'une échelle sémantique. Cependant les scènes ont été modifiées, *i. e.* privées de certaines classes de sons identifiées comme ayant un impact sur l'agrément perçu ;
4. *catégorisation libre* : les sujets doivent catégoriser les scènes sonores simulées. Au delà du problème initial de l'agrément, cette dernière expérience nous amène à considérer logiquement l'influence de la composition sémantique des scènes sur les jugements de similarités.

Le chapitre, comprend cinq sections. La première constitue notre introduction. La deuxième introduit le protocole expérimental ainsi que l'outil de simulation *SimScene*. La troisième reprend l'expérience de simulation à proprement parler, ainsi que l'épreuve *évaluation de l'agrément* qui nous semblent aller de pair. La quatrième décrit l'épreuve *évaluation de l'agrément après modification des scènes*. La cinquième et dernière détaille, elle, l'épreuve *catégorisation libre*.

### 4.1.1 *Protocole expérimental basé sur la simulation*

#### 4.1.1.1 *Organisation des sons isolés*

L'objectif de l'expérimentation est de permettre à un sujet de simuler un environnement sonore cible, à partir de sons isolés. La banque de sons suit l'organisation décrite à la section 3.3.1. Les éléments sont regroupés en classes hiérarchisées, afin de former une taxonomie. Plus le niveau d'abstraction de la classe est élevé, plus la variabilité des enregistrements appartenant à la classe est importante (cf. Figure 20).

Nous conservons la distinction observée entre les événements et les textures, en créant deux taxonomies (*i.e.* deux banques de sons) séparées.

#### 4.1.1.2 *Sélection des sons isolés*

L'objectif de la simulation est d'obtenir une image sonore de la représentation mentale que se fait un sujet d'un environnement donné. Afin que cette image soit la plus « juste » possible, il faut que le protocole limite les biais pouvant influer sur les choix du sujet.

Un de ces biais intervient dans le processus de sélection. La grande majorité des outils permettant de parcourir une banque de données propose une recherche textuelle sur la base de mots clefs. L'efficacité de ce principe repose avant tout sur la structure typologique, et la nomenclature de la base de données. Dans le cadre d'une expérience sensorielle visant à objectiver la représentation interne du sujet, cette approche pose plusieurs problèmes :

- les sons peuvent ne pas être annotés d'une manière satisfaisante. En effet, sémantiquement, un son peut être décrit de plusieurs façons. Nous pouvons en désigner la source (une portière de voiture), comme nous pouvons désigner l'action de la source (le claquement d'une portière de voiture) ou encore son environnement (le claquement d'une portière de voiture dans un garage). Concevoir un système de recherche par mots clefs efficace suppose une description précise de chaque son, qui plus est, adaptable à la représentation que s'en fait chaque sujet, ce qui est difficilement réalisable ;
- lors d'une recherche par mots clefs, le sujet doit objectiver un nom décrivant l'objet recherché. Or cette objectivation dépend des connaissances collectives du sujet, connaissances liées à sa sphère socioculturelle, et en particulier à sa langue. L'expérience visant une diffusion internationale, cette contrainte est difficilement surmontable ;
- la description verbale du son, si elle est accessible au sujet, peut potentiellement influencer sa sélection. Dans les faits, pour

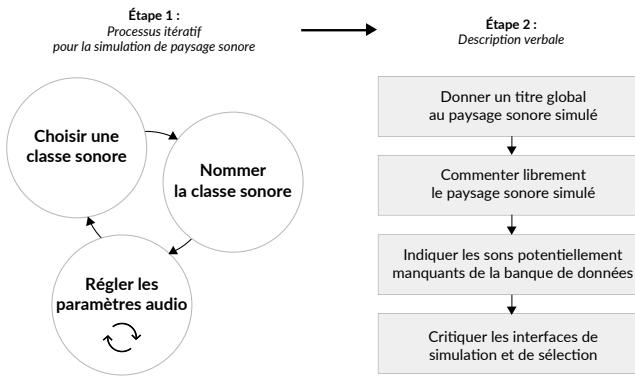


FIGURE 23 : Etape de processus de simulation pour l'analyse sensorielle

construire une scène environnementale « calme », le sujet sélectionne a priori les sons référencés sous le vocable *parc*. Cette réalité constitue encore une difficulté.

Imposer au sujet une terminologie à travers les labels décrivant les classes est un risque. Nous pensons que la sélection doit s'éloigner le plus possible d'un ancrage sémantique, et s'effectuer à l'aveugle, *i.e.* sur la base uniquement de l'écoute. Une interface développée spécialement dans ce but est présentée à la section 4.1.1.6.

Enfin, il est important de noter que le sujet ne peut accéder qu'aux classes du niveau d'abstraction le plus bas, classes qui ne possèdent pas de sous-classes, et sont directement liées à une collection de samples.

L'organisation hiérarchique sert alors deux buts :

- faciliter le parcours, par les sujets, des banques de sons isolés (cf. Section 4.1.1.6) ;
- faciliter le travail d'analyse de l'expérimentateur, en lui permettant d'observer la composition en terme de sources sonores des scènes, suivant différents niveaux d'abstraction.

#### 4.1.1.3 Processus de simulation

Trois étapes composent le processus de simulation (cf. Figure 23) :

- *sélection* d'une classe de sons. Une fois une classe sélectionnée, une piste est générée ;
- *identification* de la classe de sons sélectionnée. Le sujet nomme la classe de sons qu'il a sélectionnée ;
- *paramétrisation* de la piste liée à la classe de sons. Le sujet fixe les paramètres de la piste (pour plus de détails sur les paramètres proposés cf. Section 4.1.1.4).

Ces étapes peuvent être répétées, et dans n'importe quel ordre, le sujet pouvant agir rétroactivement sur les pistes déjà créées. A la fin de la simulation, et afin d'accumuler le maximum de connaissances sur la scène simulée, le sujet peut :

- nommer l'environnement simulé ;
- fournir un commentaire libre décrivant son processus de création, ainsi que le paysage sonore qu'il a voulu illustrer.

#### 4.1.1.4 *Paramètres de contrôle*

Les paramètres du modèle permettent au sujet de contrôler la structure de chaque piste. Ils agissent sur tous les samples à la fois, et non sur un en particulier.

Parmi ces paramètres, on retrouve ceux introduits pour le modèle initial de scène sonore (cf. Section 3.3.3 et 3.3.4), à savoir :

- *niveau sonore* (dB) : pour chaque sample, les niveaux sont tirés aléatoirement à partir d'une distribution normale, paramétrée par le sujet en terme de moyenne et de variance ;
- *espacement inter-onset* (seconde) : (piste d'événements seulement) comme pour les niveaux, les espacements sont tirés aléatoirement à partir d'une distribution normale, paramétrée par le sujet en terme de moyenne et de variance ;
- *début et fin* (seconde) : le sujet fixe le début et la fin de chaque piste.

Afin de faciliter la simulation, deux paramètres supplémentaires sont proposés :

- *fondu par événement* (seconde) : (piste d'événements seulement) le sujet fixe une durée de fondu (entrée et sortie), appliquée à chaque sample d'une piste d'événements ;
- *fondu global* (seconde) : le sujet fixe les durées de fondus pour l'entrée et la sortie de la piste. Ces fondus s'appliquent ainsi à l'ensemble des samples de la piste.

Deux de ces paramètres ne s'appliquent que pour les pistes d'événements (*fondu par événement* et *espacement inter-onset*), les samples des textures étant séquencés sans espacement (cf. Section 3.3.3)

#### 4.1.1.5 *Données produites par le processus de simulation*

Ce protocole de simulation peut potentiellement produire un grand nombre de données. Ces dernières sont décrites à la figure 24. Nous les résumons dans la liste suivante :

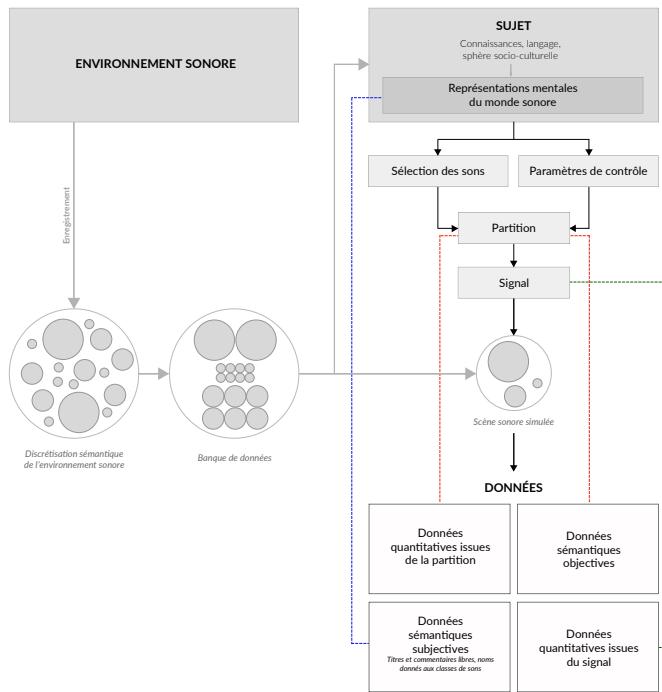


FIGURE 24 : Paradigme du protocole expérimental basé sur la simulation.

- données sémantiques objectives : la banque de données nous permet d'obtenir une information objective quant aux sources sonores présentes dans la scène. Les données sémantiques objectives sont les labels des classes sélectionnées ;
- données sémantiques subjectives : il s'agit des noms donnés par le sujet 1) à la scène simulée, 2) aux classes de sons sélectionnées ;
- données quantitatives issues de la partition : il s'agit de toutes les données relatives à la partition, *i.e.* pour chaque piste, le positionnement des samples et les paramètres (cf. Section 3.3.3) ;
- données quantitatives issues du signal : il s'agit d'indicateurs acoustiques extraits du signal, *e.g.* le niveau sonore global. Comme nous possédons les samples isolés utilisés pour la synthèse, il est possible de calculer ces descripteurs pour une classe, ou un ensemble de classes, en particulier.

Le protocole nous permet de caractériser avec précision une scène simulée, sur la base de données sémantiques, subjectives ou objectives, ainsi que quantitatives. Considérant l'ensemble des données générées, les potentiels d'analyse sont vastes.

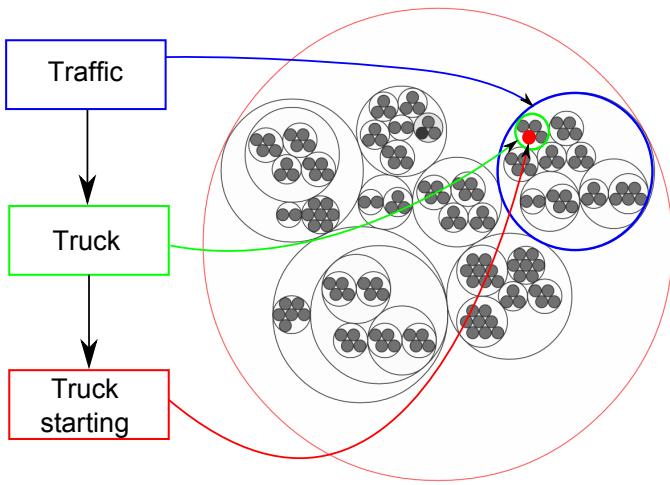


FIGURE 25 : L’interface de sélection aveugle de l’outil de simulation *Simscene*.

#### 4.1.1.6 Interface de sélection aveugle des sons isolés

Pour limiter l’influence de l’interface sur le sujet, il nous paraît nécessaire de libérer sa recherche de toute information textuelle. Nous proposons à l’utilisateur une interface graphique lui permettant d’explorer la banque de sons exclusivement à partir de l’écoute.

Visuellement, les classes du dernier niveau (les seules accessibles par le sujet) sont représentées par des cercles, et positionnées sur un plan. La disposition des cercles dans l’espace dépend de l’organisation hiérarchique de la base de données : les sous-classes appartenant à une même classe sont proches les unes des autres, et ainsi de suite jusqu’à atteindre les classes des niveaux d’abstraction élevés.

La figure 25 présente l’interface pour la banque de données d’événements sonores. Cette organisation visuelle a été pensée afin de :

1. faciliter le parcours de la banque de données, les sons similaires (au sens des classes) étant proches les uns des autres. L’organisation hiérarchique se fonde, en effet, sur des principes cognitifs. Les classes ont été établies à partir de la littérature traitant des catégories de sources sonores (cf. Section 3.3.1).
2. permettre aux sujets de rapidement appréhender toute l’étendue de la banque de données, *i.e.* l’ensemble des sons disponibles.

Chaque classe possède un son prototype. Ces sons ont été choisis par les expérimentateurs. Lorsqu’on clique sur un cercle, le prototype associé à la classe est joué. Le sujet parcourt la banque de sons en cliquant sur les cercles. Cette interface a fait l’objet d’une étude approfondie dont les résultats sont publiés dans (Lafay et al., 2016b).

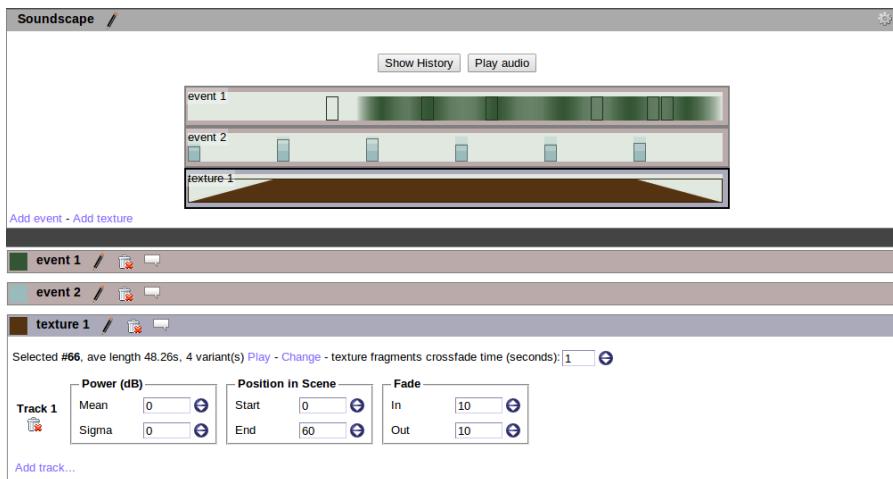


FIGURE 26 : L’outil de simulation *Simsscene*.

#### 4.1.1.7 Interface de simulation : l’outil Simscene

*Simsscene* est un environnement de travail audio-numérique dont la première version a été développée dans le cadre du projet HOULE<sup>1</sup>. Il permet de simuler des paysages sonores à partir d’un corpus de sons. Il est prévu pour fonctionner sur les navigateurs internet *Chrome* et *Firefox*. L’outil a été développé en javascript à l’aide de la bibliothèque *angular.js*<sup>2</sup> et du standard *web-audio*<sup>3</sup>. L’interface de sélection (cf. Section 4.1.1.6) a été développée à l’aide de la bibliothèque *D3.js* (Bostock et al., 2011).

*Simsscene* est présenté en détail dans (Rossignol et al., 2015b). Nous résumons ici ses fonctionnalités d’importance pour notre étude.

Le fonctionnement de *Simsscene* se rapproche de celui d’un séquenceur audio. Chaque utilisateur choisit une classe de sons via l’interface de sélection (cf. Section 4.1.1.6). Une fois la classe de sons sélectionnée, une piste audio, liée à cette classe, est créée. L’utilisateur peut alors modifier certaines propriétés de la piste via un groupe de paramètres de contrôle propre à chacune (cf. Section 4.1.1.4). Des champs de texte sont prévus afin de permettre à l’utilisateur 1) de nommer chaque piste, 2) de donner un titre à la scène simulée 3) de commenter la scène simulée.

L’interface propose un rendu graphique schématisé de la scène en cours de création (cf. Figure 26). La piste est représentée par une bande possédant un axe temporel. Sur cette bande, chaque sample est représenté par un rectangle. L’espacement entre les rectangles est relatif à l’espacement entre les samples. De même, la hauteur des rectangles est proportionnelle au niveau sonore des samples. Dans le cas d’une piste de texture, un unique rectangle apparaît sur toute la

<sup>1</sup> Projet HOULE : <http://houle.ircam.fr/>

<sup>2</sup> angular.js : <https://angularjs.org/>

<sup>3</sup> web-audio : <http://www.w3.org/TR/webaudio/>

longueur de la piste, un son de texture ne pouvant être entrecoupé de silences. Les caractéristiques des rectangles évoluent en fonction des changements de paramètres de la piste.

L'utilisateur a la possibilité, à tout moment, d'écouter la scène simulée.

## 4.2 L'IMPACT DE LA COMPOSITION SÉMANTIQUE DES SCÈNES SUR LA PERCEPTION DE L'AGRÉMENT

### 4.2.1 *Objectif*

L'objectif est d'étudier les influences spécifiques des différentes sources sonores qui composent les environnements urbains, sur la perception de l'agrément, en utilisant la simulation. Pour ce faire, nous planifions nos deux premières expériences (cf. Figure 27) :

- *expérience de simulation* : au cours de cette expérience, les sujets doivent simuler des environnements sonores urbains, en utilisant l'outil et le protocole de simulation décrits à la section 4.1.1.7. Chacun compose deux environnements sonores, le premier idéal/agréable et le deuxième non-idéal/désagréable. Cette épreuve de simulation a fait l'objet d'une expérience pilote (Lafay, 2013; Lafay et al., 2014);
- *expérience d'évaluation* : à l'issue de la simulation, nous n'avons, de fait, qu'une connaissance binaire des propriétés affectives des scènes simulées : idéale (i) et non-idéale (ni). Cette seconde étape a pour but d'affiner notre connaissance sur l'agrément. Pour ce faire, nous demandons à un deuxième groupe de sujets d'évaluer, à partir d'une échelle sémantique, l'agrément de chacune des scènes simulées. L'expérience d'évaluation sert deux buts :
  1. évaluer l'influence respective des différentes sources sur l'agrément, pour chaque type d'environnement (i ou ni);
  2. détecter la présence de cas extrêmes ou ambigus (*outlier*) dans les scènes simulées. Pour le reste de notre étude, les qualités hédoniques imposées (i et ni) servent de référence, de vérité terrain. Il nous faut donc garantir qu'il n'y ait pas d'ambiguïté entre les cas extrêmes des i- et ni-scènes, *i.e.* que la note d'agrément la plus basse des i-scènes reste supérieure à la note la plus haute des ni-scènes.

Notre analyse s'appuie sur les données produites par les deux expériences.

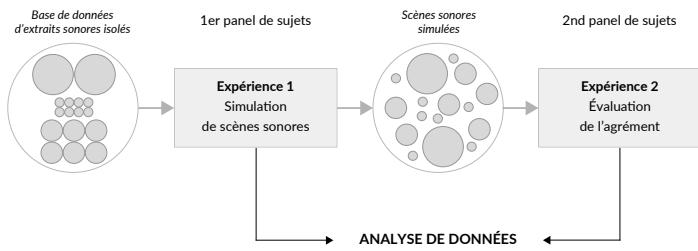


FIGURE 27 : Planification expérimentale des expériences de simulation et d'évaluation de l'agrément

#### 4.2.2 Banque de données de sons isolés

Dans cette partie, nous présentons les processus de sélection et d'acquisition des sons utilisés comme matériau de base lors de la simulation des environnements sonores urbains. La banque de données est identique à celle utilisée dans le cadre de l'expérience pilote (Lafay, 2013; Lafay et al., 2014).

Pour plus de détails sur l'organisation interne de la banque de données, ainsi que sur l'interface graphique permettant de sélectionner ces dernières, se référer aux sections 4.1.1.1 et 4.1.1.6.

#### 4.2.3 Typologie des sources sonores présentes dans l'environnement urbain

Afin de créer un corpus de sons isolés de référence pour la simulation, nous réalisons une typologie des sons environnementaux urbains.

Pour ce faire, une étude bibliographique est effectuée, afin d'identifier les sources et ambiances sonores les plus souvent citées dans les publications. Cette étude porte sur 16 articles ou thèses. Chacun d'eux traite de la manière dont nous discriminons les paysages sonores urbains. Il ressort que plusieurs approches sont possibles :

- 9 articles abordent le problème par une approche perceptive, soit en identifiant ou répertoriant des catégories de sources sonores, soit en étudiant l'impact de classes de sons spécifiques sur la perception de l'environnement : Defréville et al., 2004; Devergie, 2006; Dubois et al., 2006; Guastavino, 2003, 2006; Maffioli, 1999; Niessen et al., 2010; Rimbault, 2002; Rimbault and Dubois, 2005
- 3 articles proposent une classification morpho-typologique, divisant l'environnement sonore urbain en « zones sonores » possédant une identité acoustique forte, selon la configuration et la pratique du site : Beaumont et al., 2004; Maffioli, 1999; Polack et al., 2008

- 2 articles répertorient et classifient les sources sonores d'un point de vue expert : Brown et al., 2011; Leobon, 1986

La nature des classes est établie par rapport aux catégories perceptives, ou classes de sons, émergeant de cette littérature. À partir des éléments relevés, nous établissons deux taxonomies, une pour les événements (cf. Figures 28a et 28b), une autre pour les textures (cf. Figures 28c et 28d). Comme évoqué à la section 3.2.3, la structure taxonomique de ces deux ensembles s'inspire grandement de l'axe vertical de l'organisation catégorielle proposée par E. Rosch (cf. Section 2.3.3.3), *i.e.* plus le niveau d'abstraction de la classe est élevé, plus la description de la classe est précise, et plus les sources sonores incluses dans cette classe sont semblables (cf. Figure 20). Pour les événements, nous considérons quatre niveaux d'abstraction allant des classes les plus globalisantes (niveau d'abstraction 0), aux classes les plus spécifiques (niveau d'abstraction 3). Pour les textures, nous ne considérons que trois niveaux d'abstraction.

Pour les événements, les regroupements se font essentiellement par rapport à la source, et sont d'ordre sémantique. Pour les textures, nous considérons également la nature des lieux hébergeant ces sources (*e.g. parc, rue*). La typologie des classes d'événements suit la nomenclature source-action introduite à la section (cf. Section 3.2.3). Elle est assez proche d'une autre typologie des sources sonores urbaines introduite ultérieurement (Salamon et al., 2014).

Les réactions à la musique étant trop subjectives, et les jugements esthétiques ne pouvant qu'altérer les données d'évaluation, nous choisissons, dans cette étude, de ne pas considérer les sources musicales de type musiciens de rue, radios de voitures, d'appartements, *etc..*

#### 4.2.4 Acquisition des sons isolés

Sur la base des typologies précédemment établies, 483 sons ont été collectés, dont 381 événements, et 102 textures.

Parmi les événements :

- 260 sont issus d'enregistrements effectués pour l'étude ;
- 89 sont issus de la banque de sons SoundIdeas<sup>4</sup> ;
- 32 sont issus de la banque de sons Universal SoundBank<sup>5</sup>.

Parmi les textures :

- 72 sont issues d'enregistrements effectués pour l'étude ;
- 23 sont issues de la banque de sons SoundIdeas ;

<sup>4</sup> SoundIdeas : <http://www.sound-ideas.com/>

<sup>5</sup> Universal SoundBank : <http://www.universal-soundbank.com/>

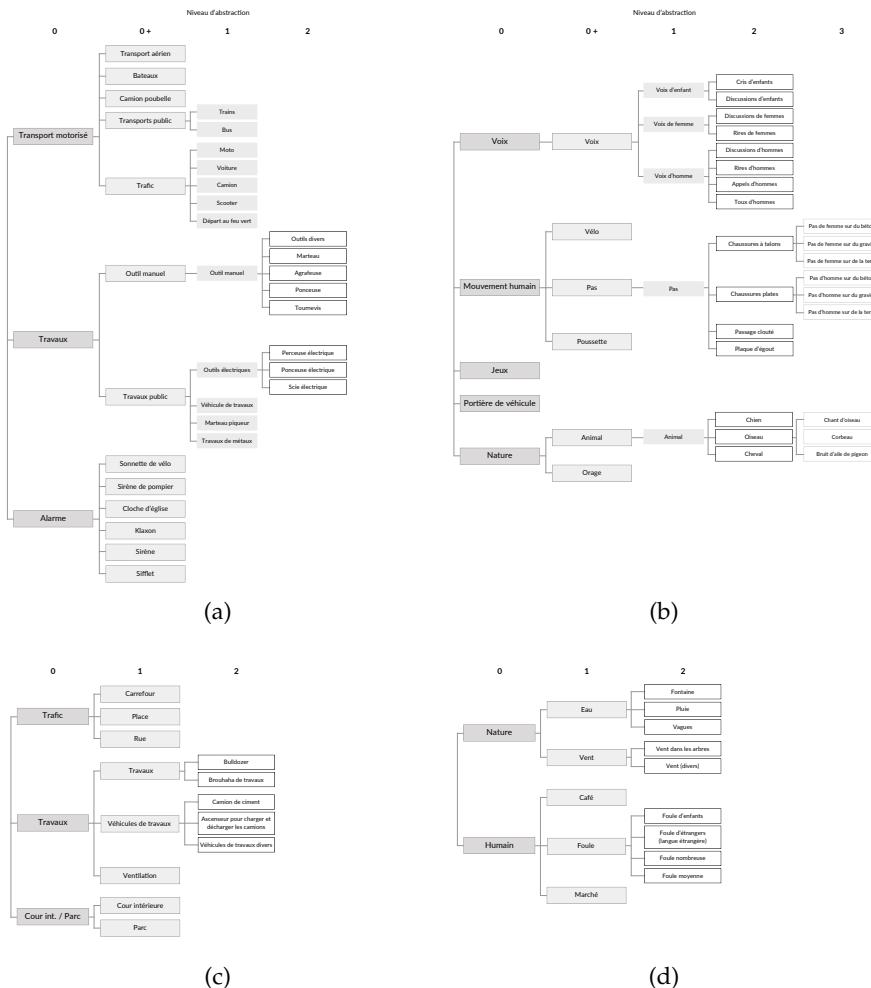


FIGURE 28 : Taxonomies des classes de sons utilisées pour la simulation des environnements sonores urbains pour (a,b) les événements sonores et (c,d) les textures sonores.

- 7 sont issues de la banque de sons *Universal SoundBank*.

Tous les enregistrements ont été effectués à l'aide d'un micro canon *AT8035*<sup>6</sup> relié à un enregistreur *ZOOM H4n*<sup>7</sup>. L'utilisation du micro canon nous permet d'isoler les événements sonores du brouhaha urbain. Pour les textures, il nous permet d'éviter les événements sonores proches du preneur de son. Nous pouvons ainsi pointer des "zones sonores", en nous tenant à une certaine distance de ces dernières, afin de capter uniquement le brouhaha émanant de la zone ciblée.

6 Micro canon *AT8035* : <http://eu.audio-technica.com/fr/products/product.asp?catID=1&subID=6&prodID=1845>

7 Enregistreur *ZOOM H4n* : <http://www.zoom.co.jp/english/products/h4n/>

Tous les sons ont été normalisés au même niveau RMS<sup>8</sup> de -12 dB (FS)<sup>9</sup>.

#### 4.2.5 Planification expérimentale

##### 4.2.5.1 Épreuve de simulation

Nous nommons cette expérience : *expérience 1.a.*

#### Procédure

Les sujets doivent simuler deux environnements sonores urbains, chacune des scènes devant durer 1 minute. Pour ces simulations, les sujets doivent se conformer aux consignes suivantes :

- première simulation : simuler un paysage sonore **urbain plausible** qui, selon vous, est idéal (où vous aimerez vivre) ;
- deuxième simulation : simuler un paysage sonore **urbain plausible** qui, selon vous, est non-idéal (où vous n'aimerez pas vivre).

Tous les sujets commencent par simuler l'environnement idéal. Les sujets ne prennent connaissance de la deuxième consigne qu'à la fin de la première simulation.

Les sujets sont totalement libres dans le choix des sons, et des paramètres (pour plus de détails sur les paramètres, se référer à la section 4.1.1.4). Ils doivent cependant se soumettre à deux contraintes :

- le sujet doit prendre le point de vue d'un auditeur fixe ;
- le paysage sonore doit être réaliste, au sens de physiquement plausible. Autrement dit, le sujet a tout à fait le droit de placer 10 chiens dans son paysage sonore, mais il n'a pas le droit de placer un chien aboyant toutes les 10 millisecondes.

Ces contraintes font partie de la consigne. Aucun contrôle n'est fait *a priori* dans l'interface de simulation.

Chaque processus de simulation comprend deux parties :

1. la réalisation de la simulation : cette étape peut, elle même, se décomposer en trois actions (cf. Section 4.1.1.3) :

<sup>8</sup> Le niveau RMS, de l'anglais *Root Mean Square* qui désigne la valeur efficace d'un signal. Formellement, le niveau RMS  $x_{RMS}$  d'un signal  $x = (x_1, x_2, \dots, x_n)$  s'obtient en calculant la moyenne quadratique de ce dernier  $x_{RMS} = \sqrt{\frac{1}{n} \sum_i x_i^2}$ .

<sup>9</sup> dB (FS) est le sigle anglais désignant une valeur en décibels relative à la pleine échelle (*relative to Full Scale*), *i.e.* le rapport entre le niveau du signal et sa valeur maximale. Dans notre cas, ce niveau pleine échelle est de 1 Volt.

Index	Tâche	Durée (min)
1	Présentation de l'expérience Lecture de la consigne	10
2	Tutoriel (Réalisation d'une scène test)	20
3	Première simulation : scène idéale	40
4	Commentaire de la scène idéale	15
3	Deuxième simulation : scène non-idéale	40
4	Commentaire de la scène non-idéale	15
5	Critique de l'interface de simulation et de l'interface de sélection	10

TABLE 4 : Résumé des étapes de l'expérience de simulation.

- sélectionner les classes de sons
  - nommer les classes de sons sélectionnées
  - paramétrier les pistes (cf. Section 3.3.2) relatives aux classes de sons sélectionnées
2. la production d'un commentaire libre du paysage sonore simulé

En complément, et une fois les deux scènes sonores réalisées, les sujets sont invités à :

- indiquer les sources sonores qu'ils voulaient mettre, mais qu'ils n'ont pas trouvées ;
- commenter l'ergonomie du logiciel de simulation ;
- commenter l'ergonomie de l'interface de sélection.

Avant de commencer la première simulation, un tutoriel de 20 minutes est proposé aux sujets, afin qu'ils se familiarisent avec le logiciel de simulation, et la banque de données. Le tableau 4 résume les étapes de l'expérience, ainsi que leurs durées respectives. L'expérience est prévue pour durer 2h30.

### Dispositif expérimental

Tous les sujets passent l'expérience sur des machines identiques. L'audio est diffusé en monophonie, par le biais de casques audio. Pendant le tutoriel, les sujets doivent ajuster le niveau sonore à un volume confortable. Ils ne peuvent le modifier par la suite.

Tous les sujets réalisent l'expérience simultanément. Ils sont répartis de manière égale dans trois pièces identiques, toutes possédant un environnement calme. Ils n'ont pas le droit de s'adresser la parole pendant l'expérience.

Trois expérimentateurs, un dans chaque pièce, sont présents durant la totalité de l'expérience, afin de contrôler le bon déroulement de cette dernière, et de répondre aux éventuelles questions des sujets.

### **Participants**

44 étudiants (14 femmes) de L'École Centrale de Nantes ont participé à l'expérience. Ils ont tous sensiblement le même âge (moyenne : 21.6, écart-type : 2). Tous les sujets sont Nantais, et vivent dans cette ville depuis deux ans ou plus.

Sur les 44 sujets, 40 réalisent l'expérience avec succès, produisant au final 80 scènes sonores simulées, dont 40 scènes idéales, et 40 scènes non idéales. 4 sont éliminés pour non respect ou incompréhension des consignes, d'une part, et dépassement du temps, d'autre part.

#### *4.2.5.2 Épreuve d'évaluation de l'agrément*

Nous nommons cette expérience : *expérience 1.b.*

### **Procédure**

En raison de contraintes temporelles, les sujets n'évaluent que des séquences de 30 secondes des scènes simulées, chacune de ces séquences commençant à la seconde 15, et finissant à la seconde 45 de la scène d'une durée d'origine de 1 minute.

L'évaluation s'effectue sur une échelle sémantique bipolaire de 7 points, allant de -3 (non-idéale/très désagréable) à +3 (idéale/très agréable). Avant de noter une scène, les sujets doivent obligatoirement écouter les 20 premières secondes de cette dernière. Après la notation, ils sont libres de passer à la scène suivante.

Pour chaque sujet, les scènes sont présentées dans un ordre aléatoire. Les 10 premières scènes permettent au sujet de calibrer ses notes. Elles sont obligatoirement composées de 5 scènes idéales, et de 5 scènes non-idéales. Ces 10 premières scènes sont rejouées à la fin de l'expérience, et seules les notes données à la deuxième occurrence sont prises en compte.

L'expérience est prévue pour durer 30 minutes. Les sujets ne connaissent pas la nature des scènes.

### **Dispositif expérimental**

Tous les sujets passent l'expérience sur des machines identiques. L'audio est diffusé en monophonie, par le biais de casques audio semi-ouverts *Beyer-Dynamic DT 990 Pro*. Toutes les scènes sonores ont été re-simulées sur la base des partitions obtenues lors de l'expérience

de simulation. Le niveau sonore de sortie est identique pour tous les sujets.

Tous les sujets réalisent l'expérience simultanément, dans un environnement calme. Ils n'ont pas le droit de s'adresser la parole pendant l'expérience.

Un expérimentateur est présent durant la totalité de l'expérience, afin de contrôler le bon déroulement de celle-ci, et de répondre aux éventuelles questions des sujets.

## Participants

10 étudiants (2 femmes) de L'École Centrale de Nantes ont participé à l'expérience. Aucun d'entre eux n'a réalisé l'expérience de simulation. Tous les sujets ont sensiblement le même âge (moyenne : 23.1, écart-type : 1.8). Tous les sujets sont Nantais, et vivent dans cette ville depuis deux ans ou plus.

Tous les sujets ont réalisé l'expérience avec succès.

### 4.2.6 *Données et méthodes d'analyses*

#### 4.2.6.1 *Nature des données analysées*

A partir des données produites par l'épreuve de simulation, nous analysons :

- les partitions des scènes simulées ;
- les signaux des scènes simulées ;
- les commentaires sur les sons manquants, et l'ergonomie des interfaces de simulation et de sélection.

Chaque scène est décrite par un groupe de descripteurs. C'est sur ces descripteurs que nous pratiquons l'analyse. Un résumé des descripteurs, ainsi que des acronymes les désignant, est présenté dans le Tableau 5. Afin de rester cohérent avec l'épreuve d'évaluation, les descripteurs issus des partitions, ou des signaux des scènes, ne sont pas calculés sur la durée totale de celles-ci, mais sur une version réduite de 30 secondes (cf. Section 4.2.5.2).

Pour chaque scène sonore, trois types de descripteurs sont considérés :

- *perceptif* : il s'agit de l'agrément perçu de la scène simulée, évalué sur une échelle sémantique de 7 points. Nous notons  $\mathcal{A}_{scène}$  l'agrément moyen d'une scène, obtenu en moyennant les notes de tous les sujets. De même, nous notons  $\mathcal{A}_{sujet}$  l'agrément par sujet, en moyennant l'ensemble de ses notes. Compte tenu du faible nombre de sujets, nous faisons le choix, dans cette étude, de ne pas normaliser les notes d'agrément ;

- *sémantique* : il s'agit d'un vecteur booléen noté  $S = (x_1, x_2, \dots, x_n)$ , indiquant les classes de sons présentes dans la scène. Chaque point  $x$  de ce vecteur correspond à une classe de sons particulière :  $x = 1$  si la classe est présente dans la scène, et  $x = 0$  autrement. La dimension  $n$  des vecteurs dépend du niveau d'abstraction considéré, *e.g.* pour le niveau d'abstraction 1, qui comprend 44 classes de sons, cette dimension sera de  $n = 44$ .
- *structurel* : les descripteurs structurels sont calculés à partir des partitions et des signaux des scènes simulées. Trois descripteurs structurels sont envisagés :
  - *diversité* (DIV) : il s'agit d'un scalaire représentant la diversité des classes sonores utilisées pour simuler une scène. Nous calculons DIV en comptant le nombre de classes de sons distinctes utilisées pour une simulation. Ce nombre dépend du niveau d'abstraction considéré. Par exemple, considérant les deux sous classes du niveau d'abstraction 2 *passage de voiture* et *démarrage de voiture*, toutes deux appartenant à la classe *voiture* du niveau d'abstraction 1, nous comptons deux classes pour la diversité des niveaux d'abstraction 2 et 1, et seulement une pour les niveaux d'abstraction 0 et 1 ;
  - *densité* (D) : il s'agit d'un scalaire représentant le nombre de sources sonores présentes en moyenne. Pour obtenir D, nous calculons le logarithme du nombre d'éléments sonores par fenêtre de 125 millisecondes (sans recouvrement), et moyennons au cours du temps. Le calcul de D peut inclure toutes les sources sonores de la scène, ou seulement une partie. Dans ce cas, les fenêtres ne contenant pas de sources sonores ne sont pas prises en compte. Nous notons D(E) et D(T) les densités calculées en considérant séparément les sources d'événements et de textures sonores ;
  - *niveau Sonore* (L) : pour représenter le niveau sonore, nous nous inspirons de la mesure  $L_{Aeq}$ . Dans notre cas, il s'agit d'un scalaire, calculé sur le signal en volts, et non en pression, et donné en décibels, en prenant un référentiel de 1 Volt. Le niveau est obtenu en calculant, toutes les secondes, la moyenne quadratique du signal, et en moyennant sur la durée de la scène. Un filtrage de type A est opéré avant le calcul des moyennes quadratiques. D'autres descripteurs, inspirés eux aussi de descripteurs acoustiques classiques ( $L_{Amin}$ ,  $L_{Amax}$ ,  $L_{A10-90}$ ), et utilisant un opérateur autre que la moyenne (minimum, maximum, les 10-90ème quantiles) pour intégrer les fenêtres de 1 seconde, ont été testés. Mais, ces derniers présentant tous une corrélation élevée avec L ( $r_{pearson} \geq 0.76$ ,  $p < 0.01$ ), nous conservons le sca-

Descripteurs	Acronymes	Descripteurs	Acronymes
Densité	D	Diversité	DIV
Densité (événements)	D(E)	Diversité (événements)	DIV(E)
Niveau	L	Diversité (textures)	DIV(T)
Niveau (événements)	L(E)	Agrément moyen (par scène)	$\mathcal{A}_{scène}$
Niveau (textures)	L(T)	Agrément moyen (par sujet)	$\mathcal{A}_{sujet}$

Termes	Acronymes
idéale/agréable	i
non-idéale/désagréable	ni
scène idéale/agréable	i-scène
scène non-idéale/désagréable	ni-scène

TABLE 5 : Acronyme des variables utilisées dans le cadre des expériences sensorielles.

laire ci-devant mentionné comme unique descripteur objectif du niveau sonore.

#### 4.2.6.2 Méthodologie et Outils statistiques

Afin d'évaluer l'impact spécifique des différentes sources sonores sur l'agrément perçu, nous soumettons nos travaux aux 6 tests/études de significativité présentés ci-après :

- *étude qualitative* : afin de vérifier la validité écologique de 1) la banque de données et 2) l'interface de sélection, nous réalisons une étude qualitative des critiques ergonomiques effectuées par les sujets ;
- *Vérification de l'agrément des scènes simulées* : afin de vérifier que la distinction affective imposée entre les i- et ni-scènes se retrouve au niveau de l'agrément perçu, nous observons s'il existe des différences entre les deux types de scènes au niveau de  $\mathcal{A}_{scène}$  et  $\mathcal{A}_{sujet}$ . La significativité est évaluée par un test de Student à deux populations indépendantes pour  $\mathcal{A}_{scène}$ , et par un test de Student à deux populations appariées pour  $\mathcal{A}_{sujet}$  (cf. Annexe A.1) ;

- *étude comparative entre les descripteurs structurels* : afin d'évaluer si la distinction affective imposée entre les i- et ni-scènes impacte de manière significative la nature des scènes, *i. e.* s'il existe des différences significatives entre les descripteurs structurels et/ou l'agrément perçu, nous évaluons cette significativité à partir d'un test de Student à deux populations (cf. Annexe A.1) ;
- *étude de l'influence des descripteurs structurels sur l'agrément perçu* : afin d'évaluer l'impact potentiel des descripteurs structurels sur l'agrément perçu, nous étudions l'existence de corrélations linéaires entre ces deux types de descripteurs. Pour mesurer la corrélation, nous utilisons le coefficient de Pearson (cf. Annexe A.4). Nous adoptons ici une méthodologie couramment utilisée dans l'approche dimensionnelle ;
- *étude comparative entre les descripteurs sémantiques* : afin d'apprécier si la distinction affective imposée a eu un impact sur la composition des scènes en terme de sources sonores, ou, pour être plus précis, s'il existe des classes de sons qui ont été particulièrement utilisées pour simuler un type d'environnement, nous utilisons le V-test. Nous vérifions si la présence d'une classe de sons est typique d'un environnement (i ou ni). Le test est effectué pour chaque niveau d'abstraction, et séparément, pour les classes d'événements et de textures. Pour chaque classe  $j$  et chaque type d'environnement  $k$  ( $k = \{i, ni\}$ ), la valeur  $V_{jk}$  du V-test se calcule comme suit :

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

où  $c$  est le nombre de classes utilisées,  $c_k$  le nombre de classes utilisées pour un type d'environnement  $k$ ,  $c_j$  le nombre de classes  $j$  utilisées, et  $c_{jk}$  le nombre de classes  $j$  utilisées pour un type d'environnement  $k$ . Le V-test teste l'hypothèse nulle que la proportion  $\frac{c_{jk}}{c}$  ne diffère pas significativement de la proportion  $\frac{c_j}{c_k}$ . Si pour un environnement  $k$ , et une classe  $j$ , l'hypothèse est rejetée, la classe  $j$  est alors typique de l'environnement  $k$ . Les classes typiques sont nommées **marqueurs sonores** ;

- *étude des espaces de représentation induits par les descripteurs sémantiques* : afin d'étudier si une représentation basée uniquement sur la présence ou l'absence de classes de sons permet de séparer les deux types d'environnement, nous considérons l'espace induit par les descripteurs sémantiques  $S$ .  $S$  étant un vecteur booléen, nous calculons les distances entre les scènes à partir de la distance de Hamming. Considérant les deux vecteurs  $S_1 = (x_{1,1}, x_{1,2}, \dots, x_{1,n})$ , et  $S_2 = (x_{2,1}, x_{2,2}, \dots, x_{2,n})$  de

dimension  $n$ , avec  $x = \{0, 1\}$ , la distance de Hamming  $d_{ham}$  mesure le pourcentage de coordonnées qui diffèrent entre les deux vecteurs :

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

où  $\oplus$  désigne l'opérateur du *ou-exclusif*. Plus la composition des deux scènes est similaire, et plus ces deux scènes sont proches. L'utilisation de la distance de Hamming permet de prendre en compte de manière égale les classes présentes et absentes. Pour mesurer la capacité intrinsèque de l'espace à séparer les  $i$ - et  $n$ -scènes, nous utilisons une métrique de *clustering* nommée précision au rang  $k$  ( $P@k$ ). La  $P@k$  mesure la précision obtenue après que  $k$  items ont été retrouvés. Formellement, pour chaque scène  $s_i$ , nous calculons le rapport entre le nombre de scènes  $s_j$  prises parmi les  $k$  plus proches voisines de  $s_i$ , et partageant le même label que  $s_i$ , sur le nombre d'items à retrouver ( $k$ ). La  $P@k$  est alors la moyenne des rapports pour tous les items ;

- *étude de l'influence spécifique des marqueurs sonores sur l'agrément perçu* : afin d'évaluer les contributions spécifiques de certaines sources sonores, nous évaluons une nouvelle fois l'impact potentiel des descripteurs structurels sur l'agrément perçu, mais en ne tenant compte, cette fois, que des marqueurs sonores pour calculer ces descripteurs.

Excepté le V-test, tous les tests de significativité sont effectués avec un seuil critique  $\alpha = 0.05$ . Pour le V-test, étant donné que nous testons beaucoup de classes, une correction de Bonferroni (cf. Annexe A.3) est appliquée. Pour les valeurs  $p$ , dans le cas où  $p \geq 0.05$ , nous indiquons sa valeur. Dans le cas où  $0.01 \leq p < 0.05$ , nous indiquons seulement  $p < 0.05$ . Dans le dernier cas nous indiquons  $p < 0.01$ .

Concernant l'interprétation du coefficient de corrélation de Pearson adoptée dans ce document, nous invitons le lecteur à se référer à l'annexe A.4.

#### 4.2.7 Validité écologique de l'expérience

##### 4.2.7.1 Diversité de la banque de sons

Afin de vérifier que la diversité des classes de sons proposées est suffisante pour pouvoir simuler un environnement sonore, nous analysons les commentaires des sujets sur la banque de données. 63% d'entre eux indiquent avoir été, au moins une fois, dans l'incapacité de trouver un son, avec un maximum de 4 sons par sujet. Parmi les sons manquants relevés, nous identifions 26 classes de sons dont :

- 16 sont pourtant présentes dans la banque de données, l'incapacité des sujets à les trouver n'étant donc pas imputable à la diversité de la base ;
- 1 fait référence à des sons de musique, que nous avons choisi délibérément d'occulter ;
- 9 sont effectivement absentes.

Concernant ces dernières, nous observons qu'il s'agit de classes très spécifiques (*e.g. voiture de sport ou voix d'adolescent*), et qui peuvent être remplacées par des classes similaires (*e.g. voiture, voix d'enfant ou voix d'adulte*). Nous en concluons que la diversité proposée par la banque de sons est satisfaisante, et suffisante, dans le cadre de notre étude.

#### 4.2.7.2 Ergonomie de l'interface de sélection

Nous voulons vérifier l'efficience de l'interface de sélection. Nous analysons les retours des sujets. 32.5% d'entre eux indiquent spontanément que l'interface est un moyen « simple et efficace » de sélectionner des sons sans l'aide de texte. 57.5% ne font pas mention de difficultés particulières, 10% signalent enfin avoir rencontré des difficultés avec l'interface, sans toutefois que la simulation en ait été affectée.

Nous en concluons que l'interface de sélection, sans texte, ne perturbe pas les sujets outre mesure. Un même constat avait été tiré de l'expérience pilote (Lafay, 2013; Lafay et al., 2014).

#### 4.2.7.3 Ergonomie de l'interface de simulation

Aucun sujet n'a rapporté de problèmes majeurs concernant l'interface de simulation. Plusieurs sujets l'ont par ailleurs spontanément décrite comme étant ludique.

#### 4.2.8 Vérification de l'agrément des scènes simulées

Nous analysons ici l'agrément perçu des 80 scènes sonores simulées. La Figure 29a affiche l'agrément moyen  $\mathcal{A}_{scène}$  pour les i- et ni-scènes.

Dans un premier temps, et afin de garantir la cohérence de nos données, nous voulons nous assurer qu'aucune ni-scène n'ait un  $\mathcal{A}_{scène}$  supérieur à celui d'une i-scène. Quatre des scènes ne respectent pas la contrainte. Elles et leurs correspondantes i ou ni sont retirées. 36 i-scènes et 36 ni-scènes restent donc dans le champ de l'analyse.

Dans un deuxième temps, nous voulons tester si les sujets ont bien perçu une différence d'agrément entre les i- et ni-scènes. Pour ce faire,

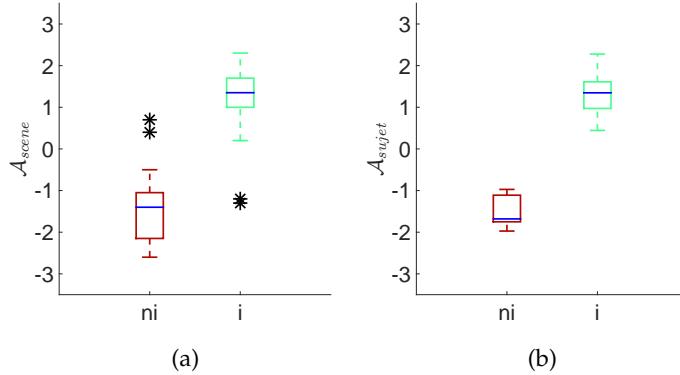


FIGURE 29 : Dispersion des notes données par les sujets lors de l’expérience 1.b moyennées suivant les sujets ( $A_{\text{scene}}$  : a), et suivant les scènes ( $A_{\text{sujet}}$  : b), en fonction du type de scènes (i ou ni).

nous observons l’agrément moyen de chaque sujet  $A_{\text{sujet}}$ , calculé séparément, pour chaque type d’environnement (cf. Figure 29b). Il apparaît que les i-scènes ont bien été perçues comme significativement plus agréables ( $p < 0.01$ ) que les ni-scènes.

#### 4.2.9 Étude comparative entre les descripteurs structurels

En premier lieu, nous nous concentrons sur le niveau sonore. Les figures 30a, 30b et 30c affichent les distributions des niveaux L, L(E) et L(T). Il existe bien une différence de niveau significative entre les i- et ni-scènes ( $L : p < 0.01$ ), avec un écart des moyennes de -7 dB. Cette différence affecte aussi bien les événements ( $L(E) : p < 0.01$ , écart moyen : -7 dB) que les textures ( $L(T) : p < 0.01$ , écart moyen : -6 dB).

Nous vérifions, sans surprise, que le niveau des sources sonores est bien un indicateur d’agrément, les ni-scènes ayant tendance à être plus fortes, fait rapporté dans un grand nombre d’études. Nous constatons encore que cette différence de niveau s’observe de manière égale pour les événements et les textures sonores.

Il apparaît que ce sont les événements qui impactent le plus le niveau global des scènes, l’écart entre L et L(E) n’étant que de 1 dB pour les i-scènes et les ni-scènes. Cette observation fait écho aux résultats obtenus par Kuwano *et al.* (Kuwano et al., 2003). Au cours de leur expérience, les auteurs demandent à leurs sujets d’évaluer une série d’environnements sonores de manière globale, dans un premier temps, puis d’en évaluer le niveau aux instants où chacun identifie une source sonore. L’étude montre qu’il n’y a pas de différences significatives entre les jugements globaux et les moyennes des jugements instantanés. Pour en revenir à notre expérience, c’est comme si nos sujets avaient inconsciemment tenu compte de cette réalité percep-

tive lors de la simulation, en faisant porter le niveau sonore global par des sons courts et bien identifiés, *i.e.* les événements.

Nous observons, enfin, que le niveau seul ne permet pas de clairement faire la distinction entre les différents types d'environnement. En effet, 20% des i-scènes ont un niveau supérieur au niveau minimal des ni-scènes, alors qu'il n'y a pas de recouvrement, si l'on considère l'agrément perçu  $\mathcal{A}_{scène}$ .

En second lieu, nous nous penchons sur les densités de sources sonores. Les Figures 31a et 31b affichent les distributions de D et D(E). Que l'on prenne en compte toutes les sources, ou uniquement les événements, la densité est significativement plus élevée pour les ni-scènes ( $D : p < 0.01$ ,  $D(E) : p < 0.05$ ). Nous observons un écart moyen de +0.36 pour D (soit en moyenne 2.3 sources sonores par fenêtre de plus pour les ni-scènes), et de +0.32 pour D(E) (soit en moyenne 2.1 sources sonores par fenêtre de plus pour les ni-scènes). Si ces écarts sont très similaires, c'est que la densité des textures D(T) ne varie pas de manière significative entre les i- et ni-scènes ( $D(T) : p = 0.15$ ), l'écart des moyennes étant de +0.17 (soit en moyenne 0.7 sources sonores par fenêtre de plus pour les ni-scènes), et l'écart médian étant, quant à lui, nul. Considérant ce résultat, nous ne tenons plus compte de D(T) dans la suite de l'analyse.

Nous constatons ici que la densité peut être un indicateur global de qualité, que l'on considère toutes les classes de sons, ou uniquement les événements sonores. Comme pour les niveaux sonores, la densité ne permet pas de clairement séparer les i- et ni-scènes, 43% des i-scènes ayant un D(E) supérieur à la densité d'événements minimale des ni-scènes.

En dernier lieu, nous nous intéressons à la diversité. Nous affichons sur la figure 32 DIV(E) et DiV(T), en séparant les différents niveaux d'abstractions. Excepté pour le niveau d'abstraction 0, la diversité des classes d'événements sonores est plus élevée pour les ni-scènes (DIV(E) niveaux 1,2 et 3 :  $p < 0.01$ ), avec en moyenne 2 classes présentes en plus. Aucune différence significative n'est observée pour les textures.

Les tendances globales observées montrent, d'une part, qu'un environnement sonore non-idéal est plus fort, plus dense, et composé d'une plus grande variété d'événements sonores —en un mot, plus « chargé »— qu'un environnement sonore idéal. Elles révèlent, d'autre part, que ce sont les caractéristiques des événements, plus que celles des textures, qui semblent porter la distinction entre les i- et ni-scènes. Cependant, aucun des descripteurs ne permet, à lui seul, de faire une distinction nette entre les deux types d'environnement, distinction pourtant perçue de manière non ambiguë par les sujets.

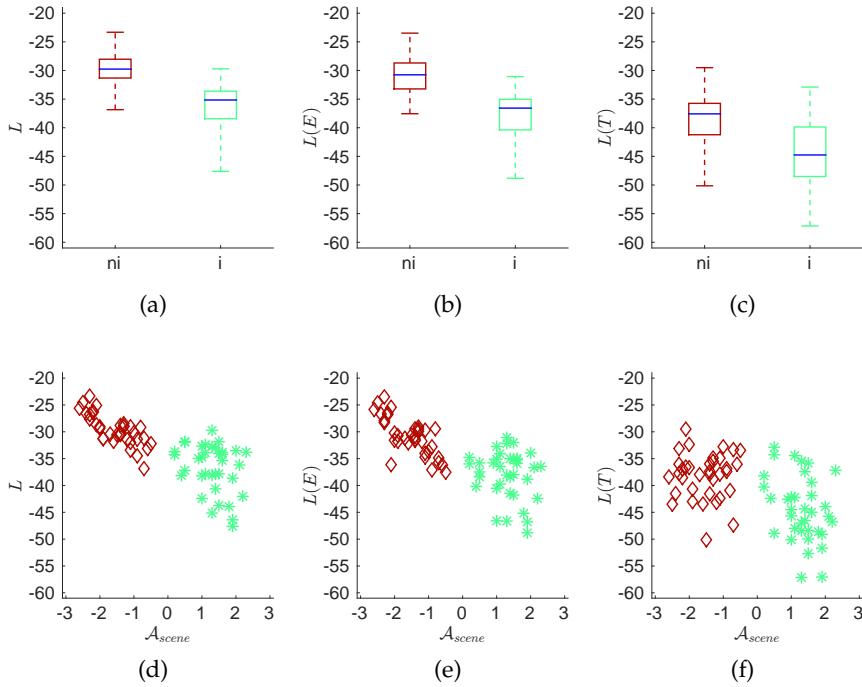


FIGURE 30 : Dispersion des descripteurs structurels de niveaux sonores  $L$  (a, d),  $L(E)$  (b, e) et  $L(T)$  (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{\text{scene}}$  de l'expérience 1.b (d, e, f).

#### 4.2.10 Influence des descripteurs structurels sur l'agrément perçu

Dans cette partie, nous analysons les relations fines qui peuvent exister entre les descripteurs structurels, d'une part, et l'agrément perçu, d'autre part. Contrairement à la section précédente, où la qualité affective des scènes est représentée de manière binaire (*i* vs. *ni*), nous considérons ici l'agrément moyen  $\mathcal{A}_{\text{scene}}$  comme descripteur perceptif. Il s'agit d'étudier l'existence de potentielles corrélations entre les descripteurs structurels et  $\mathcal{A}_{\text{scene}}$ . Les coefficients de corrélations linéaires calculés entre  $\mathcal{A}_{\text{scene}}$  vs.  $L$ ,  $L(E)$ ,  $L(T)$ ,  $D$ ,  $D(E)$  et  $DIV(E)$  sont présentés dans le tableau 6. Les relations entre  $\mathcal{A}_{\text{scene}}$  et les descripteurs structurels sont illustrées par les figures 30d, 30e et 30f, pour les niveaux sonores, et les figures 31c et 31d, pour les densités.

Concernant  $L$ , on observe une forte corrélation négative ( $r = -0.77$ ,  $p < 0.01$ ) avec  $\mathcal{A}_{\text{scene}}$ , indiquant que plus le niveau sonore est élevé, plus la scène est désagréable. Cependant, la figure 30d suggère que cette relation ne s'opère pas de la même manière pour les *i*- et *ni*-scènes. En effet, la corrélation entre  $L$  et  $\mathcal{A}_{\text{scene}}$ , pour les *ni*-scènes, reste élevée ( $r = -0.78$ ,  $p < 0.01$ ), mais est inexisteante pour les *i*-scènes.

Cette corrélation élevée, considérant l'ensemble des scènes, résulte du fait que les *i*-scènes ont tendance à être moins fortes que les *ni*-scènes.

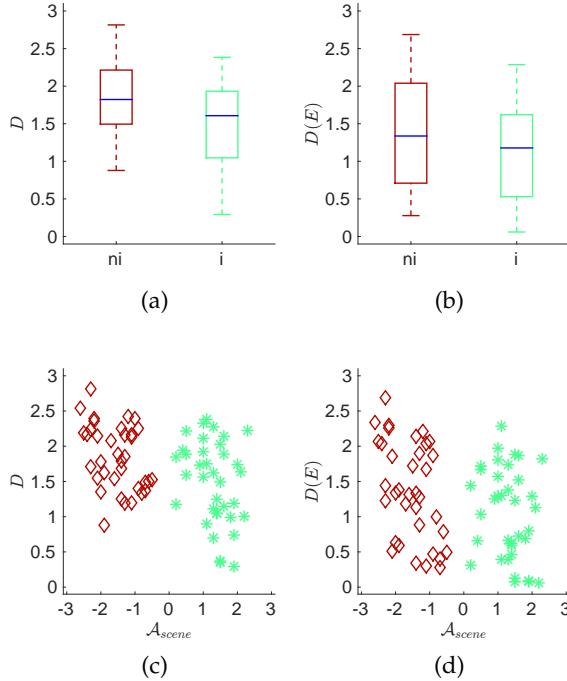


FIGURE 31 : Dispersion des descripteurs structurels de densité  $D$  (a, c) et  $D(E)$  (b, d), en fonction du type de scènes (a, b) et de l’agrément perçu  $A_{scene}$  de l’expérience 1.b (c, d).

scènes, donnant ainsi l’illusion de prolonger la corrélation négative observée pour les ni-scènes.

Nous en concluons que L :

- permet bien de faire la distinction entre les i- et ni-scènes,
- permet de finement caractériser l’agrément perçu des ni-scènes,
- n’est pas un indicateur pertinent de l’agrément perçu pour des environnements a priori agréables.

Les mêmes observations sont faites concernant  $L(E)$  (cf. 3oe). Pour  $L(T)$  (cf. 3of), bien que, à considérer l’ensemble des scènes, on observe une corrélation modérée, cela n’est pas vérifié quand on regarde séparément les i-scènes ( $r = -0.33$ ,  $p = 0.05$ ) et les ni-scènes ( $r = -0.00$ ,  $p = 0.99$ ). Là encore on peut penser que la corrélation négative observée pour l’ensemble des scènes est un artefact résultant du fait que le niveau des textures des i-scènes a tendance à être plus bas que celui des ni-scènes. Ainsi, si les événements sonores conservent une certaine capacité de prédiction de l’agrément pour les ni-scènes, le niveau des textures n’apporte, lui, que peu d’informations, quel que soit l’environnement.

Considérant l’ensemble des scènes, nous observons une corrélation négative faible pour  $D$  ( $r = -0.43$ ,  $p < 0.01$ ) et  $D(E)$  ( $r = -0.34$ ,

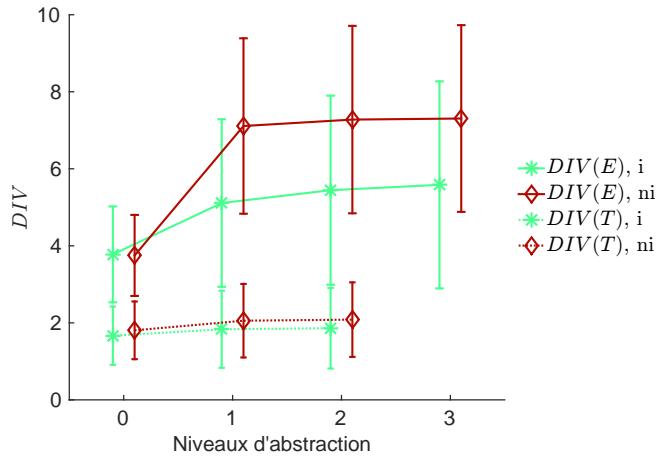


FIGURE 32 : Moyenne et écart type de la diversité des classes utilisées en considérant l'ensemble des classes (DIV), les classes d'événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i- et ni-scènes ainsi que les différents niveaux d'abstraction.

$p < 0.01$ ). Une relation semblable est observée pour les ni-scènes ( $D : r = -0.38, p < 0.05 ; D(E) : r = -0.46, p < 0.01$ ), mais aucune corrélation n'est observée pour les i-scènes. La densité des sources sonores semble donc avoir un faible impact sur l'agrément perçu, s'agissant des ni-scènes, et n'en a aucun, s'agissant des i-scènes.

En ce qui concerne la diversité des classes d'événements, une corrélation négative faible est observée pour les niveaux d'abstraction 1, 2 et 3, en tenant compte de l'ensemble des scènes. Si l'on considère les i- et ni-scènes séparément, aucune corrélation significative n'est trouvée. Les conclusions sont similaires à celles faites pour  $L(T)$  : la diversité permet uniquement de faire la distinction entre les deux types d'environnement, mais ne permet pas de caractériser précisément l'agrément perçu.

En résumé, en présence d'un environnement désagréable, les niveaux sonores, en particulier ceux des événements, ainsi que, dans une moindre mesure, la densité de sources présentes, ont un impact négatif sur l'agrément. En présence d'un environnement agréable, en revanche, aucun des descripteurs structurels considérés ici ne semble influer sur la perception de l'agrément.

Ces premiers résultats pourraient montrer qu'il existe deux modes de perception, mobilisant chacun des descripteurs indépendants, modes qui s'activent en fonction de la nature de l'environnement (i ou ni).

Le fait qu'aucun des descripteurs globaux ne permette de caractériser l'agrément des i-scènes peut nous amener à penser que toutes les sources sonores ne contribuent pas de manière égale à la perception de l'agrément, mais que seules les caractéristiques de certaines d'entre elles ont une réelle influence. Afin d'approfondir ce point,

	ensemble	i-scènes	ni-scènes
L	<b>-0.77</b> ( $p < 0.01$ )	-0.32 ( $p = 0.06$ )	<b>-0.78</b> ( $p < 0.01$ )
L(E)	<b>-0.75</b> ( $p < 0.01$ )	-0.20 ( $p = 0.24$ )	<b>-0.75</b> ( $p < 0.01$ )
L(T)	<b>-0.53</b> ( $p < 0.01$ )	-0.33 ( $p = 0.05$ )	-0.00 ( $p = 0.99$ )
D	<b>-0.43</b> ( $p < 0.01$ )	-0.31 ( $p = 0.07$ )	<b>-0.38</b> ( $p < 0.05$ )
D(E)	<b>-0.34</b> ( $p < 0.01$ )	-0.22 ( $p = 0.21$ )	<b>-0.46</b> ( $p < 0.01$ )
DIV(E) 0	-0.07 ( $p = 0.52$ )	-0.25 ( $p = 0.15$ )	-0.23 ( $p = 0.23$ )
DIV(E) 1	<b>-0.47</b> ( $p < 0.01$ )	-0.25 ( $p = 0.14$ )	-0.26 ( $p = 0.13$ )
DIV(E) 2	<b>-0.41</b> ( $p < 0.01$ )	-0.21 ( $p = 0.22$ )	-0.25 ( $p = 0.14$ )
DIV(E) 3	<b>-0.37</b> ( $p < 0.01$ )	-0.18 ( $p = 0.30$ )	-0.18 ( $p = 0.16$ )

TABLE 6 : Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen  $\mathcal{A}_{scène}$  de l'expérience 1.b et les descripteurs structurels.

nous analysons, dans la section suivante, les scènes d'un point de vue sémantique, *i.e.* en nous intéressant à la nature des sources qui les composent.

#### 4.2.11 Étude comparative entre les descripteurs sémantiques

##### 4.2.11.1 Analyse qualitative

Nous analysons la composition des scènes en comptant le nombre de sujets ayant utilisé une classe de sons pour simuler un type d'environnement. Les résultats sont présentés à la figure 33a pour les événements, et à la figure 33b pour les textures. Par souci d'espace, nous choisissons un niveau d'abstraction intermédiaire entre les niveaux 0 et 1, noté 0+, pour représenter les classes (cf. Figure 28).

Nous observons une différence notable dans le choix des classes entre les i- et ni-scènes. La répartition des classes est très proche de celle obtenue dans une étude similaire sur les environnements sonores urbains idéaux (Guastavino, 2006), *i.e.* les classes suggérant la présence humaine et la nature sont très présentes dans les i-scènes, à contrario, les classes désignant des sons mécaniques et/ou de travaux sont principalement utilisées pour les ni-scènes.

Ces résultats confirment un fait déjà observé : la nature sémantique des sources sonores joue un rôle prédominant dans l'appréciation de l'environnement (Dubois et al., 2006; Rimbault and Dubois, 2005).

Nous notons quelques différences avec (Guastavino, 2006) : les résultats obtenus par Guastavino montrent que les sons de *transports publics* sont caractéristiques des environnements sonores urbains idéaux. Les auteurs attribuent cela au fait que la perception de l'agrément est, entre autre, soumise à un contexte socio-culturel. Dans notre repré-

sentation du monde, les sons de transports publics sont positivement connotés, et ont ainsi tendance à être mieux acceptés que les sons de véhicules privés.

Dans une certaine mesure, nos résultats contredisent ce fait. La figure 33a montre, en effet, que les classes d'événements de *transports publics* (*bus* et *train*, cf. Figure 33c) ont été utilisées par les sujets, pour des i-scènes, dans 28% des cas, et pour des ni-scènes, dans 42% des cas. Les résultats ne remettent pas en question le fait que les sons de *transports publics* soient bien acceptés : 25% des sujets ont utilisé la classe *bus* pour les i-scènes, un chiffre comparable à celui de la classe *Vélo*, et bien supérieur à celui de toute autre classe de véhicules privés. Cependant les classes *transports publics* sont également bien présentes dans les ni-scènes, plus que les classes *voiture* ou *camion* par exemple. La classe *transports publics* ne peut donc pas être considérée comme typique d'un environnement sonore urbain idéal.

Cette différence peut s'expliquer par la nature des deux protocoles expérimentaux utilisés. Comme nous l'avons fait, Guastavino demande à ses sujets de décrire un environnement. Mais ces derniers travaillent de mémoire, alors que nos propres sujets disposent de supports sonores. Le fait que nos sujets soient confrontés à la réalité acoustique des sons, pour recréer leurs environnements, peut avoir pour effet de diminuer l'impact du contexte socio-culturel. D'autres études utilisant des sons comme stimuli montrent que la classe *bus* peut avoir un effet négatif sur l'appréciation de l'environnement (Lavandier and Defréville, 2006).

#### 4.2.11.2 Marqueurs sonores

Nous avons mis en évidence que, qualitativement, la composition des sources sonores des scènes diffère selon les types d'environnement (i ou ni). Nous essayons de voir maintenant si, parmi ces classes, certaines sont typiques d'un environnement en particulier. Pour ce faire, nous utilisons le V-test (cf. Section 4.2.6.2), en considérant séparément chaque niveau d'abstraction. Les résultats sont présentés dans le tableau 7.

Concernant les événements sonores, 9 marqueurs sont identifiés sur l'ensemble des niveaux d'abstraction. Comme la figure 33 le laissait présager, les classes relatives à la présence humaine (*pas homme béton*, *sonnette vélo*), et à la nature (*animaux*, *oiseaux*, *chants d'oiseaux*) sont des marqueurs de i-scènes. Nous notons également la présence de la classe *cloche* dans les marqueurs d'un environnement idéal. Ce fait est probablement dû au *background* socio-culturel des sujets, dans leur grande majorité, des citoyens européens. En effet, selon Schafer, un son reconnu par un individu comme faisant partie intégrante de son environnement est bien accepté. Les marqueurs de ni-scènes sont des classes faisant référence à des sons de travaux (*travaux*), ou suggérant un trafic dense (*klaxon*, *sirène*).

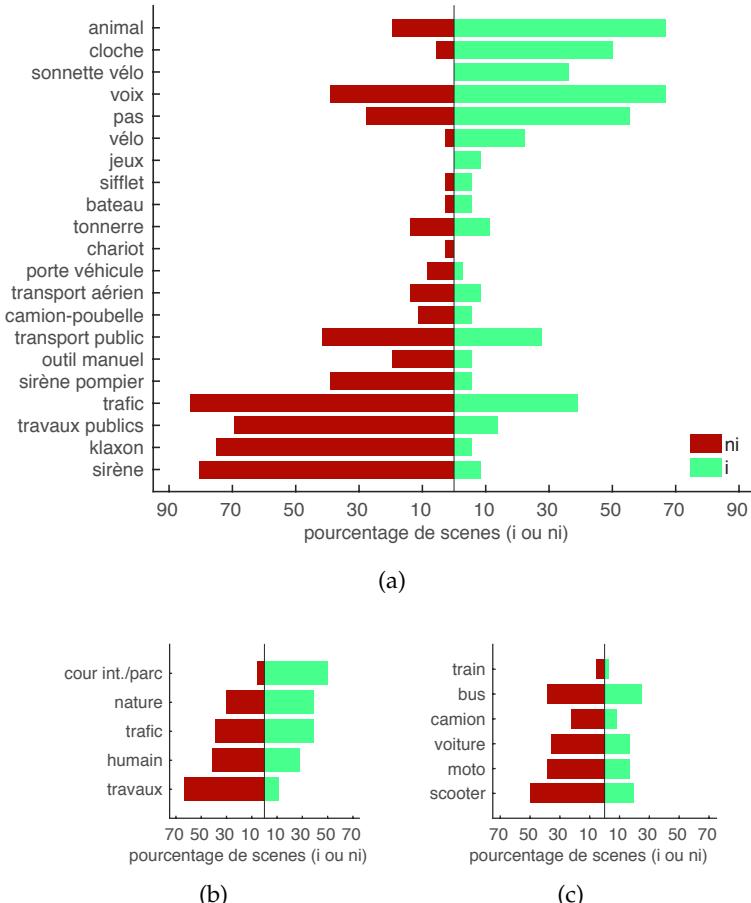


FIGURE 33 : Pourcentage de scènes simulées (i ou ni) comportant une classe de son particulière : (a) classes d'événements du niveau d'abstraction 0+, (b) classes de textures du niveau d'abstraction 0, (c) sous classes d'événements du niveau d'abstraction 1 appartenant aux classes *trafic* et *transport public* du niveau d'abstraction 0.

Concernant les textures sonores, 5 marqueurs sont identifiés. Pour les i-scènes, il s'agit de classes faisant référence à des ambiances amorphes, calmes, (*cour-intérieur/parc* et *parc*). Pour les ni-scènes, il s'agit, comme pour les événements, de classes faisant référence à des bruits de travaux (*travaux* et *véhicule de travaux*), ainsi que d'une classe faisant référence au trafic (*carrefour*).

Bien que l'ensemble des marqueurs identifiés soient intuitifs, aucune des classes d'événements faisant directement référence aux bruits de véhicules motorisés n'est un marqueur, exception faite de la classe de textures *carrefour*. Pour représenter un trafic désagréable, les sujets ont porté leurs choix sur les classes *klaxon* et *sirène*. On peut supposer que les sons isolés de véhicules sont compris comme faisant partie intégrante de l'environnement urbain, et ne sont donc pas particuliè-

Niveau d'abstraction	Marqueurs sonores événements	
	i-scènes	ni-scènes
0		travaux (3.78)
	cloche (4.5)	klaxon (3.9)
1	sonnette vélo (4.3)	sirène (3.9)
	animal (4.2)	
	oiseau (4.8)	klaxon (4.0)
2	cloche (4.4)	sirène (4.0)
	sonnette vélo (4.2)	
	chant oiseau (4.8)	klaxon (4.1)
3	cloche (4.3)	sirène (4.0)
	sonnette vélo (4.2)	
	pas chaussure (3.6)	

	Marqueurs sonores textures	
	i-scènes	ni-scènes
0	cour int./parc (4.1)	travaux (3.9)
1	parc (3.65)	carrefour (3.6)
		travaux véhicule (3.3)
2	parc (3.64)	carrefour (3.56)

TABLE 7 : Classes d'événements et de textures identifiées comme étant des marqueurs sonores. Dans chaque cellule, les marqueurs sont ordonnés par ordre décroissant de valeur V.

rement associés à un environnement désagréable.

#### 4.2.12 Étude des espaces de représentation induits par les descripteurs sémantiques

Dans cette partie, nous évaluons la capacité d'une représentation sémantique à séparer les deux types d'environnement. Pour ce faire, nous calculons une précision au rang 5 ( $p@5$ ) sur l'espace induit par les descripteurs sémantiques  $S$ , et ce pour chaque niveau d'abstraction (cf. Section 4.2.6.2). Les vecteurs  $S$  sont construits en utilisant toutes les classes (ET), les classes d'événements (E), les classes de textures (T), les classes d'événements ne considérant que les marqueurs sonores ( $E_m$ ), les classes d'événements ne considérant pas les marqueurs sonores  $E_{w/o,m}$ . Nous ne considérons pas les classes de marqueurs de textures, ces dernières étant trop peu nombreuses. Pour les mêmes raisons nous ne considérons pas les classes de marqueurs d'événements du niveau d'abstraction 0. Les résultats sont affichés sur la figure 34.

En ce qui concerne ET, la p@5 est de 76% pour le niveau d'abstraction 0, et reste supérieure à 86% à partir du niveau d'abstraction 1. Ces résultats confirment qu'il est possible de clairement distinguer les deux types d'environnement en se basant seulement sur la présence ou l'absence des classes de sons. Nous notons également que, plus le niveau d'abstraction est élevé, plus la capacité de séparer les environnements est importante. En d'autres termes, plus nous sommes précis dans notre description de la composition des scènes, plus nous sommes à même d'établir une distinction claire entre les i- et ni-scènes.

En considérant séparément E et T, il apparaît 1) que la p@5 obtenue avec E est similaire à celle obtenue avec ET, 2) que la p@5 obtenue avec T est systématiquement inférieure d'environ 10 à 15% à celle de E. Ces résultats indiquent que l'information sémantique permettant de séparer les deux environnements est principalement portée par les événements. Ces résultats font, par ailleurs, écho aux travaux de (Maffioli, 1999), qui montrent que nous analysons de manière descriptive (en identifiant les sources) les scènes événementielles, *i.e.* composées d'événements sonores (cf. Section 2.7.4.2).

Enfin, il apparaît que la p@5 obtenue avec  $E_m$  est égale, voire supérieure à celles obtenues avec E et ET, et ce bien qu'une information partielle soit utilisée dans ce cas pour décrire les scènes. La dimension des vecteurs de description S pour  $E_m$  est en effet inférieure à la dimension des vecteurs S pour E, qui est elle-même inférieure à celle obtenue dans le cas où toutes les classes sont utilisées (ET). De plus, dans le cas où les marqueurs ne sont pas pris en compte pour la description ( $E_{w/o,m}$ ), les résultats chutent, passant même en dessous de ceux obtenus en ne considérant que les textures. Cela confirme que la majorité de l'information sémantique permettant de faire la distinction entre i-scènes et ni-scènes est incluse dans les marqueurs.

En résumé, nous déduisons de cette analyse les points suivants :

1. contrairement à ce que nous avions constaté avec les descripteurs structurels, une description sémantique de la composition des scènes, en terme de présence/absence de sources sonores, permet de bien distinguer les deux types d'environnement (i ou ni);
2. l'information sémantique est majoritairement portée par les classes d'événements sonores ;
3. parmi les classes d'événements, seule une partie, *i.e.* les marqueurs sonores, est nécessaire pour faire la distinction entre les i- et ni-scènes.

Maintenant que nous avons isolé les classes typiques des i- et ni-scènes, et vérifié que la distinction entre ces environnements dépendait de la présence de ces classes, il reste à voir si une description

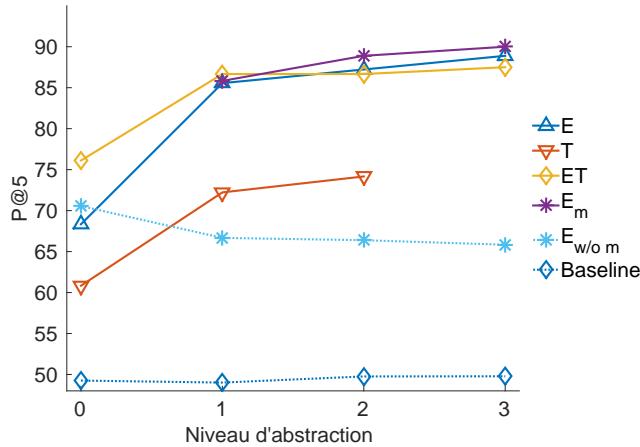


FIGURE 34 : P@5 obtenues en considérant la matrice de dissimilarité résultant des distances par paires de Hamming calculées entre les vecteurs des descripteurs sémantiques des scènes. Les vecteurs sont construits en utilisant toutes les classes (ET), les classes d'événements (E), les classes de textures (T), les classes d'événements ne considérant que les marqueurs sonores ( $E_m$ ), les classes d'événements ne considérant pas les marqueurs sonores  $E_{w/o,m}$ .

structurelle des scènes, basée uniquement sur ces marqueurs sonores, permet de caractériser l'agrément perçu, mieux qu'une description structurelle globale.

#### 4.2.13 L'influence spécifique des marqueurs sonores sur l'agrément perçu

Comme pour la section 4.2.10, nous évaluons les corrélations entre  $A_{scène}$  et les descripteurs structurels. Pour cette section, les descripteurs structurels sont calculés en tenant compte des marqueurs sonores précédemment identifiés. Nous définissons  $X_m$  le descripteur X calculé en ne prenant en compte que les sons des marqueurs. A l'inverse, nous définissons  $X_b$  (b : pour « bruit ») le descripteur X calculé en prenant en compte toutes les classes de sons, excepté les marqueurs. Lorsque le descripteur caractérise une i-scène (idem pour une ni-scène), nous ne considérons, pour le calcul, que les marqueurs identifiés pour les i-scènes (ou pour les ni-scènes), que nous nommons i-marqueurs (ou ni-marqueurs). Les résultats sont affichés sur le tableau 8.

Considérons, dans un premier temps, les densités (cf. Figures 35). Les résultats pour  $D_m$  et  $D(E)_m$  sont similaires à ceux observés précédemment pour D et D(E), à l'exception de  $D_m$  qui ne présente plus une corrélation significative pour les ni-scènes. Ces résultats tendent à confirmer que la densité est un indicateur d'agrément de faible importance, qu'on la considère globalement, ou en prenant en compte

	i-scenes	ni-scenes
$L_m$	0.03 ( $p = 0.88$ )	<b>-0.75</b> ( $p < 0.01$ )
$L(E)_m$	0.08 ( $p = 0.66$ )	<b>-0.71</b> ( $p < 0.01$ )
$L(T)_m$	-0.11 ( $p = 0.66$ )	-0.17 ( $p = 0.37$ )
$L_b$	<b>-0.52</b> ( $p < 0.01$ )	-0.32 ( $p = 0.06$ )
$L(E)_b$	<b>-0.51</b> ( $p < 0.01$ )	-0.30 ( $p = 0.07$ )
$L(T)_b$	-0.32 ( $p = 0.05$ )	<b>-0.73</b> ( $p < 0.01$ )
$L_m - L_b$	<b>0.67</b> ( $p < 0.01$ )	-0.31 ( $p = 0.07$ )
$L(E)_m - L(E)_b$	<b>0.66</b> ( $p < 0.01$ )	-0.28 ( $p = 0.10$ )
$L(T)_m - L(T)_b$	0.16 ( $p = 0.54$ )	0.21 ( $p = 0.28$ )
$D_m$	0.03 ( $p = 0.87$ )	-0.31 ( $p = 0.07$ )
$D(E)_m$	0.14 ( $p = 0.41$ )	<b>-0.44</b> ( $p < 0.01$ )

TABLE 8 : Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen  $A_{scène}$  de l'expérience 1.b et les descripteurs structurels relatifs à la présence des marqueurs sonores.

les contributions séparées de différentes sources.

Concernant les niveaux sonores (cf. Figures 36), là encore les mêmes tendances sont observées entre  $L_m$ ,  $L(E)_m$  et  $L(T)_m$ , d'une part, et  $L$ ,  $L(E)$  et  $L(T)$ , d'autre part. Que l'on considère uniquement les marqueurs, ou l'ensemble des classes, il s'avère que :

1. il existe une différence significative entre les niveaux des i- et ni-scènes ( $L_m$ ,  $L(E)_m$  et  $L(T)_m$  :  $p < 0.01$ ) ;
2. le niveau sonore des scènes est majoritairement porté par les événements sonores, comparé aux textures sonores ;
3. le niveau sonore des événements a une influence sur la perception de l'agrément pour les ni-scènes, mais pas pour les i-scènes ;
4. le niveau sonore des textures ne joue aucun rôle dans la perception de l'agrément.

En conclusion, le niveau des ni-marqueurs a une influence négative sur l'agrément pour les ni-scènes. En revanche le niveau des i-marqueurs n'impacte pas l'agrément perçu pour les i-scènes.

En considérant maintenant les classes non marqueurs (cf. Figures 37), nous remarquons, sur les i-scènes, une corrélation négative modérée/-faible pour  $L_b$  ( $r = -0.52$ ,  $p < 0.01$ ) et  $L(E)_b$  ( $r = -0.51$ ,  $p < 0.01$ ). C'est la première fois qu'un indicateur objectif nous permet de préciser l'agrément des environnements agréables. Ceci nous amène à

conclure que le niveau des classes de sons n'étant pas typiques d'un environnement agréable a un impact négatif sur l'agrément.

Par ailleurs, alors que  $L(T)$  ne présentait pas de corrélation pour les ni-scènes, une corrélation négative forte est observée pour  $L(T)_b$  ( $r = -0.73$ ,  $p < 0.01$ ). Ce fait indique que les niveaux des classes de textures n'étant pas des marqueurs n'affectent pas l'agrément perçu de la même manière pour les i- et ni-scènes. Les niveaux semblent avoir un effet négatif pour les ni-scènes, alors que pour les i-scènes, aucun effet n'est relevé.

Pour finir, nous considérons un dernier groupe de descripteurs, nommément  $L_m - L_b$ ,  $L(E)_m - L(E)_b$  et  $L(T)_m - L(T)_b$  (cf. Figures 38). Ces descripteurs expriment la différence entre les niveaux des marqueurs, et ceux des autres classes de sons. Ils traduisent l'émergence des marqueurs par rapport à la mixture sonore.

Pour les i-scènes, une corrélation modérée et positive est observée pour  $L_m - L_b$  ( $r = 0.67$ ,  $p < 0.01$ ) et  $L(E)_m - L(E)_b$  ( $r = 0.66$ ,  $p < 0.01$ ). Pour les ni-scènes, aucune corrélation n'est observée. Dans le cas des i-scènes, ce n'est donc pas le niveau absolu des marqueurs qui importe, mais leur niveau relatif, par rapport aux autres sons qui composent la scène. On observe donc pour les environnements idéaux un double mécanisme perceptif :

- plus le niveau absolu des sons n'étant pas des i-marqueurs est élevé, plus l'agrément est faible,
- plus le niveau relatif des i-marqueurs, par rapport aux autres sons, est élevé, plus l'agrément est élevé.

Pour les ni-scènes, le fait que nous observions des corrélations pour  $L_m$  et  $L(E)_m$ , et aucune pour  $L_m - L_b$  et  $L(E)_m - L(E)_b$ , montre que c'est bien le niveau absolu qui importe.

#### 4.2.14 Discussions

Dans cette expérience, nous identifions 6 indicateurs structurels globaux permettant de distinguer les environnements sonores idéaux et non-idéaux.

- niveau sonore : calculé sur tous les sons  $L$ , les événements  $L(E)$  et les textures  $L(T)$  ;
- densité : calculée de manière globale  $D$  et sur les événements  $D(E)$  ;
- diversité : calculée uniquement sur les événements  $DIV(E)$ .

Parmi ces indicateurs structurels, seuls  $L$  et  $L(E)$  permettent de prédire l'agrément. Nous notons cependant que cette prédiction ne vaut que pour les ni-scènes.

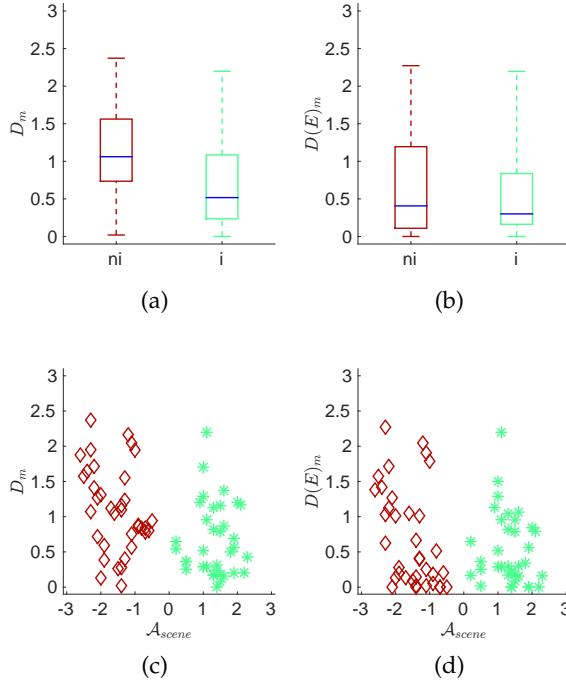


FIGURE 35 : Dispersion des descripteurs structurels de densité relatifs à la présence des marqueurs  $D_m$  (a, c) et  $D(E)_m$  (b, d), en fonction du type de scènes (a, b) et de l'agrément perçu  $\mathcal{A}_{scene}$  de l'expérience 1.b (c, d).

Nous observons qu'une description sémantique des scènes, basée sur la présence/absence des classes de sons, permet de bien prédire la nature de l'environnement. Par ailleurs, il apparaît qu'il est possible d'obtenir une prédiction similaire, voire meilleure, en ne considérant qu'un sous groupe de classes d'événements, *i.e.* les marqueurs sonores.

Parmi les descripteurs structurels spécifiques, calculés en tenant compte des marqueurs sonores, plusieurs permettent maintenant de faire la distinction entre les i-scènes et les ni-scènes :

- niveau sonore :  $L_m$ ,  $L_b$ ,  $L(E)_m$ ,  $L(E)_b$ ,  $L(T)_m$ ,  $L(T)_b$ ,  $L_m - L_b$  et  $L(E)_m - L(E)_b$  ;
- densité :  $D_m$ .

Parmi ces descripteurs, 8 semblent impacter l'agrément perçu :

- $L_b$  et  $L(E)_b$  ont un impact négatif sur les i-scènes ;
- $L(E)_m - L(E)_b$  et  $L_m - L_b$  ont un impact positif sur les i-scènes ;
- $L_m$ ,  $L(E)_m$ ,  $L(T)_b$  et  $D(E)_m$  ont un impact négatif sur les ni-scènes.

De cette analyse, nous retenons les points suivants :

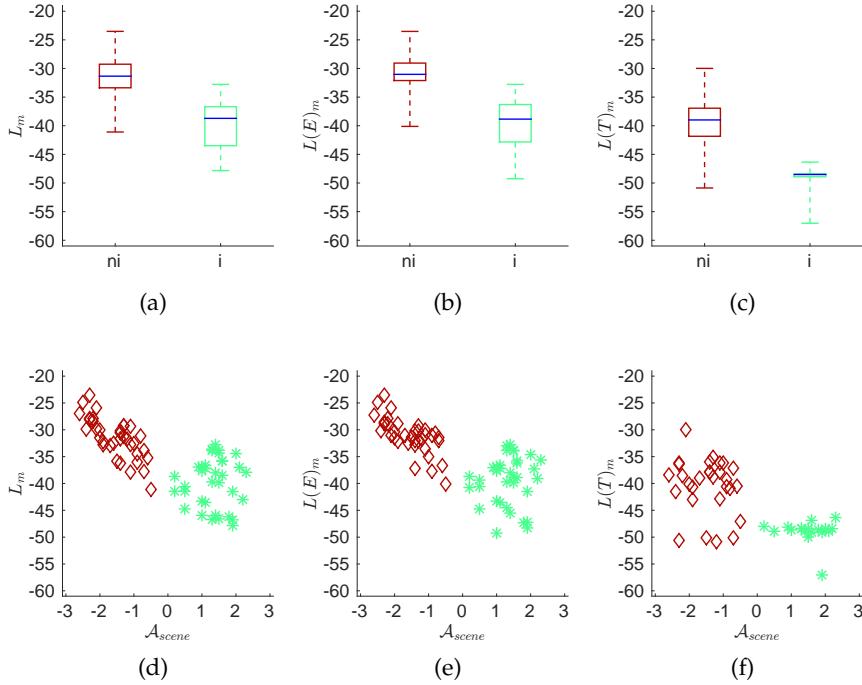


FIGURE 36 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs  $L_m$  (a, d),  $L(E)_m$  (b, e) et  $L(T)_m$  (c, f), en fonction du type de scènes (a, b, c) et de l’agrément perçu  $A_{scene}$  de l’expérience 1.b (d, e, f).

- *distinguer les i- et ni-scènes* : les descripteurs sémantiques, ainsi que certains descripteurs structurels globaux, permettent de faire la distinction entre les i-scènes et les ni-scènes. La description sémantique semble être plus performante ;
- *événements ou textures* : que ce soit pour les descripteurs sémantiques ou structurels, ce sont majoritairement les événements qui permettent de distinguer les deux types d’environnement, les textures n’apportant, au mieux, qu’une information limitée ;
- *prédir l’agrément* : si l’on considère une description fine de l’agrément, il semble que la manière de percevoir la qualité de l’environnement diffère en fonction de la nature de ce dernier (i ou ni). Il n’apparaît pas envisageable de considérer un même jeu de descripteurs pour prédire, à la fois, l’agrément des i-scènes, et l’agrément des ni-scènes. Pour les ni-scènes, ce sont le niveau global ( $L$  et  $L(E)$ ), la densité globale ( $D$  et  $D(E)$ ), ou le niveau des marqueurs sonores ( $L_m$  et  $L$ ), qui impactent négativement l’agrément. On note ici que prendre en compte les contributions de différentes sources n’améliore pas la capacité de prédiction de l’agrément, par rapport à une analyse holistique de l’environnement. Pour les i-scènes, par contre, prédire l’agrément requiert d’étudier, de manière séparée, les caractéris-

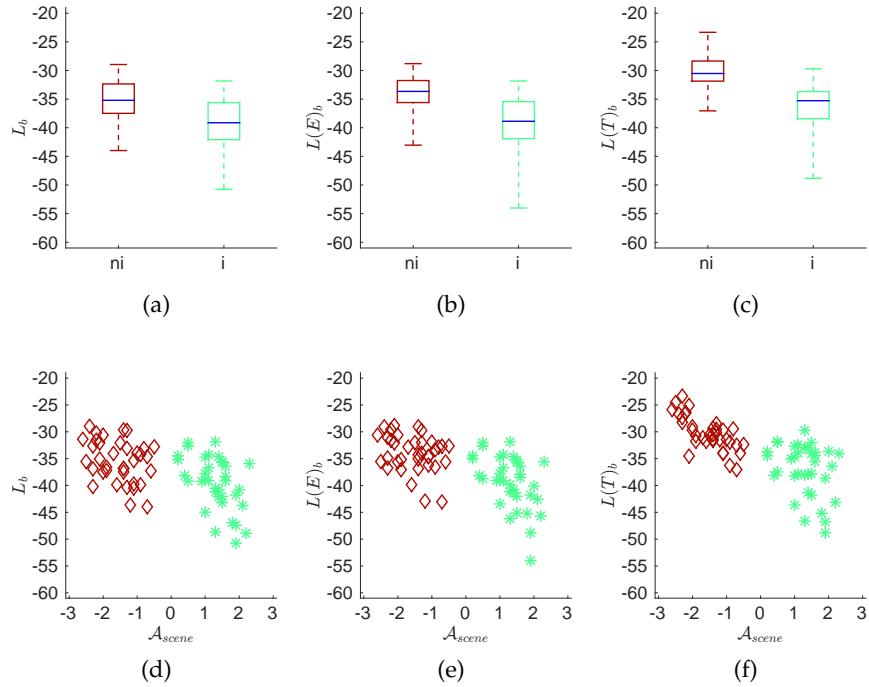


FIGURE 37 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs  $L_b$  (a, d),  $L(E)_b$  (b, e) et  $L(T)_b$  (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{\text{scene}}$  de l'expérience 1.b (d, e, f).

tiques des marqueurs sonores, et celles de l'ensemble des autres sons. Ainsi, le niveau des marqueurs relatifs au bruit est positivement corrélé à l'agrément, alors que le niveau du bruit est, lui, négativement corrélé.

Le fait que l'agrément des scènes agréables ne se corrèle pas avec des descripteurs physiques holistiques, contrairement à l'agrément des scènes désagréables, a récemment été déjà observé (Gozalo et al., 2015).

L'existence de deux modes de perception, mobilisant différents types de descripteurs, et dépendant de la nature (dans notre cas hédonique) des stimuli, est un phénomène qui se rapproche de celui observé pour la perception des textures (cf. Section 2.8). Le cerveau adapte sa manière de traiter l'information (résumé statistique pour les textures, description fine pour les événements) suite à une prise de décision antérieure quant à la nature du stimulus (à savoir « est-ce un événement ou une texture ? »). De la même manière, les indicateurs actifs dans le jugement de l'agrément dépendent, eux aussi, d'une identification préalable de la nature hédonique globale de l'environnement (idéale ou non idéale).

Ces résultats peuvent potentiellement influer sur les stratégies à adopter pour améliorer la qualité de l'environnement sonore :

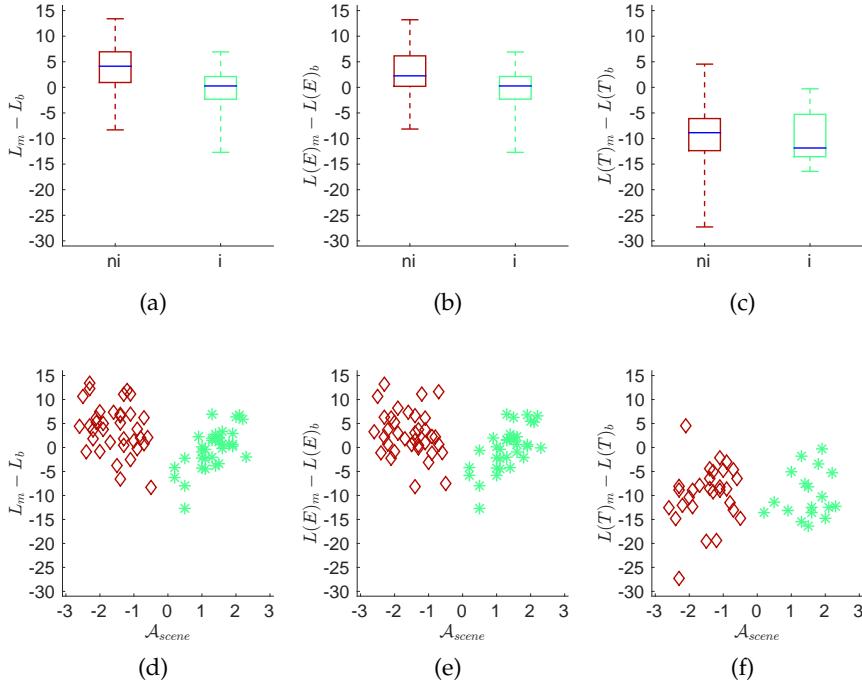


FIGURE 38 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs  $L_m - L_b$  (a, d),  $L(E)_m - L(E)_b$  (b, e) et  $L(T)_m - L(T)_b$  (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{scene}$  de l'expérience 1.b (d, e, f).

- dans le cadre de scènes non-idiéales, il s'agit de diminuer le niveau sonore, soit de manière globale, soit en agissant sur certaines sources (*sirène, klaxon*) ;
- dans le cadre de scènes idéales, il s'agit 1) d'identifier les sons agréables, *i.e.* les marqueurs sonores, 2) de baisser le niveau des autres sons, 3) voire, en restant dans la limite du raisonnable, d'augmenter le niveau des marqueurs par rapport aux autres sons.

Nous montrons que les descripteurs à utiliser dépendent de la nature de l'environnement, et que cette nature est elle-même dépendante de la composition sémantique, *i.e.* des sources sonores présentes. Dans une certaine mesure, nous pouvons donc dire que les descripteurs dépendent des sources sonores présentes.

### 4.3 AGIR SUR L'AGRÉMENT PERÇU EN MODIFIANT LA COMPOSITION SÉMANTIQUE

#### 4.3.1 *Objectif*

L'expérience précédente a montré que, parmi les classes de sons peu plant le monde sonore, celles regroupant les marqueurs sont caractéristiques de certains types d'environnement. Ces marqueurs sonores semblent avoir un impact particulier sur la perception de leurs environnements. C'est ce dernier point qui est étudié dans cette expérience.

Afin de vérifier que l'agrément des scènes idéales et non-idéales dépend de la présence des marqueurs, les scènes sonores précédemment simulées sont régénérées, sans les classes de marqueurs. Pour les i-scènes, les i-marqueurs sont retirés. Pour les ni-scènes, les ni-marqueurs. Une épreuve d'évaluation de l'agrément, dont le protocole se rapproche de celui de l'expérience 1.b, est alors conduite.

L'objectif est de vérifier si l'absence des marqueurs a un impact sur l'agrément perçu. Deux hypothèses sont formulées :

- pour les ni-scènes nous faisons l'hypothèse que l'absence des ni-marqueurs va **augmenter** la valeur de l'agrément perçu ;
- pour les i-scènes nous faisons l'hypothèse que l'absence des i-marqueurs va **diminuer** la valeur de l'agrément perçu.

Si la première hypothèse est intuitive, la deuxième l'est moins. En effet, il n'apparaît pas évident que la suppression des i-marqueurs, bien que s'agissant de sons positivement connotés, diminue la qualité globale d'un environnement. Cette suppression aura, de surcroît, pour effet de diminuer le niveau sonore global de la scène.

Néanmoins, comme nous l'avons vu, le niveau global n'est qu'un indicateur partiel de l'agrément pour les environnements sonores idéaux. Qui plus est, cet indicateur, lorsque qu'il décrit le niveau des i-marqueurs, impacte de manière positive la qualité de la scène. L'hypothèse mérite donc d'être vérifiée.

#### 4.3.2 *Planification expérimentale*

Nous nommons cette expérience : *expérience 2*.

#### Banque de données

La banque de données de stimuli compte 144 séquences de 30 secondes. Ces 144 séquences comprennent :

- 72 *am-scènes* : les 72 scènes précédemment simulées, incluant les classes de marqueurs (am). Nous notons i/am-scènes, les 36

scènes idéales avec marqueurs, et ni/am-scènes les 36 scènes non-idéales avec marqueurs ;

- 72 *sm-scènes* : les 72 scènes précédemment simulées, régénérées sans les classes de marqueurs (sm). Nous notons i/sm-scènes, les 36 scènes idéales sans marqueurs, et ni/sm-scènes les 36 scènes non-idéales sans marqueurs.

Nonobstant l'absence des marqueurs, les am- et sm-scènes sont en tout point semblables.

Nombre de am-scènes sont composées, en majorité, de samples de marqueurs. Afin de pas abusivement dénaturer ces scènes, en créant notamment des temps de « vide », *i.e.* ne comprenant aucun sample, nous ne supprimons que les marqueurs des classes d'événements du premier niveau d'abstraction (cf. Tableau 7). Ces classes sont :

- *cloche, sonnette de vélo, animaux* pour les i/sm-scènes ;
- *sirène, klaxon* pour les ni/sm-scènes.

Il est important de noter ici que tous les i- et ni-marqueurs ne sont donc pas supprimés dans les sm-scènes.

### **Procédure**

Les sujets évaluent les 144 scènes. L'évaluation s'effectue sur une échelle sémantique bipolaire de 11 points allant de -5 (non-idéale/très désagréable) à +5 (idéale/très agréable). Avant de noter une scène, les sujets doivent obligatoirement en écouter les 20 premières secondes. Après la notation, ils sont libres de passer à la scène suivante.

Pour chaque sujet, les scènes sont présentées dans un ordre aléatoire. Les 10 premières scènes permettent au sujet de calibrer ses notes. Elles sont obligatoirement composées de 5 i/am-scènes et de 5 ni/am-scènes. Ces 10 premières scènes sont rejouées à la fin de l'expérience, et seules les notes données à la deuxième occurrence sont prises en compte.

L'expérience est prévue pour durer 1 heure. Les sujets ne connaissent pas la nature des scènes.

### **Dispositif expérimental**

Tous les sujets passent l'expérience sur des machines identiques. L'audio est diffusé en monophonie, par le biais de casques audio semi-ouverts *Beyer-Dynamic DT 990 Pro*. Toutes les scènes sonores ont été re-simulées sur la base des partitions obtenues lors de l'expérience de simulation. Le niveau sonore de sortie est identique pour tous les sujets.

Tous les sujets réalisent l'expérience simultanément, dans un environnement calme. Ils n'ont pas le droit de s'adresser la parole pendant l'expérience.

Un expérimentateur est présent durant la totalité de l'expérience, afin de contrôler le bon déroulement de cette dernière, et de répondre aux éventuelles questions des sujets.

## Participants

12 sujets (4 femmes) participent à l'expérience. Aucun d'entre eux n'a réalisé l'expérience de simulation, ni la première expérience d'évaluation. Les sujets sont âgés de 22 à 61 ans (moyenne : 29.5, écart-type : 14). Tous les sujets vivent dans un milieu urbain.

Tous les sujets ont réalisé l'expérience avec succès.

### 4.3.3 *Données et méthodes d'analyses*

#### 4.3.3.1 *Nature des données analysées*

Les données analysées sont les mêmes que pour la première expérience. Nous invitons le lecteur à se référer à la section [4.2.6.1](#) pour plus de détails.

#### 4.3.3.2 *Méthodologie et Outils statistiques*

L'expérience aborde trois problématiques :

- *influence des descripteurs structurels des scènes avec marqueurs sur l'agrément perçu : une analyse comparative* : les 72 am-scènes utilisées par l'expérience 1.b étant ré-évaluées lors de cette expérience, il est donc possible de réaliser une étude comparative entre les expériences 1.b et 2, afin de vérifier la cohérence des résultats obtenus lors de l'expérience 1.b. Les méthodes d'analyse appliquées sont identiques à celles de l'expérience 1.b (cf. Section [4.2.6.2](#)) ;
- *influence de la présence des marqueurs sur l'agrément perçu* : il s'agit ici de vérifier que la suppression des i- et ni-marqueurs impacte l'agrément perçu. Pour ce faire, nous utilisons l'analyse de variance (cf. Annexe [A.2](#)). Nous considérons, comme variable dépendante,  $A_{\text{ sujet }}$ , et, comme variables indépendantes, le type d'environnement (i/ni), et la présence/absence de marqueurs (am/sm). Chaque sujet devant évaluer la totalité des stimuli, une ANOVA à mesures répétées à deux facteurs (cf. Annexe [A.2](#)) est utilisée afin vérifier s'il existe des différences significatives d'agrément perçu. Les deux variables indépendantes sont considérées comme des facteurs intra-sujet (*within-subject*,

cf. Annexe A.2). Les facteurs n'étant composés que de deux niveaux chacun (type : i/ni ; marqueur : am/sm), l'hypothèse de sphéricité n'a pas besoin d'être vérifiée. Les analyses *post hoc* sont conduites en appliquant la procédure de Tukey-Kramer (cf. Annexe A.3) ;

- *influence des descripteurs structurels des sm-scènes sur l'agrément perçu* : il s'agit là d'étudier l'agrément en fonction des indicateurs structurels des scènes. Dans un premier temps, nous vérifions que l'agrément moyen de chaque type de scènes (i, ni, am et sm) varie significativement. Une analyse de variance est pratiquée, avec, comme variable dépendante,  $A_{scene}$ , et, comme variables indépendantes, le type d'environnement (i/ni), et la présence/absence de marqueurs (am/sm). Dans cette analyse, les observations considérées sont les scènes. Comme il existe une dépendance entre les am et sm-scènes, une ANOVA à mesures répétées est utilisée (cf. Annexe A.2), comprenant, comme facteur intra-sujet (*within-subject*), la présence/absence de marqueurs, et comme facteur inter-sujet (*between-subject*), le type d'environnement. Les analyses *post hoc* sont conduites en appliquant la procédure de Tukey-Kramer (cf. Annexe A.3). Dans un second temps, comme pour l'expérience 1.b, nous étudions l'existence de relations linéaires entre les descripteurs structurels des sm-scènes et  $A_{scene}$ . Pour mesurer la corrélation, nous utilisons le coefficient de Pearson (cf. Annexe A.4).

Tous les tests de significativité sont effectués avec un seuil critique  $\alpha = 0.05$ .

#### 4.3.4 Détection de valeurs extrêmes

Considérons  $A_{sujet}$  pour les am-scènes (cf. Figure 40a). Il apparaît que les réponses du sujet 7 diffèrent des autres. Comme observé sur la figure 39 (cf. Figure 39g pour le sujet 7), ce dernier a évalué positivement près de la moitié des ni/am-scènes. Le sujet 7 a donné à 58% des ni/am-scènes une note supérieure à 0, contre une moyenne de 11% pour les autres sujets. De plus, le sujet 7 a utilisé l'ambitus maximal (-5 à 5) pour noter à la fois les i/ et ni/am-scènes. Ces faits n'ayant pas été observés pour les autres sujets, que l'on considère les expériences 2 ou 1.b, le sujet 7 est éliminé de l'analyse.

#### 4.3.5 Influence des descripteurs structurels des scènes avec marqueurs sur l'agrément perçu : une analyse comparative

Cette section présente une étude comparative entre les résultats de l'expérience 1.b, et ceux obtenus, pour les am-scènes, dans l'expérience ci-après.

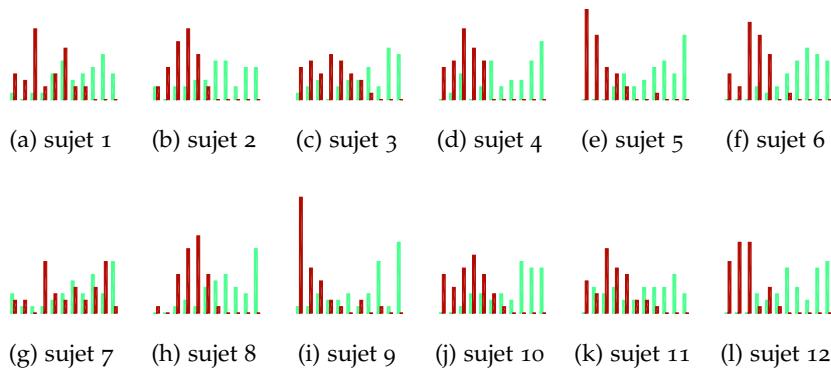


FIGURE 39 : Dispersion des notes données par les sujets lors de l'expérience 2 aux i/am-scènes (vert) et ni/am-scènes (rouge).

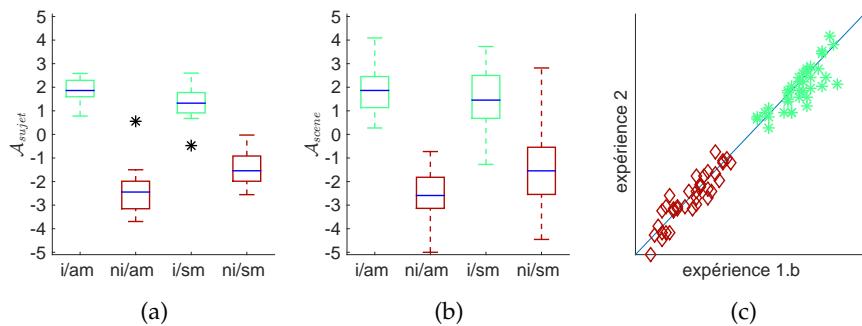


FIGURE 40 : Dispersion des notes données par les sujets lors de l'expérience 2 moyennées suivant les sujets ( $A_{sujet}$  : a), suivant les scènes ( $A_{scene}$  : b et c), en fonction du type de scènes (a et b) et des  $A_{scene}$  relevés à l'expérience 1.b.

Nous commençons par évaluer la corrélation entre les  $A_{scene}$  obtenus par les deux études. Les résultats sont affichés sur la figure 40c. La corrélation est élevée, que l'on considère l'ensemble des scènes ( $r = 0.98, p < 0.01$ ), ou séparément les i-scènes ( $r = 0.82, p < 0.01$ ) ou ni-scènes ( $r = 0.91, p < 0.01$ ).

Concernant les différences de  $A_{sujet}$  entre les i/am- et ni/am-scènes, nous observons un delta net et significatif ( $p < 0.01$ ), avec une différence moyenne des écarts de 4.5. Ces résultats sont en accord avec ceux de l'expérience 1.b.

Comme pour l'expérience 1.b, une analyse des relations entre les descripteurs structurels et  $A_{scene}$  est réalisée. Les résultats sont affichés dans le tableau 9. Ce tableau fait apparaître des différences.

Considérons les niveaux (L, L(E) et L(T)) pour les i-scènes. Une corrélation modérée négative est observée entre ces descripteurs et  $A_{scene}$ , alors qu'aucune n'était observée pour l'expérience 1.b. Il apparaît, dans l'expérience 2, que ces descripteurs ont joué un rôle (négatif) plus important dans l'évaluation des qualités affectives des

	i/am-scènes	ni/am-scènes
L	<b>-0.46*</b> ( $p < 0.01$ )	<b>-0.83</b> ( $p < 0.01$ )
L(E)	-0.33 ( $p = 0.05$ )	<b>-0.84</b> ( $p < 0.01$ )
L(T)	<b>-0.42*</b> ( $p < 0.05$ )	0.04 ( $p = 0.81$ )
D	<b>-0.42*</b> ( $p < 0.05$ )	<b>-0.47</b> ( $p < 0.01$ )
D(E)	<b>-0.36*</b> ( $p < 0.05$ )	<b>-0.57</b> ( $p < 0.01$ )
DIV(E) o	-0.26 ( $p = 0.13$ )	-0.32 ( $p = 0.06$ )
DIV(E) 1	-0.29 ( $p = 0.10$ )	-0.31 ( $p = 0.06$ )
DIV(E) 2	-0.24 ( $p = 0.17$ )	-0.32 ( $p = 0.06$ )
DIV(E) 3	-0.20 ( $p = 0.25$ )	-0.31 ( $p = 0.06$ )
L <sub>m</sub>	0.16 ( $p = 0.36$ )	<b>-0.75</b> ( $p < 0.01$ )
L(E) <sub>m</sub>	0.08 ( $p = 0.64$ )	<b>-0.73</b> ( $p < 0.01$ )
L(T) <sub>m</sub>	-0.05 ( $p = 0.86$ )	-0.06 ( $p = 0.76$ )
L <sub>b</sub>	<b>-0.64</b> ( $p < 0.01$ )	<b>-0.40*</b> ( $p < 0.05$ )
L(E) <sub>b</sub>	<b>-0.57</b> ( $p < 0.01$ )	-0.33 ( $p = 0.05$ )
L(T) <sub>b</sub>	<b>-0.46*</b> ( $p < 0.01$ )	<b>-0.83</b> ( $p < 0.01$ )
L <sub>m</sub> – L <sub>b</sub>	<b>0.60</b> ( $p < 0.01$ )	-0.25 ( $p = 0.14$ )
L(E) <sub>m</sub> – L(E) <sub>b</sub>	<b>0.56</b> ( $p < 0.01$ )	-0.27 ( $p = 0.11$ )
L(T) <sub>m</sub> – L(T) <sub>b</sub>	0.43 ( $p = 0.07$ )	0.36 ( $p = 0.05$ )
D <sub>m</sub>	-0.17 ( $p = 0.34$ )	-0.33 ( $p = 0.05$ )
D(E) <sub>m</sub>	-0.25 ( $p = 0.15$ )	<b>-0.53</b> ( $p < 0.01$ )

TABLE 9 : Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen  $\mathcal{A}_{\text{scene}}$  de l'expérience 2, et les descripteurs structurels globaux relatifs à la présence des marqueurs sonores pour les i/am-scènes et ni/am-scènes. L'indice \* indique les résultats qui diffèrent de ceux observés à l'expérience 1.

scènes, que dans l'expérience 1.b. Cependant, l'observation précédemment faite, sur le fait que ces descripteurs n'impactent pas de la même manière la perception des i et ni-scènes, se maintient. En effet les corrélations entre les niveaux et l'agrément perçu restent modérées pour les i-scènes ( $r < -0.46$ ), alors que celles observées pour les ni-scènes sont toutes élevées ( $r > -0.81$ ). Le niveau est donc bien pris en compte dans l'évaluation des i-scènes, mais moins que dans l'évaluation des ni-scènes. Cette recrudescence de l'importance du niveau est également observée sur L<sub>b</sub> pour les ni-scènes ( $r = -0.40$ ,  $p < 0.05$ ), et sur L(T)<sub>b</sub> ( $r = -0.46$ ,  $p < 0.01$ ) pour les i-scènes, mais là encore les corrélations restent modérées voire faibles.

Deux différences concernant les densités sont relevées. Nous observons une corrélation modérée sur D pour les i-scènes ( $r = -0.42$ ,  $p < 0.05$ ) et une corrélation faible pour D(E) ( $r = -0.42$ ,  $p < 0.05$ ).

	sm-scènes	i/sm-scènes	ni/sm-scènes
L	<b>-0.79</b> ( $p < 0.01$ )	<b>-0.49</b> ( $p < 0.01$ )	<b>-0.74</b> ( $p < 0.01$ )
L(E)	<b>-0.76</b> ( $p < 0.01$ )	<b>-0.44</b> ( $p < 0.01$ )	<b>-0.70</b> ( $p < 0.01$ )
L(T)	<b>-0.41</b> ( $p < 0.01$ )	-0.17 ( $p = 0.36$ )	-0.44 ( $p = 0.80$ )
D	<b>-0.49</b> ( $p < 0.01$ )	-0.31 ( $p = 0.07$ )	-0.29 ( $p = 0.08$ )
D(E)	<b>-0.45</b> ( $p < 0.01$ )	-0.29 ( $p = 0.09$ )	<b>-0.39</b> ( $p < 0.05$ )
DIV(E) o	-0.10 ( $p = 0.40$ )	-0.26 ( $p = 0.13$ )	-0.32 ( $p = 0.06$ )
DIV(E) 1	<b>-0.49</b> ( $p < 0.01$ )	-0.29 ( $p = 0.09$ )	-0.31 ( $p = 0.06$ )
DIV(E) 2	<b>-0.43</b> ( $p < 0.01$ )	-0.24 ( $p = 0.17$ )	-0.32 ( $p = 0.06$ )
DIV(E) 3	<b>-0.39</b> ( $p < 0.01$ )	-0.20 ( $p = 0.25$ )	-0.32 ( $p = 0.06$ )

TABLE 10 : Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen  $\mathcal{A}_{scène}$  de l'expérience 2 et les descripteurs structurels pour les i/sm-scènes et ni/sm-scènes.

Comme pour le niveau, la densité semble avoir une influence plus importante dans l'expérience 2.

La majorité des différences concerne les descripteurs des i-scènes. Pour tous ces descripteurs, les corrélations observées pour les ni-scènes sont plus importantes pour l'expérience 2 que pour l'expérience 1.b. Considérant les différences d'appréciation entre les i- et ni-scènes, les résultats restent donc consitants. Il apparaît que les descripteurs structurels de niveaux et de densités ont globalement plus influé sur l'agrément perçu dans l'expérience 2 que dans l'expérience 1.b.

Excepté ces points, tous les résultats observés dans les deux études concordent, notamment :

- l'effet bénéfique de l'émergence des i-marqueurs d'événements pour les i-scènes ( $L_m - L_b : r = 0.60, p < 0.01$ ;  $L(E)_m - L(E)_b : r = 0.56, p < 0.01$ );
- l'effet négatif des ni-marqueurs d'événements pour les ni-scènes ( $L_m : r = 0.75, p < 0.01$ ;  $L(E)_m : r = 0.73, p < 0.01$ );
- l'impact nul des marqueurs de textures pour les i- ( $L(T)_m : r = -0.05, p = 0.86$ ) et ni-scènes ( $L(T)_m : r = -0.06, p = 0.76$ ).

#### 4.3.6 Influence de la présence des marqueurs sur l'agrément perçu

Dans cette section nous étudions comment les sujets ont perçu les différents types de scènes, nommément : i/am-, ni/am-, i/sm- et ni/sm-scène. L'ANOVA à mesures répétées pratiquée sur  $\mathcal{A}_{sujet}$  (cf. Figure 40a) montre un effet significatif du type d'environnement (i/ni :  $F[1, 10] = 175, p < 0.01$ ), de la présence/absence des marqueurs

(am/sm :  $F[1, 10] = 7$ ,  $p < 0.05$ ), ainsi que de l'interaction entre les deux facteurs ( $F[1, 10] = 67$ ,  $p < 0.01$ ).

L'analyse *post hoc* montre, quant à elle, des différences significatives entre tous les groupes d'observations, notamment entre les i/am- et i/sm-scènes ( $p < 0.05$ ) et les ni/am- et ni/sm-scènes ( $p < 0.01$ ).

Ces résultats indiquent que la suppression des événements a effectivement modifié la perception des scènes par les sujets. Nos deux hypothèses sont ainsi vérifiées :

- la suppression des ni-marqueurs a amélioré les qualités perçues des ni-scènes ;
- la suppression des i-marqueurs a diminué les qualités perçues des i-scènes.

L'interaction significative montre que l'effet du type d'environnement influe sur l'effet de l'absence/présence des marqueurs. En effet la moyenne des écarts entre am- et sm-scènes est plus importante pour les ni-scènes (1.1) que pour les i-scènes (0.5).

#### 4.3.7 *Influence des descripteurs structurels des scènes sans marqueurs sur l'agrément perçu*

L'ANOVA à mesures répétées pratiquée sur  $\mathcal{A}_{\text{scène}}$  (cf. Figure 4ob) montre un effet significatif du type d'environnement (i/ni :  $F[1, 70] = 222$ ,  $p < 0.01$ ), de la présence/absence des marqueurs (am/sm :  $F[1, 70] = 5$ ,  $p < 0.05$ ), ainsi que de l'interaction entre les deux facteurs ( $F[1, 70] = 35$ ,  $p < 0.01$ ).

L'analyse *post hoc* montre des différences significatives entre tous les groupes d'observations, notamment, là encore, entre les i/am- et i/sm-scènes ( $p < 0.05$ ) et les ni/am- et ni/sm-scènes ( $p < 0.01$ ).

Ainsi, les quatre types de scènes, considérant  $\mathcal{A}_{\text{scène}}$  comme indicateur, forment bien quatre groupes distincts. L'interaction montre que le type d'environnement impacte l'effet provoqué par la suppression des marqueurs, les moyennes d'écart étant identiques à celles de l'analyse de la section précédente (ni-scènes : 1.1, i-scènes : 0.5, cf. Section 4.3.6).

Comparons maintenant les corrélations relevées entre les descripteurs structurels et  $\mathcal{A}_{\text{scène}}$  pour les am- et sm-scènes.

Concernant les i/sm-scènes, nous obtenons des corrélations significatives et modérées pour L ( $r = -0.49$ ;  $p < 0.01$ ) et L(E) ( $r = -0.44$ ;  $p < 0.01$ ). Ces corrélations sont plus importantes que celles relevées pour les i/am-scènes (cf. Tableau 9). Aucune corrélation n'est observée sur D et D(E), alors que c'est le cas pour les i/am-scènes.

Concernant les ni/sm-scènes, nous obtenons des corrélations significatives fortes pour L ( $r = -0.74$ ;  $p < 0.01$ ) et L(E) ( $r = -0.70$ ;  $p < 0.01$ ). Elles sont cependant plus faibles que celles observées pour

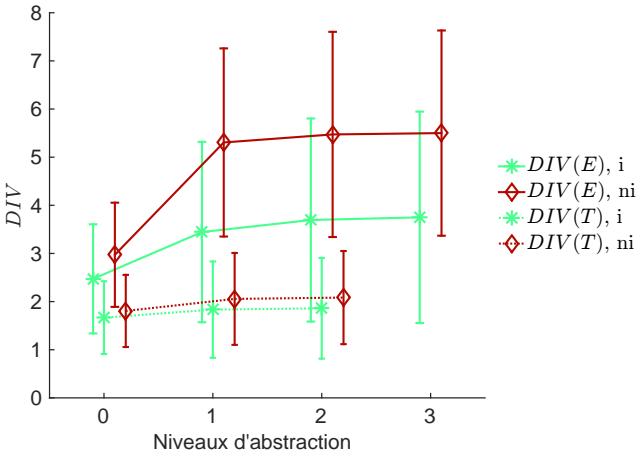


FIGURE 41 : Moyennes et écarts types de la diversité des classes utilisées en considérant l'ensemble des classes (DIV), les classes d'événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i/sm- et ni/sm-scènes ainsi que les différents niveaux d'abstraction.

les ni/am-scènes. Aucune corrélation n'est observée sur D, et une corrélation faible est observée sur D(E), alors que ces deux descripteurs montrent des corrélations modérées pour les ni/am-scènes.

Concernant les descripteurs relatifs aux densités, la suppression des marqueurs a amoindri les corrélations entre ces descripteurs et l'agrément perçu pour les i- et ni-scènes. Néanmoins, on peut raisonnablement supposer que ce fait est une conséquence directe de la réduction du nombre d'événements dans les sm-scènes.

Concernant les descripteurs de niveau sonore, on observe deux tendances contraires : pour les i-scènes, la suppression des marqueurs a renforcé les liens entre le niveau et l'agrément ; à l'inverse, pour les ni-scènes, cette suppression a amoindri les corrélations observées.

Ainsi, il apparaît que supprimer les marqueurs rapproche les manières dont sont perçues les i- et ni/sm-scènes, et ce, tant au niveau de l'agrément qu'elles suscitent, qu'au niveau des descripteurs semblant rendre compte de cet agrément.

#### 4.3.8 Discussions

De cette expérience nous retenons principalement deux points.

Premièrement, il apparaît que la présence, dans une scène, des marqueurs relevés à l'expérience 1 impacte bien l'agrément perçu. La suppression des ni-marqueurs a un effet bénéfique sur le ressenti, tandis que, plus surprenant, la suppression des i-marqueurs dégrade légèrement la qualité. Ce dernier point est d'autant plus marquant que, du fait de la suppression des marqueurs, le niveau des i/am-scènes est supérieur à celui des i/sm-scènes.

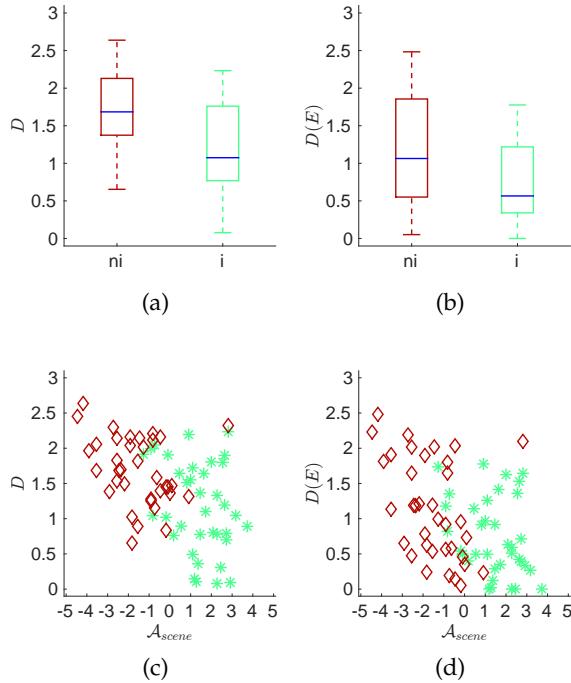


FIGURE 42 : Dispersion des descripteurs structurels de densité  $D$  (a, c) et  $D(E)$  (b, d), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b) et de l'agrément perçu  $\mathcal{A}_{scene}$  de l'expérience 2 (c, d).

Les i-marqueurs ont donc bien un effet bénéfique sur la perception d'un environnement. Le fait que leur suppression diminue  $\mathcal{A}_{scene}$  montre clairement qu'il est possible d'améliorer la qualité sonore d'un lieu en ajoutant des sons bien acceptés comme *oiseau*. Ces conclusions vont dans le sens de l'approche positive introduite par Schafer (Schafer, 1977) (cf. Section 1.2.2.1).

Deuxièmement, il apparaît que la suppression des marqueurs modifie les liens existant entre niveaux sonores et agrément perçu. Pour les i-scènes, ce lien est renforcé, tandis que pour les ni-scènes, il est dégradé. Ces résultats vont dans le sens de ceux obtenus lors de l'expérience 1, à savoir que l'influence d'un descripteur sur la perception d'une scène est conditionnée par la qualité hédonique de cette dernière. En l'occurrence, nous observons que plus l'agrément d'une scène est élevé, moins les descripteurs structurels influent sur la perception.

Enfin, si l'on considère les représentations mentales, ces observations nous incitent à conjecturer les points suivants :

- il existe un lien entre le concept concret relatif à une source sonore (*oiseau*), et le concept abstrait relatif à une qualité sonore (*agrément*) ;

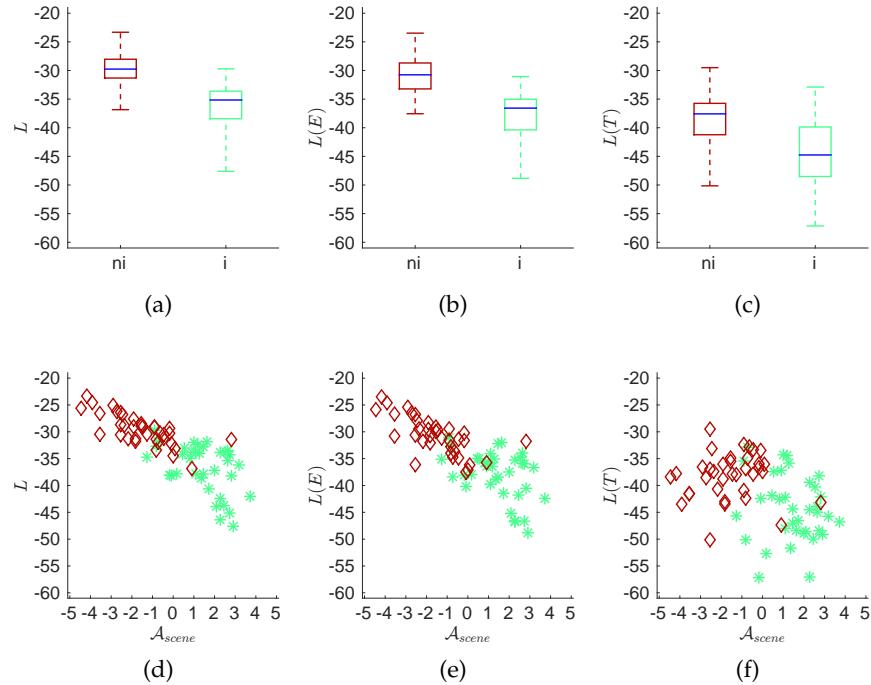


FIGURE 43 : Dispersions des descripteurs structurels de niveaux sonores L (a, d),  $L(E)$  (b, e) et  $L(T)$  (c, f), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b, c) et de l'agrément perçu  $\mathcal{A}_{scene}$  de l'expérience  $z$  (d, e, f).

- il existe un lien entre l'information modale que nous extrayons d'un environnement, et le concept abstrait relatif à sa qualité sonore (*agrément*).

#### 4.4 INFLUENCE DE LA COMPOSITION SÉMANTIQUE SUR LES PROCESSUS DE CATÉGORISATION DES SCÈNES

##### 4.4.1 Objectif de l'expérience

Cette expérience s'éloigne de la problématique de l'agrément perçu pour considérer celle, plus générale, de la catégorisation des environnements urbains en général.

L'objectif est, entre autre, de montrer que l'utilisation de scènes simulées permet d'aboutir à des résultats similaires à ceux obtenus avec des scènes enregistrées.

Nous considérons notamment les résultats obtenus par (Maffiolo, 1999) (cf. Section 2.7.4.2), qui montrent que des scènes événementielles<sup>10</sup> sont catégorisées suivant :

- les sources sonores présentes ;

<sup>10</sup> Les scènes simulées de l'expérience 1.a peuvent être toutes considérées comme des scènes événementielles, et non scènes amorphes

- la qualité de l'environnement.

Notons que la catégorisation est un processus dépendant du contexte sensoriel, *i.e.* de la nature des objets à catégoriser (cf. Section 2.3.5). En tenant compte du fait que les stimuli sont objectivement composés de deux sous-groupes caractérisés chacun par des agréments antinomiques (*i* et *ni*), il est raisonnable de présupposer que l'agrément perçu est considéré par les sujets comme une stratégie de catégorisation.

#### 4.4.2 Planification expérimentale

Nous nommons cette expérience : *expérience 3*.

##### Banque de données

La banque de données de stimuli est composée des 72 scènes simulées (36 *i*-scènes et 36 *ni*-scènes) lors de l'expérience 1.a.

##### Procédure

Les sujets doivent catégoriser les 72 scènes en fonction de la consigne suivante :

« Regrouper entre elles les scènes qui vous semblent similaires. »

La catégorisation est libre, les sujets peuvent former autant de groupes qu'ils le souhaitent, avec un minimum de deux.

Pour faciliter l'épreuve de catégorisation, une interface a été développée pour l'expérience. Sur cette interface, les scènes sont représentées par des points sur une surface plane. Lorsque le sujet clique sur un point, la scène est jouée.

Les sujets peuvent bouger les points, et leur affecter des couleurs afin de former des groupes. Le positionnement initial des points sur le plan est aléatoire et différent pour chaque sujet.

À la fin de l'expérience, les sujets doivent décrire les groupes ainsi formés. Afin de faciliter l'analyse lexicale des descriptions, il leur est demandé de se limiter à une liste de mots isolés, où à des phrases simples et courtes.

L'expérience est prévue pour durer entre 1h et 1h30.

##### Dispositif expérimental

Le dispositif expérimental est identique à celui de l'expérience 1.b (cf. Section 4.2.5.2)

## Participants

Les 10 sujets réalisant cette expérience sont les mêmes que ceux ayant participé à l'expérience 1.b (cf. Section 4.2.5.2). L'expérience 3 a été réalisée une semaine après l'expérience 1.b.

Le fait que les sujets soient déjà familiarisés avec les scènes à catégoriser permet de simplifier l'épreuve. En effet, catégoriser 72 stimuli peut se révéler une tâche laborieuse, surtout si ces stimuli sont des sons longs (30 secondes).

1 sujet est éliminé pour incompréhension des consignes.

### 4.4.3 *Données et méthodes d'analyses*

#### 4.4.3.1 *Nature des données analysées*

Dans un premier temps nous considérons de manière qualitative les stratégies de groupement suivies par les sujets. Ces stratégies sont objectivées en analysant les descriptions des groupes faites par les sujets.

Pour décrire chaque scène, nous considérons les trois descripteurs utilisés pour les expériences 1 et 2 (cf. Section 4.2.6.1) à savoir :

- *descripteur perceptif* ;
- *descripteur sémantique (objectif)* ;
- *descripteur structurel*.

À partir des descriptions verbales utilisées spontanément par les sujets pour décrire les groupes, nous isolons 2 nouveaux descripteurs, subjectifs par définition, susceptibles d'entrer dans la composition des scènes :

- *descripteur sémantique subjectif* : s'appliquant aux descriptions des sources sonores ;
- *descripteur de qualité subjectif* : s'appliquant aux descriptions des qualités affectives des scènes.

Afin d'établir un corpus de labels génériques pouvant désigner ces groupes, il s'avère nécessaire de procéder à une analyse lexicale des différents vocables utilisés par les sujets :

1. identifier les différents types de descriptions (*e.g.* sources, qualité affective, intensité sonore ...);
2. rassembler les labels dont le sens est proche sous une seule appellation (*e.g.* *trafic* et *car*  $\Rightarrow$  *trafic* cf. Section 4.4.5).

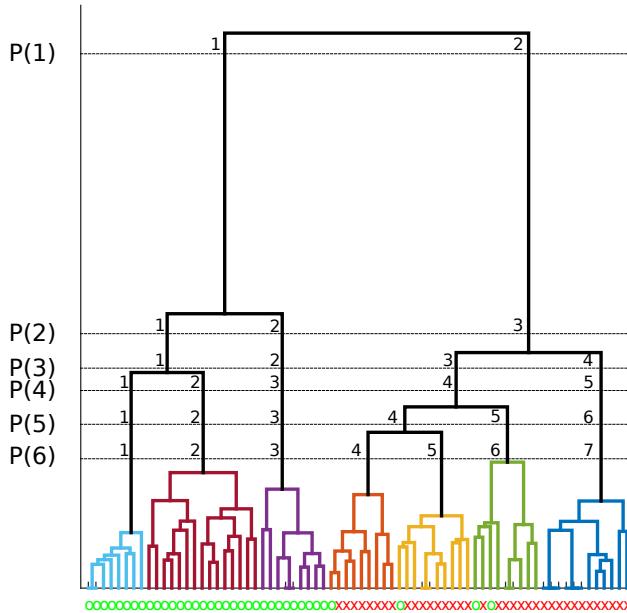


FIGURE 44 : Partitions  $P$  établies suivant la classification ascendante hiérarchique pratiquée sur la matrice de similarité  $\Theta$ , en utilisant un critère de Ward ; les ronds verts indiquent les i-scènes, et les croix rouges les ni-scènes.

La liste des labels liés à une scène forme le descripteur subjectif de cette scène. Afin d'affecter un label à une scène, celle-ci hérite des labels génériques des groupes auxquels elle a été affectée par l'ensemble des sujets.

#### 4.4.3.2 Méthodologie et outils statistiques

Dans un premier temps, il s'agit d'obtenir une vision globale des groupements effectués par les sujets. Pour ce faire, pour chaque sujet  $n$  une matrice de cooccurrence  $\delta_{i,j}^n$  est obtenue comme suit :

- $\delta_{i,j}^n = 0$  si le sujet  $n$  a groupé les scènes  $i$  et  $j$  ;
- $\delta_{i,j}^n = 1$  autrement.

Une matrice de dissimilarité globale  $\Theta_{i,j}$  est alors obtenue en moyennant les matrices  $\delta_{i,j}^n$  :

$$\Theta_{i,j} = \frac{1}{N} \sum_{n=1}^N \delta_{i,j}^n \quad (6)$$

avec  $N$  le nombre total de sujets. Il est à noter que les dissimilarités contenues dans  $\Theta$  respectent l'inégalité triangulaire ( $\Theta_{a,c} \leq \Theta_{a,b} + \Theta_{b,c}$ )

$\Theta_{b,c}$ ). Ces dissimilarités peuvent donc être considérées comme des métriques au sens mathématique du terme (Parizet and Koehl, 2012).

Une représentation arborée est utilisée afin de représenter les dissimilarités ainsi obtenues. Une Classification Ascendante Hiérarchique (CAH), utilisant le critère d'agrégation de Ward est pratiquée sur  $\Theta$  afin de faire émerger les tendances de groupement globales. Le dendrogramme résultant est affiché sur la Figure 44.

Plusieurs partitionnements peuvent être effectués à partir du dendrogramme. Au lieu de considérer une partition particulière, 6 partitions distinctes sont analysées. Ces partitions sont nommées respectivement  $P(1), P(2) \dots P(6)$  et sont composées respectivement de 2, 3 ... 7 groupes. Nous notons  $X_{P(Y)}$  le groupe  $X$  de la partition  $Y$  (cf. Figure 44).

Pour chaque partition, il s'agit d'observer les descripteurs rendant compte des groupements effectués.

La capacité des classes (respectivement labels) de descripteurs sémantiques objectifs (respectivement subjectifs) à caractériser un groupe est évaluée via un V-test en appliquant une correction de Bonferroni pour tenir compte du nombre de classes (respectivement labels) élevé (cf. Section 4.2.6.2). Contrairement à l'expérience 1, le seuil de significativité est fixé à  $\alpha = 0.01$ .

Pour les descripteurs structurels et perceptifs, l'existence de différences significatives entre les groupes est testée à l'aide d'une ANOVA à 1 facteur (cf. Annexe A.2), le nombre de niveaux du facteur étant égal au nombre de groupes. Le seuil de significativité est fixé à  $\alpha = 0.01$ . L'analyse *post hoc* est effectuée en suivant la procédure de Tukey-Kramer (cf. Annexe A.3). Notons que pour la partition 1, composée des groupes ( $1_{P(1)}$  et  $2_{P(1)}$ ), le test se ramène à un test de Student à deux populations.

#### 4.4.4 Stratégie de catégorisation

En fonction des descriptions des sujets, nous relevons 4 stratégies de catégorisation (cf. Tableau 11), opérant respectivement suivant les sources, les qualités affectives, l'intensité sonore (« fort »/« faible » ou « silence »), et enfin le contenu fréquentiel (« haute fréquence »/« basse fréquence »).

Les termes « parc » et « marché » ont également été utilisés. Ces derniers font référence à des lieux plutôt qu'à des sources. Néanmoins, ils sont les seuls dans ce cas. De plus, il est possible de les relier à un groupe de sources. Enfin, ils correspondent directement à deux classes de sons de textures des scènes simulées. Ainsi, nous considérons ici ces termes comme des descriptions de sources sonores.

La stratégie la plus utilisée est celle des sources (6 sujets). Viennent ensuite dans l'ordre la qualité (4 sujets), l'intensité (2) et le contenu

sujet	source	qualité	intensité	fréquence	# Groupes
1	x				5
2*			x	x	7
3	x	x			6
4	x	x			11
5		x			6
6	x		x		8
7		x			6
8	x				7
9	x				7

TABLE 11 : Stratégies de catégorisation et nombre de groupements effectués.  
L'indice \* indique les sujets étant supprimés de l'analyse.

fréquentiel (1). Ces résultats concordent avec ceux de Maffiolo, 1999 (cf. Section 4.4.1).

Avant d'aller plus loin dans l'analyse, nous notons que la stratégie de groupement adoptée par le sujet 2 est singulière. Il est le seul à avoir employé le contenu fréquentiel, et n'a utilisé ni les sources présentes, ni les qualités affectives. Afin de conserver des résultats cohérents, il est supprimé de l'analyse<sup>11</sup>.

#### 4.4.5 Analyse lexicale des descriptions

L'intensité n'ayant été considérée que par un seul sujet (sujet 6), seules les stratégies basées sur les sources et les qualités affectives perçues sont considérées plus avant.

Pour chacune de ces stratégies, nous relevons les termes utilisés par les sujets pour décrire les groupes. En ce qui concerne les sources, plusieurs termes semblent faire référence à une même entité. Ces termes sont regroupés sous une seule appellation. Les termes relevés, ainsi que les groupements effectués suite à l'analyse des champs lexicaux des sources sonores identifiées, sont affichés sur le tableau 12.

Comme évoqué à la section 4.4.3.1, chaque scène hérite des labels utilisés pour décrire le groupe auquel elle appartient. La liste de l'ensemble des labels attachés à une scène, en considérant l'ensemble des sujets, forme un descripteur subjectif. Au vu des stratégies de catégorisation utilisées par les sujets, 2 descripteurs subjectifs sont considérées :

- descripteurs sémantiques subjectifs (S) : les sources ;

<sup>11</sup> La matrice de similarité globale  $\Theta$  ainsi que le dendrogramme résultant de la CAH présenté à la figure 44 ne tiennent pas compte du sujet 2

labels des sources		labels des qualités	
originaux	génériques	originaux	génériques
alarme (2)	—	très calme	—
klaxon (2)	—	calme (4)	—
sirène (3)	—	moyennement calme	—
parc	—	bucolique	—
nature (3)	—	supportable	—
marché	—	oppressant	—
oiseau (4)	—	imprévisible	—
bruit* (2)	—	agité	—
pas (2)	—	fatiguant	—
foule	—	énervant (2)	—
bruit de fond*	background*	déplaisant (2)	—
fond sonore*	—	très déplaisant	—
cloche (4)	cloche	insupportable	—
église (2)	—	apaisant	relaxant
mécanique	—	reposant	—
travaux publiques (4)	travaux	normal	habituel
outils	—	habituel	—
humain (4)	humain		
personne	—		
trafic (2)	trafic		
voiture (2)	—		
eau	eau		
pluie (2)	—		

TABLE 12 : Labels relevés sur les descriptions verbales des groupements effectués par les sujets, en considérant séparément ceux relatifs aux descripteurs de qualité subjectifs, et ceux relatifs aux descripteurs sémantiques subjectifs.

- descripteurs de qualité subjectifs (Q) : les qualités affectives perçues.

La figure 45 affiche le nombre de scènes décrites par chaque label, en considérant respectivement S (cf. Figure 45a et 45c) et Q (cf. Figure 45b et 45d).

Considérons en premier lieu les labels des sources sonores (cf. Figure 45a). Les sujets décrivent les sources en considérant plusieurs niveaux d'abstractions, allant du plus concret (« pas », « oiseau ») au plus abstrait (« humain », « nature »). 2 labels font référence à des classes de sons ambiguës, ne pouvant être directement liées à une source en particulier : « bruit » et « fond sonore ». 2 labels sont utilisés pour décrire plus de 50% du corpus : « trafic » et « travaux ».

Les labels de sources font clairement la distinction entre les i- et ni-scènes (cf. Figure 45d). Nous notons cependant quelques différences entre ces labels et la répartition des classes de sons utilisées pour simuler les scènes (cf. Figure 33). En effet les labels « humain » et « pas » sont majoritairement utilisés pour décrire les i-scènes, alors les classes *pas* et *voix* sont bien présentes dans les ni-scènes, et que ces dernières ne présentent pas de différences notables entre les i- et ni- scènes au niveau de leurs descripteurs structurels.

Il est remarquable de constater que 44% des labels utilisés (50% si l'on fait l'association entre le label « trafic » et la classe *carrefour*)

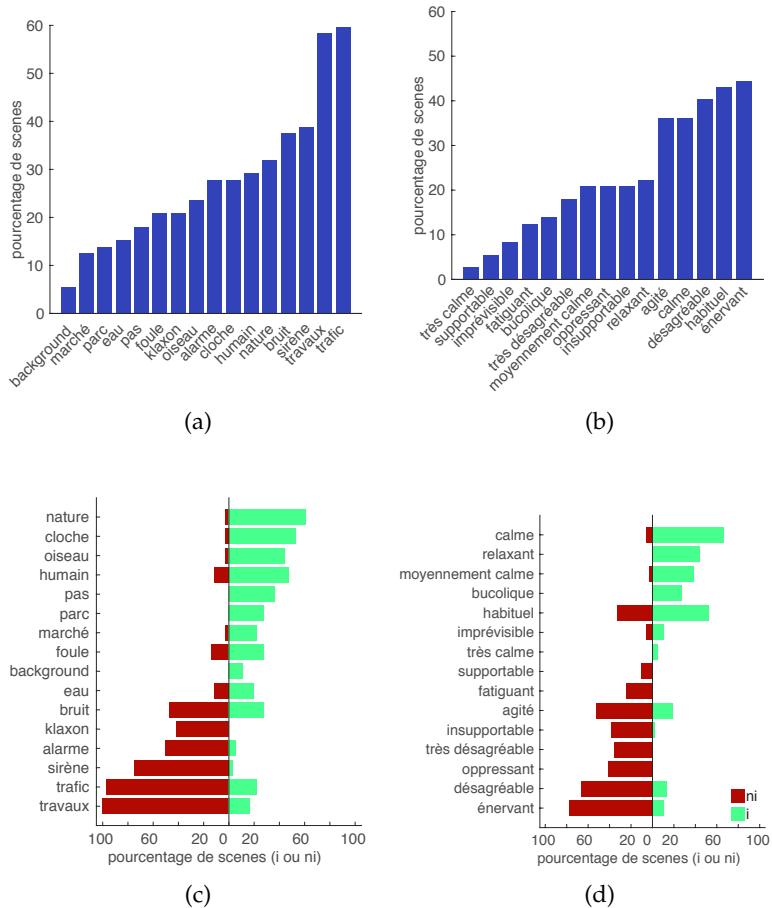


FIGURE 45 : Pourcentage de scènes étant décrites par un label sémantique subjectif donné (a, c), un label de qualité subjectif donné (b, d), en considérant l'ensemble des scènes (a, b) ou les i- et ni-scènes séparément (c, d).

font directement référence aux marqueurs sonores établis dans l’expérience 1 (cf. Section 4.2.11.2). Les marqueurs étant les classes de sons les plus représentées dans les i- et ni-scènes, ce fait était prévisible.

Considérons maintenant les labels des qualités affectives perçues (cf. Figure 45b). Ces derniers font directement référence aux indicateurs perceptifs habituellement utilisés par la communauté travaillant sur les paysages sonores pour caractériser les environnements sonores (cf. Section 2.7.3), à savoir :

- l’agrément (désagréable, très désagréable) ;
- le calme / la tranquillité (moyennement calme, calme, très calme, relaxant) ;
- la gêne (supportable, énervant, fatigant, oppressant, insupportable).

$1_{P(1)}$		$2_{P(1)}$			
nature		travaux			
cloche		sirène			
oiseau		trafic			
pas		alarme			
parc		klaxon			
marché					
humain					
$1_{P(2)}$		$2_{P(2)}$	$3_{P(2)}$		
cloche		parc	travaux		
pas		oiseau	sirène		
marché		nature	trafic		
humain			alarme		
foule			klaxon		
$1_{P(3)}$		$2_{P(3)}$	$3_{P(3)}$		
cloche		parc	travaux		
pas		oiseau	trafic		
marché		nature			
humain					
foule					
$1_{P(4)}$	$2_{P(4)}$	$3_{P(4)}$	$4_{P(4)}$		
cloche	pas	parc	travaux		
	marché	oiseau	trafic		
	nature	nature	sirène		
humain					
$1_{P(5)}$	$2_{P(5)}$	$3_{P(5)}$	$4_{P(5)}$	$5_{P(5)}$	
cloche	pas	parc	travaux		
	marché	oiseau	trafic		
	nature	nature			
humain					
$1_{P(6)}$	$2_{P(6)}$	$3_{P(6)}$	$4_{P(6)}$	$5_{P(6)}$	$6_{P(6)}$
cloche	pas	parc	klaxon		
	marché	oiseau	alarme		
	nature	nature			
humain					

TABLE 13 : Répartition des labels relatifs aux sources sonores relevées par les sujets en fonction des partitions établies par la classification ascendante hiérarchique (cf. Figure 44).

Deux derniers groupes de termes semblent faire référence à la nature prédictible de l'environnement (« habituel », « imprévisible »), ainsi qu'à sa structure temporelle (« agité »).

Nous notons que les labels liés à l'agrément, au calme et à la gêne font clairement la distinction entre les i- et ni-scènes (cf. Figure 45d). Cette distinction est moins marquée en ce qui concerne les labels « habituel » (5.5% de ni-scènes vs. 11% de i-scènes) et « imprévisible » (33% de ni-scènes vs. 53% de i-scènes).

#### 4.4.6 Descripteurs sémantiques subjectifs

Les résultats de la CAH pratiquée sur  $\Theta$  sont affichés à la figure 44. Pour chaque partition ( $P(1), P(2) \dots P(6)$ ) un V-test est pratiqué en considérant séparément les deux descripteurs sémantiques subjectifs

$1_{P(1)}$	calme relaxant moyennement calme bucolique	$2_{P(1)}$	énervant désagréable oppressant insupportable très désagréable
$1_{P(2)}$	calme moyennement calme	$2_{P(2)}$	$3_{P(2)}$ énervant désagréable oppressant insupportable très désagréable
$1_{P(3)}$	calme moyennement calme	$2_{P(3)}$	$3_{P(3)}$ désagréable agité
$1_{P(4)}$	$2_{P(4)}$ moyennement calme calme très calme	$3_{P(4)}$	$4_{P(4)}$ désagréable agité
$1_{P(5)}$	$2_{P(5)}$ moyennement calme calme très calme	$3_{P(5)}$	$4_{P(5)}$ désagréable agité
$1_{P(6)}$	$2_{P(6)}$ moyennement calme calme très calme	$3_{P(6)}$	$4_{P(6)}$ supportable désagréable
			$5_{P(6)}$ désagréable habituel
			$6_{P(6)}$ oppressant
			$7_{P(6)}$ très désagréable insupportable fatigant oppressant

TABLE 14 : Répartition des labels relatifs aux qualités affectives perçues en fonction des partitions établies par la classification ascendante hiérarchique (cf. Figure 44).

(S et Q). Les résultats sont affichés sur le tableau 13 pour S et le tableau 14 pour Q.

Concernant les groupements effectués, on constate que P(1) correspond largement à la distinction entre les i et ni-scènes.  $1_{P(1)}$  ne comporte que des i-scènes, tandis que  $2_{P(1)}$  est majoritairement composé de ni-scènes.

Considérons le descripteur S (cf. Tableau 13). La première partition P(1) fait la distinction entre, d'un coté, les sons humains, les sons naturels et les sons de cloches ( $1_{P(1)}$ ), et de l'autre, les sons de travaux, de trafic et les alarmes ( $2_{P(1)}$ ). Dans P(2), le groupe des i-scènes  $1_{P(1)}$  est subdivisé entre, d'un coté, les sons humains et les sons de cloches ( $1_{P(2)}$ ), et de l'autre, les sons naturels ( $2_{P(2)}$ ). Dans P(3), c'est le groupe des ni-scènes  $3_{P(2)}$  (similaire à  $2_{P(1)}$ ) qui est subdivisé entre, d'un coté, les sons de travaux et de trafic ( $3_{P(3)}$ ), et de l'autre, les sons de sirène et d'alarme ( $3_{P(4)}$ ). Dans P(4), le groupe  $1_{P(3)}$  (similaire à  $1_{P(2)}$ ) est subdivisé entre les sons de cloches et les sons humains.

Pour P(5) et P(6), seul le groupe  $4_{P(4)}$ , relatif aux ni-scènes, et représenté par les labels « travaux » et « trafic », est subdivisé. Cependant, aucune de ces subdivisions ne semble s'appuyer sur un label en particulier. En effet, aucun label n'est reconnu par le V-test comme étant caractéristique des groupes  $5_{P(5)}$ ,  $5_{P(6)}$  et  $5_{P(6)}$ . Seul le groupe  $4_{P(6)}$  semble dépendre du label « klaxon ».

Le descripteur sémantique subjectif relatif aux sources ne rend ainsi compte du partitionnement que jusqu'à P(4).

Considérons maintenant le descripteur Q (cf. Tableau 13). P(1) fait la distinction entre, d'un coté, les qualités positives, dominées par les labels « calme » et « relaxant » ( $1_{P(1)}$ ), et de l'autre, les qualités négatives, dominées par les labels « énervant » et « désagréable » ( $2_{P(1)}$ ). La partition P(2) subdivise les qualités positives ( $1_{P(1)}$ ) en faisant la distinction entre celles relatives au calme et à la tranquillité ( $1_{P(2)}$ ), et celles relatives à l'aspect régénératif (relaxant, reposant) des scènes ( $2_{P(2)}$ ). Par ailleurs le terme bucolique de  $2_{P(2)}$  fait directement écho aux labels de sources de ce même groupe, labels décrivant des sons naturels. La distinction opérée à la partition P(3) est principalement fonction de l'agrément, les scènes de  $3_{P(3)}$  étant caractérisées par « désagréable », et celles de  $3_{P(4)}$  par « très désagréable ».  $3_{P(4)}$  conserve par ailleurs les qualités « insupportable », « fatiguant » et « oppressant ». Enfin, c'est à ce niveau que l'impression suscitée par la structure temporelle des scènes fait son apparition, « agité » étant caractéristique de  $3_{P(3)}$ . Au niveau P(4), la qualité « calme » des scènes de  $1_{P(3)}$  (similaire à  $1_{P(2)}$ ) se subdivise entre, d'un coté, les scènes « moyennement calmes/calmes » ( $1_{P(4)}$ ), et de l'autre, les scènes « calmes/très calmes » ( $2_{P(4)}$ ). Le label « relaxant » est également associé au groupe  $2_{P(5)}$ .

Au niveau P(5), le groupe  $4_{P(4)}$  se subdivise entre les groupes  $4_{P(5)}$  et  $5_{P(5)}$ . Les qualités de  $4_{P(4)}$  (« désagréable » et « agité ») sont cependant entièrement conservées par  $4_{P(5)}$ ,  $5_{P(5)}$  étant caractérisé par « oppressant ».

Nous pouvons synthétiser ces résultats comme suit :

- P(1) fait la distinction entre les qualités positives et celles négatives ;
- les qualités positives se séparent ensuite suivant le caractère régénératif des scènes, et enfin suivant la notion de calme ;
- les qualités négatives se séparent d'abord en fonction de l'agrément, entre, d'un coté, les scènes « déplaisantes », et de l'autre, les scènes « très déplaisantes ». Une distinction est enfin opérée sur les scènes « déplaisantes » suivant qu'elles sont « supportables », « usuelles » ou « oppressantes ». Notons enfin que la notion d'agitation est liée au vocable « déplaisant » plutôt qu'au vocable « très déplaisant », alors que les qualités « insup-

portable » et « fatigante » sont toutes deux attachées à « très déplaisant ».

Comparons maintenant Q et S. Contrairement à S, le descripteur Q permet de caractériser l'ensemble des groupes,  $5_{P(5)}$ ,  $5_{P(6)}$  et  $6_{P(6)}$  n'étant décrits que par les qualités perçues. Au vu de la co-occurrence des labels, il est possible de faire des liens entre les qualités perçues, et les sources sonores utilisées pour décrire les scènes :

- « calme » ( $1_{P(2)}$ ) vs. « relaxant » ( $2_{P(2)}$ ) : la qualité « relaxant » semble liée aux sons naturels et la qualité « calme » aux sons d'origine humaine, ainsi qu'aux sons de cloches. Notons que la qualité « relaxant » apparaît également dans  $2_{P(4)}$ , simultanément à l'apparition du label « nature » ;
- « moyennement calme »/« calme » ( $1_{P(4)}$ ) vs. « calme »/« très calme » ( $2_{P(4)}$ ) : la qualité « moyennement calme »/« calme » semble liée aux sons de cloches, et la qualité « calme »/« très calme » aux sons d'origine humaine ;
- « désagréable » ( $3_{P(3)}$ ) vs. « très désagréable » ( $4_{P(3)}$ ) : les sources « travaux » et « trafic » semblent liées à « désagréable », alors que les sources « alarme » et « sirène » semblent correspondre à « désagréable ». Notons cependant que les sons « klaxon » et « alarme » ( $4_{P(6)}$ ) sont également liés aux qualités « désagréable » et « supportable ».

#### 4.4.7 Descripteurs sémantiques objectifs

Nous analysons les descripteurs sémantiques objectifs, *i.e.* les classes utilisées pour simuler les scènes. Le tableau 16 affiche les résultats des V-tests pratiqués sur les descripteurs sémantiques objectifs pour chaque partition.

Sans surprise, la distinction opérée par P(1) dépend des i et ni-marqueurs relevés à l'expérience 1 (cf. Section 4.2.11.2), exception faite de la classe marqueurs d'événements *pas chaussure* qui n'apparaît dans aucun groupe. Notons par ailleurs la présence de la classe d'événements *vélo* dans le groupe  $1_{P(1)}$ , classe n'étant pas un marqueur.

Considérons en premier les groupes relatifs aux i-scènes. S'agissant des classes d'événements, le partitionnement entre  $1_{P(2)}$  et  $2_{P(2)}$  sépare, d'un côté, les classes de sons de cloches et d'origine humaine ( $1_{P(2)}$ ), et de l'autre, les classes de sons naturels, de *bateau* et de *sonnette vélo* ( $2_{P(2)}$ ). Notons l'apparition du marqueur *pas chaussure* dans  $1_{P(2)}$ . S'agissant des textures, les classes *parc* (sons naturels) et *cour int.* sont caractéristiques de  $1_{P(2)}$ . Le partitionnement entre  $1_{P(4)}$  et  $2_{P(4)}$  sépare, lui, les classes d'événements de cloches des classes d'évé-

nements d'origine humaine. Les classes textures *cour int. parc* et *foule étrangère* sont caractéristiques de  $2_{P(4)}$ .

Considérons maintenant les groupes relatifs aux ni-scènes. Seule la classe de textures *travaux* semble déterminer les partitionnements  $P(3)$  et  $P(4)$ , les autres classes étant équitablement réparties entre les groupes ( $3_{P(3)}/4_{P(3)}$  et  $4_{P(4)}/5_{P(4)}$ ). Considérant les partitionnements  $P(5)$  et  $P(6)$ , seuls les groupes  $5_{P(5)}$  et  $6_{P(6)}$  semblent être caractérisés par la présence de classes, ces dernières étant relatives à des outils mécaniques (*outils électriques* et *perceuse*), pour les événements, et *café* pour les textures.

Comparons maintenant ces résultats à ceux obtenus par les descripteurs sémantiques subjectifs. Pour les groupes relatifs aux i-scènes, il existe une correspondance remarquable entre les labels donnés par les sujets et les classes utilisées pour simuler les scènes. Les différences sont :

- les classes *bateau* et *sonnette vélo* ( $2_{P(2,3)}$  et  $3_{P(4,5,6)}$ ) qui n'apparaissent pas dans les labels ;
- le label « marché » ( $1_{P(1,2,3)}$  et  $2_{P(4,5,6)}$ ) qui n'apparaît pas dans les classes.

Notons que si le label « nature » ( $2_{P(4,5,6)}$ ) peut faire référence aux classes de textures *cour int./parc* et *parc*, la classe correspondant au label « parc » ( $2_{P(2,3)}, 3_{P(4,5,6)}$ ) est moins évidente. On peut penser que ce label est le résultat de l'interprétation de la présence de la classe *oiseau*.

Pour les groupes relatifs aux ni-scènes, les correspondances sont évidentes pour les partitions  $P(1)$  et  $P(2)$ . Cependant à partir de la partition  $P(3)$ , les relations entre classes et labels sont plus ténues. Pour les labels « alarme » et « sirène » ( $4_{P(3)}, 5_{P(4)}, 6_{P(5)}$  et  $7_{P(6)}$ ), aucune classe n'est présente. A l'inverse, aucun label n'est présent pour les classes « outils électriques », « perceuse » et « café » ( $5_{P(5)}$  et  $6_{P(6)}$ ). Cependant, il est possible maintenant de faire le lien entre ces classes et la qualité « oppressant » ( $5_{P(5)}$  et  $6_{P(6)}$ ).

Notons enfin le groupe  $5_{P(6)}$  pour lequel aucune classe et aucun label ne sont relevés, le groupe étant seulement caractérisé par les qualités « désagréable » et « habituel ».

Le fait que les labels et les classes concordent jusqu'à la partition  $P(2)$  nous permet de conclure que les premiers groupements s'opèrent bien suivant la présence des sources. Cette stratégie perdure pour les groupes relatifs aux i-scènes. Pour ces dernières, les qualités affectives ne semblent qu'être une conséquence de la présence des classes (sons naturels  $\Rightarrow$  environnement relaxant et bucolique). Pour les groupes relatifs aux ni-scènes, les résultats obtenus soulèvent trois questions :

1. présence des labels et absence des classes ( $4_{P(3)}, 5_{P(4)}, 6_{P(5)}$  et  $7_{P(6)}$ ) : pourquoi les sujets ont-ils relevé des classes qui ne

sont pas caractéristiques des scènes à décrire, *i.e.* des classes qui sont également présentes sur d'autres scènes ? Une hypothèse est que les caractéristiques structurelles de ces classes (intensité sonore, densité d'événements), sont saillantes pour les scènes du groupe décrit ;

2. présence des classes et absence des labels ( $5_{P(5)}$  et  $6_{P(6)}$ ) : pourquoi les sujets n'ont-ils pas relevé les classes caractéristiques des scènes à décrire, *i.e.* majoritairement présentes sur ces dernières ? Ce cas n'apparaît que pour les classes *outils électriques*, *perceuse* et *café*. Pour les deux premières, on peut supposer que le niveau d'abstraction des classes est trop précis, les sujets interprétant « outils électriques » et « perceuse » comme appartenant à la classe travaux, sans faire de distinction particulière. Cette classe n'étant plus caractéristique des groupes relatifs aux niscènes à partir de la partition  $P(3)$ , il est normal que les sujets ne s'appuient plus sur la présence des classes *outils électriques*, *perceuse* pour décrire des groupes ;
3. présence des qualités et absence des labels et des classes ( $5_{P(6)}$ ) : à partir de quels descripteurs les sujets ont pu reconnaître des qualités affectives particulières aux scènes du groupe  $5_{P(6)}$  ? Comme pour la question 1, on peut supposer que cette distinction s'opère suivant les descripteurs structurels de ces scènes. Cependant, notons que parmi les deux qualités affectées au groupe  $5_{P(6)}$ , la première, « désagréable », est également caractéristique des classes des partitions supérieures de  $5_{P(6)}$  ( $4_{P(5)}$  et  $4_{P(4)}$ ), ainsi que d'une des classes du même niveau ( $4_{P(6)}$ ), soulignant *de facto* la faible capacité de ce descripteur à particulieriser  $5_{P(6)}$ . La deuxième, « habituel » illustre l'aspect normal, voire standard des scènes urbaines du groupe  $5_{P(6)}$ . Il est potentiellement difficile de rendre compte d'une telle qualité neutre à partir de descripteurs structurels.

Enfin, nous relevons que, comme observé pour les expériences 1 et 2, la perception des scènes est fonction de leur qualité (i/ni).

#### 4.4.8 Descripteurs perceptifs et structurels

##### 4.4.8.1 Descripteur perceptif

Nous considérons dans un premier temps le descripteur perceptif  $A_{scène}$ . Les résultats des ANOVA pratiquées sur les différentes partitions sont affichés sur le tableau 15. Nous ne considérons que les groupements effectués à partir de la partition  $P(3)$ , les partitions supérieures étant entièrement déterminées par les descripteurs subjectifs sémantiques  $S$  et  $Q$ .

	P(3)	P(4)	P(5)	P(6)
$\mathcal{A}_{\text{scene}}$	$F[3, 68] = 88$ $p < 0.01$	$F[4, 67] = 65$ $p < 0.01$	$F[5, 66] = 51$ $p < 0.01$	$F[6, 65] = 42$ $p < 0.01$
L	$F[3, 68] = 16$ $p < 0.01$	$F[4, 67] = 12$ $p < 0.01$	$F[5, 66] = 10$ $p < 0.01$	$F[6, 65] = 8$ $p < 0.01$
L(E)	$F[3, 68] = 15$ $p < 0.01$	$F[4, 67] = 12$ $p < 0.01$	$F[5, 66] = 10$ $p < 0.01$	$F[6, 65] = 8$ $p < 0.01$
L(T)	$F[3, 68] = 8$ $p < 0.01$	$F[4, 67] = 6$ $p < 0.01$	$F[5, 66] = 5$ $p < 0.01$	$F[6, 65] = 4$ $p < 0.01$
D	$F[3, 68] = 3$ $p < 0.05$	$F[4, 67] = 3$ $p < 0.05$	$F[5, 66] = 2.5$ $p < 0.05$	$F[6, 65] = 2.1$ $p = 0.07$
D(E)	$F[3, 68] = 2$ $p = 0.13$	$F[4, 67] = 2.5$ $p = 0.05$	$F[5, 66] = 2.1$ $p = 0.07$	$F[6, 65] = 1.8$ $p = 0.12$

TABLE 15 : Résultats des ANOVA à mesures répétées pratiquées sur les différents descripteurs structurels en tenant compte du partitionnement des scènes.

Pour les partitions P(3), P(4), P(5) et P(6), les ANOVA montrent un effet significatif des groupes sur  $\mathcal{A}_{\text{scene}}$ . Concernant l'analyse *post hoc* :

- groupes relatifs aux i-scènes : aucune différence significative n'est relevée entre ces groupes pour toutes les partitions ;
- groupes relatifs aux ni-scènes : une différence significative est relevée entre  $3_{P(3)}$  vs.  $4_{P(3)}$  (*idem* pour P(4) :  $4_{P(4)}$  vs.  $5_{P(4)}$ ) ;
- groupes relatifs aux i- et ni-scènes : toutes les comparaisons entre les groupes relatifs aux i- et ni-scènes sont significatives, et ce pour toutes les partitions.

Les résultats montrent que la distinction observée sur Q, pour les partitions P(3) et P(4), entre les qualités « désagréable » ( $3_{P(3)}$  et  $4_{P(4)}$ ) et « très désagréable » ( $4_{P(3)}$  et  $5_{P(4)}$ ), se retrouve si l'on considère l'agrément perçu.

#### 4.4.8.2 Descripteurs structurels

Nous considérons 5 descripteurs structurels, respectivement L, L(E), L(T), D et D(E). Les résultats des ANOVA sont présentés sur le tableau 15. Pour les niveaux sonores (L, L(E) et L(T)), les ANOVA montrent un effet significatif des groupes sur les variables considérées, pour toutes les partitions. Pour D, on observe un effet significa-

tif jusqu'à la partition P(5). Pour D(E), aucun effet significatif n'est trouvé, nous ne considérons ainsi plus cette variable.

Concernant l'analyse *post hoc* :

- L, L(T) et D : on observe des différences significatives uniquement en comparant des groupes relatifs aux i-scènes à des groupes relatifs aux ni-scènes ;
- L(E) : deux différences significatives sont relevées entre deux paires de groupes relatifs aux ni-scènes, à savoir  $3_{P(3)}$  vs.  $4_{P(3)}$  et  $4_{P(5)}$  vs.  $6_{P(5)}$ .

Ainsi, L(E) est le seul descripteur structurel à véhiculer une information permettant de discriminer des groupes relatifs à un même type d'environnement. Cette discrimination n'est cependant effective que pour les groupes relatifs aux ni-scènes. Elle s'accompagne d'un changement de qualité, les groupes ayant les niveaux sonores les plus faibles possédant les labels « désagréable », « agité » et « supportable » ( $3_{P(3)}$  et  $4_{P(5)}$ ), ceux ayant les niveaux plus élevés ( $4_{P(3)}$  et  $6_{P(5)}$ ) possédant les labels « très désagréable », « insupportable », « fatiguant » et « oppressant ».

Concernant le groupe  $5_{P(6)}$ , qui n'est pas caractérisé par une classe de sons ou un label de sources, il apparaît également qu'aucun descripteur structurel global ne semble rendre compte de ce groupement.

#### 4.4.9 Discussions

L'expérience permet de confirmer les résultats de (Maffiolo, 1999). Les groupements effectués dépendent bien, à la fois, des sources présentes, et des qualités affectives perçues.

Les qualités perçues font référence à l'agrément, le calme et la gêne. La première partition P(1) fait ainsi la distinction entre les i- et ni-scènes. Les qualités permettent de caractériser l'ensemble des partitions.

Les labels de sources permettent de caractériser les partitions jusqu'à P(4). A ce niveau, la distinction entre les groupes s'opère autour des sons de cloche, des sons d'origine humaine, des sons naturels, des sons de travaux et de trafic, ainsi que des sons d'alarme et de sirène.

On constate une correspondance forte entre les labels et les classes pour les groupes relatifs aux i-scènes. Pour ceux relatifs aux ni-scènes, la correspondance ne tient que jusqu'à P(2).

Les descripteurs structurels présentent des différences significatives si l'on considère deux groupes composés respectivement de i- et ni-scènes. Dans le cas où les deux groupes comprennent des scènes du même type, aucune différence significative n'est relevée. Seul le niveau sonore des événements L(E) semble présenter des différences

pour deux paires de groupes relatifs aux ni-scènes. Le rôle joué par les descripteurs structurels dans la ressemblance inter-scènes paraît ainsi minime.

En conclusion, le processus de catégorisation des scènes semble s'appuyer essentiellement sur l'identification de catégories de concepts concrets (les sources), et l'identification de catégories de concepts abstraits (les qualités). Les concepts concrets permettent de décrire entièrement les groupements relatifs aux i-scènes, mais ne décrivent que partiellement (jusqu'à P(4)) ceux relatifs aux ni-scènes. Les concepts abstraits, eux, caractérisent l'ensemble des partitions.

#### 4.5 CONCLUSION

Les expériences 1 et 2 ont permis de mettre en avant plusieurs points importants sur notre manière de percevoir l'agrément dans un milieu urbain :

- la distinction entre une scène idéale et une scène non-idéale est fonction :
  1. de la plupart des descripteurs structurels, nommément le niveau sonore, la densité de sources, la diversité de sources ;
  2. des descripteurs sémantiques. Dans ce cas, seul un groupe restreint de sources, les marqueurs, suffit à caractériser les deux types d'environnement ;
- aucun des descripteurs structurels ne permet de caractériser l'agrément de l'ensemble des scènes ;
- caractériser l'agrément des i-scènes demande de prendre en compte les contributions spécifiques des i-marqueurs. C'est l'émergence des i-marqueurs qui présente la corrélation la plus élevée avec l'agrément, cette dernière étant positive ;
- caractériser l'agrément des ni-scènes ne requiert pas de considérer séparément les sources. C'est le niveau sonore global qui présente la corrélation la plus élevée avec l'agrément, cette dernière étant négative ;
- la valeur hédonique de la scène est fonction de la présence des marqueurs ;
- la capacité des descripteurs structurels à caractériser l'agrément est fonction de la valeur hédonique de la scène.

Ces résultats nous permettent de conjecturer quant à la nature des représentations mentales des concepts « environnement sonore urbain agréable » (EA) et « environnement sonore urbain désagréable » (ED).

Premièrement, le fait que les informations sémantiques (sources sonores présentes) et structurelles soient différentes pour les i et niscènes nous porte à croire que ces deux types d'informations caractérisent les concepts EA et ED.

Deuxièmement, le fait que la suppression des marqueurs sonores modifie l'agrément perçu nous porte à croire que le concept abstrait lié à l'agrément dépend de l'activation d'un réseau de concepts concrets liés aux sources.

Troisièmement, le fait que les descripteurs structurels corrélés à l'agrément diffèrent en fonction du type de scènes (i ou ni) nous porte à croire que les caractéristiques structurelles permettant de distinguer deux instances d'un même concept abstrait lié à l'agrément ne sont pas les mêmes pour EA et ED.

Nous pensons que le fonctionnement prédictif du système auditive (cf. Section 2.6.6) permet d'apporter une explication simple et intuitive quant aux phénomènes observés.

Lorsqu'il doit se prononcer sur la qualité sonore d'un environnement, le système auditif commence par enregistrer une information globale. Il identifie les sources émettrices et intègre de manière holistique la structure de la scène.

Ces informations permettent d'activer des concepts concrets liés aux sources présentes. L'activation de ces concepts participe alors à la génération d'une image mentale de l'environnement perçu. Cette image, bâtie entre autre à partir des connaissances de l'individu, permet au système 1) de situer globalement la nature hédonique de la scène et 2) d'inférer ses propriétés non perçues.

Le système peut alors adapter le traitement de l'information sensorielle. Cette adaptation a deux conséquences : 1) elle favorise le traitement de certaines sources, 2) elle affine l'information structurelle enregistrée. Plus le système est confiant quant à l'image qu'il se fait de l'environnement perçu (EA et ED), et plus il est à même de rechercher une information dédiée, ce afin de spécifier plus avant son jugement hédonique.

L'expérience 3 a permis, elle, de montrer que la ressemblance entre deux scènes sonores dépend avant tout de leurs compositions sémantiques, et de leurs valeurs affectives intrinsèques. Les caractéristiques structurelles y jouent un rôle secondaire. Ce point ayant été déjà observé, nous pensons qu'il tend à montrer la validité écologique des scènes simulées.

$1_{P(1)}$	$2_{P(1)}$
sonnette vélo $_{1,2,3}^E$ cloche $_{1,2,3}^E$ vélo $_{1,2,3}^E$ voix enfant $_{\bar{1}}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$	travaux $_{\bar{0}}^E$ sirène $_{1,2,3}^E$ klaxon $_{1,2,3}^E$
cour int./parc $_{\bar{0}}^T$ parc $_{\bar{1},2}^T$	travaux $_{\bar{0}}^T$ carrefour $_{\bar{1},2}^T$
$1_{P(2)}$	$2_{P(2)}$
cloche $_{\bar{1},2}^E$ rire homme $_{2,3}^E$ pas $_{\bar{1}}^E$ pas chaussure $_{\bar{3}}^E$	bateau $_{1,2,3}^E$ sonnette vélo $_{1,2,3}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$
cour int./parc $_{\bar{0}}^T$ parc $_{\bar{1}}^T$ cour int. $_{\bar{1}}^T$	travaux $_{\bar{0}}^T$ carrefour $_{\bar{1},2}^T$ carrefour $_{\bar{1},2}^T$
$1_{P(3)}$	$2_{P(3)}$
cloche $_{1,2,3}^E$ rire homme $_{2,3}^E$ pas $_{\bar{1}}^E$ pas chaussure $_{\bar{3}}^E$	sonnette vélo $_{1,2,3}^E$ bateau $_{1,2,3}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$
cour int./parc $_{\bar{0}}^T$ parc $_{\bar{1}}^T$ cour int. $_{\bar{1}}^T$	travaux $_{\bar{0}}^T$
$1_{P(4)}$	$2_{P(4)}$
cloche $_{1,2,3}^E$	pas $_{\bar{1}}^E$ pas talon $_{\bar{2}}^E$ rire homme $_{\bar{2}}^E$
cour int./parc $_{\bar{0}}^T$ foule étrangère $_{\bar{2}}^T$	sonnette vélo $_{1,2,3}^E$ bateau $_{1,2,3}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$
$3_{P(4)}$	$4_{P(4)}$
$1_{P(5)}$	$2_{P(5)}$
cloche $_{1,2,3}^E$	pas $_{\bar{1}}^E$ pas talon $_{\bar{2}}^E$ rire homme $_{\bar{2}}^E$
cour int./parc $_{\bar{0}}^T$ foule étrangère $_{\bar{2}}^T$	sonnette vélo $_{1,2,3}^E$ bateau $_{1,2,3}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$
$3_{P(5)}$	$4_{P(5)}$
$1_{P(6)}$	$2_{P(6)}$
cloche $_{1,2,3}^E$	pas $_{\bar{1}}^E$ pas talon $_{\bar{2}}^E$ rire homme $_{\bar{2}}^E$
cour int./parc $_{\bar{0}}^T$ foule étrangère $_{\bar{2}}^T$	sonnette vélo $_{1,2,3}^E$ bateau $_{1,2,3}^E$ animal $_{\bar{1}}^E$ oiseau $_{\bar{2}}^E$ chant oiseau $_{\bar{3}}^E$
$3_{P(6)}$	$4_{P(6)}$
$5_{P(4)}$	$6_{P(5)}$
$5_{P(5)}$	$6_{P(5)}$
$6_{P(6)}$	$7_{P(6)}$

TABLE 16 : Répartition des classes de sons d'événements ( $E$ ) et de textures ( $T$ ) en fonction des partitions établies par la classification ascendante hiérarchique (cf. Figure 44). Les indices  $x$  indiquent le niveau d'abstraction de la classe considérée.

## DONNÉES SIMULÉES EN ANALYSE AUTOMATIQUE

### 5.1 INTRODUCTION

Ce chapitre présente deux expériences d'évaluation des performances de systèmes de détection automatique d'événements sonores. Ces expériences ont nécessité l'utilisation de corpus de scènes simulées, corpus à partir desquels la capacité de généralisation des algorithmes a pu être testée. Pour les besoins de la cause, dans chacun de ces corpus nous avons fait varier 1) les samples d'événements sélectionnés pour simuler les scènes, 2) les niveaux sonores des événements, 3) leur densité d'apparition.

Ces travaux s'inscrivent dans le cadre des éditions 2013 et 2016 du challenge international de Détection et Classification de Scènes et Événements sonores (DCASE). La première expérience est une réévaluation des systèmes de détection proposés lors la première édition du challenge DCASE (2013)<sup>1</sup>. La seconde fait partie intégrante du challenge DCASE (2016)<sup>2</sup>. Elle en constitue la deuxième tâche<sup>3</sup>.

Le chapitre comprend trois sections. La première présente les deux sessions du challenge DCASE. Elle fait notamment état des métriques utilisées pour évaluer les algorithmes. La deuxième présente l'expérience de réévaluation du challenge DCASE 2013. La troisième présente les résultats de la deuxième tâche du challenge DCASE 2016.

### 5.2 LE CHALLENGE DCASE

#### 5.2.1 Présentation

Le challenge DCASE est un challenge international d'évaluation des performances de systèmes d'analyse automatique des environnements sonores soutenu par les organisations IEEE et AASP<sup>4</sup>. Il comprend quatre tâches :

- classification de scènes environnementales : l'objectif est de regrouper des scènes sonores enregistrées en fonction de leur classe d'appartenance. Cette classe est définie à partir du lieu d'enregistrement des scènes (*e.g.* parc, métro, bureaux, *etc.*) ;

<sup>1</sup> Challenge DCASE 2013 : <http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/>

<sup>2</sup> Challenge DCASE 2016 : <http://www.cs.tut.fi/sgn/arg/dcase2016/>

<sup>3</sup> Deuxième tâche du challenge DCASE 2016 : <http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-synthetic-audio>

<sup>4</sup> IEEE et AASP : <https://signalprocessingsociety.org/get-involved/audio-and-acoustic-signal-processing/aasp-challenges>

- détection d'événements sonores sur un corpus de scènes réelles : l'objectif est de détecter et de classifier des événements sonores sur des scènes réelles enregistrées ;
- détection d'événements sonores sur des scènes simulées : l'objectif est de détecter et de classifier des événements sonores sur des scènes simulées ;
- marquage d'événements audio (uniquement DCASE 2016) : l'objectif est d'identifier les classes d'événements présents sur de courts extraits sonores (4 secondes).

La première édition du challenge DCASE (Giannoulis et al., 2013a,b; Stowell et al., 2015) s'est déroulée en 2013. Elle compte déjà une tâche d'évaluation des algorithmes de détection sur des corpus de scènes simulées. Cependant, cette dernière a été réalisée à partir d'une banque de sons isolés très réduite, extraite des scènes réelles de la première tâche.

La deuxième édition du challenge DCASE a lieu en 2016. La tâche d'évaluation sur corpus simulés des algorithmes de détection est cette fois plus ambitieuse. Elle bénéficie notamment d'une banque de sons isolés dédiée importante (cf. Section 5.3.2.1), banque dont l'enregistrement fait partie intégrante des travaux de la présente thèse. Les processus de simulation utilisés ont par ailleurs été sensiblement améliorés.

### 5.2.2 Évaluer la détection automatique d'événements sonores

L'objectif des systèmes de détection automatique d'événements (*Sound Event Detection* : SED) est de localiser et d'identifier, dans une scène sonore donnée, les événements qui la peuplent.

Par scène sonore, on entend ici un enregistrement d'environnement sonore composé de deux éléments :

- les événements sonores, événements qui occurrent ponctuellement, à différents moments de la scène. Ils sont regroupés en classes (*e.g.* clef, porte, clavier, voix, *etc.*). Ce sont ces classes que les systèmes doivent identifier ;
- le *background*, ou fond sonore, fond qui est présent durant toute la durée de la scène, et qui fait office de perturbateur pour les systèmes, ces derniers ne devant pas le détecter.

Un ensemble de scènes sonores à évaluer est appelé un corpus.

La détection d'événements sonores est une tâche dite supervisée : le système doit d'abord apprendre les caractéristiques de chaque classe, avant d'en détecter les événements. De plus, ces systèmes possèdent tous des paramètres qu'il convient de régler avant de procéder à l'évaluation des algorithmes.

L'évaluation de tels systèmes requiert la constitution de trois corpus :

- corpus d'entraînement : un corpus composé d'enregistrements de sons isolés, représentants toutes les classes à détecter. C'est sur ce corpus que le système va s'entraîner, apprendre les caractéristiques de chaque classe ;
- corpus de développement : un corpus composé de scènes sonores. C'est sur ce corpus que vont être réglés les paramètres des systèmes, afin d'optimiser les performances de détection ;
- corpus d'évaluation : un corpus composé de scènes sonores différentes de celles utilisées pour le corpus de développement. C'est sur ce corpus que vont être évalués les algorithmes, en conservant malgré tout les paramètres obtenus sur le corpus de développement.

Par ailleurs, chaque challenge considère un système de référence, appelé *Baseline*. Cette *Baseline* permet, notamment, aux auteurs des systèmes soumis, de se situer lors de l'étape de développement, par rapport aux performances d'un algorithme générique correspondant à l'état de l'art des connaissances dans le domaine.

### 5.2.3 Métrique

Les performances des algorithmes en SED sont évaluées suivant différentes métriques. Deux d'entre elles sont particulièrement utilisées. Nous les détaillons ici.

La première métrique est la F-mesure (Giannoulis et al., 2013a,b; Stowell et al., 2015), que l'on note F dans ce document. La F-mesure se calcule comme suit :

$$F = \frac{2 \times P \times R}{P + R} \quad (7)$$

où P et R représentent respectivement la précision et le rappel. La précision rend compte du rapport entre le nombre d'événements correctement détectés, c, par le nombre d'événements effectivement détectés par l'algorithme, e, tandis que le rappel rend, lui, compte du rapport entre le nombre d'événements correctement détectés, c, par le nombre d'événements à détecter (le nombre d'événements présents dans la scène), z.

$$P = \frac{c}{e} \quad , \quad R = \frac{c}{z} \quad , \quad e = c + fp \quad , \quad z = c + fn \quad (8)$$

avec fp le nombre de faux positifs, et fn le nombre de faux négatifs.

La deuxième métrique est le taux d'erreurs acoustiques (Poliner and Ellis, 2007; Stiefelhagen et al., 2007), que l'on note ER dans ce document. Ce taux se calcule comme suit :

$$ER = \frac{D + I + S}{N} \quad (9)$$

avec N le nombre d'événements à détecter, D le nombre d'événements manqués (fn), I le nombre d'événements faussement détectés (fp), et S le nombre d'événements substitués, que l'on définit comme  $S = \min[D, I]$ .

F et ER peuvent être calculées de deux manières, suivant que l'on tient compte :

- du nombre de trames correctement identifiées (*sb* : *segment based*);
- du nombre d'événements correctement identifiés (*eb* : *event based*).

Considérer le nombre d'événements plutôt que les trames permet, entre autre, d'obtenir une mesure de performance indépendante de la durée des événements. On estime habituellement qu'un événement est correctement identifié si son *onset* et/ou son *offset* sont correctement identifiés. La détection d'une frontière (*onset - offset*), est toujours considérée avec un seuil de tolérance ( $\pm 100\text{ms}$  et  $\pm 200\text{ms}$  pour les challenges DCASE 2013 et 2016).

Ainsi nous notons  $F_{sb}$  et  $ER_{sb}$ , les F-mesures et taux d'erreurs acoustiques calculés en tenant compte des trames, et  $F_{eb}$  et  $ER_{eb}$  les F-mesures et taux d'erreurs acoustiques calculés en tenant compte des événements.

La détection de l'*offset* d'un événement sonore restant une tâche difficile, aussi bien pour des algorithmes que pour nous, humains, nous ne considérons dans ce document que des mesures de  $F_{eb}$  et  $ER_{eb}$  calculées en fonction du nombre d'événements dont les *onsets* ont été correctement identifiés.

Ces métriques, si elles sont calculées sans précaution, sont susceptibles de donner des valeurs biaisées, dans l'évaluation des scènes, entre les classes bien représentées (beaucoup d'événements), et celles peu représentées (peu d'événements). Afin de parer ce biais, il est possible de calculer les métriques séparément, pour chaque classe, avant de les moyennner. On note ainsi  $F_{cw}$  et  $ER_{cw}$ , les versions alternatives de F et ER normalisées par classe :

$$F_{cw} = \frac{1}{C} \sum_{i=1}^C F^i \quad , \quad ER_{cw} = \frac{1}{C} \sum_{i=1}^C ER^i \quad (10)$$

avec  $C$  le nombre de classes à détecter, et  $F^i$  et  $ER^i$  la F-mesure et le taux d'erreurs acoustiques obtenus par un système en ne considérant que la classe d'événements  $i$ .

Au final, 8 métriques sont donc disponibles pour évaluer les algorithmes en SED, nommément  $F_{sb}$ ,  $F_{eb}$ ,  $F_{cw_{sb}}$ ,  $F_{cw_{eb}}$ ,  $ER_{sb}$ ,  $ER_{eb}$ ,  $ER_{cw_{sb}}$  et  $ER_{cw_{eb}}$ .

## 5.3 APPLICATION AU CHALLENGE DCASE 2013

### 5.3.1 Objectif

L'objectif, dans cette étude, est de ré-évaluer les algorithmes soumis dans le cadre de la tâche 2 de détection d'événements sonores (SED) du challenge DCASE 2013<sup>5</sup> (cf. Section 5.2).

Plus précisément, nous voulons tester la capacité de généralisation de ces algorithmes, *i.e.* leur aptitude à maintenir des performances de détection similaires sur plusieurs corpus de scènes présentant des conditions expérimentales différentes.

La capacité de généralisation est considérée suivant deux angles :

- robustesse à la diversité structurelle : évaluer la capacité de généralisation sur des corpus de scènes composés des mêmes samples, mais dont les caractéristiques structurelles (intensité sonore des samples, positionnement/espacement moyen des samples) diffèrent ;
- robustesse à la diversité des samples : évaluer la capacité de généralisation sur des corpus de scènes possédant les mêmes caractéristiques structurelles (intensité sonore des samples, positionnement des samples), mais composés d'une sélection de samples différents. Par « samples différents », nous entendons des enregistrements d'événements sonores différents, mais appartenant à la même classe (*e.g.* claquement de porte). En effet, quand on considère une tâche de classification, le problème est de savoir si le système évalué est capable de généraliser ses capacités de classification à des données non-observées, mais qui correspondent aux classes considérées dans les corpus d'entraînement et de développement.

La méthode suivie consiste à utiliser le modèle de scènes sonores proposé (cf. Section 3.3.4) afin de générer de nouveaux corpus de scènes simulées, scènes dont nous contrôlons les structures internes, ainsi que la nature des samples utilisés.

---

<sup>5</sup> Par souci de lisibilité, nous ne précisons plus, par la suite, la tâche du challenge DCASE à laquelle nous faisons référence. Le lecteur comprendra que pour la totalité de ce chapitre, l'appellation « challenge DCASE » désigne la tâche 2 de détection d'événements (SED) dudit challenge.

<b>Index</b>	<b>Nom</b>	<b>Description</b>
1	porte-frapper	Fraper à la porte
2	porte-cliquer	Cliqueter la porte
3	parole	Personne prononçant une phrase
4	rire	Personne riant
5	gorge	Personne se raclant la gorge
6	toux	Personne toussant
7	tiroir	Ouverture/fermeture d'un tiroir
8	imprimante	Bruit d'une imprimante
9	clavier	Bruit des touches d'un clavier
10	souris	Bruit d'un clic de souris
11	stylo	Poser un stylo sur une table
12	bouton	Bouton permettant d'allumer la lumière
13	clefs	Poser un jeu de clefs sur une table
14	téléphone	Sonnerie de téléphone
15	alerte	bruit d'une alerte électronique (ordinateur, mobile)
16	page	Tourner une page

TABLE 17 : Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2013.

Afin d'évaluer la robustesse des algorithmes à la présence de nouveaux samples, nous considérons les performances obtenues sur des corpus de scènes simulées avec une banque de données de sons isolés dont les samples (événements et *background*) ont été enregistrés dans des environnements acoustiques différents de ceux du corpus d'évaluation d'origine du challenge DCASE 2013. L'enregistrement de cette nouvelle banque de données a été effectué dans le cadre de cette étude.

Il s'agit alors d'éprouver les algorithmes sur ces nouveaux corpus, et de comparer leurs performances avec celles obtenues sur le corpus d'origine. Les différences nous permettent de conclure quant à la capacité de généralisation des algorithmes considérés.

### 5.3.2 Génération des corpus

Cette partie décrit les différents corpus de scènes simulées utilisés lors de l'expérience.

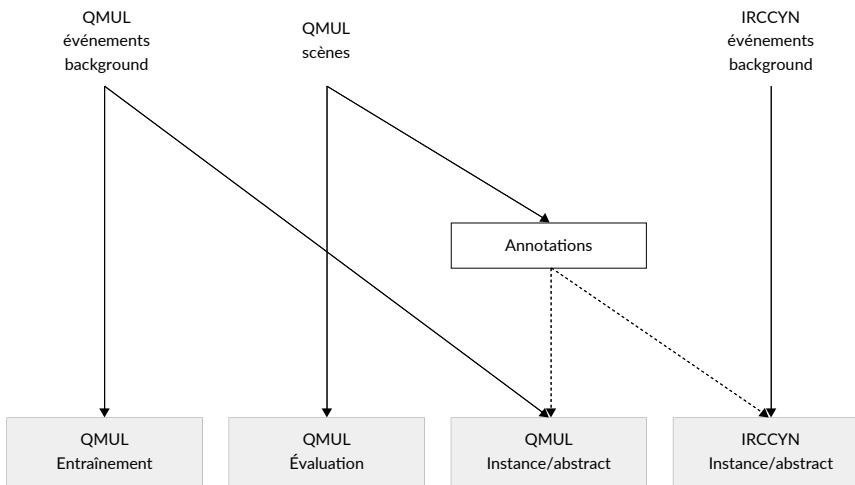


FIGURE 46 : Processus de génération des corpus de scènes simulées utilisés dans l'évaluation du challenge DCASE 2013.

Tous les corpus de scènes simulées sont générés à partir des scènes enregistrées du corpus *test-QMUL*, le corpus de *test* de la tâche de détection d'événements (SED) du challenge DCASE 2013 (Giannoulis et al., 2013b).

*test-QMUL* a été enregistré à l'université *Queen Mary University of London*. Il est composé de 11 enregistrements d'ambiances de bureau, tous d'une durée proche de la minute. Chaque scène est une séquence d'événements sonores non enchevêtrés. Ces événements sont repartis en 16 classes de sons, classes détaillées dans le tableau 17. Les enregistrements ont été effectués dans 5 environnements acoustiques différents. Les scènes sont annotées par deux sujets différents. Pour chaque scène, et à chaque événement entendu, l'annotateur indique la classe de l'événement, son *onset* (la position du début de l'événement) et son *offset* (la position de fin de l'événement). Toutes les annotations sont utilisées. Les 22 couples scène-annotateur permettent de composer une vérité terrain.

À partir des annotations de *test-QMUL*, quatre corpus de scènes simulées sont générés, mettant en œuvre deux banques de données de sons isolés, ainsi que deux processus de simulation distincts. Les banques de données de sons isolés, ainsi que les processus de simulation sont détaillés dans les sections suivantes (cf. Sections 5.3.2.1, 5.3.2.2 et 5.3.2.3).

### 5.3.2.1 Banque de données de sons isolés QMUL et IRCCYN

Deux banques de sons isolés sont utilisées pour générer les scènes isolées. Elles sont respectivement nommées *QMUL* et *IRCCYN*. Toutes deux sont composées de deux types de sons :

- les événements : les enregistrements de sons isolés devant être détectés et identifiés par les algorithmes ;
- les *backgrounds* : les enregistrements de fonds sonores, *i.e.* des scènes amorphes (textures ne possédant pas d'événement saillant, cf. Section 3.2.4) rendant compte de l'environnement acoustique naturel du lieu d'occurrence des événements.

Les sons isolés de la banque *QMUL* sont extraits de scènes enregistrées à l'université *Queen Mary University of London* (*QMUL*), dans le cadre de la préparation du challenge SED DCASE-2013, mais n'ayant pas été utilisées lors de l'évaluation, *i.e.* ne faisant pas partie des corpus d'évaluation (*test-QMUL*) et de développement. Ces sons isolés ont donc été enregistrés dans les mêmes conditions que les scènes du corpus *test-QMUL* (Giannoulis et al., 2013a). Le nombre d'événements par classe varie de 3 à 23. Les enregistrements de *backgrounds* ont été réalisés sur les mêmes environnements acoustiques que ceux utilisés pour le corpus *test-QMUL*, avec, là encore, les mêmes conditions d'enregistrements.

La banque *IRCCYN* est une nouvelle banque de sons isolés, enregistrés à l'Institut de Recherche en Cybernétique de Nantes (*IRCCyN*). Cette banque comprend les mêmes classes que celles présentes dans le corpus *test-QMUL* (cf. Tableau 17). Les enregistrements ont été effectués dans un environnement calme, à l'aide d'un micro canon *AT8035* connecté à un enregistreur *ZOOM H4n*. Chaque classe est composée de 20 événements sonores, ce qui correspond au nombre d'événements disponibles dans le corpus d'entraînement du challenge DCASE-2013 (Giannoulis et al., 2013a,b). Les *background* ont été enregistrés de nuit, dans les bureaux de l'*IRCCyN*, afin qu'ils ne soient pas pollués par des bruits non souhaités.

### 5.3.2.2 Processus de simulation instance

Pour le processus de simulation *instance*, l'objectif est de générer des scènes simulées qui ressemblent le plus possible aux scènes du corpus *test-QMUL*. Cette ressemblance s'entend sous deux aspects :

- *la structure temporelle* : le positionnement temporel en termes d'*onsets* des événements sonores ;
- *les niveaux sonores des événements* : la puissance du ratio entre l'énergie de l'événement et celle du *background*, notée EBR (*event*

*to Background power Ratios*). L'EBR d'un événement de N échantillons est obtenu en calculant le ratio en décibels entre la valeur efficace (niveau RMS, cf. Section 4.2.4) du signal de l'événement ( $E_{rms}$ ), et celle du *background* ( $B_{rms}$ ).

$$EBR = 20 \log_{10} \left( \frac{E_{rms}}{B_{rms}} \right) \quad (11)$$

$$x_{rms} = \sqrt{\frac{1}{N} \sum_{n=1}^N x(n)^2} \quad (12)$$

$x(n)$  peut être remplacé par  $e(n)$  ou  $b(n)$ , respectivement les valeurs des signaux de l'événement et du *background* en volt à l'échantillon  $n$ .

Pour chaque événement, et chaque couple scène-annotateur du corpus *test-QMUL*, nous extrayons les positions d'*onsets* et d'*offsets*, et calculons une approximation de l'EBR. Comme il n'est pas possible d'isoler le signal du *background* des scènes de *test-QMUL*,  $B_{rms}$  est obtenu à partir des périodes dénuées d'événements.

Les positions *onsets* et les EBRs ainsi extraits sont utilisés pour simuler un nouveau corpus de scènes. Pour chaque scène simulée, à chaque *onset* d'une annotation (couple scène-annotateur), nous plaçons un événement de la même classe, choisi aléatoirement parmi la banque de sons isolés (*QMUL* ou *IRCCYN*). Afin de garantir que les samples ne soient pas trop longs, ces derniers sont coupés s'ils dépassent d'au moins 0.5 seconde la durée de l'annotation. Les niveaux des événements des scènes simulées sont fixés par rapport aux EBRs calculés sur les scènes enregistrées.

Le processus de simulation *instance* ne s'appuie donc pas sur le modèle introduit à la section 3.3.4. L'objectif ici est d'obtenir des scènes simulées possédant des samples différents des scènes enregistrées, mais dont les structures temporelles et les EBRs sont aussi proches que possible de ceux des scènes du corpus *test-QMUL*.

### 5.3.2.3 Processus de simulation abstract

L'objectif du processus de simulation *abstract* est de capturer les paramètres haut-niveaux régissant la structure de la scène enregistrée, et de les utiliser afin de régénérer cette dernière. Le processus *abstract* s'appuie sur le modèle introduit à la section 3.3.4. Concrètement, le modèle est instancié suivant des paramètres  $\mu_a^i$ ,  $\sigma_a^i$ ,  $\mu_t^i$  et  $\sigma_t^i$  (cf. Équation. 4 et 5) estimés sur la scène enregistrée. Pour chaque couple scène-annotateur du corpus *test-QMUL*, ces paramètres sont

estimés à partir de l'annotation ( $\mu_t^i$  et  $\sigma_t^i$ ) et du signal ( $\mu_a^i$  et  $\sigma_a^i$ ). Les EBRs et les espacements inter-onsets de la scène simulée sont alors obtenus à partir des distributions normales  $\mathcal{N}(\mu_a^i, \sigma_a^i)$  et  $\mathcal{N}(\mu_t^i, \sigma_t^i)$ . Pour chaque classe, le début et la fin des pistes des scènes simulées sont les mêmes que ceux des scènes enregistrées.

Comme pour le processus de simulation *instance*, les événements sont choisis aléatoirement. Afin de garantir que les durées des événements des scènes simulées ne soient pas trop longues par rapport à celles des scènes enregistrées, la durée D d'un sample d'une classe i est seuillée si :

$$D - \mu_d^i - \sigma_d^i > 5 \quad (13)$$

avec,  $\mu_d^i$  et  $\sigma_d^i$  les moyennes et écarts types des durées des samples appartenant à la classe i pour une annotation donnée. La limite de 5 secondes permet de minimiser l'impact d'une telle opération de seuillage sur les sons impulsifs.

#### 5.3.2.4 Banque de données de scènes simulées

Cinq corpus sont considérés pour l'évaluation (cf. Figure reffig :databasesDCASE2013Simu), à savoir, le corpus de scènes enregistrées *test-QMUL*, et quatre corpus de scènes simulées :

- *instance-QMUL* (insQ) ;
- *abstrait-QMUL* (absQ) ;
- *instance-IRCCYN* (insI) ;
- *abstrait-IRCCYN* (absI).

Les labels « QMUL » et « IRCCYN » font référence aux banques de données de sons isolés utilisées pour générer les scènes simulées. Les labels « instance » et « abstract » désignent, eux, les processus de simulation utilisés.

Afin d'évaluer l'influence du niveau relatif des événements par rapport au *background* sur les performances des algorithmes, le corpus *instance-QMUL* est composé de quatre sous-corpus appelés respectivement *insQ-EBR(6)*, *insQ-EBR(0)*, *insQ-EBR(-6)* et *insQ-EBR(-12)*. Pour *insQ-EBR(0)*, les EBRs estimés sur *test-QMUL* sont préservés. Pour *insQ-EBR(6)*, *insQ-EBR(-6)* et *insQ-EBR(-12)*, des compensations de +6dB, -6dB, -12dB sont ajoutées, lors de la simulation, aux EBRs d'origine. À noter que pour ces sous-corpus, seul l'EBR est modifié. Les positions temporelles des événements, ainsi que les samples sélectionnés, sont strictement identiques entre les quatre sous-corpus.

Pour tous les corpus, (*abstract-QMUL*, *instance-IRCCYN* et *abstract-IRCCYN*), ainsi que les sous-corpus de *instance-QMUL*, une simulation est réalisée pour chaque couple scène-annotateur de *test-QMUL* ( $11 \times 2 = 22$  couples).

De plus, chaque simulation est répliquée 10 fois. A chaque réplication, la sélection aléatoire des samples varie. Pour les corpus générés suivant les processus de simulation *abstract* (*abstract-QMUL* et *abstract-IRCCYN*), les EBRs et espacements *inter-onsets* des samples obtenus à partir des distributions normales  $\mathcal{N}(\mu_a^i, \sigma_a^i)$  et  $\mathcal{N}(\mu_t^i, \sigma_t^i)$  sont également re-tirés d'une réplication à une autre.

Chaque corpus/sous-corpus est ainsi composé de 220 scènes simulées ( $11 \times 2 \times 10$ ). Les corpus ont été simulés à l'aide d'une version alternative de l'outil de simulation *SimScene* (version implémentée dans l'environnement MATLAB TM) développée dans le cadre de cette thèse.

### 5.3.2.5 Analyse du réalisme des scènes simulées

Afin d'évaluer le réalisme des scènes acoustiques simulées, une expérience sensorielle d'analyse sémantique différentielle est conduite.

#### Procédure

22 stimuli doivent être notés, comprenant 11 scènes enregistrées de *test-QMUL* et 11 scènes simulées de *instance-IRCCYN*. Les sujets doivent évaluer le réalisme de chaque scène suivant une échelle graduée de 7 points, allant de 1 (non réaliste) à 7 (très réaliste).

L'ordre de présentation est différent pour chaque sujet. Les sujets doivent écouter la totalité d'une scène avant de se prononcer.

À la fin de l'expérience, les sujets sont invités à commenter librement leurs notations.

#### Dispositif expérimental

L'audio est diffusé en monophonique. Au début de l'expérience, il est demandé aux sujets d'utiliser un casque audio, et de régler le volume sonore à un niveau confortable.

#### Participants

15 sujets ont participé à l'étude. Tous ont réalisé l'expérience avec succès.

#### Résultats

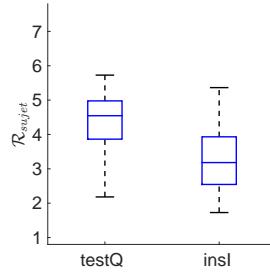


FIGURE 47 : Distributions des notes de réalisme  $\mathcal{R}_{sujet}$  pour les scènes enregistrées *test-QMUL* et les scènes simulées *instance-IRCCYN*.

Nous notons  $\mathcal{R}_{sujet}$  les notes de réalisme par sujet, notes moyennées en considérant séparément les scènes de *test-QMUL* et celles de *instance-IRCCYN*.

Les  $\mathcal{R}_{sujet}$  des scènes enregistrées et simulées sont respectivement de 4.4 et 3.3 (cf. Figure 47). Les deux échantillons présentent une différence significative (t-test appariées :  $p < 0.01$ ). D'après les commentaires des sujets, il semble que les scènes enregistrées n'aient pas été perçues comme très réalistes à cause de leur caractère scripté, les sujets ayant reconnu le fait qu'il s'agissait de scènes jouées.

En ce qui concerne les scènes simulées, les sujets ont rapporté que :

- « le fond sonore semble synthétique/artificiel », bien qu'il ait été enregistré ;
- « certains événements sont coupés ». Ce dernier point est en effet avéré. La coupe de certains événements est due à un choix de conception du corpus *instance-IRCCYN* discuté à la section 5.3.2.2. Ce choix est pris dans le but de minimiser la différence entre la scène simulée, et celle de référence.

Nous notons cependant que plusieurs participants ont donné à des scènes simulées une note de réalisme plus élevée qu'à des scènes naturelles, ce qui tend à montrer que les différences entre les unes et les autres n'ont qu'une importance relative dans la « vérité » acoustique perçue des scènes.

### 5.3.3 Métrique

La métrique considérée dans cette analyse est  $F_{cw_{eb}}$  (cf. Section 5.2.3), *i.e.* la moyenne des F-mesures calculées séparément pour chaque classe, en tenant compte des *onsets* des événements, et avec une fenêtre de tolérance de  $\pm 100\text{ms}$ . Cette métrique a l'avantage :

- d'être facilement interprétable ;
- de ne favoriser aucune classe.

### 5.3.4 Données et analyses

Pour le calcul des métriques, nous suivons la méthodologie adoptée par le challenge DCASE 2013 :

- *test-QMUL* : pour chaque scène, les performances mesurées sont moyennées suivant les deux annotateurs ;
- *instance-QMUL, abstract-QMUL, instance-QMUL* et *abstract-QMUL* : pour chaque scène, les performances sont moyennées :
  1. suivant les réplications (10 réplications par scène) ;
  2. suivant les deux annotateurs.

Nous obtenons ainsi 11 observations (*i.e.* 11 mesures de performances) par condition expérimentale (corpus et systèmes)<sup>6</sup>.

L'analyse s'effectue en deux temps, suivant le corpus de sons isolés considéré :

- corpus de sons isolés *QMUL* : cette analyse considère les trois corpus *test-QMUL, instance-QMUL* et *abstract-QMUL*. Elle a deux objectifs :
  1. évaluer s'il existe des différences significatives entre les performances des algorithmes observées sur les corpus *test-QMUL, insQ-EBR(0)* et *abstract-QMUL*. Il s'agit ici de vérifier que les algorithmes sont capables de généraliser les performances obtenues sur *test-QMUL* pour des corpus simulés avec des sons isolés enregistrés dans les mêmes conditions, et dont les scènes possèdent une structure (EBR et positions des *onsets*) identique (*insQ-EBR(0)*) ou similaire (*abstract-QMUL*). Pour apprécier la significativité des différences observées, nous considérons une ANOVA à mesures répétées comportant 1 facteur intra-sujet (les systèmes) et 1 facteur inter-sujet (les corpus). L'analyse *post hoc* s'effectue suivant une procédure de Tukey-Kramer ;
  2. évaluer s'il existe des différences significatives entre les performances des algorithmes observées sur les corpus *insQ-EBR(6), insQ-EBR(0), insQ-EBR(-6)* et *insQ-EBR(-12)*. Il s'agit ici de vérifier que les algorithmes sont capables de généraliser les performances obtenues sur *insQ-EBR(0)* pour des corpus simulés identiques, mais possédant des niveaux de

---

<sup>6</sup> À noter que la méthode adoptée dans ce document pour intégrer les performances des algorithmes est plus avancée que celle utilisée dans la publication correspondante (Lafay et al., 2016a). Les différences observées sont cependant minimes, et ne changent en rien les résultats et les conclusions.

bruit différents. Pour apprécier la significativité des différences observées, nous considérons une ANOVA à mesures répétées comportant 2 facteurs intra-sujet (les systèmes et les EBR). L'analyse *post hoc* s'effectue suivant une procédure de Tukey-Kramer ;

- corpus de sons isolés IRCCYN : cette analyse considère les trois corpus *test-QMUL*, *instance-IRCCYN* et *abstract-IRCCYN*. Elle a pour objectif de vérifier que les algorithmes sont capables de généraliser les performances obtenues sur *test-QMUL* pour des corpus simulés avec des sons isolés enregistrés dans des conditions différentes, et dont les scènes possèdent une structure (EBR et positions des *onsets*) identique (*instance-IRCCYN*) ou similaire (*abstract-IRCCYN*). Pour apprécier la significativité des différences observées, nous considérons une ANOVA à mesures répétées comportant 1 facteur intra-sujet (les systèmes) et 1 facteur inter-sujet (les corpus). L'analyse *post hoc* s'effectue suivant une procédure de Tukey-Kramer.

Pour les ANOVA à mesures répétées, la sphéricité est évaluée à l'aide d'un test de Mauchly. Si l'hypothèse de sphéricité est violée, la valeur *p* est calculée à l'aide d'une correction de Greenhouse-Geisser (cf. Annexe A.2). Dans ce cas, nous notons  $p_{gg}$  la valeur *p* ainsi corrigée. L'analyse *post hoc* est conduite en suivant la procédure de Tukey-Kramer, celle de Bonferroni étant jugée trop sévère pour le cadre de notre étude. Le seuil de significativité est fixé à  $\alpha = 0.05$  pour toutes les analyses.

Sur le plan méthodologique, il peut être utile de relever que notre analyse statistique, ici appliquée à la détection d'événements, s'éloigne légèrement de celles pratiquées habituellement en apprentissage machine pour des tâches de classification (Demšar, 2006). Dans ces dernières, le corpus est un ensemble d'objets qu'il convient de classer. On a donc une valeur de métrique par corpus, autrement dit, chaque corpus correspond à une observation pour l'analyse statistique. Dans le cas de la détection d'événements sonores, un corpus est un ensemble de scènes, et il convient alors de détecter les événements sur chacune d'elles. On a donc une valeur de métrique par scène<sup>7</sup>, et à chaque scène correspond une observation. Notons enfin que nous faisons le choix ici d'utiliser des méthodes d'analyses paramétriques (ANOVA), mais qu'il est tout à fait possible de considérer leurs pendants non-paramétriques (test de Friedman) (Demšar, 2006).

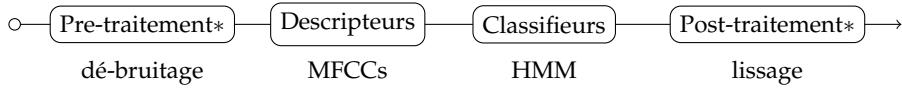


FIGURE 48 : Vision schématisée des systèmes de détection d'événements du challenge DCASE 2013 ; \* indique que le nœud n'est pas systématiquement utilisé ; les choix état de l'art sont donnés en exemple sous les nœuds.

Système	Descripteur	Classifieur	Gestion du bruit	
			réduction	apprentissage
CPS (Chauhan et al., 2013)	fusion	Seuil vraisemblance	(D) (C)	
DHV (Diment et al., 2013a,b)	MFCC	HMM	(D, C)	x
GVV (Gemmeke et al., 2013a,b)	mel	NMF HMM	(D, C) (P)	
NR (Nogueira et al., 2013; Roma et al., 2013)	MFCC	SVM	(C)	x
NVM (Niessen et al., 2013a,b)	fusion	HMM hiérarchique	(C)	
SCS (Schröder et al., 2013a,b)	GF	RF	(C)	
VVK (Gemmeke et al., 2013a; Vugelen et al., 2013)	MFCC	GMM	(D, C)	x
Baseline (Giannoulis:2013a)	CQT	NMF	(D, C)	

TABLE 18 : Description synthétique des systèmes soumis dans le cadre de la tâche 2 de challenge DCASE 2013 ; (D) indique une étape de détection, (C) de classification et (P) de post-traitement.

### 5.3.5 Système de détection

Tous les algorithmes ayant été évalués lors de la tâche 2 (SED) du challenge DCASE 2013 sont considérés dans cette étude (cf. Tableau 18). Un total de 8 algorithmes ont été soumis, auxquels nous rajoutons la *baseline* fournie par les organisateurs du challenge.

La majorité des systèmes suivent la chaîne de traitement illustrée à la figure 48, incluant parfois une étape pré-traitement de dé-bruitage.

Le classifieur de choix est un modèle de Markov caché (HMM : *Hidden markov model*) (Rabiner, 1989) à 2 couches, dont la première modélise l'événement, et la seconde, la transition entre les événements. D'autres classificateurs incluant les forêts d'arbres décisionnels (RF : *Random Forests*), les machines à vecteurs de support (SVM : *Support Vector Machines*), la factorisation en matrices non négatives (NMF : *Non-negative Matrix Factorization*), ainsi que des modèles de mélanges gaussiens (GMM : *Gaussian mixture model*) sont également utilisés. Nous invitons le lecteur à se référer à (Stowell et al., 2015) ou

<sup>7</sup> Formellement, chaque scène peut être considérée comme un corpus d'objets à détecter puis classer.

aux publications indiquées dans le tableau 18 pour une description détaillée des algorithmes.

Au niveau des descripteurs, on distingue 5 groupes :

- *mel* : une représentation temps-fréquence, où l'axe fréquentiel a été projeté sur une échelle de Mel ;
- *CQT* : une représentation temps-fréquence ;
- *MFCC* : une représentation basée sur des coefficients cepstraux calculés sur une échelle de Mel (MFCCs : *Mel-Frequency Cepstral Coefficients*) (Davis and Mermelstein, 1980) ;
- *GF* : une représentation temps-fréquence filtrée par un banc de filtres de Gabor (GF : *Gabor filterbank*) ;
- *fusion* : les algorithmes utilisant simultanément plusieurs descripteurs. NVM et CPS utilisent des jeux de descripteurs allant d'indicateurs scalaires, rendant compte des caractéristiques temporelles (*e.g. flatness*) et fréquentielles (*e.g. loudness*, centroïde spectral) du signal, à des descripteurs multidimensionnels (*e.g. bandes Mel, MFCC*).

Tous les algorithmes sont entraînés et paramétrés sur les corpus d'entraînement et de développement fournis par les organisateurs du challenge DCASE 2013.

### 5.3.6 Résultats

#### 5.3.6.1 Corpus QMUL

Avec la permission des auteurs des différents systèmes proposés (cf. Tableau 18), ces derniers sont testés sur les corpus de scènes simulées, en utilisant les mêmes serveurs de calcul que ceux utilisés pour la tâche 2 (SED) du challenge DCASE 2013. Les systèmes ont, par ailleurs, été re-testés sur le corpus *test-QMUL* (corpus de *test* du challenge SED DCASE 2013), afin de vérifier la réplicabilité des résultats précédemment publiés (Stowell et al., 2015).

Le Tableau 19 affiche les  $F_{cw_{eb}}$  en pourcentages pour les corpus *test-QMUL*, *insQ-EBR(0)* et *abstract-QMUL*. Le système CPS, tel que soumis au challenge DCASE 2013, présente un problème d'implémentation l'empêchant de fonctionner correctement. Ce problème est à l'origine des faibles résultats obtenus pour *test-QMUL*, résultats qui se retrouvent sur *insQ-EBR(0)* et *abstract-QMUL*. Pour ces raisons nous ne considérons pas plus avant ce système.

L'ANOVA montre un effet significatif du corpus ( $F[2, 30] = 6$ ,  $p < 0.01$ ), des systèmes ( $F[6, 180] = 173$ ,  $p_{gg} < 0.01$ ) et de l'interaction ( $F[12, 180] = 10$ ,  $p_{gg} < 0.01$ ). Il semble, au premier abord, que le

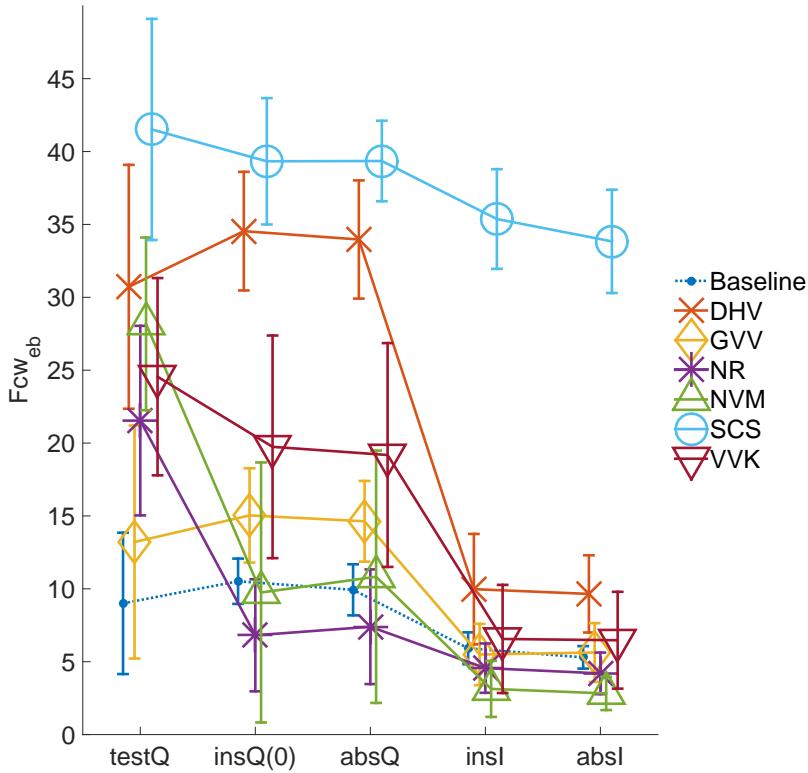


FIGURE 49 : Performances des systèmes évalués dans le cadre du challenge DCASE 2013 sur les corpus QMUL et IRCCYN en considérant  $F_{cweb}$ .

changement de corpus a bien provoqué une modification des performances, et ce bien que les sons utilisés pour simuler les corpus *abstract-QMUL* et *insQ-EBR(0)* soient similaires à ceux que l'on trouve dans *test-QMUL*.

Cependant, l'analyse *post hoc* au niveau des corpus révèle que la *Baseline*, DHV, GVV, SCS et VVK ne présentent pas de différences significatives entre les performances observées pour *test-QMUL* d'une part, et celles relevées pour *abstract-QMUL* et *insQ-EBR(0)* d'autre part. Seuls les résultats de NVM et NR décroissent significativement entre *test-QMUL* et les deux corpus simulés.

Ainsi, exception faite de NR et NVM, les classements des systèmes établis par rapport à leurs performances sont égaux pour les 3 corpus. Ces résultats permettent de conclure deux points quant aux performances de DHV, GVV, SCS et VVK :

- comparaison entre *test-QMUL* et *insQ-EBR(0)* : les performances comparables montrent que les algorithmes sont robustes au changement d'événements. À noter que les samples proviennent

Système	testQ	insQ-EBR(0)	absQ
Baseline	9.0±4.8	10.5±3.0*	9.9±3.5
CPS	0.7±0.8	0.8±1.3	0.8±1.4*
DHV	30.7±8.4	34.5±7.5*	34.0±7.9
GVV	13.2±8.0	15.0±6.4*	14.6±6.2
NR	21.5±6.5*	6.8±5.7	7.4±5.8
NVM	28.2±5.9*	9.7±9.6	10.8±9.9
SCS	41.5±7.6*	39.3±8.2	39.4±8.2
VVK	24.6±6.8*	19.7±8.7	19.2±9.2

TABLE 19 : Résultats mesurés par  $F_{cw_{eb}}$  pour chacun des systèmes évalués dans le cadre du challenge DCASE 2013 en considérant les corpus *test-QMUL*, *insQ-EBR(0)* et *abstract-QMUL*. Les résultats en gras présentent des différences significatives par ligne (procédure de Tukey-Kramer) avec le résultat obtenu pour *test-QMUL*. Le meilleur résultat de la ligne est indiqué par (\*).

tous des enregistrements de QMUL, *i. e.* ont été enregistrés dans les mêmes conditions ;

- comparaison entre *test-QMUL* et *abstract-QMUL* : les performances comparables montrent que les algorithmes sont robustes à un changement de positions temporelles des samples, si les paramètres structuraux des scènes (EBRs et espacements *onsets*) sont conservés.

Nous examinons maintenant les raisons pouvant expliquer les chutes de performances des systèmes NVM et NR dans le cas des scènes simulées. Le phénomène peut être dû soit à l'incapacité des algorithmes à généraliser sur d'autres corpus, soit à un artefact produit par les processus de simulation.

Pour chacun des algorithmes, la première étape consiste à extraire des descripteurs sur l'ensemble des trames du signal. La seconde consiste à classifier ces trames.

Considérons dans un premier temps les descripteurs extraits. Les valeurs minimales et maximales ne varient pas entre *test-QMUL* et les corpus de scènes simulées. Les distributions des valeurs des descripteurs entre les deux types de corpus présentent certes une différence, mais cette dernière se révèle faible et non-significative.

Une inspection des matrices de confusion inter-classes révèle que c'est, pour les deux systèmes, l'étape de classification qui serait responsable de la dégradation des performances. Le tableau 20 affiche, pour tous les systèmes, le plus grand nombre de faux positifs moyennés sur l'ensemble des scènes pour les trois corpus, ainsi que la classe correspondante. Pour NVM et NR, une classe en particulier (NVM :

Système	testQ	insQ-EBR(0)	absQ
Baseline	3.14 (tiroir)	8.63 (tiroir)	7.40 (tiroir)
CPS	2.66 (porte-frapper)	9.04 (porte-claquer)	7.84 (porte-claquer)
DHV	8.44 (tiroir)	6.88 (tiroir)	8.01 (clavier)
GVV	3.08 (page)	3.78 (page)	3.55 (page)
NR	4.33 (clavier)	<b>25.35</b> (porte-claquer)	<b>20.68</b> (porte-claquer)
NVM	1.26 (rire)	<b>22.48</b> (toux)	<b>19.22</b> (toux)
SCS	1.18 (alerte)	2.70 (tiroir)	1.72 (porte-claquer)
VVK	1.81 (alerte)	8.73 (porte-claquer)	8.20 (porte-claquer)

TABLE 20 : Nombre maximum de faux positifs pour chaque système évalué et pour chaque corpus. Les résultats sont moyennés suivant les enregistrements. Les classes de sons correspondantes sont indiquées entre parenthèses.

toux, NR : porte-claquer) semble être détectée de manière abusive, augmentant drastiquement le nombre de faux positifs, et diminuant *de facto* les résultats.

Nous concluons que, pour ces deux systèmes, la diminution des performances n'est probablement pas un artefact dû au processus de simulation, mais plutôt à un phénomène de sur-apprentissage de l'étape de classification. Considérant que ces systèmes sont les seuls à faire usage d'une approche de classification discriminative (NR : SVMs; NVM : RFs), nous conjecturons que le cadre d'entraînement proposé par le challenge DCASE, et notamment le faible nombre de samples disponibles pour l'apprentissage (20 par classe), n'est pas adapté pour ces deux algorithmes. Le processus de simulation peut ainsi être considéré comme sain.

### 5.3.6.2 Corpus instance-QMUL en considérant différents niveaux de bruit

Nous considérons maintenant l'influence de l'EBR sur les performances des algorithmes. Les résultats obtenus pour les corpus *insQ-EBR*(−12), *insQ-EBR*(−6), *insQ-EBR*(0) et *insQ-EBR*(6) (cf. Section 5.3.2.4) sont présentés sur la figure 50.

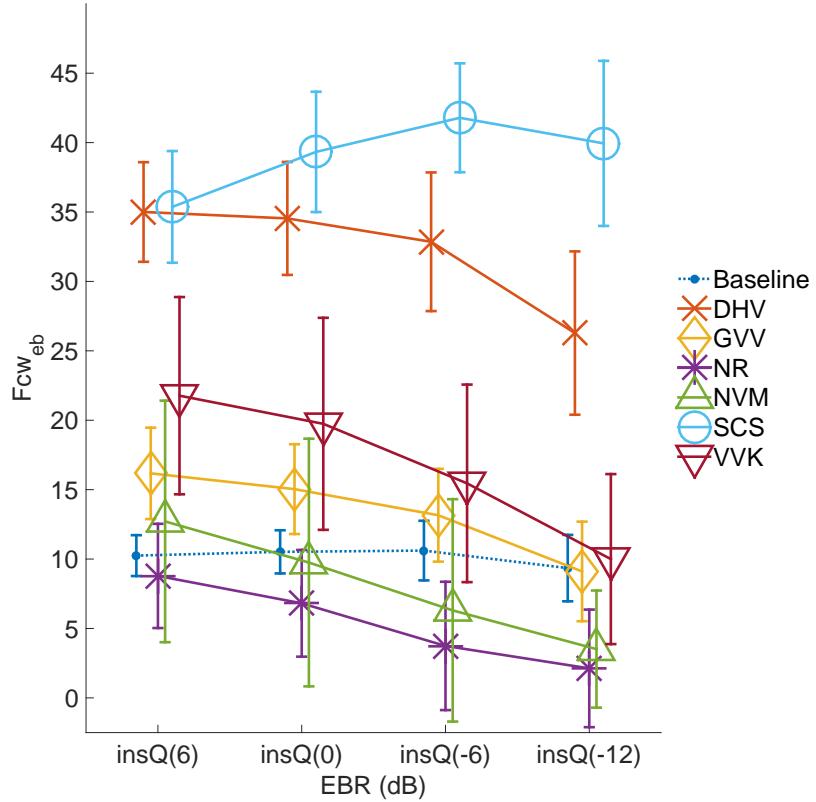


FIGURE 50 : Performances des systèmes évalués dans le cadre du challenge DCASE 2013 sur les corpus *instance-QMUL* simulés avec différents EBR (6, 0, -6 et -12dB).

L'ANOVA montre un effet significatif de l'EBR ( $F[3, 30] = 63, p_{gg} < 0.01$ ), des systèmes ( $F[6, 60] = 128, p_{gg} < 0.01$ ) et de l'interaction ( $F[18, 180] = 16, p_{gg} < 0.01$ ).

Concernant l'analyse *post hoc* sur l'EBR, tous les systèmes affichent une dégradation de performances significative lorsque l'on passe d'un EBR de 0dB à un EBR de -12dB, ainsi qu'une amélioration significative lorsque l'on passe d'un EBR de 0dB à un EBR de +6dB (excepté DHV pour lequel l'amélioration n'est pas significative). Ainsi, et sans surprise, plus l'EBR est faible, et plus les performances diminuent. Par ailleurs, plus l'EBR est faible, et plus les écarts entre les algorithmes se réduisent. Le seul système qui ne suit pas cette tendance est SCS, qui maintient des performances stables pour les différents EBRs, et améliore même significativement ces dernières pour des EBRs allant de 6 à -6dB. Ces résultats montrent l'efficacité de l'étape de dé-bruitage dont bénéficie SCS, un pré-traitement central de cet algorithme (Schröder et al., 2013a).

Système	testQ	insI	absI
Baseline	$9.0 \pm 4.8^*$	<b><math>5.9 \pm 2.9</math></b>	<b><math>5.6 \pm 2.9</math></b>
DHV	$30.7 \pm 8.4^*$	<b><math>10.0 \pm 5.8</math></b>	<b><math>9.5 \pm 5.6</math></b>
GVV	$13.2 \pm 8.0^*$	<b><math>5.6 \pm 3.7</math></b>	<b><math>5.5 \pm 3.6</math></b>
NR	$21.5 \pm 6.5^*$	<b><math>4.6 \pm 3.4</math></b>	<b><math>5.4 \pm 4.5</math></b>
NVM	$28.2 \pm 5.9^*$	<b><math>3.1 \pm 3.1</math></b>	<b><math>3.2 \pm 3.0</math></b>
SCS	$41.5 \pm 7.6^*$	<b><math>35.4 \pm 7.2</math></b>	<b><math>34.0 \pm 6.7</math></b>
VVK	$24.6 \pm 6.8^*$	<b><math>6.6 \pm 5.7</math></b>	<b><math>7.3 \pm 6.3</math></b>

TABLE 21 : Résultats mesurés par  $\text{Fcw}_{\text{eb}}$  pour chacun des systèmes évalués dans le cadre du challenge DCASE 2013 en considérant les corpus *test-QMUL*, *instance-IRCCYN* et *abstract-IRCCYN*. Les résultats en gras présentent des différences significatives par ligne (procédure de Tukey-Kramer) avec le résultat obtenu pour *test-QMUL*. Le meilleur résultat de la ligne est indiqué par (\*).

L'influence de l'EBR est cependant cohérente, le classement en terme de performances entre les algorithmes étant maintenu pour les différents corpus.

Concernant l'analyse *post hoc* sur les systèmes, seuls DHV et SCS présentent des performances significativement supérieures à celles de la *baseline*, et ce quel que soit l'EBR considéré. VVK et GVV arrivent à surpasser la *baseline* uniquement pour des EBR de 0dB et 6dB, *i.e.* des niveaux de bruit faibles. Concernant NR et NVM, ces systèmes n'améliorent jamais les résultats de la *baseline*, et affichent même des performances significativement inférieures à celle-ci pour des EBR de -12 (NR et NVM) et de -6dB (NR), montrant ainsi leur faible capacité de généralisation pour des niveaux de bruit élevés.

Ces résultats nous amènent à conclure que, à part SCS, aucun des systèmes considérés n'est robuste à la variation du niveau de bruit.

### 5.3.6.3 Corpus IRCCYN

Concernant l'évaluation menée sur les corpus simulés utilisant la nouvelle banque de sons isolés enregistrés à l'IRCCyN, les résultats sont affichés sur le tableau 21 et la figure 49. L'ANOVA montre un effet significatif du corpus ( $F[2,30] = 89$ ,  $p < 0.01$ ), des systèmes ( $F[6,180] = 249$ ,  $p_{gg} < 0.01$ ) et de l'interaction ( $F[12,180] = 17$ ,  $p_{gg} < 0.01$ ).

Alors que la plupart des systèmes ont obtenu des performances comparables entre *test-QMUL* et les corpus *abstract-* et *instance-QMUL*, l'analyse *post hoc* relative aux corpus montre que tous les algorithmes voient leurs résultats diminuer de manière significative pour les corpus *abstract-* et *instance-IRCCYN*.

De plus, l'analyse *post hoc* des systèmes révèle que, à l'exception du système SCS, tous les systèmes ont des résultats équivalents ou significativement inférieurs (NVM) à ceux de la *baseline* pour les deux corpus *IRCCYN*. C'est en particulier le cas du système DHV, qui, pourtant, montre de bons résultats pour les corpus *QMUL*.

L'ensemble de ces résultats nous permet de conclure que, pour les systèmes DHV, GVV, NR, NVM et VVK, le gain de performance observé sur le corpus *test-QMUL*, par rapport à la baseline, n'est dû qu'à une sur-adaptation des systèmes aux données d'entraînement (corpus d'entraînement et de développement du challenge DCASE 2013).

Comme on peut le voir sur la figure 49, seul le système SCS (gagnant du challenge SED DCASE 2013) arrive à maintenir des performances stables entre tous les corpus considérés. Cette capacité de généralisation est par ailleurs cohérente, le système parvenant en effet à généraliser quelle que soit la condition expérimentale que l'on fait varier :

- les samples sélectionnés (en considérant deux banques de sons isolés différentes) ;
- les positions temporelles des samples ;
- les EBRs.

### 5.3.7 Discussion

En résumé, l'utilisation des scènes simulées à partir du modèle de scènes sonores proposé nous a permis de :

1. reproduire le classement des systèmes dans les mêmes conditions d'enregistrement pour 5 d'entre eux. Les deux systèmes posant problème (NR et NVM) présentent des performances dégradées. Nous montrons que cette dégradation est probablement due à un sur apprentissage de leurs classifieurs discriminants respectifs ;
2. évaluer les capacités de généralisation des systèmes dans de nouvelles conditions d'enregistrement. A cet égard, le système SCS est le seul à généraliser correctement ;
3. évaluer la robustesse des systèmes devant traiter des niveaux de bruits de fond différents. Une nouvelle fois, le système SCS est le seul à présenter des performances stables pour les différents EBR considérés, et ce, probablement en raison d'une étape efficace de pré-traitement du bruit.

L'ensemble de ces résultats montre la pertinence des processus de simulation proposés (cf. Sections 5.3.2.3 et 5.3.2.2), ces derniers per-

mettant et de répliquer, et d'aller plus loin dans l'analyse des performances des algorithmes.

## 5.4 APPLICATION AU CHALLENGE DCASE 2016

### 5.4.1 *Objectifs*

Nous présentons dans cette partie les résultats de la tâche 2 du challenge DCASE 2016, nommée « détection d'événements sonores dans des environnements simulés ». Cette tâche 2 a été réalisée dans le cadre de notre thèse.

L'objectif est ici d'évaluer les performances d'algorithmes en SED sur des corpus de scènes simulées, scènes dont nous contrôlons :

- l'intensité des événements sonores ;
- le nombre d'événements sonores par scène.

Par ailleurs nous faisons la distinction entre deux types de scènes, à savoir :

- les scènes autorisant le recouvrement entre les événements de classes différentes ;
- les scènes n'autorisant pas le recouvrement.

### 5.4.2 *Génération des corpus*

#### 5.4.2.1 *Banque de sons isolés*

La banque de données de sons isolés IRCCYN (cf. Section 5.3.2.1) a été utilisée pour simuler les scènes. 11 classes d'événements sont considérées dans le cadre de cette tâche. Les classes sont décrites dans le tableau 22. Comparé au challenge DCASE 2013, 5 classes ont été supprimées :

- alerte : la classe a été supprimée en raison de sa définition trop « vague ». En effet la diversité des sons pouvant appartenir à la classe alerte électronique est importante ;
- bouton, souris, stylo : ces classes ont été supprimées après analyse des résultats du challenge DCASE 2013. En effet, lors de ce dernier, ces classes :
  1. ont souvent été confondues entre elles ;
  2. ont souvent été mal détectées.
- imprimante : cette classe a été supprimée en raison de son aspect singulier par rapport aux autres classes. En effet, la classe

Index	Nom	Description
1	porte-frapper	Fraper à la porte
2	porte-cliquer	Cliqueter la porte
3	parole	Personne prononçant une phrase
4	rire	Personne riant
5	gorge	Personne se raclant la gorge
6	toux	Personne toussant
7	tiroir	Ouverture/fermeture d'un tiroir
8	clavier	Bruit des touches d'un clavier
9	clefs	Poser un jeu de clefs sur une table
10	téléphone	Sonnerie de téléphone
11	page	Tourner une page

TABLE 22 : Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2016.

imprimante est composée de sons significativement plus longs que ceux des autres classes. Un tel déséquilibre rend difficile un contrôle équitable du nombre d'événements par classe, pour chaque scène, particulièrement dans les cas où le recouvrement entre les événements est interdit.

#### 5.4.2.2 *Simulation des scènes sonores*

Deux paramètres sont considérés pour contrôler la simulation des scènes sonores :

- EBR : le rapport moyen entre les niveaux des événements et du *background* (cf. Section 5.3.2.2);
- nec : le nombre d'événements présents pour chaque classe.

Contrairement à ce qui s'était fait pour les scènes simulées du Challenge DCASE 2013, nous ne considérons plus l'espacement moyen entre les *onsets* des événements pour contrôler la densité d'événements présents, mais directement le nombre d'événements (nec). De fait, nous garantissons que chaque classe est représentée par le même nombre d'événements, quelle que soit la durée de ces derniers.

Par ailleurs, les EBRs des événements sont constants, *i.e.* fixer un EBR de  $-6\text{dB}$  pour une scène revient à fixer un EBR de  $-6\text{dB}$  pour chaque événement de cette scène.

Enfin, deux types de scènes sont considérés. Les scènes polyphoniques, scènes où les événements de différentes classes peuvent se recouvrir temporellement, et les scènes non-polyphoniques, scènes où un seul événement peut être actif à un moment donné.

Pour chaque scène, la position des *onsets* des événements est tirée aléatoirement, suivant une distribution uniforme. Dans le cas des scènes non-polyphoniques, une étape de post-traitement assure qu'aucun recouvrement n'ait lieu entre les événements.

Nous considérons trois niveaux pour les paramètres EBR et nec. Les valeurs de ces niveaux dépendent de la nature polyphonique de la scène :

- EBR : -6, 0 et +6dB ;
- nec :
  - non-polyphonique : 1, 2 et 3 ;
  - polyphonique : 3, 4 et 5.

L'ensemble des valeurs des paramètres nous donne 18 conditions expérimentales ( $3 \text{ EBR} \times 3 \text{ nec} \times 2 \text{ polyphonie}$ ). À noter cependant que, comme pour les scènes simulées du challenge DCASE 2013, une modification de l'EBR n'affecte que ce dernier, *i.e.* les positions des *onsets* et les samples sélectionnés ne sont pas modifiés.

#### 5.4.2.3 Banque d'entraînement

La banque d'entraînement comprend 220 sons isolés, 20 pour chacune des classes considérées.

#### 5.4.2.4 Corpus de développement

Les scènes du corpus de développement sont simulées à partir des sons isolés d'événements de la banque d'entraînement, auxquels nous ajoutons 1 son de *Background*.

Nous simulons une scène pour chacune des 18 conditions expérimentales définies par les paramètres de contrôle (cf. Section 5.4.2.2), nous dotant ainsi d'un corpus de 18 scènes sonores simulées.

Concernant la sélection des samples d'événements, ces derniers sont différents pour chaque valeur de nec et de polyphonie. Autrement dit, un changement de nec, ou de nature polyphonique, modifie l'ensemble des samples de la scène.

Un même sample de *background* est utilisé pour simuler l'ensemble des scènes de la banque de développement.

#### 5.4.2.5 Corpus d'évaluation

Les scènes du corpus d'évaluation sont simulées à partir d'une banque de sons isolés composée de 440 sons d'événements, 40 pour chacune des classes considérées, et 3 sons de *background*.

Pour chacune des 18 conditions expérimentales définies par les paramètres de contrôle (cf. Section 5.4.2.2), la simulation est répliquée

trois fois, nous donnant ainsi un corpus de 54 scènes sonores simulées ( $18 \times 3$ ). Pour chaque réplication, la position des *onsets* diffère aléatoirement. Les graines des générateurs aléatoires sont différentes de celles employées pour simuler le corpus de développement.

Concernant les samples d'événements sélectionnés, ces derniers sont différents pour chaque valeur de *nec*, de polyphonie et chaque réplication.

Concernant les samples de *background*, ces derniers sont différents pour chaque réplication.

Les scènes ont toutes une durée de 2 minutes. La durée totale du corpus est ainsi de 108 minutes.

#### 5.4.3 Métrique

Parmi les 4 métriques utilisées dans le cadre du challenge DCASE 2016 (cf. Section 5.2.3), nous en choisissons une, la même que celle utilisée lors de l'analyse relative au challenge DCASE 2013 (cf. Section 5.3.3) :

- $F_{CwEB}$  : la F-mesure, calculée en prenant en compte les *onsets* des événements, et en normalisant les résultats par classe.

L'identification des *onsets* des événements est effectuée avec une fenêtre de tolérance de  $\pm 200\text{ms}$ .

#### 5.4.4 Données et analyses

Les métriques sont calculées séparément sur chacune des scènes du corpus d'évaluation, et moyennées en fonction des conditions expérimentales considérées. De ce fait, nous nous éloignons de l'évaluation officielle réalisée pour le challenge 2016, où le calcul des métriques est effectué sur l'ensemble des scènes, *i.e.* en considérant que toutes les scènes n'en forment qu'une.

L'analyse s'effectue en trois temps :

- dans un premier temps, nous considérons les résultats sans tenir compte des différentes conditions expérimentales (*nec* et EBR). Il s'agit ici d'apprécier les performances globales des algorithmes. Les différences entre les systèmes sont appréciées à l'aide d'une ANOVA à mesures répétées (cf. Annexe A.2) à 1 facteur (les différents systèmes) ;
- dans un second temps, nous considérons les résultats entre les scènes polyphoniques et les scènes non-polyphoniques, sans toutefois prendre en compte *nec* et EBR. Les différences entre les systèmes sont appréciées à l'aide d'une ANOVA à mesures répétées (cf. Annexe A.2) à 1 facteur intra-sujet (*within subject*) :

les différents systèmes) et 1 facteur inter-sujet (*between subject* : la polyphonie) ;

- dans un troisième temps, nous évaluons l'impact des différentes conditions expérimentales (nec, EBR) sur les performances des algorithmes, en considérant séparément les scènes polyphoniques, et les scènes non-polyphoniques. Pour nec, les différences entre les systèmes sont appréciées à l'aide d'une ANOVA à mesures répétées (cf. Annexe A.2) à 1 facteur intra-sujet (*within subject* : les différents systèmes) et 1 facteur inter-sujet (*between subject* : nec). Pour EBR cependant, les différences sont appréciées à l'aide d'une ANOVA à mesures répétées à 2 facteurs intra-sujets (les différents systèmes et EBR). En effet, les samples et les positions des *onsets* n'étant pas modifiés lors d'un changement d'EBR, il existe clairement une relation de dépendance entre les différents niveaux d'EBR. Nous n'évaluons jamais les deux conditions expérimentales nec et EBR en même temps, les observations disponibles (au nombre de trois) n'étant pas jugées suffisantes.

Pour les ANOVA à mesures répétées, la sphéricité est évaluée à l'aide d'un test de Mauchly. Si l'hypothèse de sphéricité est violée, la valeur p est calculée à l'aide d'une correction de Greenhouse-Geisser (cf. Annexe A.2). Dans ce cas, nous notons  $p_{gg}$  la valeur p ainsi corrigée. L'analyse *post hoc* est conduite en suivant la procédure de Tukey-Kramer. Un seuil de significativité de  $\alpha = 0.05$  est choisi.

#### 5.4.5 Systèmes de détection

Cette partie décrit les systèmes soumis à la tâche 2 du challenge DCASE 2013. 10 algorithmes sont proposés, auxquels nous rajoutons la *baseline*. Une description synthétique de ces systèmes est donnée au tableau 23.

Concernant les descripteurs, on peut regrouper les algorithmes en 4 groupes :

- *mel/bark* : une représentation temps-fréquence, où l'axe fréquentiel a été projeté sur une échelle particulière, soit de Bark (*Pikrakis*) soit de Mel (*Choi, Hayashi 1, Hayashi 2, Giannoulis et Kong*) ;
- *VQT/CQT* : une représentation temps-fréquence ;
- *MFCC* : une représentation basée sur des coefficients cepstraux calculés sur échelle de Mel (MFCCs : *Mel-Frequency Cepstral Coefficients*) ;
- *GTCC* : une représentation basée sur des coefficients cepstraux calculés sur échelle de Gammatone (GTCCs : *Gammatone cepstral coefficients*) .

Système	Descripteur	Classifieur	Gestion du bruit	
			réduction	apprentissage
<i>Komatsu</i> (Komatsu et al., 2016)	VQT	NMF-MLD		x
<i>Choi</i> (Choi et al., 2016)	mel	DNN	x	x
<i>Hayashi 1</i> (Hayashi et al., 2016)	mel	BLSTM-PP		x
<i>Hayashi 2</i> (Hayashi et al., 2016)	mel	BLSTM-HMM		x
<i>Phan</i> (Phan et al., 2016)	GTCC	RF		x
<i>Giannoulis</i> (Giannoulis et al., 2016)	mel	CNMF		x
<i>Pikrakis</i> (Pikrakis and Kopsinis, 2016)	Bark	Template matching	x	
<i>Vu</i> (Vu and Wang, 2016)	CQT	RNN		
<i>Gutierrez</i> (Gutierrez-Arriola et al., 2016)	MFCC	Knn		x
<i>Kong</i> (Kong et al., 2016)	mel	DNN		
<i>Baseline</i> (Benetos et al., 2016a)	VQT	NMF		

TABLE 23 : Description synthétique des systèmes soumis dans le cadre de la tâche 2 du challenge DCASE 2016.

Les classifieurs utilisés sont variés, allant du plus classique (KNN : plus proches voisins), au plus récent (DNN : réseaux de neurones profonds).

#### 5.4.6 Résultats

##### 5.4.6.1 Analyse globale

Les résultats globaux sont affichés sur la figure 51. L'ANOVA pratiquée sur  $F_{cw_{eb}}$  révèle un effet positif du type de système ( $F[10, 530] = 466$ ,  $p_{gg} < 0.01$ ). L'analyse *post hoc* nous permet d'isoler 4 groupes de systèmes, les systèmes d'un même groupe ne présentant pas de différences significatives dans leurs résultats :

1. *Komatsu, Hayashi 1, Hayashi 2 et Choi* : les performances moyennes allant de 67% (*choi*) à 71% (*Komatsu*) ;
2. *Phan, Giannoulis et Pikrakis* : les performances moyennes allant de 34% (*Pikrakis*) à 36% (*Phan*) ;

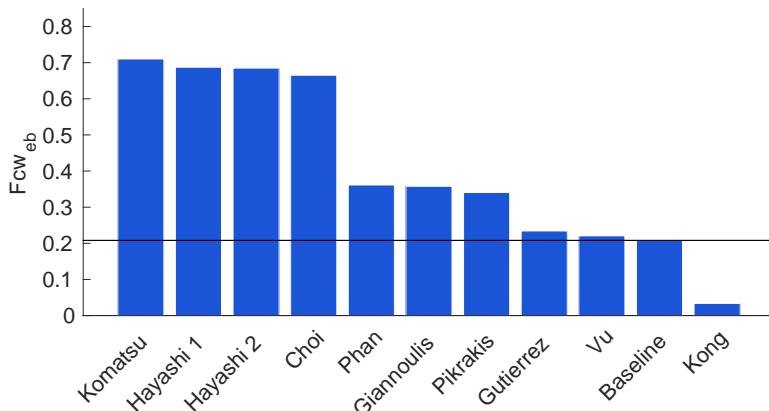


FIGURE 51 : Performances globales des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique  $F_{CW_{eb}}$ .

3. *Baseline*, *Vu* et *Gutierrez* : les performances allant de 21% (*Baseline*) à 23% (*Gutierrez*) ;
4. *Kong* : la performance moyenne étant de 22%.

Ainsi sur les 10 systèmes soumis, 7 parviennent à surpasser les résultats présentés par la *Baseline*, les systèmes du groupe 2 affichant une amélioration d'environ 15%, ceux du groupe 1 une amélioration de près de 45%.

Il est difficile de dégager l'influence d'un classifieur particulier, ceux utilisés par le groupe 1 étant variés. Pour les descripteurs cependant, 3 systèmes sur 4 du groupe 1 utilisent des bandes de Mel. Notons également l'importance, pour le système, de considérer le bruit (*background*), soit en le modélisant, soit en le réduisant au niveau des données à évaluer. En effet, les trois systèmes n'ayant pas tenu compte de l'influence du bruit présentent les trois performances les plus faibles.

Le système affichant les résultats les moins bons est *Kong*. Ce dernier est le seul à présenter des résultats systématiquement en deçà de la *baseline*. Une explication possible de ces faibles performances : la phase d'apprentissage du DNN utilisé (Kong et al., 2016). En effet, l'entraînement d'un tel classifieur nécessite un grand nombre de données afin d'être robuste, *i.e.* capable de généraliser. La banque d'entraînement proposée dans le cadre de cette tâche n'est pas dimensionnée à cet effet.

L'autre système faisant usage d'un DNN (*Choi*) applique, lui, une étape d'augmentation de données, visant à augmenter artificiellement le nombre d'items sur lesquels entraîner l'algorithme (Choi et al., 2016) (ce qui manque dans l'apprentissage de *Kong*). Nous ne considérons pas plus avant les résultats de *Kong* dans la suite de l'analyse.

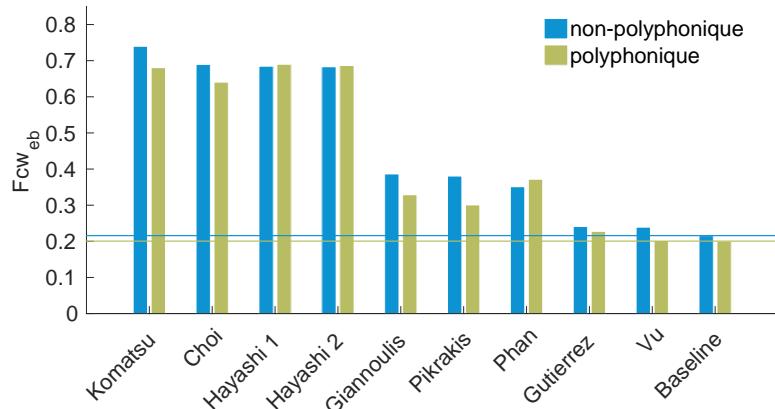


FIGURE 52 : Influence de la polyphonie sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique FcW<sub>eb</sub>.

#### 5.4.6.2 Influence de la polyphonie

Les résultats par type de scènes (polyphoniques et non-polyphoniques) sont affichés sur la figure 52. L’ANOVA appliquée à FcW<sub>eb</sub> révèle un effet significatif du type de système ( $F[9,468] = 358, p_{gg} < 0.01$ ), mais pas de la polyphonie ( $F[1,52] = 3.5, p = 0.07$ ). Un effet d’interaction est néanmoins constaté ( $F[10,520] = 2.5, p_{gg} < 0.05$ ).

Ainsi, la qualité polyphonique des scènes n’a pas significativement affecté les performances des algorithmes, ces derniers étant globalement capables de gérer de manière équivalente les deux types de scènes. L’analyse *post hoc* sur le facteur polyphonique nous indique que sur les 10 systèmes considérés, 4 affichent des performances différentes, suivant le caractère polyphonique des scènes, nommément *choi*, *Giannoulis*, *Komatsu* et *Pikrakis*. Pour ces 4 systèmes, le passage au polyphonique dégrade les performances, constat qui était déjà suggéré par l’effet significatif de l’interaction dans l’ANOVA.

L’analyse *post hoc* sur le facteur système nous permet d’isoler les trois mêmes groupes d’algorithmes (le groupe de *Kong* ayant été écarté) que ceux relevés en considérant les résultats globaux (cf. Section 5.4.6.1), s’agissant des scènes polyphoniques, ou s’agissant des scènes non polyphoniques. Un seul écart est néanmoins noté au niveau des scènes polyphoniques : les systèmes *Gutierrez* et *Pikrakis* ne présentant plus de différences significatives.

#### 5.4.6.3 Influence du niveau de bruit

Considérant les scènes non-polyphoniques, les résultats sont affichés sur la Figure 53a. l’ANOVA révèle un effet significatif du type de système ( $F[9,72] = 80, p_{gg} < 0.01$ ), et de l’EBR ( $F[2,16] = 164, p_{gg} < 0.01$ ), ainsi que de l’interaction ( $F[18,144] = 6.5, p_{gg} < 0.01$ ). Ainsi

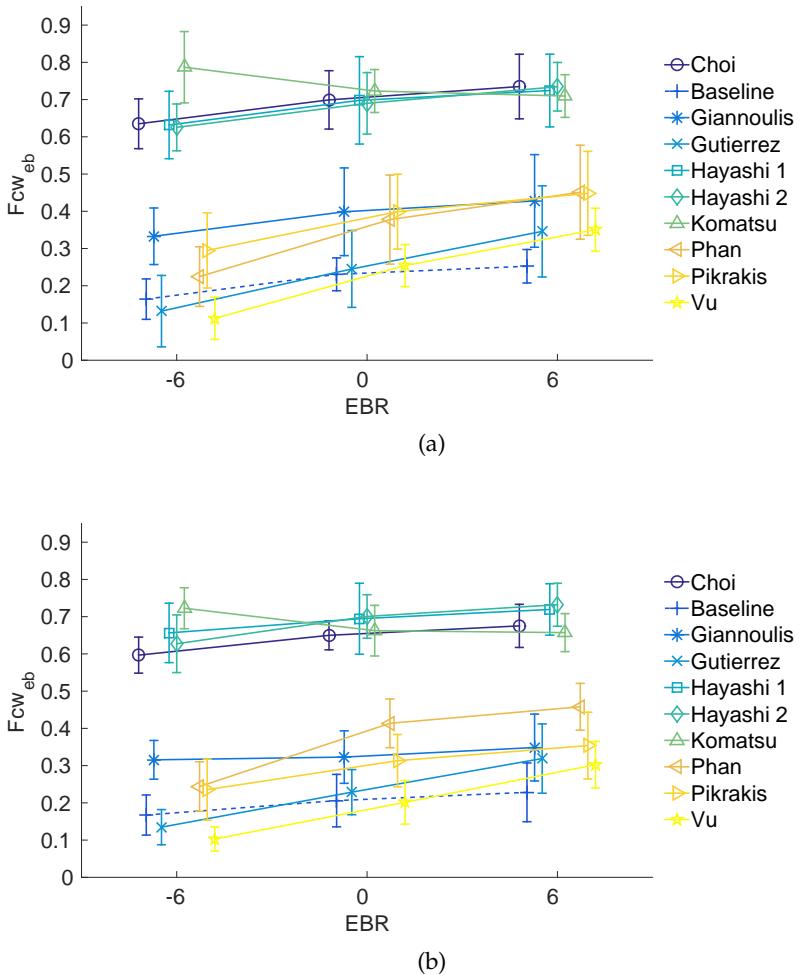


FIGURE 53 : Influence du niveau de bruit (EBR) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique  $F_{CWEb}$ ; (a) scènes non-polyphoniques, (b) scènes polyphoniques.

plus l'EBR est élevé, plus les performances augmentent, et ce, globalement, pour tous les systèmes.

Concernant l'analyse *post hoc*, nous observons si les systèmes présentent des différences significatives avec la baseline. Pour un EBR de  $-6\text{dB}$ , 4 groupes émergent :

1. *Komatsu* : performances supérieures à celles de la *Baseline* ;
2. *choi*, *Hayashi 1* et *Hayashi 2* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
3. *Gianoulis* : performances supérieures à celles de la *Baseline*, mais inférieures à celles du groupe 2 ;
4. *Gutierrez*, *Pikrakis*, *Phan* et *Vu* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

Pour un EBR de 0dB, trois groupes sont isolés :

1. *Komatsu, choi, Hayashi 1 et Hayashi 2* : performances supérieures à celles de la *Baseline* ;
2. *Pikrakis et Phan* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
3. *Gutierrez, Vu et Gianoulis* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

Enfin, pour un +6dB, seulement trois groupes émergent :

1. *Komatsu, choi, Hayashi 1 et Hayashi 2* : performances supérieures à celles de la *Baseline* ;
2. *Gianoulis, Pikrakis et Vu* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
3. *Gutierrez et Phan* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

S'agissant des scènes non polyphoniques, il apparaît que le système *Komatsu* permet d'obtenir les meilleures performances, notamment dans les situations de niveau de bruit élevé (EBR = -6dB). À noter que seul cet algorithme voit ses performances décroître avec l'EBR, ceci étant dû, sans doute, à l'attention particulière portée par ses auteurs à la modélisation du *Background*.

Les systèmes *choi, Hayashi 1 et Hayashi 2* égalent les performances de *Komatsu* pour des EBR de 0 et +6dB. Ils dépassent systématiquement celles des autres systèmes. Cependant, l'augmentation du niveau de bruit (6dB → -6dB) provoque, pour les trois, une chute de performances d'environ 10%.

Concernant les autres systèmes évalués, *Vu, Gianoulis, Pikrakis et Phan* surpassent la *Baseline* pour certains EBR seulement. Tous ces systèmes semblent souffrir du niveau de bruit, leurs performances diminuant sensiblement avec ce dernier, de -10 à -20% entre un EBR de 6dB et un de -6dB. Seul *Gutierrez* reste systématiquement au même niveau que la *Baseline*.

Considérant les scènes polyphoniques, les résultats sont affichés sur la Figure 53b. l'ANOVA révèle un effet significatif du type de système ( $F[9, 72] = 113, p_{gg} < 0.01$ ), et de l'EBR ( $F[2, 16] = 127, p_{gg} < 0.01$ ), ainsi que de l'interaction ( $F[18, 144] = 15, p_{gg} < 0.01$ ). Encore une fois, plus l'EBR est élevé, plus les performances augmentent, et ce pour tous les systèmes, sauf *Komatsu*.

Pour un EBR de -6dB, l'analyse *post hoc* met en évidence 4 groupes de systèmes :

1. *Komatsu* : performances supérieures à celles de la *Baseline* ;

2. *choi, Hayashi 1 et Hayashi 2* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
3. *Gianoulis* : performances supérieures à celles de la *Baseline*, mais inférieures à celles du groupe 2 ;
4. *Gutierrez, Pikrakis, Phan et Vu* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

Pour des EBR de 0 et  $-6\text{dB}$ , trois groupes sont isolés :

1. *Komatsu, choi, Hayashi 1 et Hayashi 2* : performances supérieures à celles de la *Baseline* ;
2. *Phan* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
3. *Pikrakis, Gutierrez, Vu et Gianoulis* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

Les résultats, pour les scènes polyphoniques, sont similaires à ceux obtenus pour les scènes non-polyphoniques. Deux différences sont cependant notées :

- *Vu, Pikrakis et Gutierrez* présentent des résultats équivalents à ceux de la *Baseline* quel que soit l'EBR considéré ;
- *Phan* semble clairement améliorer ses performances par rapport à celles de la *Baseline* pour des EBR de 0 et  $+6\text{dB}$ . Ce dernier système semble souffrir d'une mauvaise prise en compte du bruit, ce qui est particulièrement pénalisant dans le cas de scènes polyphoniques ( $0\text{dB} \rightarrow -6\text{dB}$  : 41%  $\rightarrow$  24%).

#### 5.4.6.4 Influence du nombre d'événements

Considérant les scènes non-polyphoniques, les résultats sont affichés sur la figure 54a. l'ANOVA révèle un effet significatif du type de système ( $F[9, 216] = 264, p_{gg} < 0.01$ ), mais pas de nec ( $F[2, 24] = 0.5, p = 0.6$ ). Une interaction sensible est néanmoins observée ( $F[18, 216] = 3, p_{gg} < 0.01$ ).

Les mêmes résultats sont obtenus pour les scènes polyphoniques (cf. Figure 54a ; système :  $F[9, 216] = 170, p_{gg} < 0.01$  ; nec :  $F[2, 24] = 0.1, p = 0.9$  ; interaction :  $F[18, 216] = 3, p_{gg} < 0.01$ ). Ainsi, il nous est difficile de conclure quant à l'influence de nec sur de potentielles différences significatives entre les systèmes.

Malgré tout, nous pouvons isoler certaines tendances. Concernant les scènes non-polyphoniques, l'augmentation du nombre d'événements par classe s'accompagne d'une amélioration systématique des performances pour 2 systèmes (nec : 1  $\rightarrow$  3 ; *Hayashi 1* : 62%  $\rightarrow$  75% ; *Hayashi 2* : 63%  $\rightarrow$  73%) et d'une dégradation pour 1 système (nec :

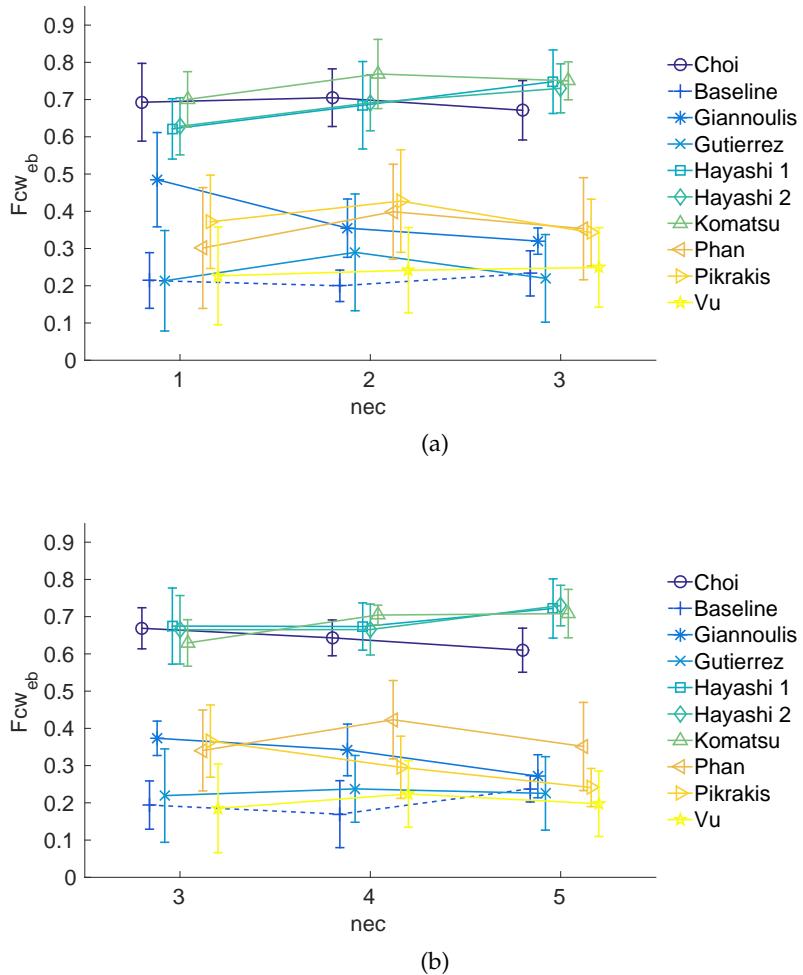


FIGURE 54 : Influence du nombre d'événements (nec) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique  $F_{CWEb}$ ; (a) scènes non-polyphoniques, (b) scènes polyphoniques.

$1 \rightarrow 3$ ; *Gianoulis* : 48%  $\rightarrow$  32%). Concernant les scènes polyphoniques, une augmentation est constatée pour 3 systèmes (nec : 3  $\rightarrow$  5; *Hayashi 1* : 67%  $\rightarrow$  72%; *Hayashi 2* : 66%  $\rightarrow$  73%; *Komatsu* : 63%  $\rightarrow$  70%) et une dégradation pour 3 autres (nec : 3  $\rightarrow$  5; *Gianoulis* : 37%  $\rightarrow$  27%; *Pikrakis* : 36%  $\rightarrow$  24%; *Choi* : 67%  $\rightarrow$  60%).

Alors que l'augmentation du niveau de bruit a globalement tendance à diminuer les performances des algorithmes, il apparaît que la réaction aux nombres d'événements à détecter varie d'un système à l'autre. Quelle que soit la nature polyphonique des scènes, les systèmes *Hayashi 1* et *Hayashi 2* réagissent systématiquement positivement à l'augmentation du nombre d'événements. En revanche, le système *Gianoulis* voit, lui, ses performances systématiquement décroître.

#### 5.4.7 Discussion

Des trois conditions expérimentales testées, seule celle intéressant l'EBR semble avoir une influence significative sur les performances des algorithmes. Aucun impact réel n'est détecté pour la polyphonie et le nombre d'événements sonores par classe (EBR). Ces résultats indiquent qu'il est capital, afin d'améliorer les performances des algorithmes, de proposer des systèmes de gestion du bruit efficaces.

Quel que soit l'EBR considéré, quatre algorithmes (*Hayashi 1*, *Hayashi 2*, *Komatsu* et *choi*) présentent systématiquement des performances supérieures à celles de la *Baseline*. Parmi ces quatre algorithmes, *Komatsu* est le seul à présenter des performances significativement supérieures aux trois autres. Ce détachement se produit pour un EBR de  $-6\text{dB}$ .

À ce titre, *Komatsu* peut être considéré comme le système présentant les meilleures capacités de généralisation, et ce, toujours, grâce à sa gestion efficace des niveaux de bruit.

## 5.5 CONCLUSION

Concernant le challenge DCASE 2013 (cf. Section 5.3), seul un algorithme (le vainqueur) arrive à maintenir des résultats stables pour toutes les conditions expérimentales (sélection des samples, EBR et espacement moyen). Tous les autres voient leurs résultats chuter jusqu'à atteindre le niveau de la *baseline*.

Concernant le challenge DCASE 2016 (cf. Section 5.4), nous observons des variations significatives des métriques relevées en fonction de l'EBR, variations pouvant atteindre les 20%. La polyphonie et le nombre d'événements présents par classe n'impactent cependant que modérément le fonctionnement des algorithmes.

Dans les deux challenges, les algorithmes semblent particulièrement sensibles à l'EBR, ce dernier ayant, dans la grande majorité des cas, un effet négatif sur les performances. Ce résultat montre l'importance de considérer un pré-traitement permettant de gérer l'influence du bruit. Les systèmes présentant les meilleurs résultats le prennent en compte de manière explicite dans leurs structures algorithmiques.

Les conditions d'enregistrement des événements paraissent, par ailleurs, impacter significativement les résultats du challenge DCASE 2013<sup>8</sup>. Dans ce challenge, les algorithmes ont, pour la plupart, souffert d'un phénomène de sur-apprentissage, leurs performances étant conditionnées à des scènes présentant des caractéristiques très semblables à celles des sons de la banque d'entraînement.

---

<sup>8</sup> À noter que dans le cas du challenge DCASE 2016, les corpus sont tous enregistrés dans les mêmes conditions.

En résumé, il apparaît que parmi toutes les conditions expérimentales testées, deux semblent être primordiales pour évaluer les performances des algorithmes : l'EBR et les conditions d'enregistrement.

À noter cependant que nous n'avons considéré qu'une seule métrique dans cette analyse, et qu'il conviendrait d'en considérer d'autres, notamment celles basées sur le nombre de segments détectés (et non le nombre d'*onsets*), afin de conclure définitivement quant aux performances globales des algorithmes.

À la lumière de ces résultats, nous pensons que, tenir compte de données soigneusement simulées est utile afin d'acquérir plus de connaissances sur les propriétés et les comportements des systèmes en cours d'évaluation, connaissances pouvant aider les chercheurs dans leurs choix algorithmiques avec une plus grande efficacité, et à moindre coût.

Les facteurs influant sur les performances tels que le niveau de bruit, le niveau de polyphonie, la diversité intra-classe (différence acoustique entre les données d'entraînement et les données de test.) peuvent ainsi être évalués de façon indépendante, sans qu'il soit nécessaire :

1. d'enregistrer des scènes présentant les propriétés désirées ;
2. d'annoter manuellement les données.

Nous pensons que, de même que l'usage exclusif de données simulées pour valider une approche algorithmique est insuffisant, l'utilisation seule de données réelles ne permet pas d'acquérir des connaissances sur l'impact de certains problèmes de conception et de paramétrisation rencontrés dans la mise en œuvre d'un système d'ingénierie. Les données réelles sont, la plupart du temps, des ressources rares, la conception de grands ensembles de données d'évaluation étant une tâche prenante et exigeante. Par ailleurs, l'annotation *a posteriori* des événements, suppose chez les sujets un relatif consensus, qui n'est pas forcément garanti. L'utilisation de données de simulation est un entre deux, qui, combiné à une validation sur données réelles, permet d'obtenir une meilleure compréhension des systèmes en cours d'évaluation. Il est ainsi possible de dépasser le simple objectif de « notation » de l'évaluation, qui devient un guide précieux pour le développement et l'amélioration des systèmes évalués, en pointant précisément leurs forces et faiblesses.

Quatrième partie  
**CONCLUSIONS ET PERSPECTIVES**



# 6

## CONCLUSIONS ET PERSPECTIVES

---

Ce chapitre, organisé en trois parties, dresse un bilan des travaux effectués dans le cadre de cette thèse.

La première partie aborde l'expérience sur la perception de l'agrément des environnements sonores urbains. La deuxième traite des campagnes d'évaluation des algorithmes de détection d'événements sonores. Nous concluons alors quant aux expériences menées, et présentons plusieurs pistes nouvelles d'investigation. La troisième et dernière propose un résumé des contributions et efforts de valorisation, qu'il s'agisse des publications, des programmes informatiques, ou encore des corpus mis à disposition.

### 6.1 ANALYSE SENSORIELLE

#### 6.1.1 *Agrément des paysages sonores urbains*

En ce qui concerne l'agrément des paysages sonores urbains, les expériences ont montré que la majorité des descripteurs utilisés, qu'ils soient sémantiques ou structurels, permettent de faire la distinction entre une scène idéale et une scène non-idéale.

Cependant, nous observons que les caractéristiques physiques corrélatées à l'agrément diffèrent clairement suivant la nature hédonique des scènes. Dans le cas des scènes idéales, c'est avant tout l'émergence de marqueurs sonores qui détermine la qualité perçue, alors que dans le cas des scènes non-idéales, c'est le niveau sonore global qui influe sur l'agrément.

Ces résultats, déjà suggérés par d'autres études (cf. Section 2.7.6), tendent à confirmer que la perception des qualités d'une scène dépend avant tout des sources sonores qui la composent, les caractéristiques structurelles mobilisées dans le processus perceptif semblant varier d'une source à l'autre, et d'un type d'environnement à l'autre. Ce fait montre qu'il est illusoire d'envisager qu'un descripteur physique holistique puisse rendre compte, de manière pertinente, des qualités affectives de tous types d'environnements.

#### 6.1.2 *Simulation et cognition*

Des études portant sur les processus perceptifs et cognitifs (cf. Section 2.2), nous retenons que demander à un sujet de simuler une scène est un moyen efficace d'accéder à la représentation mentale qu'il se fait de celle-ci. De fait, si l'on se place dans une vision an-

crée de la cognition, le sujet se construit une image intérieure modale de la représentation qu'il a de l'environnement (cf. Section 2.2.3). La simulation est alors un processus d'objectivation qui lui permet de traduire les modalités physiques de cette image mentale en données sonores.

Ainsi, la scène simulée apparaît comme une donnée physique et objective, mais ancrée dans une réalité cognitive. Ce qui en fait une ressource idéale permettant de faire le lien entre la structure de l'environnement sonore et les qualités affectives qu'il suscite.

Afin de faciliter la matérialisation de l'image mentale, les paramètres de contrôle du simulateur ont tous trait à des attributs reconnus par la communauté (cf. Section 2.7.7) pour leur importance dans la perception des scènes, à savoir, les classes de sons présentes, les niveaux sonores, la structure des séquences événementielles.

Ces réflexions ont mené à l'élaboration d'un outil de simulation opérationnel, accessible (aux fonctionnalités rapidement maîtrisées), et permettant à un sujet même non-initié de générer sans peine un environnement complet.

Par ailleurs, l'application de la simulation à l'étude de la perception de l'agrément des environnements sonores urbains valide les qualités écologiques des scènes simulées, nombre de résultats présents dans la littérature étant retrouvés (cf. Section 2.7.7).

### 6.1.3 *Perspectives*

Au terme de notre étude, nous pensons que la simulation est un outil dont le développement pourrait permettre aux décideurs en matière d'urbanisme d'interroger toute une communauté sur ses représentations propres des environnements sonores auxquels elle est exposée, et, pourquoi pas, sur les représentations des environnements sonores auxquels elle voudrait être exposée.

Dans la continuité des travaux réalisés, il conviendrait de multiplier les expériences de simulation, en faisant varier les qualités affectives (calme, confortable, gênante, etc.), mais aussi en spécifiant des lieux particuliers (parc, place, rue, etc.), afin d'élaborer des corpus entiers de scènes cognitivement renseignées de paysages sonores.

Il serait par ailleurs intéressant d'utiliser la simulation afin d'étudier plus avant les effets provoqués par la modification volontaire d'une caractéristique d'une scène, comme lors de la suppression des marqueurs sonores pratiquée dans l'expérience 2 (cf. Section 4.3).

Enfin, l'on pourrait encore étudier l'influence des contextes socioculturels sur la perception. Dans les faits, si le son de cloche est le plus souvent un marqueur d'environnement de qualité pour un occidental, cela ne se vérifie pas nécessairement auprès de sujets de culture orientale, moyen-orientale ou autre.

Outre les possibilités déjà évoquées, la simulation présente encore dans ce cas deux avantages :

- le simulateur peut être déployé à large échelle via internet ;
- les scènes simulées peuvent être analysées sans avoir à tenir compte des différentes langues maternelles des sujets, la nature sémantique des classes de sons utilisées étant connue *a priori* par l'expérimentateur.

Bien évidemment, ces approches nécessitent d'accroître la taille des banques de sons isolés disponibles, un effort conséquent, mais nécessaire, qui contribuera grandement aux nombreux domaines de recherche ayant trait aux scènes sonores.

## 6.2 ANALYSE AUTOMATIQUE

### 6.2.1 Détection automatique d'événements sonores

Les campagnes d'évaluation menées dans le cadre des deux sessions des challenges DCASE ont montré que tous les algorithmes n'avaient pas les mêmes capacités de généralisation.

Ces observations ont été rendues possibles grâce à l'utilisation de corpus de scènes simulées. La simulation nous permettant de contrôler les caractéristiques structurelles des scènes, il nous est possible de finement apprécier l'effet de leurs variations respectives sur les performances des algorithmes.

Sans simulation, de telles observations seraient difficiles à déduire, l'expérimentateur ne possédant plus, en sortie, qu'un résultat brut, global, dissocié de l'état structurel des scènes sonores considérées.

L'ensemble de ces résultats valide l'intérêt d'évaluer des algorithmes sur des scènes dont l'expérimentateur maîtrise la nature.

### 6.2.2 Perspectives

L'interrogation, par la simulation de données contrôlées, de systèmes complexes comme les processus computationnels utilisés dans le cadre de l'écoute artificielle, n'est pas un fait marginal, mais bien ce que nous pensons être une phase de maturation ubiquitaire en apprentissage automatique.

De fait, dans beaucoup de domaines d'application de l'intelligence artificielle, la question est posée de la validité et de la robustesse des systèmes chargés d'effectuer des tâches, ou de prendre des décisions susceptibles d'interférer dans nos vies (O'Neil, 2016).

Contrairement aux systèmes experts (Leondes, 2002), dont l'algorithmie peut être étudiée pour prédire les comportements, les systèmes actuels, notamment les réseaux profonds, sont difficilement

approchables par la méthode analytique. La communauté privilégie l'étude de la réponse de ces systèmes à des stimuli choisis dont on contrôle *a priori* mieux les propriétés. Pour exemple, on citera, en analyse d'image, les travaux novateurs de Ian Goodfellow (Goodfellow et al., 2014) sur l'amélioration de la robustesse des réseaux neuronaux profonds par l'utilisation d'exemples adversaires, des exemples d'apprentissage peu, mais judicieusement modifiés. On peut également citer les travaux de Nguyen et consorts (Nguyen et al., 2015) montrant comment l'utilisation d'images totalement synthétiques, générées afin d'optimiser la réponse du système de détection, permet de forcer le système interrogé.

Tous ces travaux partent du postulat que, devant la complexité de l'objet d'étude, il est possible, par la génération de données précises, d'interroger cette complexité de manière pertinente. Une telle utilisation, dans le domaine de l'apprentissage automatique, de paradigmes expérimentaux courants en psychologie expérimentale et en neurosciences, est exemplifiée par l'usage, dans la communauté, du terme « *artificial neuroscience* ». Ce rapprochement entre des pratiques méthodologiques propres à l'analyse sensorielle, d'une part, et à l'apprentissage automatique, d'autre part, résonne avec les thèses défendues, à notre échelle, dans ce document, montrant leurs adéquations avec des questionnements majeurs du monde scientifique.

### 6.3 CONTRIBUTIONS

#### 6.3.1 Valorisation scientifique

Les recherches réalisées dans le cadre de cette thèse ont donné lieu à plusieurs communications.

Concernant l'évaluation des systèmes de détection d'événements sonores, la campagne de ré-évaluation des algorithmes ayant participé au challenge DCASE 2013 (cf. Section 5.3) a fait l'objet d'une publication (Lafay et al., 2016a). Cette publication a introduit également le modèle de scènes sonores proposé (cf. Section 3.2). Les résultats obtenus lors du challenge DCASE 2016 (cf. Section 5.4) seront eux aussi prochainement présentés, à l'occasion d'une communication conjointe des organisateurs des différentes tâches du challenge.

D'autres recherches ont porté sur le recouvrement automatique des similarités entre scènes sonores. Elles ont montré notamment l'intérêt de considérer la scène comme un objet composite dont les éléments constitutifs (sources sonores) n'impactent pas de la même manière les similarités inter-scènes. Deux de ces études ont été, l'une, publiée (Lagrange et al., 2015), l'autre, soumise (Lostanlen et al., 2016). Comme elles ne portent pas directement sur l'utilisation expérimentale des scènes simulées, elles ne sont pas abordées dans ce document.

Concernant l'analyse sensorielle, les fonctionnalités de l'outil de simulation *SimScene* ont fait l'objet de deux articles de conférence (Lafay et al., 2016b; Rossignol et al., 2015b). La valorisation de l'expérience pilote de simulation (Lafay et al., 2014) a également fait partie de nos travaux.

Par ailleurs, plusieurs travaux réalisés en collaboration ont permis de motiver l'utilisation de corpus de scènes simulées afin d'évaluer de manière fine les performances de différents systèmes en apprentissage machine (Benetos et al., 2016b,c; Rossignol et al., 2015a).

### 6.3.2 Programmes et banques de données

L'auteur a participé au développement de deux simulateurs d'environnements sonores.

Le premier est supporté par navigateur web, et possède une interface optimisée afin de pouvoir être utilisée facilement par des sujets dans le cadre d'expériences sur la perception des environnements sonores (cf. Section 4.1.1.7). Le deuxième, développé en MATLAB, est dédié à la génération de corpus d'évaluation en apprentissage machine. Les deux simulateurs sont accessibles au public<sup>1,2</sup>.

Par ailleurs, les campagnes d'évaluation menées dans le cadre des deux sessions des challenges DCASE ont abouti à la création de plusieurs corpus de scènes sonores simulées et complètement annotées. Ces simulations ont requis l'enregistrement d'une banque de données conséquente de sons isolés.

Les corpus de scènes simulées dans le cadre des challenges DCASE 2013<sup>3,4</sup> et 2016<sup>5</sup>, ainsi que la banque de sons isolés sont accessibles au public.

---

<sup>1</sup> Simscene Web : <http://www.irccyn.ec-nantes.fr/~lagrange/demonstrations/simScene.html>

<sup>2</sup> Simscene MATLAB : <https://bitbucket.org/mlagrange/simscene/downloads>

<sup>3</sup> DCASE 2013 instance-QMUL, abstract-QMUL : [https://archive.org/details/dcase\\_replicate\\_qmul](https://archive.org/details/dcase_replicate_qmul).

<sup>4</sup> DCASE 2013 instance-IRCCYN, abstract-IRCCYN : [https://archive.org/details/dcase\\_replicate](https://archive.org/details/dcase_replicate).

<sup>5</sup> DCASE 2016 entraînement, développement et évaluation : <http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio>.



Cinquième partie

APPENDICES



# A

## OUTILS D'ANALYSE STATISTIQUE UNI-VARIÉE

---

### A.1 TESTS PARAMÉTRIQUES À DEUX ÉCHANTILLONS

Un test paramétrique à deux échantillons est un test statistique permettant d'apprécier si les moyennes de deux échantillons présentent une différence significative, la valeur statistique calculée étant supposée émaner d'une distribution particulière, nommée distribution nulle.

Le test de Student est, lui, un test paramétrique pour lequel la valeur statistique calculée est sensée procéder d'une distribution de Student. La valeur  $p$  mesure alors la probabilité que la valeur statistique calculée émane effectivement d'une distribution de Student. La significativité de la valeur  $p$  est évaluée à partir d'un seuil critique arbitraire  $\alpha$ .

On distingue deux types de tests de Student :

- Test de Student à deux échantillons appariés : les deux échantillons sont liés points à points par une relation de dépendance. On peut citer comme exemple deux mesures de poids, prises sur un même groupe de sujets avant et après l'administration d'un traitement donné ;
- Test de Student à deux échantillons indépendants : les deux échantillons ne sont liés par aucune relation de dépendance. Pour reprendre l'exemple précédent, les deux mesures de poids sont cette fois ci prises sur deux groupes de sujets distincts, les sujets de l'un ayant reçu un traitement particulier, les sujets de l'autre non ;

#### *Test de Student à deux échantillons appariés*

Considérant  $x$  et  $y$  deux échantillons de  $n$  individus, le test de Student à deux échantillons appariés éprouve l'hypothèse nulle que l'échantillon  $x - y$  vienne d'une population déterminée par une distribution normale, de moyenne nulle, et de variance inconnue.

Il s'agit alors de calculer la valeur statistique  $t$  :

$$t = \frac{m}{s\sqrt{\frac{1}{n}}} \tag{14}$$

avec  $m$  et  $s$  respectivement la moyenne et l'écart type de l'échantillon  $x - y$ . La valeur  $p$  est obtenue en cherchant la valeur  $t$  dans

une table des valeurs de la distribution d'une loi de Student à  $n - 1$  degrés de liberté.

#### *Test de Student à deux échantillons indépendants*

Considérant  $x$  et  $y$  deux échantillons de  $n$  et  $m$  individus, le test de Student à deux échantillons indépendants éprouve l'hypothèse nulle que les échantillons  $x$  et  $y$  viennent de deux populations déterminées par deux distributions normales, ayant la même moyenne et la même variance.

Il s'agit alors de calculer la valeur statistique  $t$  :

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n} + \frac{1}{m}}} \quad (15)$$

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \quad (16)$$

avec  $\bar{x}$  et  $\bar{y}$  les moyennes des échantillons  $x$  et  $y$ , et  $s_x$  et  $s_y$  les écarts types des échantillons  $x$  et  $y$ . La valeur  $p$  est obtenue en cherchant la valeur  $t$  dans une table des valeurs de la distribution d'une loi de Student à  $n + m - 2$  degrés de liberté.

## A.2 TESTS PARAMÉTRIQUES À PLUS DE DEUX ÉCHANTILLONS

L'analyse de variance (ANOVA) est un ensemble de tests statistiques paramétriques permettant d'apprécier si les moyennes de deux échantillons, ou plus, présentent des différences significatives.

L'ANOVA considère une variable dépendante quantitative, *i.e.* la mesure considérée, et une ou plusieurs variable(s) indépendante(s) catégorielle(s). Les variables indépendantes sont appelées « facteurs ». Les valeurs (catégories) prises par la(es) facteur(s) sont appelées « niveaux ».

La valeur statistique calculée par l'ANOVA est sensée procéder d'une distribution de Fisher. La valeur  $p$  mesure alors la probabilité que la valeur statistique calculée émane effectivement d'une distribution de Fisher. La significativité de la valeur  $p$  est évaluée à partir d'un seuil critique arbitraire  $\alpha$ .

On distingue deux types d'ANOVA :

- l'ANOVA à mesures répétées : les échantillons sont liés points à points par une relation de dépendance. On peut citer comme exemple trois mesures de poids, prises sur un même groupe de sujets, après trois administrations d'un traitement donné. Dans ce cas il y une variable dépendante, la mesure de poids, et un

facteur indépendant, la prise de traitement, ce dernier ayant trois niveaux (prise 1, prise 2 et prise 3) ;

- l'ANOVA à échantillons indépendants : les échantillons ne sont liés par aucune relation de dépendance. Pour reprendre l'exemple précédent, les trois mesures de poids sont cette fois-ci prises sur trois groupes de sujets distincts, l'un ayant reçu une fois le traitement, l'autre deux fois, et le dernier trois fois.

Il est également possible de considérer une ANOVA comportant deux facteurs, un facteur à mesures répétées, et un facteur à mesures indépendantes. Dans ce cas, le facteur à mesures répétées est appelé facteur intra-sujet, et le facteur à mesures indépendantes, facteur inter-sujet.

Nous détaillons dans la suite deux exemples simples de modèles d'ANOVA : un modèle à un facteur inter-sujet, et un modèle à un facteur intra-sujet.

#### *Analyse de variance à un facteur inter-sujet*

Considérons les groupes  $y_i$  ( $i = 1, \dots, k$ ), relatifs aux  $k$  niveaux du facteur analysé. Chaque groupe  $y_i$  est composé de  $n_i$  mesures. On note  $N = \sum_i n_i$  le nombre total de mesures. L'ANOVA à mesures indépendantes teste l'hypothèse nulle que tous les échantillons sont issus de distributions normales ayant la même moyenne. L'hypothèse alternative est qu'au moins une de ces distributions présente une moyenne différente des autres.

L'ANOVA cherche à partitionner la variation totale  $S_{\text{totale}}$  en une variation inter-groupe  $S_{\text{inter-groupe}}$ , celle que l'on cherche à évaluer, et une variation intra-groupe  $S_{\text{intra-groupe}}$ , considérée comme l'erreur.

$$S_{\text{totale}} = S_{\text{inter-groupe}} + S_{\text{intra-groupe}} \quad (17)$$

$$S_{\text{totale}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad (18)$$

$$S_{\text{inter-groupe}} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \quad (19)$$

$$S_{\text{intra-groupe}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (20)$$

avec  $\bar{y}_i$  la moyenne des observations relatives au groupe  $i$ , et  $\bar{y}$  la moyenne globale des observations. Il s'agit de calculer une valeur statistique  $F$  :

$$F = \frac{S_{\text{inter-groupe}}/k - 1}{S_{\text{intra-groupe}}/(N - k)} \quad (21)$$

La valeur  $p$  est obtenue en cherchant la valeur  $F$  dans une table des valeurs de la distribution d'une loi de Fisher à  $[k - 1, N - k]$  degrés de liberté.

#### *Analyse de variance à un facteur intra-sujet*

Considérons les groupes  $y_i$  ( $i = 1, \dots, k$ ), relatifs aux  $k$  niveaux du facteur analysé. Chaque groupe  $y_i$  est composé de  $n_i$  mesures. On note  $N = \sum_i n_i$  le nombre total de mesures. L'ANOVA à mesures répétées teste l'hypothèse nulle que tous les échantillons sont issus de distributions normales ayant la même moyenne. L'hypothèse alternative est qu'au moins une de ces distributions présente une moyenne différente des autres.

Comme pour l'ANOVA à mesures indépendantes, l'ANOVA à mesures répétées cherche à partitionner la variation totale  $S_{\text{totale}}$  en une variation inter-groupe  $S_{\text{inter-groupe}}$ , et une variation intra-groupe  $S_{\text{intra-groupe}}$ . Seulement, dans le cas de l'ANOVA à mesures répétées,  $S_{\text{intra-groupe}}$  se retrouve elle-même partitionnée en une variation intra-sujet  $S_{\text{intra-sujet}}$ <sup>1</sup>, et une variation due à l'erreur  $S_{\text{erreur}}$ .

$$S_{\text{totale}} = S_{\text{inter-groupe}} + S_{\text{intra-groupe}} \quad (22)$$

$$= S_{\text{inter-groupe}} + S_{\text{intra-sujet}} + S_{\text{erreur}} \quad (23)$$

$$S_{\text{totale}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad (24)$$

$$S_{\text{inter-groupe}} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \quad (25)$$

$$S_{\text{intra-groupe}} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (26)$$

---

<sup>1</sup> Le terme sujet est ici générique et désigne l'entité sur laquelle on réalise la mesure. Il peut s'agir d'un individu, ou, comme souvent dans ce document, d'une scène sonore.

$$S_{\text{intra-sujet}} = k \sum_{j=1}^n (\bar{y}_j - \bar{y})^2 \quad (27)$$

$$S_{\text{erreur}} = S_{\text{intra-groupe}} - S_{\text{intra-sujet}} \quad (28)$$

avec  $\bar{y}_i$  la moyenne des observations relatives au groupe  $i$ ,  $\bar{y}_j$  la moyenne des observations relatives au sujet  $j$ , et  $\bar{y}$  la moyenne globale des observations. Il s'agit de calculer une valeur statistique  $F$ :

$$F = \frac{S_{\text{inter-groupe}}/k - 1}{S_{\text{erreur}}/(n - 1)(k - 1)} \quad (29)$$

La valeur  $p$  est obtenue en cherchant la valeur  $F$  dans une table des valeurs de la distribution d'une loi de Fisher à  $[k - 1, (n - 1)(k - 1)]$  degrés de liberté.

### A.3 COMPARAISONS MULTIPLES

L'ANOVA permet seulement de savoir si tous les échantillons sont issus de la même loi normale. Dans le cas où l'hypothèse nulle est rejetée, l'ANOVA seule ne permet pas de savoir quels sont les échantillons qui présentent des différences significatives.

Pour ce faire, on peut avoir recours à une analyse dite *Post hoc* de comparaisons multiples. Il s'agit d'effectuer un test statistique pour toutes les paires d'échantillons considérées, afin d'établir les couples qui présentent une différence significative au niveau de leurs moyennes.

L'analyse *Post hoc* implique d'effectuer une succession de tests statistiques. Il est nécessaire dans ce cas de corriger les valeurs  $p$  obtenues, afin de tenir compte de la multiplicité des tests opérés. En effet, analyser un même jeu de données en effectuant une série de tests statistiques entraîne un accroissement du risque d'erreur de type I, *i.e.* rejeter l'hypothèse nulle alors qu'elle est vraie.

Plusieurs procédures peuvent être considérées. Nous en détaillons deux, nommément la procédure de Tukey-Kramer, et la procédure de Bonferroni.

#### *Procédure de Tukey-Kramer*

La procédure de Tukey-Kramer s'appuie sur la statistique d'« écart studentisé », notée  $q$ , afin de corriger la valeur critique des comparaisons.

valeur	interprétation
0	corrélation inexistante
$\pm 0.3$	corrélation faible
$\pm 0.5$	corrélation modérée
$\pm 0.7$	corrélation forte
$\pm 1$	corrélation exacte

TABLE 24 : Interprétation du coefficient de corrélation de Pearson. Un signe négatif indique une corrélation négative et inversement.

Considérons  $k$  groupes  $y_i$  ( $i = 1, \dots, k$ ). Pour chaque paire  $y_i$  et  $y_j$ , on calcule la statistique  $q_{ij}$  comme suit :

$$q_{ij} = \frac{\bar{y}_i - \bar{y}_j}{s \sqrt{\frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (30)$$

avec  $\bar{y}_i$  et  $\bar{y}_j$  les moyennes des groupes  $y_i$  et  $y_j$ , et  $n_i$  et  $n_j$  le nombre d'observations des groupes  $y_i$  et  $y_j$ . La valeur  $q_{ij}$  est supposée émaner d'une distribution  $Q$  dite d'« écart studentisé » (studentized range). La table des valeurs de la distribution  $Q$  nous permet d'établir la valeur  $q_c$  critique permettant de rejeter l'hypothèse nulle avec un seuil critique  $\alpha = 0.05$ , en tenant compte du nombre d'échantillons, *i.e.* du nombre de groupes à comparer, ici  $k$ .

#### *Procédure de Bonferroni*

La procédure de Bonferroni, aussi appelée correction de Bonferroni, consiste à corriger le seuil critique  $\alpha$  en fonction du nombre de comparaisons à effectuer. Formellement, si l'on considère un seuil critique d'origine  $\alpha_0$ , et une comparaison multiple impliquant  $d$  comparaisons, chacun des tests statistiques sera mené avec un seuil  $\alpha = \frac{\alpha_0}{k}$ .

A noter que la correction de Bonferroni peut s'appliquer à n'importe quel test statistique. Dans le cas d'une comparaison multiple suivant une ANOVA, on pourra considérer un test de Student.

#### A.4 MESURES DE CORRÉLATION PARAMÉTRIQUES

On dit que deux variables sont corrélées si un changement chez l'une est suivi d'un changement chez l'autre. Le coefficient de corrélation de Pearson  $r$  permet de mesurer les relations linéaires existant entre

deux variables. Considérant deux variables  $x$  et  $y$ , le coefficient  $r$  se calcule comme suit :

$$r_{x,y} = \frac{\frac{1}{n} \sum_i (x_i - m_x)(y_i - m_y)}{s_x s_y} \quad (31)$$

avec  $m_x$  et  $m_y$  les moyennes des variables  $x$  et  $y$ , et  $s_x$  et  $s_y$  les écarts types des variables  $x$  et  $y$ . Il est possible d'évaluer la significativité de la relation  $r$  mesurée, en reportant la valeur sur une table de coefficients de corrélation de Pearson. Cette table nous renseigne sur le risque  $\alpha$  de rejeter, par erreur, l'hypothèse nulle qu'il n'y a pas de relation entre les deux variables  $x$  et  $y$ . Le risque  $\alpha$  obtenu dépend du nombre d'observations considéré. Plus ce nombre est important, plus le risque est faible.

Le coefficient  $r$  prend des valeurs comprises entre  $-1$  et  $1$ . Nous nous appuyons sur le tableau 24 pour donner une appréciation qualitative du coefficient  $r$ .



# B

## SÉQUENCE D'ÉVÉNEMENTS OU TEXTURE SONORE : L'INFLUENCE DE LA PÉRIODE D'ATTENTION.

---

### B.1 OBJECTIF DE L'EXPÉRIENCE

Nous présentons ici les résultats d'une étude sur la perception des textures sonores, étude menée dans le cadre de cette thèse, mais déconnectée du sujet principal.

Comme nous l'avons vu (cf. Section 2.8.1), la texture est un objet composite, dont les éléments constitutifs cessent d'être perçus de manière distincte, dès lors qu'ils occurrent suivant un pattern dont les caractéristiques physiques restent stables au cours du temps. La perception de ce pattern nécessite une certaine période d'attention. Nous proposons ici un protocole expérimental permettant d'étudier ces périodes d'attention.

En considérant comme stimuli une mixture d'événements du même type, nous faisons l'hypothèse qu'à partir du moment où le cerveau parvient à isoler un événement de cette mixture, il ne perçoit plus la mixture comme une texture, mais comme une succession d'événements. Inversement, s'il ne parvient pas à distinguer un événement isolé, alors la mixture est perçue comme une texture.

Nous appliquons le modèle de scènes sonores proposé (cf. Section 3.3.4) afin de simuler des textures à partir de séquences d'événements dont nous contrôlons l'espacement *inter-onsets* moyen. Il s'agit de faire varier cet espacement, afin d'identifier le seuil à partir duquel la séquence d'événements cesse d'être perçue comme une texture.

A ce titre, le protocole proposé s'inscrit complètement dans le cadre des expériences perceptives portant sur la détection du signal.

### B.2 BANQUE DE DONNÉES

Chaque stimuli est composé d'un son cible, suivi d'une séquence d'événements enchevêtrés. Le son cible est le même pour tous les stimuli, et tous les sujets. Il a été choisi par les expérimentateurs, et voulu ni trop identifiable, ni tout à fait dénué de caractéristiques saillantes.

Les événements sont tous des sons isolés ayant une durée de 1 seconde. Les séquences durent 6 secondes. Chacune donne à entendre des scènes de trafic. Ces scènes sont simulées en agglomérant des sons de voiture (isolés). La simulation est contrôlée par un paramètre réglant l'espacement temporel *inter-onsets* moyen entre les évé-

	réponse : oui	réponse : non
Présence son cible : oui	vp	fr
Présence son cible : non	fp	vr

TABLE 25 : Théorie de la détection du signal.

nements. Cinq valeurs d'espacement sont considérées : 0.1, 0.3, 0.5, 0.7 et 0.9 secondes (cf. Figure 55a, et 55b).

Pour chaque espacement, nous simulons 20 séquences de trafic. Les sujets doivent donc écouter 100 stimuli. La moitié d'entre eux sont des pièges (*catch trial*). Ils ne contiennent pas le son cible.

### B.3 PLANIFICATION EXPÉRIMENTALE

#### Procédure

L'expérience est une épreuve d'évaluation de type oui/non. Chaque sujet évalue l'ensemble des 100 stimuli. Pour chacun il indique si oui ou non il entend le son cible.

Les scènes sont présentées dans un ordre aléatoire.

#### Dispositif expérimental

Tous les sujets passent l'expérience sur des machines identiques. L'audio est diffusé en monophonique, par le biais de casques audio semi-ouverts *Beyer-Dynamic DT 990 Pro*.

Le niveau sonore de sortie est le même pour tous les sujets. Il a été préalablement fixé par les expérimentateurs afin de correspondre à un niveau d'écoute confortable.

Les sujets passent l'expérience individuellement, dans un environnement acoustique calme. Un expérimentateur est présent durant la totalité de l'expérience, afin de contrôler le bon déroulement de cette dernière, et de répondre aux éventuelles questions des sujets.

#### Participants

13 sujets (5 femmes), tous étudiants à l'École Centrale de Nantes, ont réalisé l'expérience.

#### B.4 MÉTHODOLOGIE ET OUTILS STATISTIQUES

Pour chaque condition expérimentale (espacement inter-onsets), les performances des sujets sont évaluées à l'aide de la mesure de sensibilité  $d'$ .

$$d' = z(H) - z(F) \quad (32)$$

$$H = \frac{vp}{vp + fr} \quad , \quad F = \frac{fp}{fp + vr} \quad (33)$$

avec  $H$  le taux de réponses correctes,  $F$  le taux de faux positifs,  $vp$  le nombre de vrais positifs,  $fr$  le nombre de fausses réjections,  $fp$  le nombre de fausses alarmes,  $vr$  le nombre de vraies réjections (cf. Tableau 25), et  $z(x)$  l'opérateur de transformation  $z$  Gaussien permettant d'exprimer la quantité  $x$  en nombre d'écart types par rapport à la moyenne. L'analyse se déroule en deux temps.

- Pour apprécier l'existence de différences au niveau des  $d'$  entre les conditions expérimentales, nous considérons une analyse de variance à mesures répétées comportant une variable indépendante à 5 niveaux : l'espacement inter-onsets. La sphéricité est évaluée à l'aide du test de Mauchly. Si cette dernière n'est pas vérifiée, la correction de Greenhouse-Geisser est appliquée, la valeur  $p$  est alors notée  $p_{gg}$ . Le seuil de significativité est fixé à  $\alpha = 0.05$  ;
- Afin d'établir un seuil théorique d'espacement en deçà duquel les sujets n'arrivent plus à distinguer le son cible, nous calculons une régression linéaire des  $d'$  en fonction de l'espacement, et considérons la valeur de l'espacement pour  $d' = 1$ . Ce seuil de  $d' = 1$  est couramment utilisé dans la théorie de la détection du signal (Macmillan and Creelman, 2004).

#### B.5 RÉSULTATS

L'ANOVA à mesures répétées montre un effet significatif de l'espacement sur la sensibilité  $d'$  ( $F[4, 48] = 24, p_{gg} < 0.01$ ). L'observation des évolutions des taux  $H$  et  $F$  montre que les réponses des sujets ne sont pas biaisées : le taux de réponses correctes  $H$  (*i.e.*, la capacité à distinguer le son cible) continue d'augmenter avec l'espacement, tandis que le taux de faux positifs  $F$  reste à peu près constant à partir d'une valeur d'espacement de 0.3 seconde (cf. Figure 55c).

La régression linéaire de  $d'$  en fonction de l'espacement (cf. Figure 55d) nous permet d'obtenir une limite théorique ( $d' = 1$ ) à partir de laquelle la mixture cesse d'être perçue comme une texture. Pour des

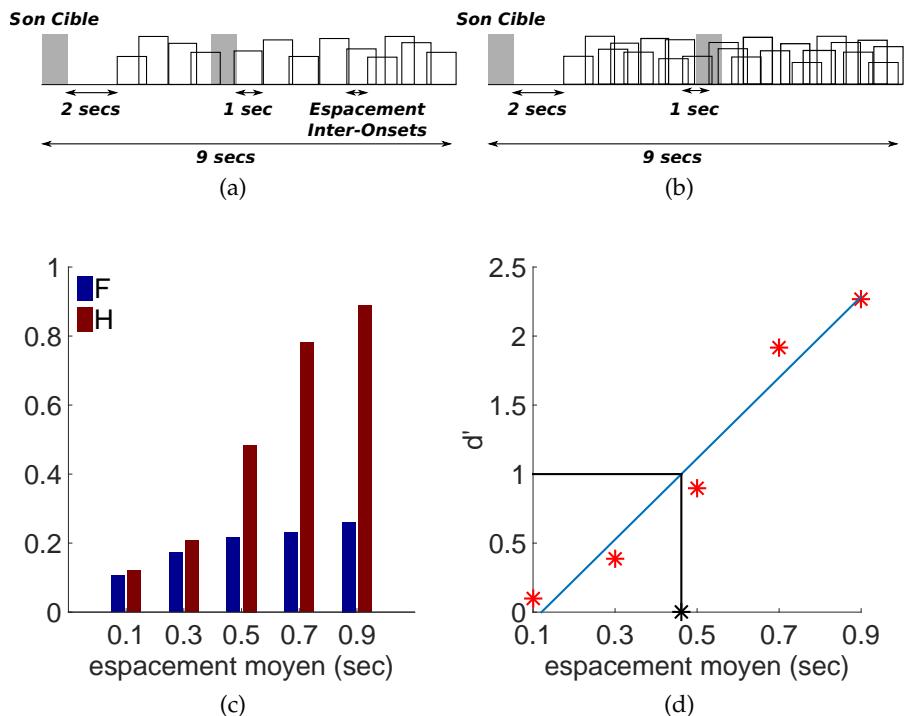


FIGURE 55 : Événement ou texture sonore : influence de la période d'attention. (a) stimulus ayant un espacement inter-onsets élevé ; (b) stimulus ayant un espacement inter-onsets faible ; (c) évolution des taux de réponses correctes H et de faux positifs F en fonction de l'espacement ; (d) régression linéaire de  $d'$  en fonction de l'espacement.

sons isolés d'une seconde, cet espacement limite est de 0.46 seconde, soit la moitié de la durée des événements utilisés.

# C

## INFLUENCE DE LA CONGRUENCE SUR LA DÉTECTION DES ÉVÉNEMENTS SONORES

---

### C.1 OBJECTIF DE L'EXPÉRIENCE

Une étude réalisée par (Gygi and Shafiro, 2011), tend à montrer que les sons incongrus dans un environnement sonore ambiant (canard dans aéroport) sont plus facilement reconnaissables que des sons naturellement adaptés (avion dans aéroport). Ils nomment cet effet, l'avantage de l'incongru (AI). Pour démontrer cela, ils ont effectué un certain nombre d'expériences sur une population élargie, en demandant aux individus d'identifier la nature d'un son court dans une scène longue, parmi un liste de réponses forcées. La seule condition expérimentale variant est le rapport noté  $\frac{S_o}{S_b}$  entre le niveau du son cible à identifier ( $S_o$ ), et celui du fond sonore ( $S_b$ ).

L'objectif de cette expérience est d'introduire une nouvelle variable à l'étude réalisée par (Gygi and Shafiro, 2011), à savoir la position du son cible à reconnaître dans la scène. Il est en effet possible que l'AI dépende du temps d'exposition du sujet à l'environnement ambiant.

Nous faisons l'hypothèse que, plus le temps d'exposition précédent le son cible est long, plus l'AI sera élevé. Cet effet pouvant dépendre de la durée totale de la scène, nous considérons 3 conditions expérimentales :

- la durée de la scène : nous considérons deux valeurs de 5 et 10 secondes ;
- la position du son cible : nous considérons trois valeurs correspondant au début, au milieu et à la fin de la scène. Le calcul des positions dépend de la durée de la scène :
  - pour une durée de 5 secondes : début= 0.5 seconde, milieu= 2.25 secondes, fin= 3.5 secondes ;
  - pour une durée de 10 secondes : début= 1 seconde, milieu= 4 secondes, fin= 7 secondes ;
- la congruence : nous considérons deux valeurs, congru et incongru, suivant que le son cible est adapté ou non à l'environnement ambiant.

### C.2 BANQUE DE DONNÉES

Nous considérons une banque de données composée de 8 enregistrements d'ambiances et de 8 enregistrements de sons isolés, ces der-

niers correspondant aux sons cibles à identifier. La moitié de ces enregistrements est prévue pour des scènes longues (ambiance : 10 secondes, sons cibles : 2 secondes), l'autre moitié pour des scènes courtes (ambiance : 5 secondes, sons cibles : 0.5 secondes).

Cibles et ambiances ont été choisies de sorte que toutes les cibles puissent être introduites et dans une scène congrue, et dans une scène incongrue. Pour chaque association cible/ambiance, trois scènes sont simulées réservant chacune une position différente au son cible (début, milieu et fin). Enfin chaque scène est simulée deux fois en considérant des ratios  $\frac{S_o}{S_b}$  de -6 et -9dB.

Cette procédure de simulation nous donne une banque de sons de 96 scènes simulées ( $8*2*3*2$ ). A noter cependant qu'un même sujet n'est soumis qu'à des scènes partageant le même  $\frac{S_o}{S_b}$ , soit 48 scènes.

### C.3 PLANIFICATION EXPÉIMENTALE

#### Procédure

Pour chacune des 48 scènes, le sujet doit identifier le son cible parmi 15 réponses proposées. Ce choix des 15 réponses, quand il n'y a que 8 cibles, a été fait de manière 1) à limiter les effets de mémoire, 2) à empêcher que le sujet procède par élimination.

Afin que le sujet puisse identifier les cibles, il est nécessaire de lui indiquer quand celles-ci surviennent dans la scène. Une interface graphique a été développée à cet effet. Elle permet d'afficher les 15 réponses possibles, ainsi qu'un voyant qui s'allume au vert lorsque le son cible est joué. Cette méthode est jugée moins invasive que celle proposée par (Gygi and Shafiro, 2011), où l'apparition du son cible dans la scène est précédée d'un court son pur.

Les scènes sont présentées dans un ordre aléatoire.

#### Dispositif expérimental

Tous les sujets passent l'expérience sur des machines identiques. L'audio est diffusé en monophonique, par le biais de casques audio semi-ouverts *Beyer-Dynamic DT 990 Pro*.

Les sujets passent l'expérience individuellement, dans un environnement acoustique calme. Un expérimentateur est présent durant la totalité de l'expérience, afin de contrôler son bon déroulement, et de répondre aux éventuelles questions des sujets.

#### Participants

15 sujets ont passé l'expérience pour un  $\frac{So}{Sb} = -6\text{dB}$ , et 15 autres pour un  $\frac{So}{Sb} = -9\text{dB}$ . Tous les sujets sont des étudiants de l'École centrale de Nantes.

#### C.4 MÉTHODOLOGIE ET OUTILS STATISTIQUES

Pour chaque condition expérimentale, et chaque sujet, nous calculons le pourcentage de réponses correctes noté  $p(c)$ .

Pour apprécier s'il existe des différences entre les conditions expérimentales au niveau des moyennes de  $p(c)$  relevées, nous considérons une ANOVA à mesures répétées à trois facteurs, nommément, la durée, la position et la congruence. La sphéricité est évaluée à l'aide du test de Mauchly. Si cette dernière n'est pas vérifiée, la correction de Greenhouse-Geisser est appliquée, la valeur  $p$  est alors notée  $p_{gg}$ . Le seuil de significativité est fixé à  $\alpha = 0.05$ .

#### C.5 RÉSULTATS

Concernant  $\frac{So}{Sb} = -6\text{dB}$ , l'ANOVA à mesures répétées montre un effet significatif de la congruence ( $F[1, 14] = 9, p < 0.05$ ), de la durée des scènes ( $F[1, 14] = 21, p < 0.01$ ), mais pas de la position du son cible ( $F[2, 28] = 2, p_{gg} = 0.17$ ) (cf. Figures 56a et 56b).

Les mêmes résultats sont observés pour  $\frac{So}{Sb} = -9\text{dB}$ , avec un effet significatif de la congruence ( $F[1, 14] = 15, p < 0.01$ ), de la durée des scènes ( $F[1, 14] = 64, p < 0.01$ ), mais pas de la position du son cible ( $F[2, 28] = 2, p_{gg} = 0.15$ ) (cf. Figures 56c et 56d).

Plusieurs points sont à retenir. Premièrement, nous retrouvons bien l'AI de (Gygi and Shafiro, 2011), et ce pour presque toutes les conditions expérimentales. Deuxièmement, nous observons un effet du rapport  $\frac{So}{Sb}$  cohérent, *i. e.* plus le niveau du son cible est fort, par rapport à celui de la scène, et plus le  $p(c)$  augmente. Troisièmement, plus la durée de la scènes est courte, et plus  $p(c)$  augmente.

Nous ne relevons pas d'effet significatif de la position. Cependant la figure 56 nous permet d'apprécier qualitativement certaines tendances. On peut voir que l'effet de la position varie suivant que l'on considère une scène courte (cf. Figures 56a et 56c), ou une scène longue (cf. Figures 56b et 56d). Dans le cas d'une scène courte, aucune tendance ne semble se dégager. En revanche, dans le cas d'une longue, plus la position est éloignée dans le temps, et plus l'AI se réduit. Ces dernières observations nous permettent de conjecturer deux points. Si l'AI dépend de la position du son cible, alors :

- il dépend également de la durée de la scène. Pour une durée courte (5 secondes), la position est négligeable ;

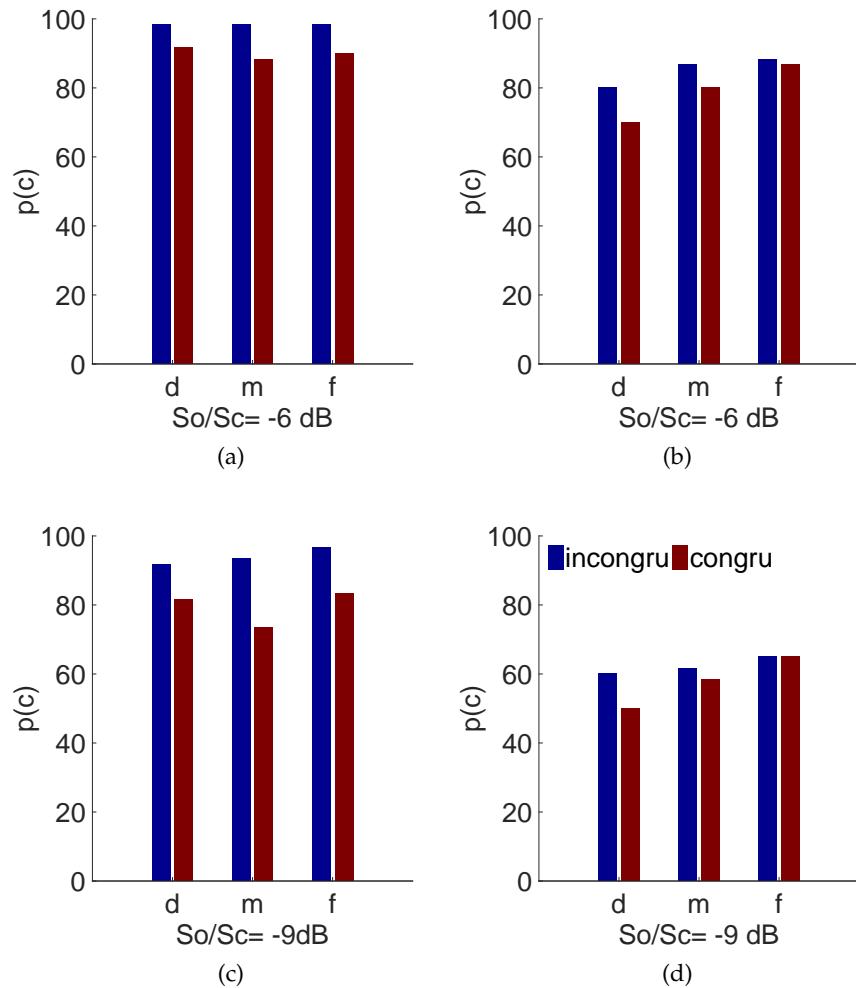


FIGURE 56 : Pourcentage de réponses correctes en fonction de la position du son cible et de la congruence ; En considérant  $\frac{So}{Sb} = -6\text{dB}$  (a, b),  $\frac{So}{Sb} = -9\text{dB}$  (c, d), des scènes longues (b, d) et des scènes courtes (a, c).

- plus la position est éloignée dans le temps et plus l'AI se réduit. On constate ainsi l'inverse de notre hypothèse de départ. Il semblerait que plus le sujet a le temps de se familiariser avec l'environnement, et moins le caractère incongru d'un son favorise son identification.

## BIBLIOGRAPHIE

---

- Adams, Mags D., Neil S. Bruce, William J. Davies, Rebecca Cain, Paul Jennings, Angus Carlyle, Peter Cusack, Ken I. Hume, and C.J. Plack (2008). "Soundwalking as a methodology for understanding soundscapes." In: *Proceedings of the Institute of Acoustics Spring Conference*. Vol. 30. 2. Reading, UK.
- Agus, Trevor R., Simon J. Thorpe, and Daniel Pressnitzer (2010). "Rapid formation of robust auditory memories : insights from noise." In: *Neuron* 66.4, pp. 610–618.
- Aletta, Francesco, Jian Kang, and Östen Axelsson (2016). "Soundscape descriptors and a conceptual framework for developing predictive soundscape models." In: *Landscape and Urban Planning* 149, pp. 65–74.
- Anderson, John R. (1991). "The adaptive nature of human categorization." In: *Psychological Review* 98.3, p. 409.
- Axelsson, Östen, Birgitta Berglund, and Mats E. Nilsson (2005). "Soundscape assessment." In: *The Journal of the Acoustical Society of America* 117.4, pp. 2591–2592.
- Axelsson, Östen, Mats E. Nilsson, and Birgitta Berglund (2010). "A principal components model of soundscape perception." In: *The Journal of the Acoustical Society of America* 128.5, pp. 2836–2846.
- Ballas, James A. and James H. Howard (1987). "Interpreting the language of environmental sounds." In: *Environment and behavior* 19.1, pp. 91–114.
- Ballas, James A. and Timothy Mullins (1991). "Effects of context on the identification of everyday sounds." In: *Human performance* 4.3, pp. 199–219.
- Barsalou, Lawrence W. (1983). "Ad hoc categories." In: *Memory & cognition* 11.3, pp. 211–227.
- (1999). "Perceptions of perceptual symbols." In: *Behavioral and brain sciences* 22.04, pp. 637–660.
- (2003a). "Abstraction in perceptual symbol systems." In: *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 358.1435, pp. 1177–1187.
- (2003b). "Situated simulation in the human conceptual system." In: *Language and cognitive processes* 18.5-6, pp. 513–562.
- (2008). "Grounded cognition." In: *Annual review of Psychology* 59, pp. 617–645.
- (2010). "Grounded cognition : Past, present, and future." In: *Topics in cognitive science* 2.4, pp. 716–724.
- (2016). "On staying grounded and avoiding quixotic dead ends." In: *Psychonomic bulletin & review*, pp. 1–21.

- Barsalou, Lawrence W. and Katja Wiemer-Hastings (2005). "Grounding cognition : The role of perception and action in memory, language, and thought." In: MIT Press, Cambridge, MA. Chap. Situating abstract concepts, pp. 129–163.
- Barsalou, Lawrence W., Janellen Huttenlocher, and Koen Lamberts (1998). "Basing categorization on individuals and events." In: *Cognitive Psychology* 36.3, pp. 203–272.
- Beaumont, Jacques, Stéphen Lesaux, Benjamin Robin, Jean-Dominique Polack, Cristina Pronello, Christine Arras, and Laurent Droin (2004). "Pertinence des descripteurs d'ambiance sonore urbaine." In: *Acoustique et techniques*.
- Bendavid, R. and M. Chasles-Parot (2014). *Les Français et les Nuisances Sonores (French and Noise Nuisances)*. Tech. rep. Paris, France: Institut français d'opinion publique (IFOP), p. 24.
- Benetos, Emmanouil, Grégoire Lafay, and Mathieu Lagrange (2016a). *DCASE2016 Task 2 Baseline*. Tech. rep. DCASE Challenge.
- Benetos, Emmanouil, Grégoire Lafay, Mathieu Lagrange, and Mark D. Plumbley (2016b). "Detection of overlapping acoustic events using a temporally-constrained probabilistic model." In: *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE. Shanghai, China, pp. 6450–6454.
- Benetos, Emmanouil, Grégoire Lafay, and Mathieu Lagrange (2016c). "Polyphonic Sound Event Tracking using Linear Dynamical Systems." In: *IEEE/ACM Transactions on audio, speech and language processing, Special issue on Sound Scene and Event Analysis*, (accepted).
- Bilger, Robert C., J.M. Nuetzel, W.M. Rabinowitz, and C. Rzeczkowski (1984). "Standardization of a test of speech perception in noise." In: *Journal of Speech, Language, and Hearing Research* 27.1, pp. 32–48.
- Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer (2011). "D<sup>3</sup> data-driven documents." In: *IEEE Transactions on Visualization and Computer Graphics* 17.12, pp. 2301–2309.
- Botteldooren, Dick and Bert De Coensel (2009). "The role of saliency, attention and source identification in soundscape research." In: *Proceedings of the 38th International Congress and Exposition on Noise Control Engineering (InterNoise)*. Ottawa, Canada.
- Botteldooren, Dick, Bert De Coensel, and Tom De Muer (2006). "The temporal structure of urban soundscapes." In: *Journal of sound and vibration* 292.1, pp. 105–123.
- Bregman, Albert S. (1994). *Auditory scene analysis : The perceptual organization of sound*. MIT press, Cambridge, MA.
- Brocolini, Laurent, Catherine Lavandier, Catherine Marquis-Favre, Matthias Quoy, and Mathieu Lavandier (2012). "Prediction and explanation of sound quality indicators by multiple linear regressions and artificial neural networks." In: *Proceedings of the Acoustics Conference*. Nantes, France.

- Brown, A.L., Jian Kang, and Truls Gjestland (2011). "Towards standardization in soundscape preference assessment." In: *Applied Acoustics* 72.6, pp. 387–392.
- Bruce, Neil S. and William J. Davies (2014). "The effects of expectation on the perception of soundscapes." In: *Applied Acoustics* 85, pp. 1–11.
- Bruce, Neil S., William J. Davies, and Mags D. Adams (2009). "Development of a soundscape simulator tool." In: *Proceedings of the 38th International Congress and Exposition on Noise Control Engineering (InterNoise)*. Ottawa, Canada.
- Caclin, Anne, Stephen McAdams, Bennett K. Smith, and Suzanne Winsberg (2005). "Acoustic correlates of timbre space dimensions : A confirmatory study using synthetic tonesa)." In: *The Journal of the Acoustical Society of America* 118.1, pp. 471–482.
- Cain, Rebecca, Paul Jennings, and John E.W. Poxon (2013). "The development and application of the emotional dimensions of a soundscape." In: *Applied Acoustics* 74.2, pp. 232–239.
- Carlyon, Robert P. (2004). "How the brain separates sounds." In: *Trends in cognitive sciences* 8.10, pp. 465–471.
- Carlyon, Robert P., John Deeks, Dennis Norris, and Sally Butterfield (2002). "The continuity illusion and vowel identification." In: *Acta Acustica United with Acustica* 88.3, pp. 408–415.
- Chauhan, Sameer, Sharang Phadke, and Christian Sherland (2013). *Event detection and classification*. Tech. rep. DCASE Challenge.
- Choi, Inkyu, Kisoo Kwon, Soo Hyun Bae, and Nam Soo Kim (2016). *DNN-Based Sound Event Detection with Exemplar-Based Approach for Noise Reduction*. Tech. rep. DCASE Challenge.
- Cusack, Rhodri, John Deeks, Genevieve Aikman, and Robert P. Carlyon (2004). "Effects of location, frequency region, and time course of selective attention on auditory scene analysis." In: *Journal of Experimental Psychology : Human Perception and Performance* 30.4, p. 643.
- Dannenbring, Gary L. (1976). "Perceived auditory continuity with alternately rising and falling frequency transitions." In: *Canadian Journal of Psychology* 30.2, p. 99.
- Davies, William J. et al. (2009). "The positive soundscape project : a synthesis of results from many disciplines." In: *Proceedings of the 38th International Congress on Noise Control Engineering (InterNoise)*. Ottawa, Canada.
- Davies, William J., Mags D. Adams, Neil S. Bruce, Rebecca Cain, Angus Carlyle, Peter Cusack, Deborah A. Hall, Ken I. Hume, Amy Irwin, Paul Jennings, et al. (2013). "Perception of soundscapes : An interdisciplinary approach." In: *Applied acoustics* 74.2, pp. 224–231.

- Davies, William J., Neil S. Bruce, and Jesse E. Murphy (2014). "Sound-scape reproduction and synthesis." In: *Acta Acustica United with Acustica* 100.2, pp. 285–292.
- Davis, Steven and Paul Mermelstein (1980). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences." In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28.4, pp. 357–366.
- Davis, Tyler and Bradley C Love (2010). "Memory for category information is idealized through contrast with competing options." In: *Psychological Science* 21.2, pp. 234–242.
- De Coensel, Bert and Dick Botteldooren (2006). "The quiet rural soundscape and how to characterize it." In: *Acta Acustica united with Acustica* 92.6, pp. 887–897.
- (2010). "A model of saliency-based auditory attention to environmental sound." In: *Proceedings of the 20th International Congress on Acoustics (ICA)*. Sydney, Australia.
- De Coensel, Bert, Annelies Bockstaal, Luc Dekoninck, Dick Botteldooren, Brigitte Schulte-Fortkamp, Jian Kang, and Mats E. Nilsson (2010). "Application of a model for auditory attention to the design of urban soundscapes." In: *Proceedings of the 1ste European Acoustics Association (EAA-EuroRegio) : Congress on Sound and Vibration*. Slovenian Acoustical Society, pp. 1–6.
- De Coensel, Bert, Michiel Boes, Damiano Oldoni, and Dick Botteldooren (2013). "Characterizing the soundscape of tranquil urban spaces." In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America, pp. 040–052.
- Defréville, Boris, Catherine Lavandier, and Marc Laniray (2004). "Activity of urban sound sources." In: *Proceedings of the 18th International Congress in Acoustics (ICA)*. Kyoto, Japan.
- Delaitre, Pauline, Catherine Lavandier, Caroline Cance, and Jean Pruvost (2012). "What is the Definition for the French Word calme in the European Directive Related to "Quiet Areas"? A Lexicographic Study from the 16th Century Until Today." In: *Acta Acustica united with Acustica* 98.5, pp. 734–740.
- Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets." In: *Journal of Machine Learning Research* 7, pp. 1–30.
- Devergie, Aymeric (2006). "Relations entre Perception Globale et Composition de Séquences Sonores." MA thesis. IRCAM, Paris VI UPMC.
- Diment, Aleksandr, Toni Heittola, and Tuomas Virtanen (2013a). "Sound event detection for office live and office synthetic AASP challenge." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.

- (2013b). *Sound event detection for office live and office synthetic AASP challenge*. Tech. rep. DCASE Challenge.
- Dubois, Danièle (1991). *Sémantique et cognition : catégories, prototypes, typicalité*. Diffusion, Presses du CNRS.
- (2000). “Categories as acts of meaning : The case of categories in olfaction and audition.” In: *Cognitive science quarterly* 1.1, pp. 35–68.
- Dubois, Danièle, Catherine Guastavino, and Manon Raimbault (2006). “A cognitive approach to urban soundscapes : Using verbal data to access everyday life auditory categories.” In: *Acta acustica united with acustica* 92.6, pp. 865–874.
- Dyson, Benjamin J. and Claude Alain (2004). “Representation of concurrent acoustic objects in primary auditory cortex.” In: *The Journal of the Acoustical Society of America* 115.1, pp. 280–288.
- Elhilali, Mounya, Juanjuan Xiang, Shihab A. Shamma, and Jonathan Z. Simon (2009). “Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene.” In: *PLoS Biol* 7.6, e1000129.
- Fiebig, André, Sandro Guidati, and Alexander Goehrke (2009). “Psychoacoustic evaluation of traffic noise.” In: *Proceedings of the NAG-/DAGA International Conference*. Rotterdam, Netherlands.
- Finney, Nathaniel and Jordi Janer (2010). “Soundscape generation for virtual environments using community-provided audio databases.” In: *Proceedings of the W3C Workshop : Augmented Reality on the Web*.
- Fishman, Yonatan I., David H. Reser, Joseph C. Arezzo, and Mitchell Steinschneider (2001). “Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey.” In: *Hearing research* 151.1, pp. 167–187.
- Fried, Lisbeth S. and Keith J. Holyoak (1984). “Induction of category distributions : A framework for classification learning.” In: *Journal of Experimental Psychology : Learning, Memory, and Cognition* 10.2, pp. 234–257.
- Galbrun, Laurent and Tahrir Ali (2012). “Perceptual assessment of water sounds for road traffic noise masking.” In: *Proceedings of Meetings on Acoustics*. Acoustical Society of America.
- García Pérez, Igone, Itziar Aspuru Soloaga, Karmele Herranz-Pascual, and Ibone García-Borreguero (2012). “Validation of an indicator for the assessment of the environmental sound in urban places.” In: *Proceedings of the Euronoise conference*. Prague, Czech Republic.
- Gaver, William W. (1993a). “How do we hear in the world ? Explorations in ecological acoustics.” In: *Ecological psychology* 5.4, pp. 285–313.
- (1993b). “What in the world do we hear ? : An ecological approach to auditory event perception.” In: *Ecological psychology* 5.1, pp. 1–29.

- Gemmeke, Jort F., Lode Vuggen, B. Vanrumste, and H. Van hamme (2013a). "An exemplar-based NMF approach to audio event detection." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.
- (2013b). *An exemplar-based NMF approach to audio event detection*. Tech. rep. DCASE Challenge.
- Giannoulis, Dimitrios, Dan Stowell, Emmanouil Benetos, Mathias Ros-signal, Mathieu Lagrange, and Mark D. Plumbley (2013a). "A da-tabase and challenge for acoustic scene classification and event detection." In: *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*. Marrakech, Morocco.
- Giannoulis, Dimitrios, Emmanouil Benetos, Dan Stowell, Mathias Ros-signal, Mathieu Lagrange, and Mark D. Plumbley (2013b). "De-tection and classification of acoustic scenes and events : An ieee aasp challenge." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.
- Giannoulis, Panagiotis, Gerasimos Potamianos, Petros Maragos, and Athanasios Katsamanis (2016). *Improved Dictionary Selection and Detection Schemes in Sparse-Cnmf-Based Overlapping Acoustic Event Detection*. Tech. rep. DCASE Challenge.
- Gibson, James J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
- (1978). "The ecological approach to the visual perception of pic-tures." In: *Leonardo* 11.3, pp. 227–235.
- Gille, Laure-Anne and Catherine Marquis-Favre (2016). "Dose-effect relationships for annoyance due to road traffic noise : Multi-level regression and consideration of noise sensitivity." In: *The Journal of the Acoustical Society of America* 139.4, pp. 2070–2070.
- Gille, Laure-Anne, Catherine Marquis-Favre, and Achim Klein (2016a). "Noise Annoyance Due To Urban Road Traffic with Powered-Two-Wheelers : Quiet Periods, Order and Number of Vehicles." In: *Acta Acustica united with Acustica* 102.3, pp. 474–487.
- Gille, Laure-Anne, Catherine Marquis-Favre, and Julien Morel (2016b). "Testing of the European Union exposure-response relationships and annoyance equivalents model for annoyance due to transpor-tation noises : The need of revised exposure-response relation-ships and annoyance equivalents model." In: *Environment Interna-tional* 94, pp. 83–94.
- Gloaguen, Jean-Rémy, Arnaud Can, Mathieu Lagrange, and Jean-François Petiot (2016). "Estimating traffic noise levels using acoustic moni-toring : a preliminary study." In: *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events*. Budapest, Hungary.
- Goldstone, Robert L. and Lawrence W. Barsalou (1998). "Reuniting perception and conception." In: *Cognition* 65.2, pp. 231–262.

- Goldstone, Robert L. and Alan Kersten (2003). "Concepts and categorization." In: *Handbook of psychology*.
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples." In: *arXiv preprint arXiv:1412.6572*.
- Gozalo, G. Rey, J. Trujillo Carmona, J.M. Barrigón Morillas, R. Vílchez-Gómez, and V. Gómez Escobar (2015). "Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes." In: *Applied Acoustics* 97, pp. 1–10.
- Guastavino, Catherine (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*). " PhD thesis. Paris, France: Université Paris VI UPMC.
- (2006). "The ideal urban soundscape : Investigating the sound quality of French cities." In: *Acta Acustica united with Acustica* 92.6, pp. 945–951.
  - (2007). "Categorization of environmental sounds." In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 61.1, p. 54.
- Guastavino, Catherine and Pascale Cheminée (2003). "Une approche psycholinguistique de la perception des basses fréquences : Conceptualisations en langue, représentations cognitives et validité écologique." In: *Psychologie française* 48.4, pp. 91–101.
- Guastavino, Catherine and Brian F.G. Katz (2004). "Perceptual evaluation of multi-dimensional spatial audio reproduction." In: *The Journal of the Acoustical Society of America* 116.2, pp. 1105–1115.
- Guastavino, Catherine, Brian F.G. Katz, Jean-Dominique Polack, Daniel J. Levitin, and Daniele Dubois (2005). "Ecological validity of soundscape reproduction." In: *Acta Acustica united with Acustica* 91.2, pp. 333–341.
- Guillén, José Domingo and Isabel López Barrio (2007). "Importance of personal, attitudinal and contextual variables in the assessment of pleasantness of the urban sound environment." In: *Proceedings of the 19th International Congress on Acoustics (ICA)*. Madrid, Spain.
- Gutierrez-Arriola, Juana M., Rubén Fraile, Alexander Camacho, Thibaut Durand, Jaime L. Jarrín, and Shirley R. Mendoza (2016). *Synthetic Sound Event Detection Based on MFCC*. Tech. rep. DCASE Challenge.
- Guyot, F., M. Castellengo, and B. Fabre (1997). "Catégorisation et Cognition : De la Perception au Discours." In: Paris, France: Édition Kimé. Chap. A study of the categorization of an everyday sound set, pp. 41–58.
- Guyot, F., Chrysanthy Nathanail, Francois Montignies, and B. Masson (2005). "Urban sound environment quality through a physical and perceptive classification of sound sources : A cross-cultural

- study." In: *Proceedings of the 4th European Congress on Acoustics*. Budapest, Hungary.
- Gygi, Brian and Valeriy Shafiro (2011). "The incongruity advantage for environmental sounds presented in natural auditory scenes." In: *Journal of Experimental Psychology : Human Perception and Performance* 37.2, p. 551.
- Gygi, Brian, Gary R. Kidd, and Charles S. Watson (2007). "Similarity and categorization of environmental sounds." In: *Perception & psychophysics* 69.6, pp. 839–855.
- Hall, Deborah A., Amy Irwin, Mark Edmondson-Jones, Scott Phillips, and John E.W. Poxon (2013). "An exploratory evaluation of perceptual, psychoacoustic and acoustical properties of urban soundscapes." In: *Applied Acoustics* 74.2, pp. 248–254.
- Hayashi, Tomoki, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda (2016). *Bidirectional LSTM-HMM Hybrid System for Polyphonic Sound Event Detection*. Tech. rep. DCASE Challenge.
- Hintzman, Douglas L (1986). "Schema abstraction in a multiple-trace memory model." In: *Psychological Review* 93.4, pp. 411–428.
- Hong, Joo Young and Jin Yong Jeon (2013). "Designing sound and visual components for enhancement of urban soundscapes." In: *The Journal of the Acoustical Society of America* 134.3, pp. 2026–2036.
- Houdé, Olivier, Daniel Kayser, Olivier Koenig, Joëlle Proust, and François Rastier (1998). *Vocabulaire de sciences cognitives*. Presses Universitaires de France, Paris.
- Houix, Olivier (2003). "Catégorisation auditive des sources sonores, (*Sound sources Categorization*)."  
PhD thesis. Le Mans, France: Université du Maine.
- Houix, Olivier, Guillaume Lemaitre, Nicolas Misdariis, Patrick Susini, and Isabel Urdapilleta (2012). "A lexical analysis of environmental sound categories." In: *Journal of Experimental Psychology : Applied* 18.1, pp. 52–80.
- Hume, Ken I. and Mujthaba Ahtamad (2013). "Physiological responses to and subjective estimates of soundscape elements." In: *Applied Acoustics* 74.2, pp. 275–281.
- ISO 12913-1 :2014 acoustics-soundscape-part 1 : definition and conceptual framework* (2013). International Organization for Standardization (ISO), Genève.
- Jeon, Jin Yong, Pyoung Jik Lee, Joo Young Hong, and Densil Cabrera (2011). "Non-auditory factors affecting urban soundscape evaluation." In: *The Journal of the Acoustical Society of America* 130.6, pp. 3761–3770.
- Jeon, Jin Yong, Joo Young Hong, and Pyoung Jik Lee (2013). "Sound-walk approach to identify urban soundscapes individually." In: *The Journal of the Acoustical Society of America* 134.1, pp. 803–812.
- Kang, Jian (2006). *Urban sound environment*. CRC Press.

- Kang, Jian and M Zhang (2010). "Semantic differential analysis of the soundscape in urban open public spaces." In: *Building and environment* 45.1, pp. 150–157.
- Kiefer, Markus, Eun-Jin Sim, Bärbel Herrnberger, Jo Grothe, and Klaus Hoenig (2008). "The sound of concepts : four markers for a link between auditory and conceptual brain systems." In: *The Journal of Neuroscience* 28.47, pp. 12224–12230.
- Klein, Achim, Catherine Marquis-Favre, Rheinard Weber, and Arnaud Trollé (2015). "Spectral and modulation indices for annoyance-relevant features of urban road single-vehicle pass-by noises." In: *The Journal of the Acoustical Society of America* 137.3, pp. 1238–1250.
- Kocsis, Zsuzsanna, István Winkler, Alexandra Bendixen, and Claude Alain (2016). "Promoting the perception of two and three concurrent sound objects : An event-related potential study." In: *International Journal of Psychophysiology* 107, pp. 16–28.
- Kohonen, T. (1995). *Self-organizing maps*. Berlin : Springer-Verlag.
- Komatsu, Tatsuya, Takahiro Toizumi, Reishi Kondo, and Yuzo Senda (2016). *Acoustic Event Detection Method Using Semi-Supervised Non-Negative Matrix Factorization with a Mixture of Local Dictionaries*. Tech. rep. DCASE Challenge.
- Kong, Qiuqiang, Iwnoa Sobieraj, Wenwu Wang, and Mark D. Plumbley (2016). *Deep Neural Network Baseline for DCASE Challenge 2016*. Tech. rep. DCASE Challenge.
- Krumhansl, Carol L. (1978). "Concerning the applicability of geometric models to similarity data : The interrelationship between similarity and spatial density." In:
- Kuwano, Sonoko, Seiichiro Namba, Tohru Kato, and Jürgen Hellbrück (2003). "Memory of the loudness of sounds in relation to overall impression." In: *Acoustics Science and Technics* 4.24.
- Lafay, Grégoire (2013). "Caractérisation sémantique des scènes sonores environnementales : Étude paramétrique et perceptive d'un paradigme de synthèse séquentielle par corpus." MA thesis. IR-CAM, Paris VI UPMC.
- Lafay, Grégoire, Mathias Rossignol, Nicolas Misdariis, Mathieu Lagrange, and Jean-François Petiot (2014). "A new experimental approach for urban soundscape characterization based on sound manipulation : A pilot study." In: *Proceedings of the International Symposium on Musical Acoustics (ISMA)*. SFA. Le Mans, France.
- Lafay, Grégoire, Mathieu Lagrange, Emmanouil Benetos, Mathias Rossignol, and Axel Roebel (2016a). "A Morphological Model for Simulating Acoustic Scenes and Its Application to Sound Event Detection." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10, pp. 1854–1864.
- Lafay, Grégoire, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol (2016b). "Semantic browsing of sound databases wi-

- thout keywords." In: *Journal of the Audio Engineering Society* 64.9, pp. 628–635.
- Lagrange, Mathieu, Grégoire Lafay, Boris Defreville, and Jean-Julien Aucouturier (2015). "The bag-of-frames approach : a not so sufficient model for urban soundscapes." In: *The Journal of the Acoustical Society of America, express letter* 138.5, pp. 487–492.
- Lavandier, Catherine and Boris Defréville (2006). "The contribution of sound source characteristics in the assessment of urban soundscapes." In: *Acta Acustica united with Acustica* 92.6, pp. 912–921.
- Lecointre, Guillaume and Hervé Le Guyader (2006). *The tree of life : a phylogenetic classification*. Vol. 20. Harvard University Press.
- Lemaitre, Guillaume, Olivier Houix, Nicolas Misdariis, and Patrick Susini (2010). "Listener expertise and sound identification influence the categorization of environmental sounds." In: *Journal of Experimental Psychology : Applied* 16.1, p. 16.
- Leobon, A. (1986). "Analyse psycho-acoustique du paysage sonore urbain, (*Psychoacoustic analysis of urban soundscape*).". PhD thesis. Strasbourg, France: Université Louis Pasteur.
- Leondes, Cornelius T. (2002). "Expert Systems." In: Burlington: Academic Press. Chap. Preface, pp. xxiii–xxiv.
- Leshinskaya, Anna and Alfonso Caramazza (2016). "For a cognitive neuroscience of concepts : Moving beyond the grounding issue." In: *Psychonomic bulletin & review* 23.4, pp. 991–1001.
- Lostanlen, Vincent, Grégoire Lafay, Joakim Anden, and Mathieu Lagrange (2016). "Auditory Scene Similarity Retrieval and Classification with Relevance-based Quantization of Scattering Features." In: *IEEE/ACM Transactions on audio, speech and language processing, Special issue on Sound Scene and Event Analysis, (AQ)*.
- Ludwig, Wittgenstein (1953). *Philosophical investigations*. New York, NY : Macmillan.
- Macmillan, Neil A. and C. Douglas Creelman (2004). *Detection theory : A user's guide*. Psychology press.
- Maffiolo, Valérie (1997). *Méthodes d'approche de l'environnement sonore urbain*. Tech. rep. Paris, France: Mairie de Paris, Direction de la protection de l'environnement, SPAAS.
- (1999). "De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain, (*Semantic and acoustical characterisation of the sound quality of urban environment*).". PhD thesis. Le Mans, France: Université du Mans.
- Marcell, Michael M., Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers (2000). "Confrontation naming of environmental sounds." In: *Journal of clinical and experimental neuropsychology* 22.6, pp. 830–864.
- Marquis-Favre, Catherine and Julien Morel (2015). "A simulated environment experiment on annoyance due to combined road traffic

- and industrial noises." In: *International journal of environmental research and public health* 12.7, pp. 8413–8433.
- Marquis-Favre, Catherine, E. Premat, D. Aubrée, and M. Vallet (2005a). "Noise and its effects : A review on qualitative aspects of sound. Part I : Notions and acoustic ratings." In: *Acta acustica united with acustica* 91.4, pp. 613–625.
- Marquis-Favre, Catherine, E. Premat, and D. Aubrée (2005b). "Noise and its effects : A review on qualitative aspects of sound. Part II : Noise and annoyance." In: *Acta acustica united with acustica* 91.4, pp. 626–642.
- Martin, Alex (2001). "Handbook of functional neuroimaging of cognition." In: MIT Press, Cambridge, MA. Chap. Functional neuroimaging of semantic memory, pp. 149–190.
- McAdams, Stephen and Emmanuel Bigand (1994). *Penser les sons : psychologie cognitive de l'audition*. Presses Universitaires de France, Paris.
- McCloskey, Michael E. and Sam Glucksberg (1978). "Natural categories : Well defined or fuzzy sets?" In: *Memory & Cognition* 6.4, pp. 462–472.
- McDermott, Josh H. and Eero P. Simoncelli (2011). "Sound texture perception via statistics of the auditory periphery : evidence from sound synthesis." In: *Neuron* 71.5, pp. 926–940.
- McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli (2013). "Summary statistics in auditory perception." In: *Nature neuroscience* 16.4, pp. 493–498.
- Medin, Douglas L. and Marguerite M. Schaffer (1978). "Context theory of classification learning." In: *Psychological review* 85.3, p. 207.
- Memoli, Gianluca, Alan Bloomfield, and Max Dixon (2008). "Soundscape characterization in selected areas of Central London." In: *Proceedings of Meetings on Acoustics*. Vol. 123. 5. Acoustical Society of America, p. 3811.
- Meng, Qi, Jian Kang, and Hong Jin (2013). "Field study on the influence of spatial and environmental characteristics on the evaluation of subjective loudness and acoustic comfort in underground shopping streets." In: *Applied Acoustics* 74.8, pp. 1001–1009.
- Mervis, Carolyn B. and Eleanor Rosch (1981). "Categorization of natural objects." In: *Annual review of psychology* 32.1, pp. 89–115.
- Miedema, Henk M.E. (2004). "Relationship between exposure to multiple noise sources and noise annoyance." In: *The Journal of the Acoustical Society of America* 116.2, pp. 949–957.
- Miedema, Henk M.E. and C.G. Oudshoorn (2001). "Annoyance from transportation noise : relationships with exposure metrics DNL and DENL and their confidence intervals." In: *Environmental health perspectives* 109.4, p. 409.
- Misra, Ananya, Perry R. Cook, and Ge Wang (2006). "A new paradigm for sound design." In: *Proceedings of the 9th Internatio-*

- nal Conference on Digital Audio Effects (DAFx)*. Montreal, Canada, pp. 319–324.
- Misra, Ananya, Ge Wang, and Perry Cook (2007). “Musical Tapestry : Re-composing Natural Sound.” In: *Journal of New Music Research* 36.4, pp. 241–250.
- Moore, Brian C.J. (1973). “Frequency difference limens for short-duration tones.” In: *The Journal of the Acoustical Society of America* 54.3, pp. 610–619.
- Morel, Julien, Catherine Marquis-Favre, and L-A Gille (2016). “Noise annoyance assessment of various urban road vehicle pass-by noises in isolation and combined with industrial noise : A laboratory study.” In: *Applied Acoustics* 101, pp. 47–57.
- Müller, Meinard (2007). *Information Retrieval for Music and Motion*. Se-caucus, NJ, USA: Springer-Verlag New York, Inc.
- Neisser, Ulric (1967). *Cognitive psychology*. (Reprinted as *Cognitive psychology : Classic edition*. Psychology Press, 2014). New York: Appleton-Century-Crofts.
- (1976). *Cognition and reality : principles and implications of cognitive psychology*. W.H. Freeman.
- Nelken, Israel (2004). “Processing of complex stimuli and natural scenes in the auditory cortex.” In: *Current opinion in neurobiology* 14.4, pp. 474–480.
- Nelken, Israel and Omer Bar-Yosef (2008). “Neurons and objects : the case of auditory cortex.” In: *Frontiers in neuroscience* 2.1, p. 107.
- Nelken, Israel and Alain de Cheveigné (2013). “An ear for statistics.” In: *Nature neuroscience* 16.4, pp. 381–382.
- Ness, Steven R., Helena Symonds, Paul Spong, and George Tzanetakis (2013). “The Orchieve : Data mining a massive bioacoustic archive.” In: *Proceedings of the 1st International Workshop on Machine Learning for Bioacoustics*. Atlanta, USA.
- Nguyen, Anh, Jason Yosinski, and Jeff Clune (2015). “Deep neural networks are easily fooled : High confidence predictions for unrecognizable images.” In: *Proceedings of the 1st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, USA, pp. 427–436.
- Nielbo, Frederik L., Daniel Steele, and Catherine Guastavino (2013). “Investigating soundscape affordances through activity appropriateness.” In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America, pp. 040–059.
- Niessen, Maria E., Leendert Van Maanen, and Tjeerd C. Andringa (2008). “Disambiguating sound through context.” In: *International Journal of Semantic Computing* 2.03, pp. 327–341.
- Niessen, Maria E., Caroline Cance, and Danièle Dubois (2010). “Categories for soundscape : toward a hybrid classification.” In: *Proceedings of the 39th International Congress and Exposition on Noise*

- Control Engineering (InterNoise)*. Vol. 2010. 5. Lisbon, Portugal, pp. 5816–5829.
- Niessen, Maria E., Tim L.M. Van Kasteren, and Andreas Merentitis (2013a). "Hierarchical modeling using automated sub-clustering for sound event recognition." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.
- (2013b). *Hierarchical sound event detection*. Tech. rep. DCASE Challenge.
- Nilsson, Mats E. (2007). "Soundscape quality in urban open spaces." In: *Proceedings of the 36th International Congress and Exposition on Noise Control Engineering (InterNoise)*. Istanbul, Turkey.
- Nilsson, Mats E. and Birgitta Berglund (2006). "Soundscape quality in suburban green areas and city parks." In: *Acta Acustica united with Acustica* 92.6, pp. 903–911.
- Nilsson, Mats E., Dick Botteldooren, and Bert De Coensel (2007). "Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas." In: *Proceedings of the 19th International Congress on Acoustics (ICA)*. Madrid, Spain.
- Nogueira, Waldo, Guido Roma, and Perfecto Herrera (2013). *Automatic event classification using front end single end channel noise reduction, MFCC features and support vector machine classifier*. Tech. rep. DCASE Challenge.
- Nosofsky, Robert M. (1986). "Attention, similarity, and the identification–categorization relationship." In: *Journal of experimental psychology : General* 115.1, p. 39.
- (1992). "Similarity scaling and cognitive process models." In: *Annual review of Psychology* 43.1, pp. 25–53.
- O'Neil, Cathy (2016). *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group (NY).
- Oldoni, Damiano, Bert De Coensel, Michiel Boes, Timothy Van Renterghem, and Dick Botteldooren (2012). "A computational auditory attention model for urban soundscape design." In: *Proceedings of the 41st International Congress and Exposition on Noise Control Engineering (InterNoise)*. New York, USA.
- Oldoni, Damiano, Bert De Coensel, Michiel Boes, Michaël Rademaker, Bernard De Baets, Timothy Van Renterghem, and Dick Botteldooren (2013). "A computational model of auditory attention for use in soundscape research." In: *The Journal of the Acoustical Society of America* 134.1, pp. 852–861.
- Palmeri, Thomas J. and Robert M. Nosofsky (1995). "Recognition memory for exceptions to the category rule." In: *Journal of Experimental Psychology : Learning, Memory, and Cognition* 21.3, p. 548.

- Parizet, Etienne and Vincent Koehl (2012). "Application of free sorting tasks to sound quality experiments." In: *Applied Acoustics* 73.1, pp. 61–65.
- Park, Tae Hong, Johnathan Turner, Michael Musick, Jun Hee Lee, Christopher Jacoby, Charlie Mydlarz, and Justin Salomon (2014). "Sensing Urban Soundscapes." In: *Proceedings of the EDBT/ICDT Workshops*, pp. 375–382.
- Payne, Sarah R. (2013). "The production of a perceived restorativeness soundscape scale." In: *Applied Acoustics* 74.2, pp. 255–263.
- Phan, Huy, Lars Hertel, Marco Maass, Philipp Koch, and Alfred Mertins (2016). *Car-Forest : Joint Classification-Regression Decision Forests for Overlapping Audio Event Detection*. Tech. rep. DCASE Challenge.
- Pheasant, R.J., G.R. Watts, and K.V. Horoshenkov (2009). "Validation of a tranquillity rating prediction tool." In: *Acta Acustica united with Acustica* 95.6, pp. 1024–1031.
- Pheasant, Robert, Kirill Horoshenkov, Greg Watts, and Brendan Barrett (2008). "The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places ?" In: *The Journal of the Acoustical Society of America* 123.3, pp. 1446–1457.
- Pijanowski, Bryan C., Almo Farina, StuartH. Gage, Sarah L. Dumyahn, and Bernie L. Krause (2011). "What is soundscape ecology ? An introduction and overview of an emerging new science." In: *Landscape Ecology* 26.9, pp. 1213–1232.
- Pikrakis, Aggelos and Yannis Kopsinis (2016). *Dictionary Learning Assisted Template Matching for Audio Event Detection (Legato)*. Tech. rep. DCASE Challenge.
- Poeppel, David (2003). "The analysis of speech in different temporal integration windows : cerebral lateralization as asymmetric sampling in time." In: *Speech communication* 41.1, pp. 245–255.
- Polack, Jean-Dominique, Jacques Beaumont, Christine Arras, Mikael Zekri, and Benjamin Robin (2008). "Perceptive relevance of soundscape descriptors : a morpho-typological approach." In: *Journal of the Acoustical Society of America* 123.5, p. 3810.
- Poliner, Graham E. and Daniel P.W. Ellis (2007). "A discriminative model for polyphonic piano transcription." In: *EURASIP Journal on Applied Signal Processing* 2007.1, pp. 154–154.
- Rabiner, Lawrence R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." In: *Proceedings of the IEEE* 77.2, pp. 257–286.
- Rabiner, Lawrence R. and Biing-Hwang Juang (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.
- Raimbault, Manon (2002). "Simulation des ambiances sonores urbaines : intégration des aspects qualitatifs, *Urban soundscape simulation* :

- focusing on qualitative aspect).*" PhD thesis. Nantes, France: Université de Nantes - Ecole polytechnique de Nantes.
- (2006). "Qualitative judgements of urban soundscapes : Questioning questionnaires and semantic scales." In: *Acta acustica united with acustica* 92.6, pp. 929–937.
  - Raimbault, Manon and Daniele Dubois (2005). "Urban soundscapes : Experiences and knowledge." In: *Cities* 22.5, pp. 339–350.
  - Reed, Stephen K. (1972). "Pattern recognition and categorization." In: *Cognitive psychology* 3.3, pp. 382–407.
  - Ribeiro, Carlos, Celine Anselme, Fanny Mietlicki, Bruno Vincent, Raphael Da Silva, and Piotr Gaudibert (2013). "At the heart of Harmonica project : the Common Noise Index (CNI)." In: *Proceedings of the 42nd International Congress on Noise Control Engineering (InterNoise)*. Innsbruck, Austria.
  - Ricciardi, Paola, Pauline Delaitre, Catherine Lavandier, Francesca Torschia, and Pierre Aumont (2015). "Sound quality indicators for urban places in Paris cross-validated by Milan data." In: *The Journal of the Acoustical Society of America* 138.4, pp. 2337–2348.
  - Roma, Guido, Waldo Nogueira, and Perfecto Herrera (2013). "Recurrence quantification analysis features for environmental sound recognition." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.
  - Rosch, Eleanor (1975). "Cognitive representations of semantic categories." In: *Journal of experimental psychology : General* 104.3, p. 192.
  - Rosch, Eleanor and Barbara B Lloyd (1974). *Human communication : Theoretical perspectives*. Halsted Press, New York.
  - Rosch, Eleanor and Barbara B. Lloyd (1978). *Cognition and categorization*. Hillsdale, New Jersey.
  - Rosch, Eleanor and Carolyn B. Mervis (1975). "Family resemblances : Studies in the internal structure of categories." In: *Cognitive Psychology* 7, pp. 573–605.
  - Rosch, Eleanor, Carol Simpson, and R. Scott Miller (1976). "Structural bases of typicality effects." In: *Journal of Experimental Psychology : Human perception and performance* 2.4, p. 491.
  - Rossignol, Mathias, Mathieu Lagrange, Grégoire Lafay, and Emmanouil Benetos (2015a). "Alternate level clustering for drum transcription." In: *Proceedings of the European Signal Processing Conference (EUSIPCO)*. IEEE. Nice, France, pp. 2023–2027.
  - Rossignol, Mathias, Grégoire Lafay, Mathieu Lagrange, and Nicolas Misdariis (2015b). "SimScene : a web-based acoustic scenes simulator." In: *Proceedings of the Web Audio Conference (WAC)*. IRCAM. Paris, France.
  - Rychtáriková, Monika and Gerrit Vermeir (2013). "Soundscape categorization on the basis of objective acoustical parameters." In: *Applied Acoustics* 74.2, pp. 240–247.

- Saint-Arnaud, Nicolas (1995). "Classification of sound textures." MA thesis. Massachusetts Institute of Technology.
- Salamon, Justin, Christopher Jacoby, and J. P. Bello (2014). "A Dataset and Taxonomy for Urban Sound Research." In: *Proceedings of the 22st ACM International Conference on Multimedia*. Orlando, FL, USA.
- Schafer, R.M. (1969). *The New Soundscape : A Handbook for the Modern Music Teacher*. Ontario : Berandol Music Limited.
- (1977). *The Tuning of the World*. Borzoi book. (Reprinted as *Our Sonic Environment and the Soundscape : The Tuning of the World*. Destiny Books, 1994). New York: Knopf.
- Schirosa, Mattia, Jordi Janer, Stefan Kersten, and Gerard Roma (2010). "A system for soundscape generation, composition and streaming." In: *Proceedings of the XVII CIM-Colloquium of Musical Informatics*.
- Schröder, J., B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze (2013a). *Acoustic event detection using signal enhancement and spectro-temporal feature extraction*. Tech. rep. DCASE Challenge.
- (2013b). "On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events'." In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA.
- Schulte-Fortkamp, Brigitte (2013). "Soundscape-focusing on resources." In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America, pp. 040–117.
- Schulte-Fortkamp, Brigitte and André Fiebig (2006). "Soundscape analysis in a residential area : An evaluation of noise and people's mind." In: *Acta Acustica united with Acustica* 92.6, pp. 875–880.
- Schulte-Fortkamp, Brigitte and Jian Kang (2010). "Soundscape research in networking across countries : COST Action TD0804." In: *The Journal of the Acoustical Society of America* 127.3, pp. 1801–1801.
- Schulte-Fortkamp, Brigitte, Bennett M. Brooks, and Wade R. Bray (2007). "Soundscape : An Approach to Rely on Human Perception and Expertise in the Post-Modern Community Noise Era." In: *Acoustics Today* 3.1, pp. 7–15.
- Schwartz, Jean-Luc, Nicolas Grimault, Jean-Michel Hupé, Brian C.J. Moore, and Daniel Pressnitzer (2012). "Multistability in perception : binding sensory modalities an overview." In: *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 367.1591, pp. 896–905.
- Schwarz, Diemo (2011). "State of the art in sound texture synthesis." In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*. Paris, France.

- Schyns, Philippe G. (1998). "Diagnostic recognition : task constraints, object information, and their interactions." In: *Cognition* 67.1, pp. 147–179.
- Snyder, Joel S. and Claude Alain (2007). "Toward a neurophysiological theory of auditory stream segregation." In: *Psychological bulletin* 133.5, p. 780.
- Southworth, Michael (1969). "The sonic environment of cities." In: *Environment and behavior* 1.1, p. 49.
- Spellman, Barbara A. (2015). "A short (personal) future history of Revolution 2.0." In: *Perspectives on Psychological Science* 10.6, pp. 886–899.
- Stansfeld, Stephen A., Birgitta Berglund, Charlotte Clark, Isabel Lopez-Barrio, Peter Fischer, Evy Öhrström, Mary M. Haines, Jenny Head, Staffan Hygge, Irene Van Kamp, et al. (2005). "Aircraft and road traffic noise and children's cognition and health : a cross-national study." In: *The Lancet* 365.9475, pp. 1942–1949.
- Stiefelhagen, Rainer, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan (2007). "Multimodal Technologies for Perception of Humans." In: vol. 4122. Lecture Notes in Computer Science. Springer Berlin Heidelberg. Chap. The CLEAR 2006 Evaluation, pp. 1–44.
- Stowell, Dan and Mark D. Plumbley (2013a). "Large-scale analysis of frequency modulation in birdsong databases." In: *Methods in Ecology and Evolution* 11.
- (2013b). "Segregating Event Streams and Noise with a Markov Renewal Process Model." In: *Journal of Machine Learning Research* 14, pp. 2213–2238.
- Stowell, Dan, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley (2015). "Detection and classification of acoustic scenes and events." In: *IEEE Transactions on Multimedia* 17.10, pp. 1733–1746.
- Sturm, Bob L. (2014). "A Simple Method to Determine if a Music Information Retrieval System is a "Horse"." In: *IEEE Transactions on Multimedia* 16.6, pp. 1636–1644.
- Sueur, Jérôme, Almo Farina, C. Bobryk, D. Llusia, J. McWilliam, and N. Pieretti (2014). "Ecology and acoustics : Emergent properties from community to landscape." In: Paris, France: Muséum national d'Histoire naturelle.
- Szeremeta, Bani and Paulo Henrique Trombetta Zannin (2009). "Analysis and evaluation of soundscapes in public parks through interviews and measurement of noise." In: *Science of the total environment* 407.24, pp. 6143–6149.
- Torija, Antonio J., Diego P. Ruiz, and A.F. Ramos-Ridao (2013). "Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-

- differential attributes." In: *The Journal of the Acoustical Society of America* 134.1, pp. 791–802.
- Trollé, Arnaud, Catherine Marquis-Favre, and Étienne Parizet (2015). "Perception and Annoyance Due to Vibrations in Dwellings Generated From Ground Transportation : A Review." In: *Journal of Low Frequency Noise, Vibration and Active Control* 34.4, pp. 413–457.
- Truax, Barry (1978). *Handbook for acoustic ecology*. (originally published by the world soundscape project). simon fraser university and ARC Publications.
- Tse, Man Sze, Chi Kwan Chau, Yat Sze Choy, Wai Keung Tsui, Chak Ngai Chan, and Shiu Keung Tang (2012). "Perception of urban park soundscape." In: *The Journal of the Acoustical Society of America* 131.4, pp. 2762–2771.
- Tversky, Amos (1977). "Features of similarity." In: *Psychological review* 84.4, p. 327.
- Tversky, Amos and Itamar Gati (1978). "Cognition and categorization." In: Hillsdale, New Jersey. Chap. Studies of similarity, pp. 79–98.
- Ullman, Shimon (1980). "Against direct perception." In: *Behavioral and Brain Sciences* 3.03, pp. 373–381.
- Valle, Andrea, Mattia Schirosa, and Vincenzo Lombardo (2009). "A framework for soundscape analysis and re-synthesis." In: *Proceedings of the 6th Sound and Music Computing Conference (SMC)*.
- Vanderveer, Nancy J. (1980). "Ecological acoustics : Human perception of environmental sounds." PhD thesis. ProQuest Information & Learning.
- Vincent, Emmanuel, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot (2014). "From blind to guided audio source separation : How models and side information can improve the separation of sound." In: *IEEE Signal Processing Magazine* 31.3, pp. 107–115.
- Vu, Toan H. and Jia-Ching Wang (2016). *Acoustic Scene and Event Recognition Using Recurrent Neural Networks*. Tech. rep. DCASE Challenge.
- Vuegen, Lode, B. Van Den Broeck, P. Karsmakers, Jort F. Gemmeke, B. Vanrumste, and H. Van hamme (2013). *A MFCC-GMM approach for event detection and classification*. Tech. rep. DCASE Challenge.
- Warren, Paige S., Madhusudan Katti, Michael Ermann, and Anthony Brazel (2006). "Urban bioacoustics : It's not just noise." In: *Animal behaviour* 71.3, pp. 491–502.
- Winkler, István and Erich Schröger (2015). "Auditory perceptual objects as generative models : Setting the stage for communication by sound." In: *Brain and language* 148, pp. 1–22.
- Winkler, Istvan, Susan L. Denham, and Israel Nelken (2009). "Modeling the auditory scene : predictive regularity representations and perceptual objects." In: *Trends in cognitive sciences* 13.12, pp. 532–540.

- Woloszyn, Philippe (1997). "Vers un simulateur des ambiances sonores urbaines." In: *Acoustique & Techniques* 8, pp. 17–19.
- Yabe, Hirooki, Mari Tervaniemi, Janne Sinkkonen, Minna Huotilainen, Risto J. Ilmoniemi, and Risto Näätänen (1998). "Temporal window of integration of auditory information in the human brain." In: *Psychophysiology* 35.5, pp. 615–619.
- Yang, Ming and Jian Kang (2013). "Psychoacoustical evaluation of natural and urban sounds in soundscapes." In: *The Journal of the Acoustical Society of America* 134.1, pp. 840–851.
- Yang, Wei and Jian Kang (2005). "Acoustic comfort evaluation in urban open public spaces." In: *Applied acoustics* 66.2, pp. 211–229.
- Yost, William A. (1994). *Fundamentals of hearing : An introduction*. Academic Press.
- Yu, Lei and Jian Kang (2009). "Modeling subjective evaluation of soundscape quality in urban open spaces : An artificial neural network approach." In: *The Journal of the Acoustical Society of America* 126.3, pp. 1163–1174.
- (2010). "Factors influencing the sound preference in urban open spaces." In: *Applied Acoustics* 71.7, pp. 622–633.
- Zwicker, Eberhard and Hugo Fastl (1990). *Psychoacoustics : Facts and models*. Berlin : Springer Verlag.