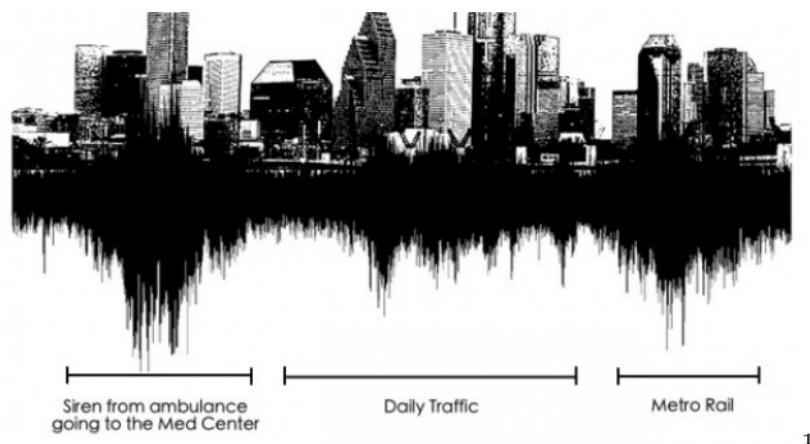


GRÉGOIRE LAFAY

MODÉLISATION DE SCÈNES SONORES
ENVIRONNEMENTALES

MODÉLISATION DE SCÈNES SONORES ENVIRONNEMENTALES

De l'analyse sensorielle à l'analyse automatique: une approche pluridisciplinaire



1

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

Décembre 2016

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

Grégoire Lafay : *Modélisation de scènes sonores environnementales, De l'analyse sensorielle à l'analyse automatique: une approche pluridisciplinaire*, © Décembre 2016

cite.
 cite.

— TOTO & TOTO

Dedicated to TOTO.

date – date

ABSTRACT

TODO

RÉSUMÉ

TODO

PUBLICATIONS

Analyse perceptive des scènes sonores environnementales

- 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*.
- Lafay, Grégoire, Mathieu Lagrange, Jean-François Petiot, Mathias Rossignol, and Nicolas Misdariis (2016a). “Investigating soundscapes perception through acoustic scenes simulation. Part I : simulation protocol presentation and case study.” In: *Acta acustica, (under revision)*.
- (2016b). “Investigating soundscapes perception through acoustic scenes simulation. Part II : complementary experiments.” In: *Acta acustica, (under revision)*.
- Lafay, Grégoire, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol (2016c). “Semantic browsing of sound databases without keywords.” In: *Journal of the Audio Engineering Society*.

Attention : This requires a separate run of `bibtex` for the `refsections`, e.g., `theseLafay20161-blx`, `theseLafay20162-blx` and `theseLafay20163-blx`.

Analyse automatique de scènes sonores environnementales

- 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, EL487–EL492.
- Lostanlen, Vincent, Grégoire Lafay, Joakim Anden, and Mathieu Lagrange (2016). “Object-based Auditory Scenes Similarity Retrieval and Classification With Wavelet Scattering.”

Simulation automatique de scènes sonores — Utilisation de scènes sonores simulées pour l'évaluation des algorithmes en détection automatique d'événements sonores

- Benetos, E., G. Lafay, M. Lagrange, and M. D. Plumbley (2016a). "Detection of overlapping acoustic events using a temporally-constrained probabilistic model." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6450–6454. DOI: [10.1109/ICASSP.2016.7472919](https://doi.org/10.1109/ICASSP.2016.7472919).
- Benetos, Emmanouil, Grégoire Lafay, and Mathieu Lagrange (2016b). "Polyphonic Sound Event Tracking using LDS."
- Lafay, G., M. Lagrange, E. Benetos, M. Rossignol, and A. Roebel (2016). "A Morphological Model for Simulating Acoustic Scenes and Its Application to Sound Event Detection." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* PP.99, pp. 1–1. ISSN: 2329-9290. DOI: [10.1109/TASLP.2016.2587218](https://doi.org/10.1109/TASLP.2016.2587218).
- Lagrange, Mathieu and Grégoire Lafay (2016). "Results of the DCASE 2016 challenge."
- Rossignol, Mathias, Mathieu Lagrange, Grégoire Lafay, and Emmanouil Benetos (2015a). "Alternate level clustering for drum transcription." In: *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, pp. 2023–2027.
- Rossignol, Mathias, Grégoire Lafay, Mathieu Lagrange, and Nicolas Misdariis (2015b). "SimScene : a web-based acoustic scenes simulator." In: *1st Web Audio Conference (WAC)*.

Cite

— *TOTO*

REMERCIEMENTS

TODO

TABLE DES MATIÈRES

(pas relue en tant que telle, corrections au fil du texte
si nécessaire)

I PRÉAMBULE	1
1 INTRODUCTION	3
2 MOTIVATION	7
II ANALYSE SENSORIELLE DE SCÈNES SONORES	9
3 ÉTAT DE L'ART	11
3.1 Le traitement de l'information auditive	11
3.1.1 La chaîne de traitement	11
3.1.2 Processus <i>Bottom-up</i> et processus <i>Top-down</i>	14
3.1.3 Perception et cognition	15
3.1.4 L'approche écologique	19
3.2 Représentation mentale de l'environnement sonore	21
3.2.1 La notion de catégorie	22
3.2.2 Le processus de catégorisation	24
3.2.3 Organisation de la structure catégorielle	25
3.2.4 Théories de la catégorisation	29
3.2.5 Catégorisation et contexte sensoriel	32
3.2.6 Similarité et catégorisation	34
3.3 Analyse de scènes acoustiques	34
3.3.1 Définition	35
3.3.2 Une approche psychoacoustique	35
3.3.3 Régularités et processus primitifs	36
3.3.4 Perception de la forme	38
3.3.5 Flux auditif et stratégie de groupement	40
3.3.6 Attention, saillance et perception	42
3.3.7 L'approche par les neurosciences	43
3.4 L'étude des paysages sonores	43
3.4.1 La notion de paysage sonore	43
3.4.2 Application à la nuisance sonore urbaine	44
3.4.3 Approches catégorielle et dimensionnelle	46
3.4.4 Descripteurs perceptifs des paysages sonores	55
3.4.5 Catégoriser les sources et paysages sonores	60
3.4.6 Classifier les sources et environnements sonores	65
3.4.7 Contributions des différentes sources sonores	67
3.5 Événements et textures sonores	69
3.5.1 Définition	69
3.5.2 Percevoir les textures	70
3.5.3 Discussions	72
4 UN MODÈLE MORPHOLOGIQUE	75
4.1 Motivations	75
4.1.1 Analyse sensorielle	75

4.1.2	Analyse automatique	78
4.2	Proposition d'un modèle de scènes sonores	78
4.2.1	Discrétiser l'environnement sonore	78
4.2.2	Description du modèle morphologique	82
4.3	Du modèle à la simulation : l'analyse sensorielle	86
4.3.1	Simulation et analyse sensorielle	86
4.3.2	Protocole expérimental basé sur la simulation	87
4.3.3	Paramètres de contrôle	89
4.3.4	Interface de sélection aveugle des sons isolés	91
4.3.5	Interface de simulation : l'outil <i>Simscene</i>	92
4.4	Du modèle à la simulation : l'analyse automatique	93
5	APPLICATION DU MODÈLE À L'ANALYSE SENSORIELLE	95
5.1	Introduction	95
5.2	Agrément perçu et composition sémantique	96
5.2.1	Objectif	96
5.2.2	Banque de données de sons isolés	97
5.2.3	Typologie des sources sonores	97
5.2.4	Acquisition des sons isolés	98
5.2.5	Planification expérimentale	100
5.2.6	Données et méthodes d'analyses	103
5.2.7	Validité écologique de l'expérience	108
5.2.8	Vérification de l'agrément des scènes simulées	109
5.2.9	Étude comparative entre les descripteurs structurels	109
5.2.10	Influence des descripteurs structurels sur l'agrément perçu	111
5.2.11	Étude comparative entre les descripteurs sémantiques	114
5.2.12	Étude des espaces de représentation induits par les descripteurs sémantiques	117
5.2.13	L'influence spécifique des marqueurs sonores sur l'agrément perçu	120
5.2.14	Discussions	122
5.3	Modification de la composition sémantique	126
5.3.1	Objectif	126
5.3.2	Planification expérimentale	127
5.3.3	Données et méthodes d'analyses	128
5.3.4	Détection de valeurs extrêmes	130
5.3.5	Influence des descripteurs structurels des scènes avec marqueurs sur l'agrément perçu : une analyse comparative	131
5.3.6	Influence de la présence des marqueurs sur l'agrément perçu	133
5.3.7	Influence des descripteurs structurels des scènes sans marqueurs sur l'agrément perçu	134
5.3.8	Discussions	135

5.4	Composition sémantique et catégorisation	135
5.4.1	Objectif de l'expérience	135
5.4.2	Planification expérimentale	136
5.4.3	Données et méthodes d'analyses	137
5.4.4	Stratégie de catégorisation	140
5.4.5	Analyse lexicale des descriptions	141
5.4.6	Partitions et descripteurs sémantiques subjectifs	143
5.4.7	Partitions et descripteurs sémantiques objectifs	147
5.4.8	Descripteurs perceptifs et structurels	149
5.4.9	Discussions	150
III	ANALYSE AUTOMATIQUE DE SCÈNES SONORES	153
6	ÉTAT DE L'ART	155
6.1	Introduction	155
6.1.1	Historique, communauté et application	155
6.1.2	Campagnes d'évaluation : le challenge DCASE	155
6.2	Descripteurs	155
6.2.1	Spectrogramme	155
6.2.2	Échelle de Bark et de Mel	155
6.2.3	Coefficients cepstraux	155
6.2.4	Transformation à Q-constant et Q-variable	155
6.2.5	Filtre de Gabor	155
6.2.6	Scattering	155
6.3	Algorithmes et classificateurs	159
6.3.1	Modèle de Markov caché	159
6.3.2	Machines à vecteurs de support	159
6.3.3	Factorisation de matrice non-négative	159
6.3.4	Autres classificateurs	159
6.4	Détection des événements sonores	160
6.4.1	Objectifs	160
6.4.2	Métrique	160
6.4.3	Tâche 3 du challenge DCASE 2013	161
6.4.4	Tâche 3 du challenge DCASE 2016	162
6.5	Classification des scènes sonores environnementales	162
6.5.1	Objectifs	162
6.5.2	Métrique	162
6.5.3	Tâche 1 du challenge DCASE 2013	162
6.5.4	Tâche 1 du challenge DCASE 2016	162
6.6	Recouvrement des similarités acoustiques	162
6.6.1	Objectifs	162
6.6.2	Métrique	162
6.6.3	Méthodes et algorithmes	162
7	APPLICATION DU MODÈLE À L'ÉVALUATION	163
7.1	Introduction	163
7.2	Application au challenge DCASE 2013	163

7.2.1	Objectif	163
7.2.2	Génération des corpus	164
7.2.3	Métrique	170
7.2.4	Données et analyses	171
7.2.5	Système de détection	173
7.2.6	Résultats	174
7.2.7	Discussion	180
7.3	Application au challenge DCASE 2016	181
7.3.1	Objectifs	181
7.3.2	Génération des corpus	182
7.3.3	Métrique	184
7.3.4	Données et analyses	185
7.3.5	Systèmes de détection	186
7.3.6	Résultats	187
7.3.7	Discussion	193
8	SIMILARITÉS ET OBJETS	195
8.1	Introduction	195
8.2	Le bag-of-frame : une approche non satisfaisante	195
8.3	Une représentation basée sur l'objet	195
8.3.1	Formation des objets	195
8.3.2	Similarité entre objets	195
8.3.3	Coefficients de Scattering	195
8.3.4	Proposition d'un algorithme	195
8.4	Évaluation de l'approche objet	198
8.4.1	Objectif	198
8.4.2	Banque de données	198
8.4.3	Descripteurs	199
8.4.4	Systèmes évalués	200
8.4.5	Paramètres	200
8.4.6	Métriques et analyse	201
8.4.7	Résultats	201
8.4.8	Discussion	203
IV	CONCLUSIONS ET PERSPECTIVES	205
9	CONCLUSIONS	207
9.1	Analyse sensorielle	207
9.2	Analyse automatique	207
9.3	Approche pluridisciplinaire	207
9.4	Délivrables	207
10	PERSPECTIVES	209
V	APPENDICES	211
A	OUTILS D'ANALYSE STATISTIQUE UNI-VARIÉE	213
A.1	Test paramétriques à deux populations	213
A.2	Test paramétriques à deux populations ou plus	213
A.3	Mesures de corrélation paramétrique	213
A.4	Régression linéaire multiple	213

B OUTILS D'ANALYSE DIMENSIONNELLE	215
B.1 Analyse discriminante	215
B.2 Analyse par composante principale	215
B.3 Positionnement multidimensionnel	215
C EXPÉRIENCE ANNEXE : PÉRIODE D'ATTENTION	217
C.1 Objectif de l'expérience	217
C.2 Banque de données	218
C.3 La théorie de la détection du signal	218
C.4 Planification expérimentale	218
C.5 Méthodologie et outils statistiques	219
C.6 Résultats	219
D EXPÉRIENCE ANNEXE : CONGRUENCE	221
D.1 Objectif de l'expérience	221
D.2 Planification expérimentale	221
D.3 Résultats	221
BIBLIOGRAPHY	223

TABLE DES FIGURES

FIGURE 1	Principaux processus de traitement de l'information auditive et leurs interactions	12
FIGURE 2	Le phénomène de bistabilité : l'illusion du canard-lapin	15
FIGURE 3	Processus cognitifs et perceptifs	16
FIGURE 4	Paradigme de la psychologie cognitive	20
FIGURE 5	Les trois niveaux d'abstraction de l'axe vertical de la structure catégorielle.	28
FIGURE 6	Prototype et caricature	33
FIGURE 7	Groupement séquentiel : proximité temporelle.	36
FIGURE 8	Groupement séquentiel : proximité fréquentielle.	37
FIGURE 9	Groupement simultané : régularité harmonique	38
FIGURE 10	Groupement ancien-plus-nouveau	39
FIGURE 11	Compétition entre groupement séquentiel et groupement simultané	41
FIGURE 12	Tâche de description et tâche de tri ou de catégorisation	51
FIGURE 13	Les dimensions de calme et de dynamisme permettant de caractériser l'environnement sonore urbain, d'après (Cain et al., 2013)	60
FIGURE 14	Catégorisation des paysages sonores urbains, d'après (Raimbault and Dubois, 2005)	64
FIGURE 15	Taxonomie des sources sonores urbaines, d'après (Brown et al., 2011)	66
FIGURE 16	Information potentielle contenue dans les séquences d'événements, les textures, et le bruit	70
FIGURE 17	Plannification expérimentale de l'expérience de discrimination de textures sonores et d'exemplaires de textures sonores	72
FIGURE 18	Taxonomie des sources sonores urbaines, d'après (Salamon et al., 2014)	73
FIGURE 19	TODO	78
FIGURE 20	Organisation hiérarchique de la banque de sons isolés utilisée pour la simulation	84
FIGURE 21	TODO	85
FIGURE 22	Etape de processus de simulation pour l'analyse sensorielle	89
FIGURE 23	TODO	90
FIGURE 24	L'interface de sélection aveugle de l'outil de simulation <i>Simscape</i>	91
FIGURE 25	L'outil de simulation <i>Simscape</i>	92

- FIGURE 26** Planification expérimentale des expériences de simulation et d'évaluation de l'agrément 97
- FIGURE 27** Taxonomies des classes de sons utilisées pour la simulation des environnements sonores urbains 99
- FIGURE 28** Dispersion des notes données par les sujets lors de l'expérience 1.b moyennées suivant les sujets ($\mathcal{A}_{\text{subject}}$: a), et suivant les scènes ($\mathcal{A}_{\text{scene}}$: b), en fonction du type de scènes (i ou ni). 109
- FIGURE 29** 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). 111
- FIGURE 30** 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). 112
- FIGURE 31** 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. 113
- FIGURE 32** Pourcentage de scènes simulées comportant une classe de son particulière. 116
- FIGURE 33** 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. 119
- FIGURE 34** Dispersions des descripteurs structurels de densité relatif 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). 122
- FIGURE 35** 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). 123

- FIGURE 36** 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). [124](#)
- FIGURE 37** 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). [125](#)
- FIGURE 38** Dispersion des notes données par les sujets lors de l'expérience 2 aux i/am-scènes (vert) et ni/am-scènes (rouge). [130](#)
- FIGURE 39** 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. [130](#)
- FIGURE 40** 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/sm- et ni/sm-scènes ainsi que les différents niveaux d'abstraction. [134](#)
- FIGURE 41** 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). [135](#)
- FIGURE 42** 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). [136](#)
- FIGURE 43** Paritions P établies suivant la classification ascendante hiérarchique pratiquée sur la matrice de similarité Θ en utilisant un critère de Ward. [139](#)
- FIGURE 44** 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). [142](#)

- FIGURE 45** Generation process of the corpora considered in this evaluation. As part of the DCASE challenge, systems were trained on QMUL Train and tested on QMUL Test during the DCASE challenge. [166](#)
- FIGURE 46** Distribution des notes de réalisme $\mathcal{R}_{\text{ sujet}}$ pour les scènes enregistrés *test-QMUL* et les scènes simulées *instance-IRCCYN*. [171](#)
- FIGURE 47** Vision schématisée des systèmes de détection d'événements du challenge DCASE 2013. [173](#)
- FIGURE 48** Performances des systèmes évaluées dans le cadre du challenge DCASE 2013 sur les corpus QMUL et IRCCYN en considérant F_{cweb} . [175](#)
- FIGURE 49** Performances des systèmes évaluées dans le cadre du challenge DCASE 2013 sur les corpus *instance-QMUL* simulés avec différents EBR (6, 0, -6 et -12dB). [178](#)
- FIGURE 50** Performances globales des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique F_{cweb} . [187](#)
- FIGURE 51** 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 F_{cweb} . [189](#)
- FIGURE 52** 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} . [190](#)
- FIGURE 53** 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} . [193](#)
- FIGURE 54** Acoustic scene similarity retrieval (ASSR) in the DCASE 2013 private dataset : precisions at rank k ($P@k$) obtained for MFCCs and scattering with logarithmic compression, as a function of the rank k. [202](#)
- FIGURE 55** Acoustic scene similarity retrieval (ASSR) in the DCASE 2013 private dataset : precisions at rank k ($P@k$) obtained for scattering coefficients, with and without logarithmic compression, as a function of the rank k. [203](#)
- FIGURE 56** Événement ou texture sonore : influence de la période d'attention. [219](#)

LISTE DES TABLEAUX

TABLE 1	GL : TODO Indicateurs acoustiques	48
TABLE 2	Indicateurs psychoacoustiques : modèles mathématiques illustrant des qualités affectives perçues	49
TABLE 3	Les catégories sonores les plus citées, d'après (Niessen et al., 2010)	61
TABLE 4	Résumé des étapes de l'expérience de simulation	102
TABLE 5	Acronyme des variables utilisées dans le cadre des expériences sensorielles.	105
TABLE 6	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 1.b et les descripteurs structurels.	114
TABLE 7	Classes d'événements et de textures identifiées comme étant des marqueurs sonores	118
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.	120
TABLE 9	Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen $A_{scène}$ de l'expérience 2 et les descripteurs structurels globaux et relatifs à la présence des marqueurs sonores pour les i/am-scenes et ni/am-scenes.	131
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.	133
TABLE 11	Stratégies de catégorisation et nombres de groupements effectués	140
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.	141
TABLE 13	Répartitions des labels relatifs aux sources sonores relevées par les sujets en fonction des partitions établies par la classification ascendante hiérarchique.	144
TABLE 14	Répartitions des labels relatifs aux qualités affectives perçues en fonction des partitions établies par la classification ascendante hiérarchique.	145

TABLE 15	Répartitions des classes de sons en fonction des partitions établies par la classification ascendante hiérarchique	151
TABLE 16	GL : TODO	152
TABLE 17	Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2013	165
TABLE 18	Description synthétique des systèmes soumis dans le cadre de la tâche 2 de challenge DCASE 2013	173
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(o)</i> et <i>abstract-QMUL</i> .	176
TABLE 20	Nombre maximum de faux positifs pour chaque système évalué et pour chaque corpus	177
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> .	179
TABLE 22	Classes d'événements sonores utilisées dans le cadre du challenge DCASE 2016	182
TABLE 23	Description synthétique des systèmes soumis dans le cadre de la tâche 2 du challenge DCASE 2016	186
TABLE 24	Classes de scènes sonores considérées dans le cadre de la tâche 1 (ASC) du challenge DCASE 2013.	199

Première partie

PRÉAMBULE

preamble text here.

INTRODUCTION

La présente thèse traite des environnements sonores. L'étude se veut pluridisciplinaire. Elle 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 permettant à une machine de mimer 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 décrite. Trier signifiant 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 à 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 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.

Les travaux présentés dans cette étude se limitent à une modalité sensorielle particulière, l'audition. Plus particulièrement, elle se focalise sur un type de stimuli : les sons environnementaux.

Donnons une définition. Habituellement on entend pas son environnemental¹ tout extrait sonore qui ne se réclame ni de la parole, ni de la musique. Il s'agit d'une définition par exclusion, musique et parole étant des stimuli étudiés depuis longtemps, bien plus que

apprentissage, compréhension, pas forcément perception

phrase sans verbe conjugué (je garde le vert pour marquer ça)

ambitieux !

informel (je garde le bleu pour marquer ça)

¹ 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.

leur pendant environnementaux. Chacun bénéficie de champs de recherche dédiés, que ce soit en perception², ou en intelligence artificielles³.

Cette définition n'est pas satisfaisante. D'un côté, elle réduit les sons environnementaux à des entités secondaires. D'un autre côté, l'opposition suggérée entre sons environnementaux et sons de paroles et/ou de musique, deux domaines où le sens donné aux sons est de première importance, peut mener à penser que l'influence de la valeur sémantique des sons environnementaux est anecdotique, induisant, *de facto*, la primauté de leurs caractéristiques physiques. Ce qui n'est pas le cas (Ballas and Howard, 1987).

Nous prenons dans ce document la définition donnée par Vanderveer, 1980 (cité par Ballas and Howard, 1987).

La définition est en quatre points. Un son environnemental :

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 stimuli 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 l'effet du contexte, et notamment celui relatif à l'individu récepteur, est de première importance : une même scène sonore peut être perçue différemment par deux individus. Nous nous proposons ainsi de renforcer le point deux de la définition comme suit :

- 2 • 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 synthèses produits en laboratoire, ainsi qu'aux sons assumant une portée communicationnelle comme la parole ou la musique.

La définition insiste sur le fait que la perception d'un environnement sonore dépend éminemment de l'interprétation sémantique des

² on parle de perception de la parole (*speech perception*) et de perception de la musique (*music perception*)

³ On parle de traitement automatique du langage naturel pour la parole (NLP : *Natural Language Processing*) et de recouvrement de l'information musicale pour la musique (MIR : *Music Information Retrieval*)

Non, en français correct c'est TAL, "traitement automatique des langues" -- puisque le français fait, contrairement à l'anglais, la distinction entre "langue" et "langage". "Traitement automatique du langage naturel" est un "anglicisme de traduction", pour ainsi dire.

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 ainsi 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, modèle 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;
- montrer l'utilité d'un tel modèle dans le cadre de l'analyse automatique; point

répétition inutile

Ces objectifs déterminent l'organisation de ce document. Il comprend 4 parties. La partie **i** est constituée du chapitre **1**, la présente introduction, et du chapitre **2**, où est motivée l'approche pluridisciplinaire adoptée dans nos travaux.

La partie **ii** traite de l'analyse sensorielle. Elle regroupe les chapitres **3**, **4** et **5**. Le chapitre **3** présente un état de l'art des connaissances sur les processus mis en œuvre par le système auditif lors de la perception des environnements sonores. Sur la base de ces considérations perceptives, le chapitre **4** présente un modèle morphologique de scènes sonores environnementales. Il introduit également les outils permettant de simuler, à partir du modèle proposé, des environnements sonores. Le chapitre **5** présente une série d'expériences montrant comment le modèle introduit permet d'étendre les possibilités des méthodologies classiquement utilisées en analyse sensorielle. Le cadre applicatif choisi par ces expériences est l'évaluation de l'agrément dans les environnements sonores urbains.

Bonne annonce, met bien les idées en place sur la structure.

Petite suggestion de forme :

"la partie ii, regroupant les chapitres 3, 4 et 5, présente..."

"la partie iii, composée des chapitres 6, 7 et 8...", etc.

La partie **iii** traite de l'analyse automatique des environnements sonores. Elle est composée des chapitres **6**, **7** et **8**. Le chapitre **6** présente un état de l'art des connaissances sur l'apprentissage machine appliqué aux sons environnementaux. Cette présentation s'emploie à détailler les différentes tâches (classification ou détection de sons isolés ou de scènes complexes), les algorithmes couramment utilisés, ainsi que les pratiques expérimentales ayant cours dans ce domaine. Par pratiques expérimentales on entend ici la manière avec laquelle sont évaluées les performances des systèmes proposés, notamment en ce qui concerne les banques de données et les métriques utilisées. Le chapitre **7** 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, ce modèle,

inutile, pas d'autre antécédent masculin possible :-P

permet de gagner en connaissance quant à la capacité de généralisation des systèmes de détection proposés. Dans le chapitre 8, nous nous éloignons légèrement du modèle introduit, et proposons un algorithme non-supervisé permettant de recouvrir les similarités existantes entre différentes scènes sonores. Cet algorithme s'appuie sur la vision composite des scènes sonores, en adoptant une approche dite objet. Utilisant une représentation *sparse* du signal, via un descripteur de type *scattering*, l'approche objet groupe, dans un premier temps, les éléments similaires d'une même scène. La similarité inter-scène est ensuite calculée sur la base des différents groupes précédemment obtenus.

La partie iv, partie conclusive, comprend les chapitres 9 et 10. Le chapitre 9 résume les différentes contributions de cette thèse, et conclue quant au travail effectué. Le chapitre 10, lui, GL : discute plus avant à la fois l'approche pluridisciplinaires adoptée, les résultats obtenus, et propose de nouvelles pistes à explorer.

sugg. : propose des éléments
de réflexion

2

MOTIVATION

GL : TODO : motivation évaluation des algorithmes : horse effect

Deuxième partie

ANALYSE SENSORIELLE DE SCÈNES SONORES

preamble text here.

3

ÉTAT DE L'ART

Avant d'aller plus loin dans l'exposé de nos travaux, il nous semble indispensable de dresser ici un état des lieux des connaissances liées à la perception des sons. Nous procérons en cinq parties.

Au cours de la première, nous proposons un tableau général des différents processus intervenant dans le traitement de l'information sonore, processus mis en œuvre dès lors que le signal atteint le tympan.

Au cours de la seconde, nous interrogeons la manière dont nous nous représentons le monde sonore perçu. Nous démontrons comment cette représentation influe sur notre perception.

Au cours de la troisième, nous nous intéressons aux processus 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.

Au cours de la quatrième, nous introduisons la notion de paysage sonore, et examinons l'impact que cette notion a sur les recherches en matière de perception des environnements. Nos travaux s'inscrivant largement dans cette approche, nous dressons un état de l'art des connaissances en matière de perception des paysages sonores, et tentons de dégager les grands axes méthodologiques suivis par ces études.

Au cours de la cinquième et dernière partie de cet état des lieux, sont précisés les concepts d'événement et de texture sonore, notions clefs qui interviennent par la suite dans le modèle de scène sonore proposé.

GL : TODO : moins de "nous"

approche ? tu n'as parlé que de notion. À la limite, utilise "paradigme", ça fait chic.

3.1 LE TRAITEMENT DE L'INFORMATION AUDITIVE

3.1.1 *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 systèmes forment une chaîne qui, au fur et à mesure des traitements, interprète le signal acoustique afin d'en extraire l'information séman-

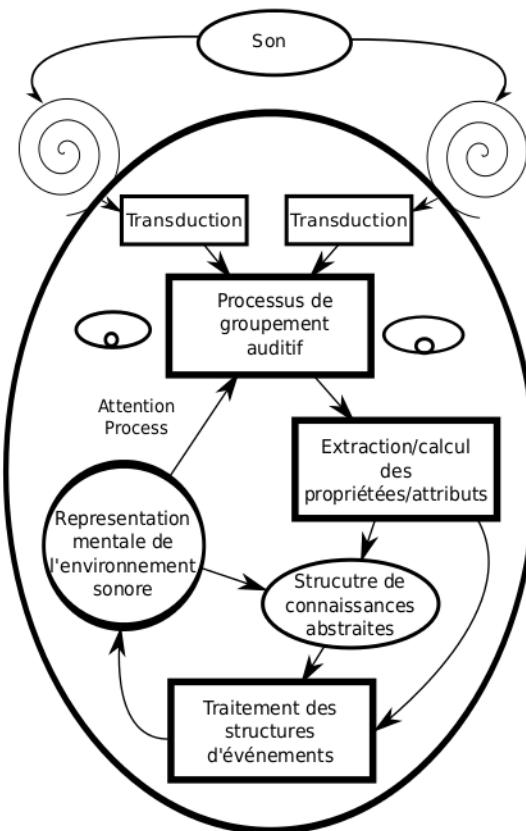


FIGURE 1 : Principaux processus de traitement de l'information auditive et leurs interactions, d'après (McAdams and Bigand, 1994)

à une étape avancée de

tique. Plus on se place loin dans la chaîne de traitement, plus on a accès à une information abstraite, potentiellement utilisable par d'autres processus de haut niveau. La figure 1 extraite de (McAdams and Bigand, 1994) nous 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 cette dernière 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 la membrane basilaire, sont transmises au cerveau via les nerfs auditifs, chaque point codant une information correspondant à une bande fréquentielle limitée; **Plusieurs phrases dans l'item -> majuscule au début et point à la fin**
- **processus de groupement auditif :** C'est une étape d'intégration temporelle au cours de laquelle 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; lui

- *d'extraction/calcul des propriétés/attributs* : le processus de groupement précède généralement la phase dite *d'extraction/calcul des propriétés/attributs*. C'est lors de cette phase que sont extraites les qualités perceptives des objets groupés, ces qualités pouvant être vues comme des propriétés cognitives de haut niveau. Pour revenir au choral, c'est à partir d'une analyse des attributs perceptifs que nous sommes capables de percevoir les mélodies comme des objets unitaires, même si celles-ci sont développées entre les différentes voix;
- *structure des connaissances abstraites* : les phases de groupement et d'extraction concernent l'élaboration et l'analyse d'entités mentales. Une fois représentées dans le cerveau, ces entités sont interprétées pendant l'étape de *structure des connaissances abstraites*. C'est lors de cette étape qu'elles sont identifiées, et qu'un sens interprétatif leur est donné. En pratique, il s'agit de déterminer si le Choral est plaisant ou non;
- *traitement des structures d'événements* : cette étape permet d'intégrer dans le processus cognitif différents contextes, comme par exemple le contexte fonctionnel (dans quel cadre ce son est-il entendu ?), ou encore le contexte sensoriel (l'information visuelle, ou la mémoire des événements sonores précédemment entendus). Pour revenir à notre mélomane, c'est cette étape qui lui permet 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* : la dernière étape du processus de traitement est l'élaboration d'une représentation mentale de l'environnement perçu. C'est au cours de cette étape que nous organisons et conservons les différentes informations extraites. GL : Ces représentations structurent les connaissances acquises par un individu, connaissances qui, par effet de rétroaction, vont influer sur les processus précédemment cités, afin

Choral (maj. pas indispensable, juste plus clair)

Fais référence au mélomane dans le dernier point si tu peux, j'aime bien le procédé de l'exemple suivi du début à la fin de la chaîne.

que ces derniers bénéficient des informations accumulées lors d'expériences passée.

3.1.2 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 un individu, d'autre part, de la rétroaction exercée par lui sur ces données.

Typiquement, si nous reprenons la figure 3.2, 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. Par exemple, si la fréquence fondamentale d'un son harmonique change au cours du temps, toutes ses harmoniques changeront également afin de maintenir la structure harmonique du son [p. 38](Bregman, 1994) (cf. Section 3.3.3).

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 3.2 et Figure 3.2). 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 de l'individu, 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 l'individualité qui explique que deux personnes ayant des capacités sensorielles semblables peuvent percevoir différemment un même environnement.

Le système auditif mobilise ainsi deux formes de traitements (cf. Figure 3) :

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

Un exemple concret, emprunté au domaine de la vision, est celui du 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 2).

anglicisme ("experienced by")
mais la version française
correcte m'échappe... Je
cherche.

de

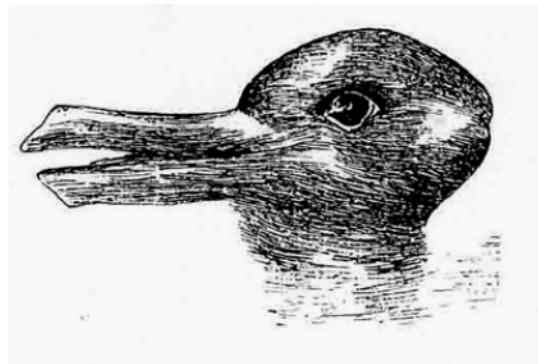


FIGURE 2 : Le phénomène de bistabilité : l'illusion du canard-lapin. Première publication dans *Fliegende Blätter*, 23 octobre 1892, p. 147

Un autre exemple, emprunté cette fois au domaine de l'audition, nous semble illustrer encore le caractère dual de l'analyse de l'information sensorielle. Il est donné 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.”

Étudier le système auditif demande de prendre en compte aussi bien l'information externe (processus ascendant) que l'information interne (processus descendant). Réduire la perception à une simple association de sensations établies à partir du signal capté, ne permet pas de rendre compte de l'éventail des processus cognitifs entrant dans le décodage de l'environnement. Cette distinction entre perception et cognition est approfondie à la section suivante (cf. Section 3.1.3).

pas de virg.

3.1.3 Perception et cognition

Perception. Le mot désigne l'ensemble des processus de traitement de l'information sensorielle. La perception du monde sonore qui nous entoure est un phénomène complexe, aujourd'hui encore mal connu.

Cette perception est à l'origine de l'interaction que nous créons avec notre environnement. Elle détermine notre capacité d'adaptation à ce dernier. Cette relation au monde *réel* ne se rompt jamais. Nous percevons des sons en permanence, et ce, même si aucune source sonore n'est présente.

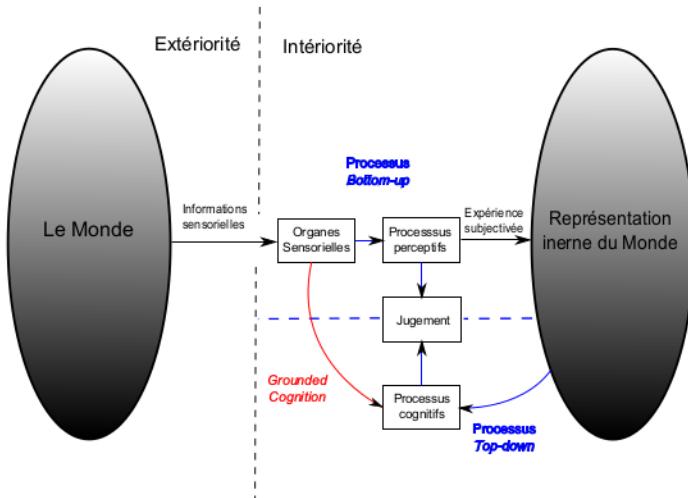


FIGURE 3 : Processus cognitifs et perceptifs

Ainsi tel mélomane dont la tête résonne encore de l'air, bien après que les instruments se soient tus. Ainsi tel usager des transports dont l'oreille anticipe (pour s'en protéger...) les crissements du métro alors que la rame n'est pas encore à quai.

Cognition. Selon U. Neisser¹ dans (Neisser, 1976, p. ??):

“Cognition is the activity of knowing : the acquisition, organisation and use of knowledge.”

Le mot désigne l'ensemble des processus d'acquisition et de développement d'une connaissance du monde.

Selon la théorie classique, perception et cognition dépendent de deux groupes de systèmes fonctionnels du cerveau distincts. La perception mobilise les systèmes de traitement dits modaux, c'est à dire supportés par les organes sensoriels (oreilles, yeux etc ...), tandis que la cognition s'appuie sur des représentations mentales des réalités externes, par essence amodales.

Cette dichotomie entre perception et cognition est aujourd'hui remise en question. Dans une approche “incarnée” de la cognition (*Grounded Cognition*), Barsalou nie le caractère amodal des représentations mentales, prônant que ces dernières dépendent également des modalités sensorielles (Barsalou, 2010). Il tente ainsi de réunir les processus perceptifs et cognitifs (Barsalou, 1999; Goldstone and Barsalou, 1998).

Les deux approches sont illustrées sur la Figure 3.

¹ 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.

3.1.3.1 Psychologie cognitive et psychoacoustique

La psychologie cognitive est un domaine de recherche dédié aux phénomènes se rapportant à la connaissance. Elle est née dans les années 50, en réaction au *Béhaviorisme*, théorie fondée sur l'étude des comportements objectivement observables de l'être humain, et négligeant, de fait, le rôle des connaissances propres du sujet. La psychologie cognitive s'interroge sur des modèles théoriques complexes rendant compte de tous les faits et de toutes les lois connus. Les chercheurs y explorent tout à la fois, la mémoire, le langage, l'intelligence, la perception.

L'approche cognitiviste, dans le domaine de la perception auditive, se distingue de celle plus traditionnelle de la psychoacoustique². Tandis que la psychoacoustique émet l'hypothèse d'une relation directe entre le stimulus et la réponse du sujet, la psychologie cognitive soutient que la réponse, elle, est entièrement corrélée au contexte, à l'expérience, et aux interactions multi-sensorielles.

La réponse tient compte non seulement des traitements perceptifs mais aussi des représentations issues et de la mémoire individuelle (*i.e.* construites en particulier à partir de la relation sensible au monde), et de la mémoire collective, à travers le développement des connaissances partagées.

La psychologie cognitive s'intéresse prioritairement à l'aspect cognitif de la perception en considérant l'individu comme un tout. Elle prend en compte la culture, l'expérience, l'activité de l'individu et ne focalise pas seulement sur la réaction des organes sensoriels comme l'oreille. Elle questionne les aspects qualitatifs plus que quantitatifs de notre compréhension du monde sonore (Maffiolo, 1997, 1999).

Elle envisage l'ensemble des étapes du traitement auditif de manière globale et permet ainsi de faire le lien entre une information sensorielle et une information abstraite (McAdams and Bigand, 1994).

En psychologie cognitive, on distingue les approches cognitivistes, qui s'intéressent plus particulièrement aux processus montants (*botttom-up*, cf. Section 3.1.2), relatifs au traitement de l'information perçue, et les approches cognitives, qui interrogent, avant tout, les processus descendants (*top-down*, cf. Section 3.1.2) liés à la mémoire du sujet ainsi qu'au contexte (Guastavino, 2003, p. 34).

GL : TODO : reprendre guastavino thèse

3.1.3.2 Paradigme de la psychologie cognitive

Nous le voyons, la psychologie cognitive ne conçoit pas le sujet comme une "boîte noire", mais reconnaît en lui un système de traitement

² La psychoacoustique est une branche de la psychophysique qui applique au domaine de l'acoustique les concepts et les méthodes de la psychophysique

de l'information. Elle admet que l'individu 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 ??). Ces présupposés sont résumés ci-après :

- le monde est discrétilisé 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 en comparaison 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 naturelle. 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* (cf. Section 3.4.3) a caractériser l'intensité perçue, et **encore plus**, la gêne occasionnée par cette intensité, pour des stimuli réels enregistrés (cf. Section 3.4.2).

plus encore

GL : TODO : a) perception dépend d'autres caractéristiques, b) construction de la *loudness* à partir de sons purs.

Concernant le point 3, bien que le paradigme suppose l'influence d'un contexte, ce dernier est considéré 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 être à juste titre considéré comme une hypothèse très 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 ?

Toutes ces questions/remarques restent ouvertes. Loin de dénigrer l'approche de la psychologie cognitive, elles doivent être vues comme des aides, et prises en compte dans l'interprétation des résultats d'expériences reposant sur ce paradigme.

GL : TODO : faire le lien ici avec les descripteurs subjectifs utilisés pour caractériser un environnement sonore

3.1.4 *L'approche écologique*

L'approche écologique a d'abord été introduite dans le domaine de la vision par Gibson (Gibson, 1966), qui se demande entre autre si les "lois structurant les objets sont porteuses d'informations, ou si cette information est tirée de comparaisons" (Gibson, 1978).

Cette approche reconnaît que la réponse au stimulus dépend à la fois de l'information perçue, et de la connaissance du monde, autrement dit, l'environnement quotidien, et le contexte habituel d'écoute du stimulus.

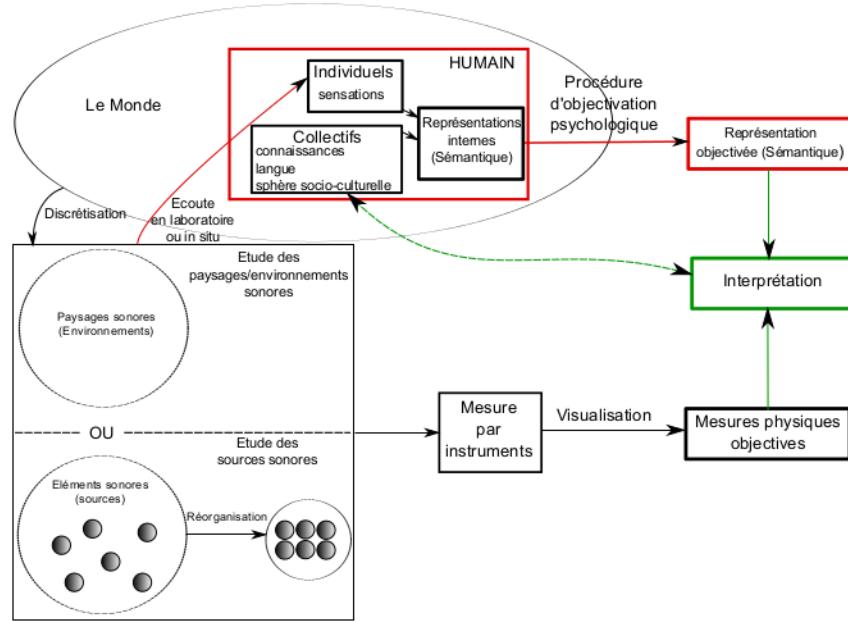


FIGURE 4 : Paradigme de la psychologie cognitive, d'après (Maffiolo, 1999)

3.1.4.1 Soundwalk

Appliquée à la perception sonore, l'approche écologique requiert 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*³ 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) ;
- d'évaluer l'environnement sonore, tout en maintenant actif les autres sens (vue, olfaction) ;
- d'éviter 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

³ *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.

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. Se pose le problème de la reproductibilité des expériences. 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.

3.1.4.2 *Reproduction écologique des environnements sonores en laboratoire*

Le problème de la 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 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 et *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.

GL : TODO : sélection des stimuli

3.2 REPRÉSENTATION MENTALE DE L'ENVIRONNEMENT SONORE

GL : TODO : être moins prétentieux, ajouter "il semble, il est utile, il est communément admis"

Le cerveau entretient un dialogue constant avec l'environnement sonore. Ce dialogue s'effectue par le biais des représentations mentales. Une définition des représentations mentales est donnée par (Houdé et al., 1998):

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

Ces représentations font office de sauvegardes de l'information. Conservées en mémoire sous une forme abstraite (McAdams and Bigand, 1994, p. 357), elles rendent compte à la fois de notre compréhension du monde, et de la manière dont nous l'abordons. Ces

Une phrase d'intro pour rappeler pourquoi tu parles de ça, peut-être ? Arrivé à ce point, avec plein de concepts à avaler, ce serait utile de prendre un peu le lecteur par la main.

connaissances subjectives, non directement observables, restent néanmoins accessibles au chercheur par le biais d'expériences d'objectivation (voir section 3.4.3.1)

Ces représentations forment une image mentale discrète du monde réel continu (Houdé et al., 1998). 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.

3.2.1 *La notion de catégorie*

3.2.1.1 *Définition*

Une des vérités de tout être humain est de segmenter son environnement, *i.e.* de se bâtir un système de classification permettant de regrouper des objets n'étant pas identiques (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 présupposée de la catégorisation. L'opération permanente consistant à isoler un objet, et à lui attribuer un nom, est l'antinomie du "geste adamique", *i.e.* d'attribution libre du nom, hors toute influence et/ou contexte (d'Adam, le premier homme révélé et nommé dans la bible). La langue est un code partagé par une communauté. Dans une certaine mesure, sa définition, *i.e.* le sens donné aux mots, peut varier suivant les groupes de cette

étiquette

communauté. Ainsi, catégoriser ne dépend pas seulement d'une réalité physique du monde, mais également d'un contexte socio-culturel. 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, et, d'autre part, la constitution d'une représentation collective pouvant être partagée par le biais d'un langage commun (Du-bois 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 d'interpréter des objets possédant des fonctions similaires comme faisant partie d'un même groupe, et ce même si ces objets ont des caractéristiques distinctes. Pour exemple : deux hommes descendant d'un camion de pompier pour éteindre un feu de forêt seront catégorisés 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.

3.2.1.2 *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 naturelles : regroupent des objets existant à l'état naturel (animaux, fleurs, *etc.*);
- catégories artificielles : regroupent des objets fabriqués par l'homme (voitures, outils) ;
- catégories de concepts : 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 similarités. Pour les catégories d'objets naturels et artificiels, ces

similarités s'établissent majoritairement à partir de leurs caractéristiques physiques. Pour les catégories de concepts, ces similarités relèvent d'attributs cognitifs de plus haut niveau. De plus, 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, *e.g.* la liste de courses. Les catégories inhérentes à de tels groupements sont nommées *ad hoc* (Barsalou, 1983):

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

3.2.2 *Le processus de catégorisation*

3.2.2.1 *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 : il suffit de voir la croupe d'un cheval pour en "déduire" l'animal. Autre exemple : le bourdonnement d'un insecte peut laisser supposer la présence d'une guêpe, et susciter le sentiment du 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.

3.2.2.2 *Catégorie et langage*

Comme nous l'avons vu (cf. Section 3.2.1.1), catégorie et langage fonctionnent de concert. 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 3.4.3.1).

L'analyse linguistique, comme outil d'approche des processus de catégorisation des sons, a été particulièrement étudiée par Dubois

et ses collègues (Dubois, 2000; Dubois et al., 2006; Guastavino, 2006; Raimbault 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 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 de sa communauté (Dubois, 2000).

GL : TODO : a compléter

3.2.2.3 *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, si un très jeune enfant voit pour la première fois un groupe de hyènes dans un zoo, il interprétera ces animaux comme faisant partie de la même espèce. Cependant, il a de très grandes chances d'identifier l'espèce comme "une sorte de chien".

Encore trop de i.e. :-P

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.

3.2.3 *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.

3.2.3.1 *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.

3.2.3.2 *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.

3.2.3.3 *Catégorie et abstraction*

Pour des catégories d'objets 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 représentant des objets abstraits ou concepts, et incluant un grand nombre de sous catégories, et les catégories de bas niveau représentant des objets concrets, incluant peu de sous catégories. Ainsi, plus le niveau d'abstraction est grand, plus les similitudes entre objets

d'une même catégorie (intra-catégorielles), ainsi que les similitudes entre objets 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 5) : superordonné, base et subordonné. Le niveau superordonné regroupe les catégories à haut niveau d'abstraction. Les périmètres de ces catégories sont larges (Mobilier, Véhicule...) *i.e.* les objets qu'elles contiennent peuvent être très distincts. Le niveau subordonné regroupe les catégories à bas niveau d'abstraction, ou catégories concrètes. Les périmètres de ces catégories sont plus précis (Chaise longue, Cabriolet...), et les objets qu'elles contiennent sont nécessairement très similaires. On notera ici que les objets de la classe Cabriolet présentent plus de propriétés communes avec les objets de la classe Berline, que n'en sauraient partager les objets des classes Mobilier et Véhicule. Au niveau de base, les objets d'une même catégorie partagent encore beaucoup de propriétés en commun, 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 objets contenus dans une même classe. Les catégories ne sont pas des objets strictement discrets, et les propriétés des objets qu'elles regroupent peuvent se trouver attribuées à d'autres objets, dans 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 représentées 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 ainsi d'obtenir le maximum d'informations au prix d'un moindre effort cognitif.

Cependant, si cette vision bidimensionnelle de la structure catégorielle est adaptée aux catégories d'objets concrets, elle est moins pertinente dans le cas des catégories de concepts, 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.

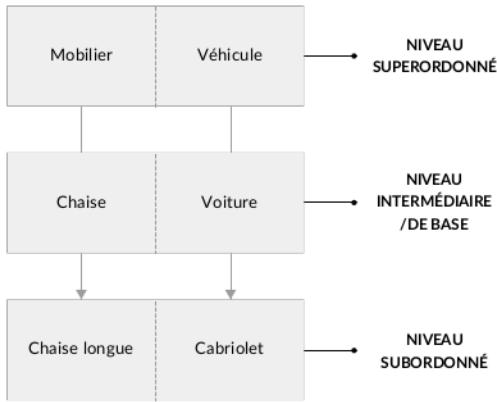


FIGURE 5 : Les trois niveaux d'abstraction de l'axe vertical de la structure catégorielle.

3.2.3.4 *La notion de typicalité*

Une notion clef, dans les processus catégoriels, est la typicalité. Tous les objets d'une catégorie ne sont pas égaux. Il y a une gradation dans l'appartenance catégorielle. Certains objets, partageant les propriétés dominantes d'une catégorie, sont considérés comme très représentatifs de cette dernière. D'autres, ne partageant que peu des attributs caractéristiques de la catégorie, 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 vérifications 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). GL : TODO : préciser en quoi l'attraction implique l'asymétrie

Est-ce vraiment nécessaire ? Toute cette partie est super intéressante, mais tu n'es pas là pour faire un cours, non plus. Ce chapitre est là pour introduire les notions utiles pour comprendre le reste du travail, c'est tout.

3.2.4 Théories de la catégorisation

3.2.4.1 Théorie classique

Suivant les principes de l'approche logique, dite aussi approche par règles, 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. La nature de ces propriétés étant inhérentes à la catégorie.

Cette approche, qui sous-tend que tous les objets 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 3.2.3.4), 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 "enfants", "veuf" ou encore "pape" qui, intuitivement, n'ont rien à voir avec "célibataire" GL : TODO : peu clair ?.

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.

3.2.4.2 Théorie prototypique

Une alternative à l'approche classique, consiste à envisager la catégorie non plus comme relevant de règles, mais comme découlant, en quelque sorte, de la ressemblance ou "air de famille" liant ses membres (Ludwig, 1953).

Partant de cette idée, la théorie prototypique a été formalisée par E. Rosch et B. B. Lloyd (Rosch and Lloyd, 1978). Dans 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). Tous les objets d'une catégorie ne sont pas également représentatifs 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 cette dernière (Rosch, 1975; Rosch and Lloyd, 1974). Les cas non ambigus peuvent être vus comme les objets les plus typiques de la catégorie. Le terme prototype, qui donne son nom à la théorie, vient de l'assertion que, parmi ces cas non ambigus, il existe un, le prototype, plus représentatif que les autres, et qui forme le noyau de la catégorie.

Les catégories sont structurées en interne, en référence à un prototype, *i.e.* l'objet 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 and Mervis, 1975), il s'agit de l'objet possédant le plus de propriétés en commun avec les objets de la catégorie, et le moins de propriétés avec les objets des catégories externes. 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, un objet défini comme étant la moyenne des attributs des objets de la catégories. GL : TODO : Préciser qu'on retrouve cette différence entre le centroïde et le medoïde

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

3.2.4.3 Théorie des exemplaires

La théorie des exemplaires nie l'existence d'un prototype. Au contraire, elle propose que la catégorie soit représentée sur la base de tous les objets (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 3.2.2.1) 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), et le modèle de contexte généralisé (Nosofsky, 1986). Dans le modèle de contexte, les objets sont représentés suivant leurs attributs, la dimension de l'espace de représentation étant alors égale aux nombres d'attributs, choisis dans un contexte particulier (Hitzman, 1986). Dans le modèle généralisé, les objets sont représentés dans un espace psychologique aux dimensions réduites, espace établi sur la base des distances inter-objets (Nosofsky, 1992). Typiquement, le positionnement multidimensionnel (*multidimensional scaling* : cf. annexe B.3) est utilisé afin de reconstruire de tels espaces.

Dans la théorie des exemplaires, l'appartenance catégorielle se fait sur la base de la somme pondérée des similarités entre l'objet à catégoriser, et les exemplaires de la catégorie. Dans le modèle de contexte, la pondération s'effectue sur les propriétés des objets, et rend compte d'un poids attentionnel, favorisant les propriétés saillantes dans un contexte donné. Dans le modèle généralisé, la pondération se fait en fonction de la distance entre les exemplaires et l'objet, et ce afin de favoriser les exemplaires proches de l'objet à catégoriser.

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

- Comment justifier que nombres d'études montrent que l'appartenance catégorielle s'effectue via une comparaison à un prototype ?
- Comment justifier que le principe d'économie cognitive reste valide, si le cerveau utilise l'ensemble des items d'une catégorie pour la représenter ?

Dans la théorie des exemplaires, la proximité entre un objet et une catégorie est la somme des similarités entretenues entre l'objet et les items de la catégorie (Nosofsky, 1986). Dans certains cas, cette opération équivaut à calculer la similarité entre l'objet et le représentant moyen des items de la catégorie. L'existence d'un prototype n'est alors qu'un artefact. Notons quand même que la théorie des exemplaires ne nie pas l'existence d'un gradient de typicalité entre les objets d'une catégorie.

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 écologiquement peu probable que, pour chaque catégorie, le cerveau sauvegarde la totalité

Pas clair si on ne connaît pas l'expé -- si tu en parles, dis-en plus.

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) GL : TODO : à nuancer, ces deux options sont triviales et les processus de mémorisation sont plus complexes.

3.2.4.4 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'objet représentatif putatif 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 objets de sa catégorie (cf. Figure 6).

(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 objets 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.

3.2.5 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. mal connue

(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 un effet bénéfique 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

"cooccurrences" ou
"co-occurrent avec lui"
(dans tous les cas, 2 "r")

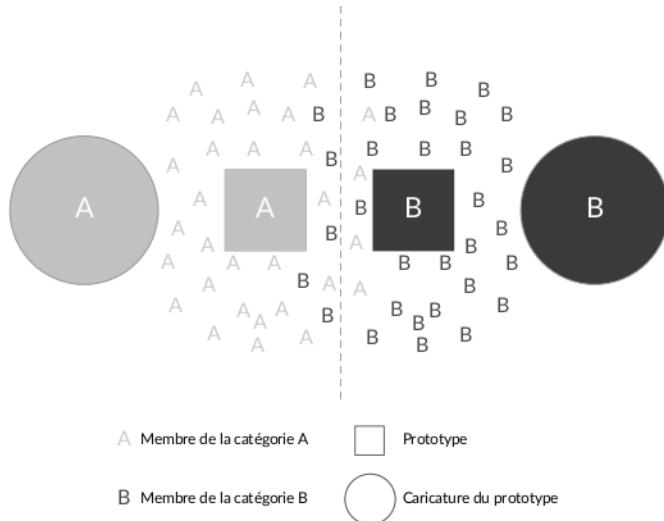


FIGURE 6 : Prototype et caricature, d'après (Davis and Love, 2010)

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 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 D.

mais tu écris juste au-dessus
"la source est plus facilement
identifiable si ce dernier est en
adéquation avec son contexte"

Conclusion, on ne peut rien
conclure ?

GL : TODO : a compléter, notamment sur le lien entre environnement et parole.

3.2.6 *Similarité et catégorisation*

Comme vu précédemment, notamment au moment d'évoquer les théories prototypiques, d'exemplaires et de 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 3.2.5), dépendent également d'un 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 similarités 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), alors que la catégorisation est 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).

GL : TODO : Questionner ici la pertinence des études perceptives portant sur des sons isolés, si la perception de ces derniers dépend autant du contexte. Rebondir alors sur la pertinence en Machine learning de tâche de classification d'objets isolés, et motivé ainsi les tâches AED.

3.3 ANALYSE DE SCÈNES ACOUSTIQUES

GL : TODO : plus insister sur la notion de flux, lien chapitre 4

Petite intro / rappel de où on est

3.3.1 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 3.1.1, 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 de connaissances (schémas) issues de notre représentation mentale du monde, représentation construite à partir des écoutes antérieures.

3.3.2 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,⁴ sont synthétisés en laboratoire.

Ces sons sont émis de manière séquentielle. Au cours de l'expérience, les paramètres d'écoute sont modifiés (intensité, fréquence, es-

⁴ 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

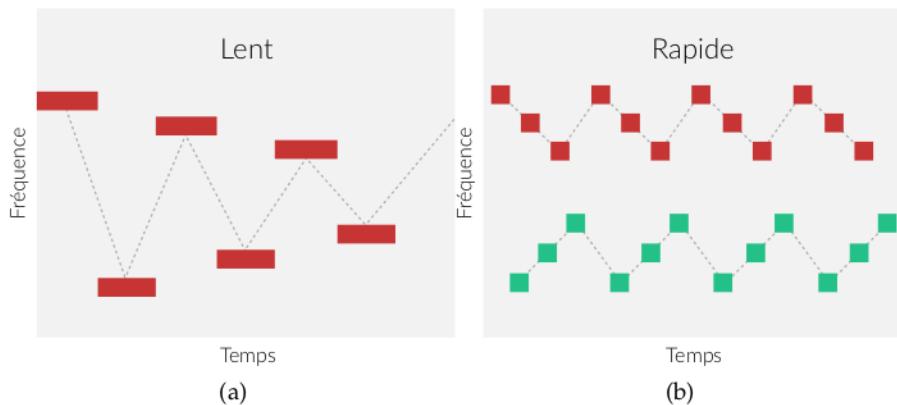


FIGURE 7 : Groupement séquentiel : proximité temporelle. Dans l'exemple (a), la durée entre les événements est importante, aucun regroupement n'est effectué, les sons sont perçus comme des événements distincts. Dans l'exemple (b) la durée entre les événements est réduite, un groupement s'opère suivant la proximité fréquentielle. Deux flux sont ainsi créés, le premier (rouge) regroupe les sons haute fréquence, le deuxième (vert), les sons basse fréquence.

pace entre les séquences...) afin d'évaluer le seuil à partir duquel la capacité du sujet à distinguer les sources sonores est altérée.

L'ASA se focalise 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 3.1.4). Il est difficile de faire un parallèle entre la notion de source sonore utilisée dans le domaine de la psychoacoustique, et celle admise dans le domaine de la psychologie cognitive. Dans l'une et l'autre approche, *a priori*, le terme source sonore s'applique 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. Cela l'est moins en psychoacoustique, 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 sont difficiles à généraliser à des applications plus incarnées.

3.3.3 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):

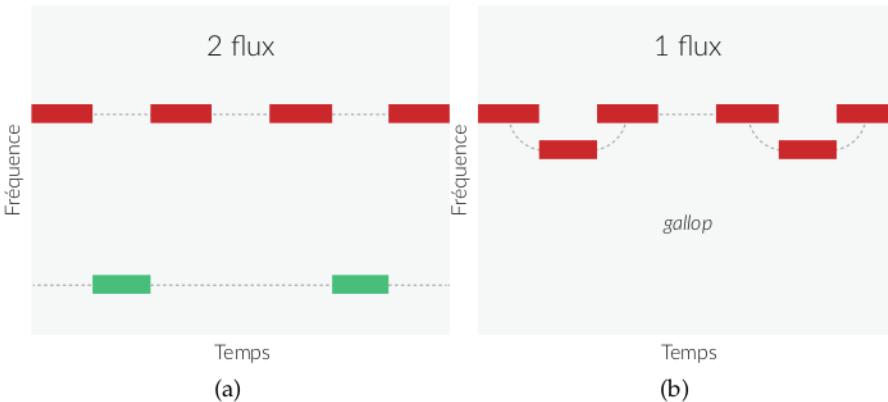


FIGURE 8 : Groupement séquentiel : proximité fréquentielle. Deux groupes de sons sont joués à deux fréquences. Dans l'exemple (a), la distance entre les fréquences des sons est importante, deux flux sont créés, le premier (rouge) regroupe les sons haute fréquence, et le deuxième (vert), les sons basse 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.

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.
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.



FIGURE 9 : 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.

3.3.4 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 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*⁵ et l'intensité.
2. *similarité* : des éléments qui se ressemblent ont tendance à être groupés ensemble.

⁵ En traitement du signal audio, on désigne par le mot anglais *onset* le début du signal. Ce terme étant couramment utilisé en français, nous ne le traduirons pas dans ce document.

Faire une note similaire pour "pattern", si tu veux l'utiliser.

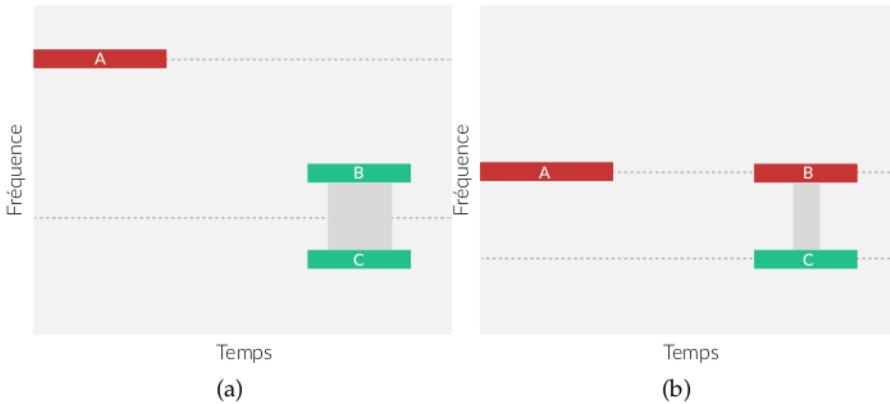


FIGURE 10 : Groupement ancien-plus-nouveau. Dans l'exemple (a), les sons A et B ont des fréquences éloignées. Le cerveau génère deux flux, le premier relatif au son A (rouge), et le deuxième comprenant les sons B et C (vert). Dans l'exemple (b), les sons A et B ont la même fréquence. Le cerveau interprète 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).

Commentaire. 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 ...), 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 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. Par ce principe, des objets distincts mais proches temporellement (en vision, proches spatialement) ont tendance à être reçus 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.

perçus

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 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. Comme évoqué à la section 3.3.1 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.

3.3.5 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 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 auditifs" (*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 :

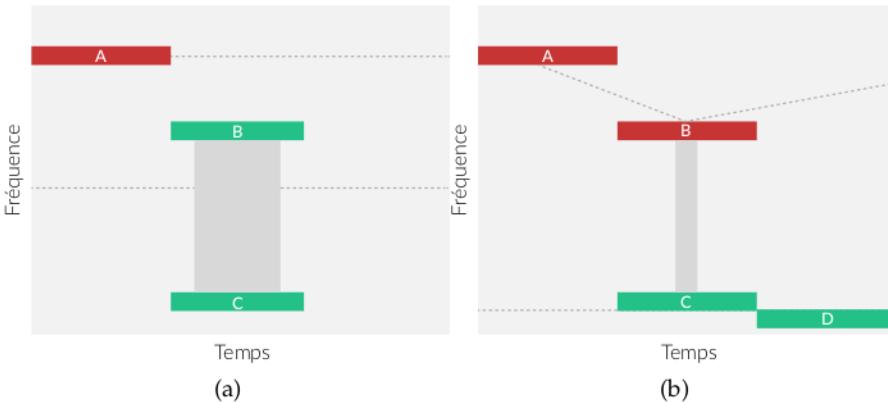


FIGURE 11 : Compétition entre groupement séquentiel et groupement simultané. Dans l'exemple (a), le cerveau perçoit deux flux, le premier regroupant le son A (rouge) et le deuxième regroupant les sons B et C (vert). Dans l'exemple (b), un quatrième son (D) est ajouté, ce dernier possédant une fréquence très proche de celle de (D). 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.

“ 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, et, à ce titre, 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 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 8). 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 7) ;

- *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 9) ;
- *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 10).

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 11). 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.

3.3.6 Attention, saillance et perception

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

faisant abstraction

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), seront 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. Il sélectionne l'information utile. Ces modèles ne , et 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). Ce modèle permet, entre autre, de composer des résumés acoustiques des environnements sonores, à partir des sons jugés typiques de ces derniers.

3.3.7 *L'approche par les neurosciences*

GL : TODO : review (Snyder and Alain, 2007)

3.4 L'ÉTUDE DES PAYSAGES SONORES

Petite intro / rappel, on est où ?

3.4.1 *La notion de paysage sonore*

La notion de paysage sonore (*soundscape*) a été introduite par 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."

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.”⁶

L'une et l'autre définition 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 3.1.3.2). 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.

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

3.4.2 Application à la nuisance sonore urbaine

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 problème⁷. Le consul Jules César interdit d'ailleurs à ces derniers 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 (Tae Hong 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

⁶ Cette définition a été publiée dans le cadre de la norme ISO-12913 (Standardization, 2013)

⁷ cf. Juvenal, Satire 3.232–238

une réduction du niveau des bruits produits, essentiellement, par les transports et l'industrie.

Mais le problème persiste, le bruit demeurant un phénomène subjectif, autrement dit dépendant de l'appréciation de l'auditeur. Le bruit est affaire de contexte et beaucoup de lieux urbains sont appréciés aussi pour leur atmosphère vivante, c'est à dire "bruyante". Ville agréable ne rime pas nécessairement avec ville silencieuse.

Corriger l'environnement sonore uniquement suivant des paramètres acoustiques, par définition objectifs, ne suffit donc pas. Il est aujourd'hui communément admis que des mesures objectives de niveaux sonores (*e.g.* L_{Aeq}) ne peuvent, seules, rendre compte du confort sonore, et que vouloir influer sur l'environnement uniquement sur la base de paramètres acoustiques, par définition objectifs, ne suffit pas (Aletta et al., 2016; Kang and Zhang, 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. Les recherches se sont ainsi portées sur la notion de paysage sonore, envisageant la nuisance sonore et à travers les aspects qualitatifs, et à travers les aspects sémantiques des phénomènes acoustiques.

De plus, si beaucoup d'efforts sont faits afin de réguler les niveaux de bruits 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 d'oiseaux, ou d'eau, à des sons de trafics 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.

par un indiv. particulier

inférieur de 3dB
(ou alors c'est supérieur :-P)

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)

Un des enjeux actuels de l'analyse des paysages sonores est de relier ces données perceptives, établies à partir d'enquêtes, 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*⁸ (Schulte-Fortkamp and Kang, 2010) et *the Positive Soundscape project* (Davies et al., 2009; Davies et al., 2013). Mais les difficultés persistent (Ribeiro et al., 2013; Schulte-Fortkamp, 2013).

Afin d'acquérir la masse de données nécessaire permettant d'évaluer la qualité de l'environnement sonore sur un temps long, les caractéristiques d'un environnement variant au cours de la journée, comme au cours des saisons, (Tae Hong Park et al., 2014) ont lancé, à New York, un vaste projet collaboratif de 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.

3.4.3 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 ?⁹
- 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 apprêhendons l'environnement sonore ?

⁸ TD0804 : *soundscape of European Cities and Landscapes* : http://www.cost.eu/COST_Actions/tud/TD0804

⁹ 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.

2. Quels indicateurs influent sur ces descripteurs perceptifs ?

Développons. 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 3.4.4 pour plus de détails sur les descripteurs couramment utilisés). Un des enjeux de l'approche dimensionnelle est 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 ($L_{A\text{eq}}$), 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 permettant 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'agrément 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, on cite 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, GL : d'autres encore au moment de la journée. Contrairement aux indicateurs acoustiques, psychoacoustiques et perceptifs, qui sont tous évalués/mesurés suivant des échelles ordonnées, discrètes ou continues, certains indicateurs extra-sonores sont eux évalués sur des échelles de catégories¹⁰

¹⁰ Le terme catégorie, employé pour décrire un type d'échelle, n'a rien à voir avec le terme catégorie relatif au représentations mentales

ici

"est ici sans rapport" serait mieux

Nom	Acoustique	Description
L_A		Niveau sonore calculé avec une pondération A
$L_{A\text{eq}}$		Moyenne des L_A
L_{A10-90}		10-90ème quantiles des L_A
$L_{A\text{min}}, L_{A\text{max}}$		minimum maximum des L_A
Facteur crête	Ratio entre la valeur de pression maximale et la valeur RMS	

TABLE 1 : GL : TODO Indicateurs acoustiques

(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. On note cependant qu'avec le temps, la communauté privilégie l'approche dimensionnelle.

3.4.3.1 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 ;
- de comprendre l'influence de ces représentations sur le traitement de l'information sonore.

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

Afin d'objectiver la nature des catégories mentales représentant des paysages sonores, ou des sources sonores, l'approche catégorielle peut avoir recourt à trois types d'expériences (cf. Figure 12) :

- *Tâche de description* : On demande au sujet de décrire l'environnement sonore auquel il a été exposé (Axelsson et al., 2005;

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, 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 délicat. 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, (Guastavino, 2007; Maffiolo, 1999), en groupes ou paquets, suivant une consigne fixée en fonction des objectifs mêmes de l'expérience. 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 sujet reste libre d'organiser les stimuli comme il l'en-

tend. Ces expériences de tri sont pratiquées en laboratoire, en utilisant 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 alors d'obtenir une matrice de similarités illustrant les ressemblances entre tous les stimuli. Via un positionnement multidimensionnel (*Multidimensional scaling*, cf. Annexe B.3), il est alors possible de retrouver l'espace rendant compte au mieux de ces similarités. GL : TODO : Citer à partir de (Gygi et al., 2007) cette méthode a été utilisée pour comprendre la notion de timbre. De la position des stimuli dans l'espace on peut alors 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.

L'avantage de ces pratiques expérimentales est double :

1. GL : 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 3.4.3.2). Présupposer des dimensions intervenant dans la comparaisons de plusieurs stimuli, c'est en effet prendre le risque que ces dimensions ne fasse 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 similarité physiques, mais également sémantiques ;
2. GL : elles profitent d'une information riche via l'utilisation de descriptions verbales (tâche de description ou de catégorisation avec verbalisation). Quelles soient libres ou rattachées à des groupes, l'analyse de ces descriptions 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 3.2.2.2). Le langage agit ici comme senseur qualitatif.

GL : TODO : ajouter : discussion sur les outils d'analyse mds et analyse discriminante

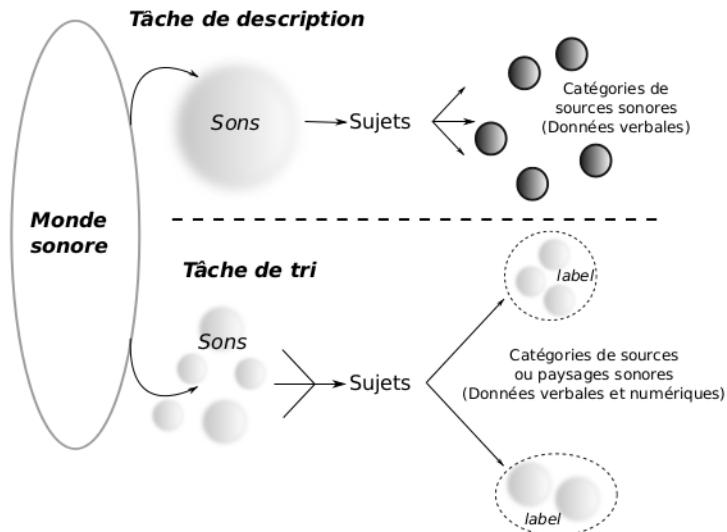


FIGURE 12 : Tâche de description et tâche de tri ou de catégorisation

3.4.3.2 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 séquentielle 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, ou dans un cadre *in situ*, par le biais 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 (Kardous and Shaw, 2014; Ricciardi et al., 2015). L'outil présente plusieurs avantages. Il permet une collecte des données sur serveur directement, et il 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 3.1.4).

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 (cf. Annexe B) 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, au sein d'un groupe de descripteurs perceptifs et/ou objectifs donné, sélectionne 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*, cf. Annexe B), 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 entre les descripteurs objectifs. Par exemple, (Torija et al., 2013) évalue les corré-

lations 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 (cf. Annexe B) 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 différents 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 l'ensemble des 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 Organized 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 pallier ces difficultés. (Hall et al., 2013) évalue ainsi l'agrément en utilisant une échelle de 9 points dont les extrémités sont décrites par les triplets désagréable-mécontent-insatisfait / agréable-content-satisfait. Ce biais peut encore être réduit en apportant un soin particulier à la sélection des termes extrêmes afin de s'assurer que ces derniers soient bien appropriés, par exemple, en menant une expérience intermédiaire sur la base d'un questionnaire libre (Guastavino and Katz, 2004), réalisable en condition *in situ* (Hong and Jeon, 2013; Kang and Zhang, 2010), ou en demandant au sujet d'expliquer verbalement sa notation (Raimbault, 2006); **items complexes : point et maj.**
- tous les sujets peuvent ne pas utiliser les échelles de la même manière. Certains sont portés à en utiliser toutes les valeurs. D'autres peuvent n'en privilégier que certaines, et notamment écarter les extrêmes (sans que, d'ailleurs, il soit possible de déterminer si ces variances entre les sujets sont involontaires ou, au contraire, décidées). 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.*) a priori non bornées de termes aux extrémités. 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 est moins pertinente s'agissant d'échelles sémantiques. S'ajoute à cela le fait 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 dis-

tributions des réponses avant de considérer des résultats moyennés.

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 décrivent ;
2. les échelles soient correctement interprétées par les sujets.

3.4.4 *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 sur :

1. la définition de ces descripteurs ;
2. les pratiques expérimentales permettant d'étudier ces descripteurs.

est-il vraiment utile de faire un \itemize ?

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).

3.4.4.1 *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 d'ailleurs 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 et d'autre part des 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).

GL : TODO : préciser ?

On évalue principalement la gêne en considérant l’influence des bruits issus des transports (routiers, aériens, ferroviaires) et/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 and Oudshoorn, 2001; Miedema, 2004), et ces derniers continus d’être revisités/améliorés (Gille et al., 2016b). Aujourd’hui, une attention particulière est portée sur l’influence de facteurs extra-sonores, comme par exemple l’activité du sujet, sa sensibilité au bruit, mais également son sentiment de peur suscité potentiellement par les sources sonores considérées (trafic, industrie) (Marquis-Favre and Morel, 2015; Morel et al., 2016). Afin d’être valide 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).

3.4.4.2 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 et al., 2007; Nilsson and Berglund, 2006) 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 relatif à la présence de différentes sources, ainsi qu'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 étendu 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 (cf. annexe A.4) 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 3.4.7) (Brocolini et al., 2012; Ricciardi et al., 2015). Il apparaît que le silence perçu, comme la qualité visuelle perçue, contribuent grandement à la qualité sonore perçue. 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 3.4.4.4).

GL : TODO : (Ozcevik and Can, 2012)

3.4.4.3 *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 ? ”

L'environnement est-il...

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 3.4.7) (García Pérez et al., 2012; Lavandier and Defréville, 2006).

Le contexte (physique, visuel, social, personnel, cf. Section 3.4.4.8) semble être d'une grande importance dans l'évaluation de l'agrément (Guillén and López Barrio, 2007).

3.4.4.4 *Confort acoustique*

GL : Un autre notion très proche de l'agrément est celle du confort acoustique. Comme pour l'agrément, le confort semble dépendre plus du type de source perçue (Yang and Kang, 2005) et du contexte d'exposition (Meng et al., 2013) que des caractéristiques physiques globales de l'environnement.

Ce paragraphe est un peu confus, on n'identifie pas tout de suite de quoi tu parles en priorité.
Je pense que tu ferais mieux de commencer par "(Brocolini et al., 2012) montrent que la régression linéaire multiple (cf annexe...)".

lesquelles ?

GL : TODO : G₂ : (Jeon et al., 2011, 2013)

GL : TODO : G₃ : (Tse et al., 2012)

GL : TODO : G₄ : (Yu and Kang, 2009) modèle du confort

3.4.4.5 Calme et tranquillité

(Delaire et al., 2012) a effectué une analyse lexicale du vocable français utilisé depuis le XVI^e 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.”

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.”

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 (Delaire et al., 2012).

Les études sur le calme sont complémentaires des études sur la gêne. 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 (Delaire et al., 2012). Une scène stable et amorphe (cf. Section 3.4.5.2 pour la définition de amorphe), composée de peu d'événements saillants, peut être vue comme un environnement très calme.

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 courrant dans l'environnement, et comment 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 de cet environnement (voiture, tracteur).

3.4.4.6 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 composante principale, (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 13), 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). Le paysage sonore est majoritairement perçu suivant son caractère agréable/calme ainsi que suivant son dynamisme/son 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 la variance. Cependant les auteurs indiquent 1) qu'il n'y a pas de relation évidente entre ces deux dimensions, et 2) que des descripteurs objectifs acoustiques seuls ne permettent pas de prédire avec précision les valeurs perceptives de ces dimensions.

tu as mis singulier ailleurs

3.4.4.7 Autre descripteurs

GL : Go : (Kang and Zhang, 2010) autres

GL : G1 : (Botteldooren et al., 2006) music

3.4.4.8 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, d'un contexte environnemental (température, chaleur, humidité), (Jeon et al., 2011; Meng et al., 2013), relatif au sujet (age, sexe) ainsi qu'à sa

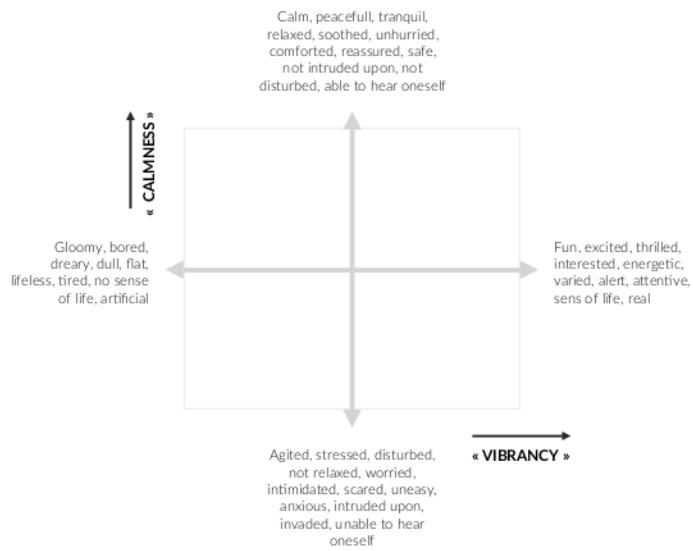


FIGURE 13 : Les dimensions de calme et de dynamisme permettant de caractériser l'environnement sonore urbain, d'après (Cain et al., 2013)

sphère socio-culturelle (Guillén and López Barrio, 2007; Hall et al., 2013; Yu and Kang, 2010), relatif à la configuration spatiale du lieu d'exposition (Hall et al., 2013), au éléments visuels (De Coensel and Botteldooren, 2006; Guillén and López Barrio, 2007), ainsi que 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.

3.4.4.9 Réponse physiologique

GL : TODO : (Hume and Ahtamad, 2013)

3.4.5 Catégoriser les sources et paysages sonores

3.4.5.1 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). Se plaçant dans le cadre de l'"écoute de tout les jours" (cf. Section ??), 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.

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)

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 à celles proposées par Gaver (Solide, Liquide, Gaz et Machine), et 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. montre que les sons s'organisent en trois clusters incluant les sons harmoniques, les sons d'impacts (discret) 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, 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 sa sémantique associée, sur les processus de catégorisation a particulièrement été étudiée. A partir d'une étude de 35 papiers traitant de catégories sonores, (Niessen et al., 2010) établissent une liste de 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

espacer

abstraits (*Événement, Background*). Ces catégories de sources ne s'expriment pas toutes au même niveau d'abstraction. Certaines sont précises (*Aboiement chien*), d'autres sont très larges (*Sons naturels*). Par ailleurs, certaines sont incluses dans d'autres (*Bruits de pas*<*Sons humain*), 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 3.4.5.2 et 3.4.7) 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 positivement connotées (*oiseaux, nature*), d'autres négativement connotées (*machine, alarme/signaux, train*), d'autres encore peuvent être jugées soit positives soit négatives comme *personne* (majoritairement positive), *trafic véhicule* (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* négativement perçues, les catégories *personne* (majoritairement positivement connotée) et *musique* ayant une connotation variable.

L'auteur 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 inhérentes aux catégories de véhicules vont dans ce sens : les catégories *trafic* (*voiture, moto/scooter, camion*) sont systématiquement négativement 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 véhicules individuels, sont largement bien acceptés.

Nous noterons cependant que l'étude de (Guastavino, 2006) est réalisée sans support sonore, les sujets n'ayant que leur mémoire pour se représenter l'environnement urbain idéal. On peut penser que dans ce cas, les attributs sémantiques sont particulièrement sollicités. Nous approfondissons ce point à la section 5.2.11.1.

3.4.5.2 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) montre l'existence de 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, et 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. du type de source présent ;
 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 montre que les descriptions des sons à basse fréquence peuvent se diviser en deux catégories appelées "événements sonores" et "bruit de fond". Dans les derniers, aucune source ne peut être identifiée.

bruits de fond

Raimbault et Dubois (Raimbault and Dubois, 2005), combinant les résultats obtenus par trois thèses (Guastavino, 2003; Maffiolo, 1999;

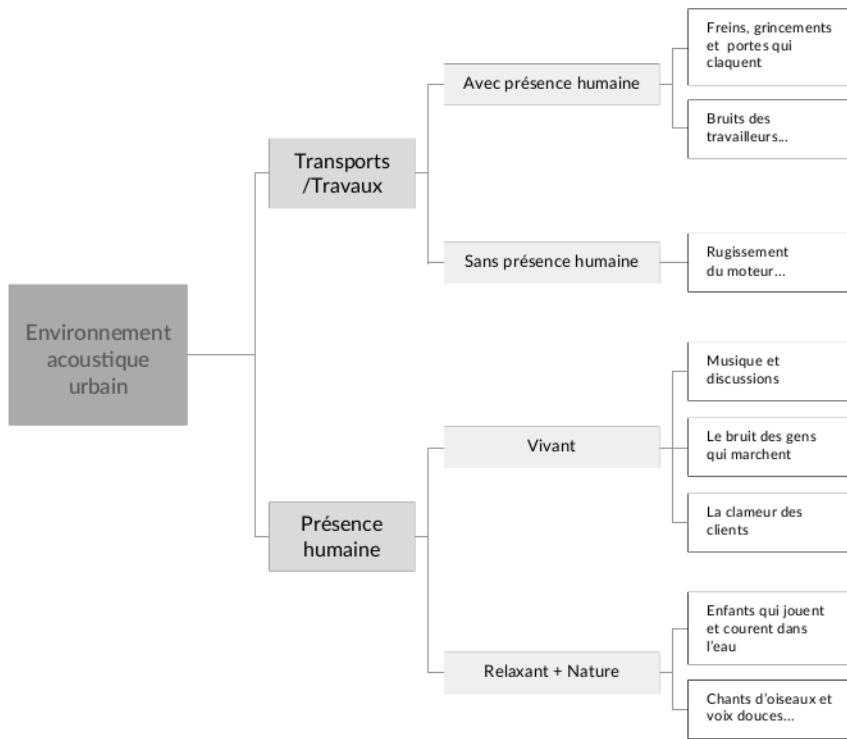


FIGURE 14 : Catégorisation des paysages sonores urbains, d'après (Rimbault and Dubois, 2005)

Rimbault, 2002), montrent que la catégorisation des paysages sonores s'opère, suivant leurs compositions, en termes de sources sonores (cf. Figure 14). Une première distinction s'opère entre d'un coté les paysages sonores comportant des sons de *transports motorisés* et/ou de *travaux*, et 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", et 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 (Schafer, 1977).

Des résultats très similaires sont obtenus par (Guastavino, 2007). Passant par une tâche de catégorisation libre avec verbalisation, Guastavino montre que la catégorisation s'opère suivant la présence/absence de sons d'origine humaine, ainsi que sur un jugement hédonique des sources. La présence de sons humains agit sur deux niveaux, 1) les environnements sont divisés entre ceux dominés par les sons humains, et ceux dominés par les sons mécaniques. 2) Les premiers se subdivisent en fonction de l'activité et du lieu (parc calme, marché actif). Les seconds se subdivisent encore à partir de la présence ou non de sons humains.

niveaux: 1) les mécaniques; 2) les

3.4.6 *Classifer les sources et environnements sonores*

Classer ! Pas de "classifer" en français :-P

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 le regroupement 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 3.4.3.2).

3.4.6.1 *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 (Raimbault 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 15. Cette classification est centrée sur l'objet. En partant de la taxonomie proposée par (Brown et al., 2011), (Salamon et al., 2014) propose une nouvelle taxonomie, plus détaillée, centrée, elle, à la fois sur l'objet et sur l'action (cf. Figure 18). Les auteurs partent de l'idée que 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 du plus bas niveau sont subdivisées en classes d'actions, labellisées par des verbes.

cette taxonomie

Tu ne préfères pas "étiquetées" ?
Là, ça fait un peu "viande de bœuf française"

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 provenant d'environnements urbains (*musique, mécaniques et trafic*) et d'environnements naturels (*eau, vent et oiseaux*). Chaque indicateur est calculé sur le signal, à l'aide d'une fenêtre glissante, et moyenné. En réalisant une analyse en composante principale sur 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).

3.4.6.2 *Classes de paysages sonores*

Beaucoup d'études analysent l'existence de similarités entre environnements sonores à partir d'indicateurs quantitatifs, qu'ils soient ob-

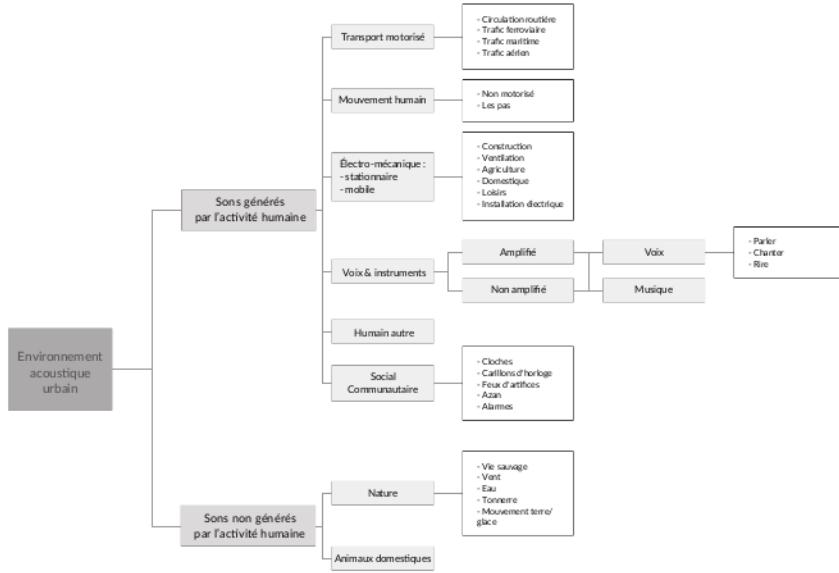


FIGURE 15 : Taxonomie des sources sonores urbaines, d'après (Brown et al., 2011)

jectifs (Rychtáriková and Vermeir, 2013), subjectifs (Jeon et al., 2013), ou 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 quatre classes comprenant respectivement, des environnements dominés par le bruit urbain, des environnements comprenant majoritairement des composantes naturelles, 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. Il apparaît que la distinction correspond à des différences 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) permettent à eux seuls d'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).

GL : TODO : contredire cette vérité

dans

(tu accordes les verbes au singulier, donc [Toto et Titi, 2010] désigne une publication, pas des auteurs)

3.4.7 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, joue un rôle important dans l'évaluation perceptive des paysages, particulièrement au niveau de l'agrément perçu (Guastavino, 2006; Szeremeta and Zannin, 2009).

GL : TODO : l'écoute (Gaver, 1993b)

Continuant dans ce sens, les études adoptant l'approche dimensionnelle cherchent 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 différentes études *in situ* sur la qualité de différents environnements, (Nilsson et al., 2007; Nilsson, 2007) montrent que l'identification des sources sonores permet de mieux prévoir la qualité globale de l'environnement que le niveau sonore. En particulier les sons *technologique/mécanique* ont un impact négatif sur l'environnement alors que les sons *naturels* ont un impact positif. Les sons *humains* restent cependant neutre. De plus, l'étude montre que dans le cas d'une exposition modérée au bruit de trafic, l'ajout de sons positivement perçus (*naturels* dans leur cas) peut potentiellement améliorer la qualité de l'environnement, une observation déjà effectuée par d'autres études (Galbrun and Ali, 2012; Hong and Jeon, 2013). Cependant, pour une exposition élevée au bruit, 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 relatifs au niveau (objectif : *loudness de Zwicker*; subjectif : niveau perçu) sont également pris en compte. La ré-

gression linéaire multiple est utilisée afin de mesurer l'influence des indicateurs sur l'agrément.

Que l'on considère les indicateurs subjectifs ou objectifs, l'utilisation combinée de l'indicateur de niveau global avec les indicateurs spécifiques aux différentes sources permet d'augmenter la capacité de prédiction de la qualité sonore, comparé à l'utilisation de l'indicateur de niveau global seul. Là encore les auteurs montrent que dans le cas où les environnements sont peu exposés au trafic, les sons d'*oiseaux* et d'*humain* ont un effet positif, la qualité augmentant en fonction de leur présence. Ils notent également que l'appréciation des *voitures* diffère en fonction du type d'environnement : dans un parc, elles ont un effet négatif alors que dans une rue, elles sont comprises comme faisant partie de l'environnement et n'influencent pas (de manière individuelle) 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 tous 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. Cette corrélation décroît à 58% si l'on supprime les indicateurs visuels, et tombe à 19% si l'on ne considère 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 sont mises à jours, les regroupements étant encore une fois relatifs à 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 le cas des modèles par classes, mais varient significativement d'une classe à l'autre. Par exemple, l'indicateur relatif aux sources d'*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 tout type de situations et d'environnements.

3.5 ÉVÉNEMENTS ET TEXTURES SONORES

Transition

3.5.1 *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 texture sonore.

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), voire les deux ;”
- “les caractéristiques haut niveaux des textures restent constantes sur de longue 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 les 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 celle d’événement sonore et 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 éventuellement par atteindre une asymptote. A contrario, 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 16). 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 celle de “bruit de fond” / “événements de premier plan” (*background/foreground*), utilisée dans le langage courant pour discriminer l’environnement urbain. Concernant les notions de *background* et de *foreground*, nous

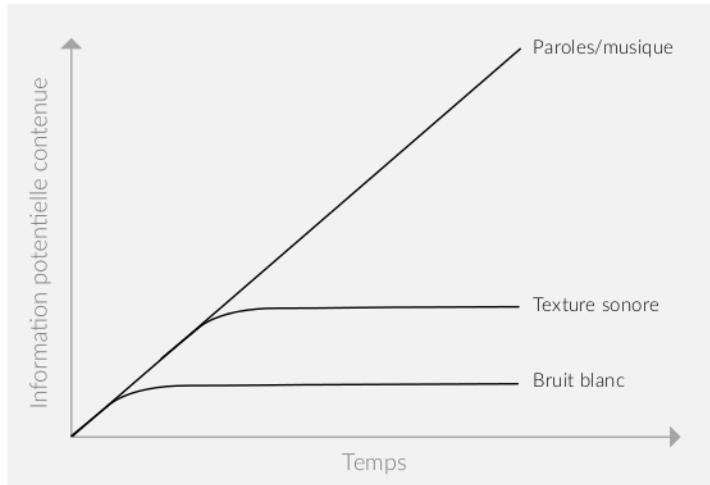


FIGURE 16 : Information potentielle contenue dans les séquences d'événements, les textures, et le bruit. D'après (Saint-Arnaud, 1995)

considérons que l'une, et l'autre peuvent être vues comme des flux auditifs, ces derniers pouvant être composés à la fois de textures et d'événements regroupés dans le but de faciliter le traitement auditif de la scène.

3.5.2 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 utile, 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 17). 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 stimuli 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, décision qui va ensuite influer sur la manière d'analyser l'information montante. L'expérience prouve que cette prise de décision n'a rien d'anodine, 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 texture (*e.g.* trois sons synthétisés de pluie), dont deux étaient produits à partir des mêmes statistiques extraites, ont du identifier le troisième, issu de statistiques différentes (Figure 17). 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. Partant du principe que le signal sonore est 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.

Une des avancées majeures de ces études est qu'elles apportent de nouvelles réponses sur la nature des représentations sonores stockées en mémoire. Dans le cas des textures, il s'agirait ainsi de descripteurs bas-niveaux, 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 en mémoire qu'une information condensée, qui lui permet pourtant de traiter des sons potentiellement longs.

Il a été montré que le cerveau peut stocker bien plus que des statistiques. (Agus et al., 2010) a mis en évidence qu'un bruit blanc, écouté de manière répétée, pouvait être reconnu encore plusieurs semaines après l'écoute, et ce parmi d'autres bruits blancs. Dans ce cas le cerveau emmagasine bien la totalité du signal acoustique.

Et alors, comment a-t-il "su" que ce son-là, il fallait le stocker in extenso parce que deux semaines plus tard il allait y avoir interro ? Le travail d'Agus remet assez profondément en cause ce que tu as dit juste avant, il faut développer... Peut-être que ça marche parce que Agus travaille avec des fragments très courts -- à contrario du travail précédent qui montrait que plus c'est long, moins on spécifie ? Ou effet d'apprentissage "par cœur" déclenché par la répétition ?

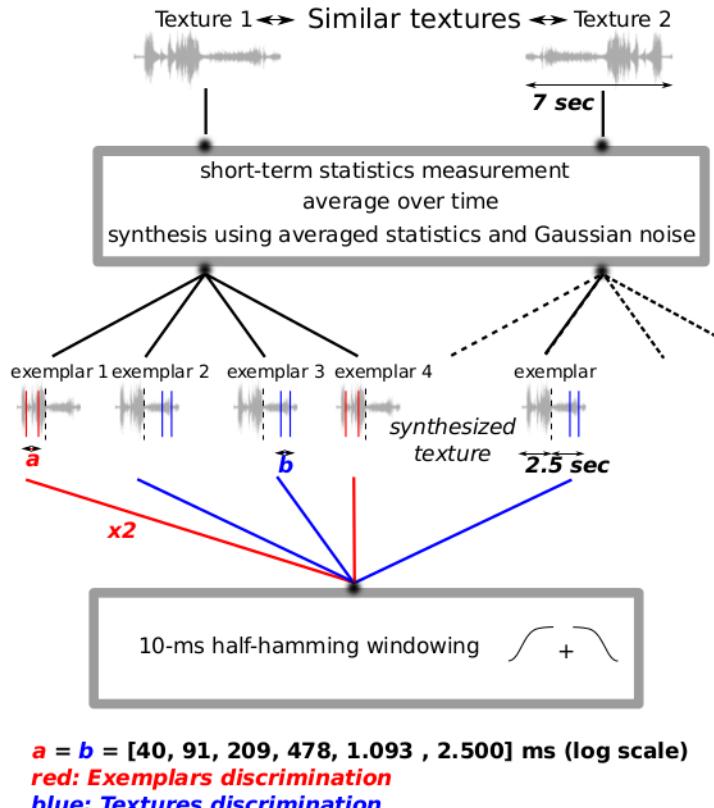


FIGURE 17 : Plannification expérimentale des expériences de discrimination de textures sonores et d'exemplaires de textures sonores menées par (McDermott et al., 2013)

3.5.3 Discussions

GL : TODO : terminer sur une discussion "texture, un groupe d'evt qui ont perdu leur signification individuelle" + "importance de l'attention span" (dire que ce point a été étudié : introduire l'expérience) + "problème avec le bruit blanc(Agus et al., 2010), texture non informatif mais pourtant stockée en totalité par le cerveau"

Ah ok, c'est peut-être là que tu voulais en parler, au temps pour moi

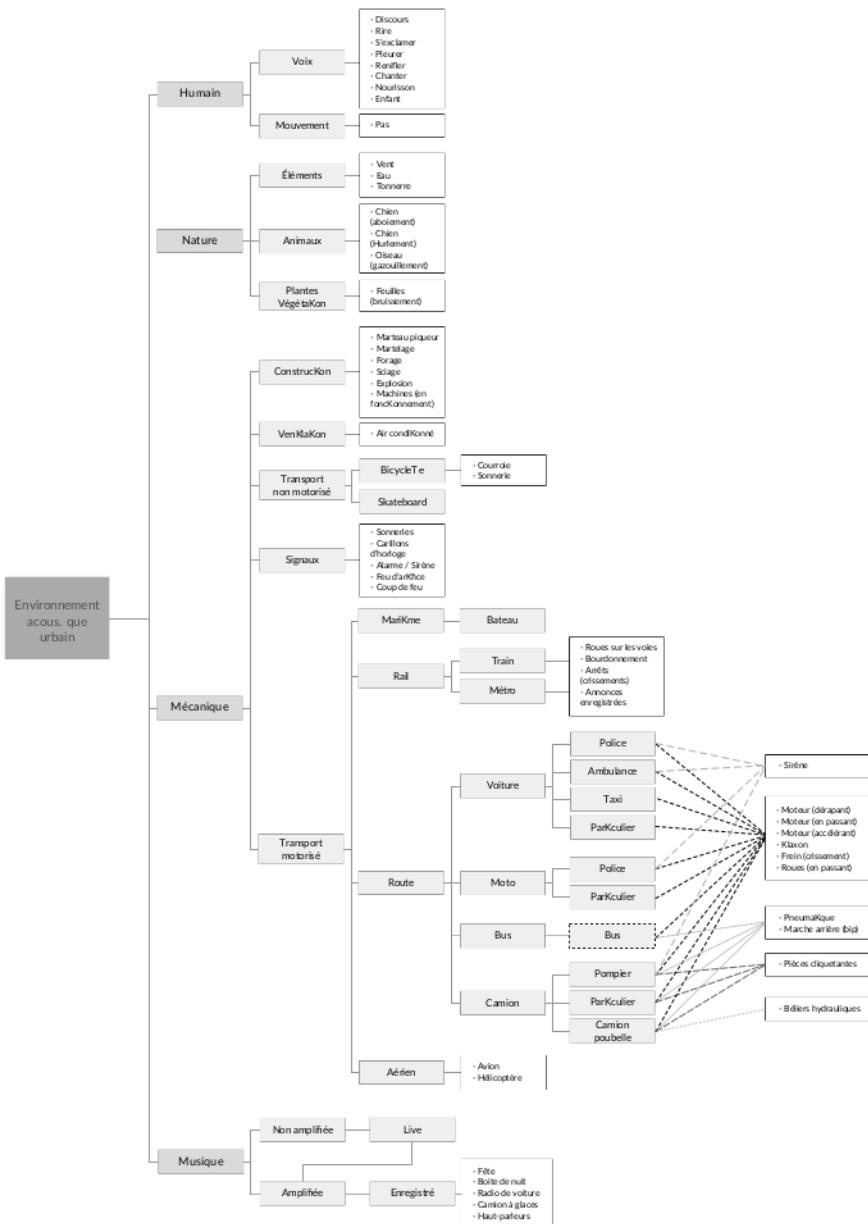


FIGURE 18 : Taxonomie des sources sonores urbaines, d'après (Salamon et al., 2014)

Souci de structure, le chapitre est censé parler de modèle morphologique, mais tu pars plein pot sur la simulation. Le choix de faire de la simulation relève plutôt d'une décision méthodologique dans tes travaux -- pour moi, tout ce qui cause de l'intérêt de l'approche simulation devrait être dans le chapitre 5...
Bien sûr, ça pose la question de "peut-on présenter le modèle indépendamment de l'usage qui en sera fait ?", et "si non, comment arranger tout ça ?". Il faut qu'on en discute, mais je pense que le modèle se tient "en soi".

4

UN MODÈLE MORPHOLOGIQUE DE SCÈNES SONORES ENVIRONNEMENTALES

4.1 MOTIVATIONS

4.1.1 Analyse sensorielle

Comme nous l'avons vu à la section [3.4.3](#), l'étude expérimentale des paysages sonores adopte deux approches :

- l'approche catégorielle : la mise en évidence des catégories d'environnements ou sources sonores ;
- l'approche dimensionnelle : l'identification des dimensions perceptives engagées dans les processus cognitifs, et des indicateurs dont elles dépendent.

Ces approches s'inscrivent dans des démarches différentes :

- l'approche catégorielle veut identifier les objets d'intérêt de l'environnement sonore ;
- l'approche dimensionnelle veut identifier les éléments caractérisant les qualités affectives perçues d'une scène.

Si l'on interroge les influences qu'ont les différents éléments constituant une scène sur les qualités affectives perçues, alors les deux approches se rejoignent naturellement. L'approche catégorielle permet d'établir la liste des éléments d'intérêt, laquelle liste peut servir de base à une annotation des stimuli utilisés par l'approche dimensionnelle afin d'étudier les contributions spécifiques de leurs éléments respectifs.

Nous pensons que la simulation offre un cadre expérimental élégant permettant de faire le lien entre les deux approches [GL : TODO : plus insister ?](#).

D'une part, comme illustré à la figure [19](#), la simulation peut être vue comme l'approche inverse des épreuves catégorielles (cf. Section [3.4.3.1](#)). Quand les épreuves catégorielles discrétisent l'environnement sonore sur la base d'un tri et/ou d'une description verbale réalisés par un sujet, la simulation, elle, recompose cet environnement, à partir d'une banque imposée de sons isolés. Ainsi, les catégories sonores, point de sortie des épreuves catégorielles, constituent-elles la banque de sons, point d'entrée de la simulation.

D'autre part, la finalité de la simulation est de produire un environnement complet, dont la partition (cf. Section 4.2.2), *i.e.* ses caractéristiques structurelles et compositionnelles, est parfaitement connue. Les stimuli ainsi produits peuvent être utilisés par l'approche dimensionnelle, afin d'étudier de manière fine les contributions des différents éléments.

La simulation se pose donc comme un outil intermédiaire, faisant le lien entre les connaissances issues des études adoptant l'approche catégorielle, et les stimuli requis par l'approche dimensionnelle.

La simulation présente d'autres intérêts :

- *intérêt pratique* : afin d'étudier l'importance relative des différentes sources, il est indispensable de disposer de stimuli dont la partition est connue. Une première solution, adoptée par (Lavandier and Defréville, 2006), a été d'annoter les stimuli. L'annotation cependant est une solution limitée :

1. l'opération est fastidieuse, longue, et difficile à mettre en œuvre sur de grandes banques de données ;
2. 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 GL : TODO : citation.

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

- *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 ; inutile d'itemiser
- de la qualité Q.

Il est possible, dans les approches classiques, de résoudre ces problèmes en étudiant de manière préalable les stimuli à enregistrer (cf. Section 3.1.4).

La simulation, en renversant la question posée ("générer un environnement qui correspond à une certaine valeur de Q"), garantit la validité écologique des stimuli, par définition connectés à la représentation sonore du sujet, GL : à condition que la

banque de données et l'outil de simulation soient suffisamment expressifs.

- *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 3.1.4), et de répartir équitablement les stimuli parmi ces zones. GL : Cependant, l'environnement d'une même zone est changeant, aussi bien si l'on considère le type de source présent, que si l'on considère 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 voiture à la fin des horaires de travail). Il est donc nécessaire de contrôler la diversité des sources qui y occurrent ainsi que la diversité structurelle de leurs séquences d'émission, d'autant plus 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 contrôlable. 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 5).

Enfin, la plupart des environnements que nous percevons sont relativement neutres, et ne provoquent pas en nous de réactions particulières. Il peut n'être pas évident d'évaluer des dimensions perceptives comme l'agrément, la gêne ou le confort, de

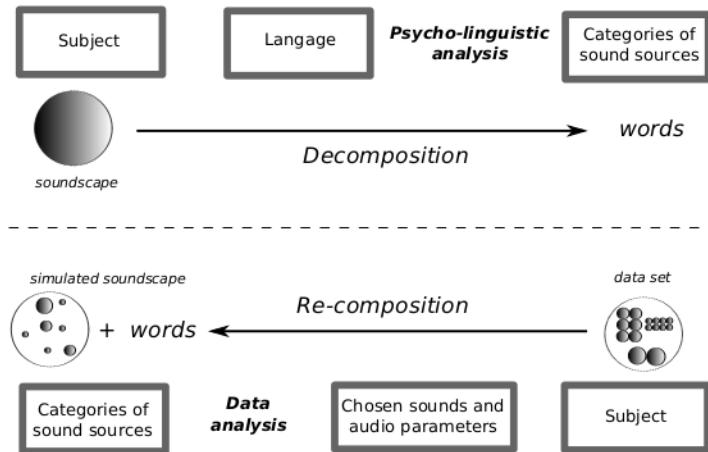


FIGURE 19 : TODO

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.

4.1.2 Analyse automatique

4.2 PROPOSITION D’UN MODÈLE DE SCÈNES SONORES

4.2.1 Discréter l’environnement sonore

4.2.1.1 L’unité de bases : la source sonore

Les études portant sur l’ASA, et plus spécifiquement sur les processus de ségrégation, montrent, d’une part, que l’humain 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.

GL : TODO : Neuroscience

Dans le même temps, 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.

Pour notre modèle, il apparaît fondé de considérer comme élément de base la source sonore. Comme vu précédemment (cf. Sec-

tion 3.2.3.3), la notion de source sonore est variable, un même objet pouvant être reconnu suivant plusieurs degrés d'abstraction.

4.2.1.2 Typologie : niveaux d'abstractions et nomenclature 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 3.2). 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. 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, au passage, devait donner naissance à la théorie de l'évolution (Lecointre and Le Guyader, 2006). Dans le domaine du son (ainsi que dans celui des odeurs), 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 3.2.2.2). 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 de 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 3.2.5), *i. e.* de la nature des sources cooccurentes 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 hiérarchisées. 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 3.4.5, 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 3.2.3.3), Guyot et al. (Guyot et al., 1997) proposent un système de catégorisation où les auditeurs identifient des groupements de sources abstraits, au niveau superordonné ("Bruit généré par une excitation mécanique"), des actions, au niveau de base ("gratter", "frotter") et des sources, 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, et ensuite, seulement, à partir d'actions.

L'association source-action semble être une base sensée sur laquelle bâtir une taxonomie, dans laquelle les classes haut-niveaux 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 choisir, il est possible de se référer aux travaux de Gaver (Gaver, 1993b), qui propose une taxonomie phénoménologique des sons, aux travaux

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, aux travaux de Salamon *et al.* (Salamon et al., 2014), qui, partant des travaux de (Brown et al., 2011) et reprenant l'association source-action, élabore une taxonomie de sons urbains.

4.2.1.3 événements, textures et scènes amorphes

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

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

Afin de contourner le problème, on peut là encore s'appuyer sur des considérations perceptives pour établir, dans un contexte expérimental donné, quels sons requièrent d'être enregistrés séparément, et quels groupes de sons peuvent être enregistrés ensemble.

D'une part, 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).

Maffioli montre à ce sujet (cf. Section 3.4.5.2) 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.* comportant des événements identifiables.

D'autre part, 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. **Voir les travaux sur les textures sonores** (cf. Section 3.5).

Ces résultats 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. A l'inverse, quand le cerveau ne parvient pas à isoler d'événement, la scène est traitée globalement, tous ces éléments étant agglomérés dans un même flux.

Ainsi, trois types de sons semblent pouvoir être isolés :

- événement sonore : un son ponctuel dont les caractéristiques physiques varient au cours du temps ;

- texture sonore : un son 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* : un son contenant une information sémantique élevée ;
- *scène amorphe* : un son contenant une faible information sémantique.

Il est concevable qu'il existe des connexions entre les notions de textures/événements, et celles de scènes amorphes/événemementielles.

Les séquences événementielles peuvent être vues comme des séquences composées soit uniquement d'événements, soit d'événements et de textures, la présence d'événements, porteurs d'une information plus riche, primant quant au choix du processus à mettre en œuvre

Les textures et les scènes amorphes 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. Un exemple de texture souvent cité est le son du "galop", qui selon le contexte peut se retrouver au premier plan de la scène.

Partant de là, il est 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 GL : TODO : citation. Cependant, l'inverse, considérer une texture comme une scène amorphe, n'est pas forcément vrai.

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

4.2.2 *Description du modèle morphologique*

4.2.2.1 *Classe et collection de samples*

Dans le modèle proposé, la scène sonore est vue comme une somme de sources sonores, ou autrement dit, "un squelette d'événements sur un lit de textures" (Nelken and Cheveigné, 2013). GL : pas d'équivalence, faire un lien entre les deux

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

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

Ces samples, regroupés en classes de sons hiérarchisées, **forment** une taxonomie. 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, regroupant des samples similaires (*e.g.* voix-adulte-cri).

sont organisés selon une taxo
(ils ne la "forment" pas, elle est
une abstraction qui leur pré-existe)

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 bas niveau utilisent la nomenclature source-action (*e.g.* voiture passe). Quant aux classes du dernier niveau, elles correspondent à des collections de samples, par définition, équivalents les uns aux autres.

4.2.2.2 Séquences de samples

Chaque classe de sons, sélectionnée pour faire partie de la scène simulée, est liée à une piste. Cette piste est une séquence temporelle où sont positionnés les différents samples. Elle est la contrepartie simulée 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 une partition.

Définition 4 *La partition 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 4.2.2.3).*

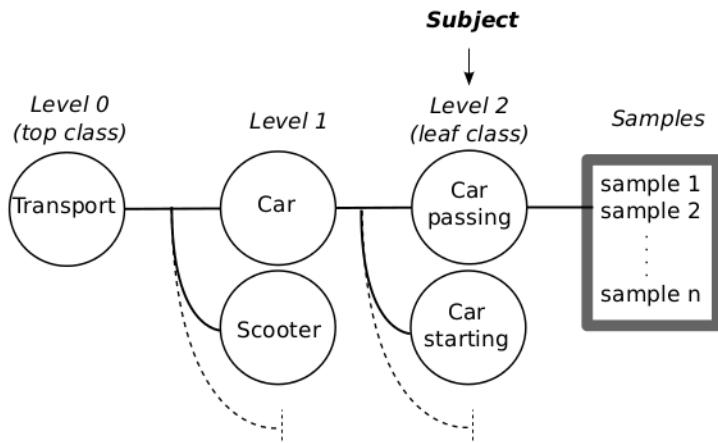


FIGURE 20 : Organisation hiérarchique de la banque de sons isolés utilisée pour la simulation

4.2.2.3 Paramètres

En suivant la terminologie ci-devant introduite, une scène sonore est vue comme une somme de pistes. Chaque piste est une séquence temporelle, dont la structure dépend d'une série de paramètres (cf. figure 21). Le modèle ne propose pas d'interagir avec un sample en particulier, mais toujours avec une séquence de samples.

Nous isolons trois attributs 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).

4.2.2.4 Formalisation du modèle

Tout au long de nos travaux, nous avons utilisé plusieurs modèles, ainsi que différents paramètres, afin de simuler des scènes sonores. Nous formalisons, dans cette partie, une version générale du modèle proposé. Les diverses modifications appliquées au modèle, en fonction de son utilisation, sont indiquées dans les sections 4.3 et 4.4.

Pas la peine de dire qu'il y a eu plein de modèles, seule la version aboutie est pertinente.

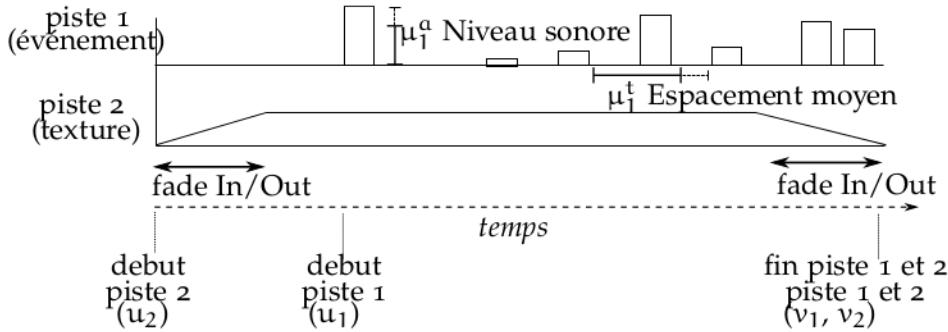


FIGURE 21 : TODO

La formalisation présentée 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 $c_{i,m}$, $1 < m < |c_i|$.

Une piste p_i est définie comme une séquence de n_i samples d'événements $e_i^k(n)$ ($k = (1, 2, \dots, n_i)$), choisis aléatoirement parmi les $|c_i|$ samples de la classe c_i . Considérant $\mathcal{U}(x, y)$, une distribution uniforme d'entiers allant de x à y ($x < y$), on a alors :

$$e_i^k = c_{i,\mathcal{U}(1,|c_i|)} \quad (2)$$

Pour chaque piste p_i , un facteur d'amplitude est tiré aléatoirement à partir d'une distribution normale de moyenne μ_i^a et de variance σ_i^a . De même, les espacements inter-onsets sont tirés d'une distribution normale de moyenne μ_i^t et de variance σ_i^t . Les indices temporels de début et de fin de chaque piste sont notés u_i et v_i respectivement. Formellement, une piste p_i se définit comme suit :

$$p_i(n) = \sum_{j=1}^{n_i} \mathcal{N}(\mu_i^a, \sigma_i^a) c_{i,\mathcal{U}(1,|c_i|)}(n - n_i^j) \quad (3)$$

$$n_i^j = n_i^{j-1} + \mathcal{N}(\mu_i^t, \sigma_i^t) \quad (4)$$

où $n_i^0 = u_i$ par convention. Le signal d'une piste est défini de telle sorte que $p_i(n) = 0$ si $n > v_i$. Les paramètres du modèle sont, μ_i^a , σ_i^a , μ_i^t , σ_i^t , u_i et v_i , et doivent être réglés pour chaque piste p_i . La figure 21 offre une illustration de l'action des paramètres introduits.

Répétition de ce que j'ai déjà dit, mais ce mélange entre modèle générique et modèle "simplifié" pour la simulation (par ex. espacements inter-onsets tirés dans une distrib normale) me pose problème.

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

1. l'amplitude du signal (μ_i^a, σ_i^a) n'est tirée qu'une seule fois, et la valeur est appliquée à tous les samples ;
2. 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.

GL : TODO : à revoir avec mathieu et mathias

4.3 DU MODÈLE À LA SIMULATION : L'ANALYSE SENSORIELLE

Dans cette section nous présentons une version du modèle précédemment introduit, afin qu'il puisse servir de base à un outils de simulation, nommé *SimScene*, utilisable dans le cadre de l'analyse sensorielle des scènes sonores.

Nous commençons par présenter différents outils existant, permettant de simuler des environnements sonores. Nous proposons, par la suite, un protocole expérimental, décrivant le cadre applicatif des épreuves perceptives basées sur la simulation. Enfin, nous relions ce protocole au modèle de scènes sonores (cf. Section 4.2.2.3), et présentons les fonctionnalités de l'outil *SimScene*.

4.3.1 *Simulation et analyse sensorielle*

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 pouvant être 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.

A notre connaissance, seuls 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 permettant au sujet d'agir sur un environnement en ajoutant ou supprimant des sources sonores spécifiques. Celui-ci peut par ailleurs 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. Les résultats montrent que l'inclusion ou l'exclusion des sources dépend 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.

GL : TODO G1 : compléter Bruce *et al.*, cf wac
GL : TODO G2 : (Davies *et al.*, 2014) montre que, lorsqu'on demande à des participants de simuler un paysage sonore, les simulations font références à ce que ces derniers s'imaginent être un environnement typique, sans tenir compte de leur propre préférence pour des sons particuliers.

4.3.2 Protocole expérimental basé sur la simulation

4.3.2.1 Organisation des sons isolées

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 4.2.2.1. 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.

4.3.2.2 Sélection des sons isolées

L'objectif de la simulation est d'obtenir une image sonore de la représentation mentale que ce 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 une représentation interne d'un sujet, cette approche pose trois problèmes majeurs :

- les sons peuvent ne pas être tagués 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 par le sujet, peut potentiellement influencer sa sélection. Dans les faits, pour 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 **GL : à travers les labels** décrivant les classes est un risque. 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 [4.3.4](#).

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 isolées (cf. Section [4.3.4](#)) ;
- 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.

4.3.2.3 *Processus de simulation*

Trois étapes composent le processus de simulation (cf. Figure [22](#)) :

- *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 ;

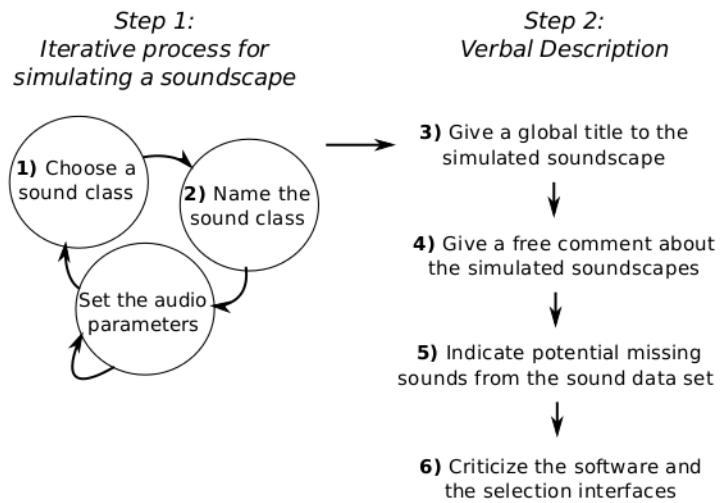


FIGURE 22 : Etape de processus de simulation pour l'analyse sensorielle

- *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 4.3.3).

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

4.3.3 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 4.2.2.3 et 4.2.2.4), 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 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 variance ;
- *début et fin* (seconde) : le sujet fixe le début et la fin de chaque piste.

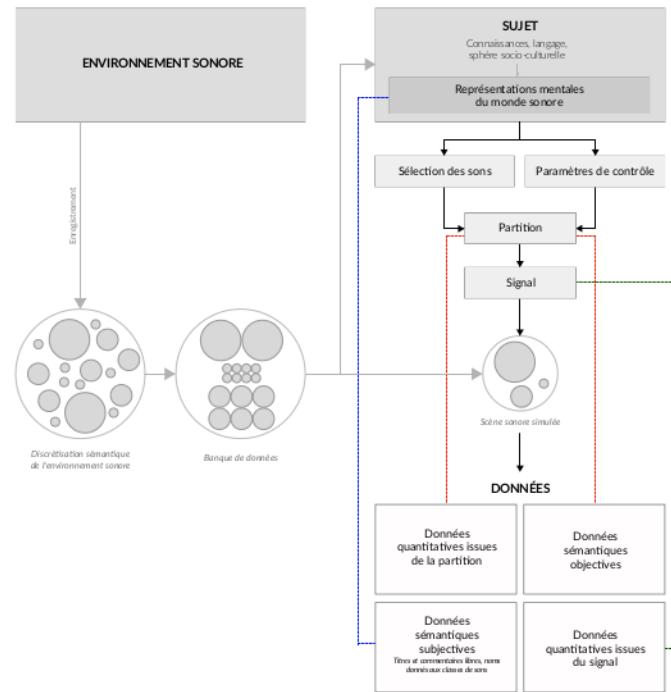


FIGURE 23 : TODO

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 séparément, 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 4.2.2.3)

4.3.3.1 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 23. Nous les résumons dans la liste suivante :

- 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 ;

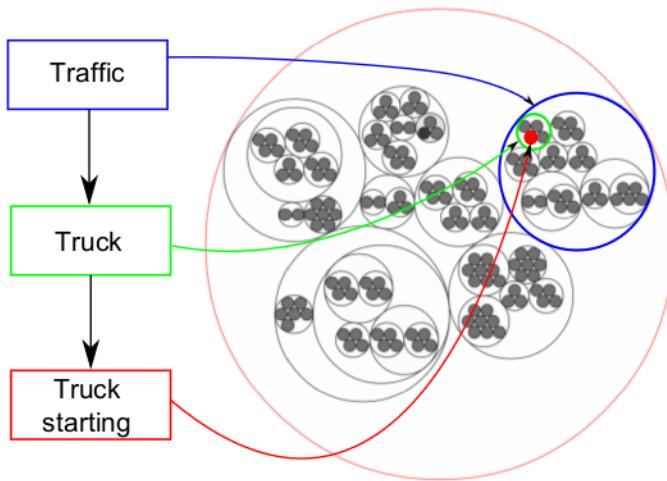


FIGURE 24 : L'interface de sélection aveugle de l'outil de simulation *Simscene*

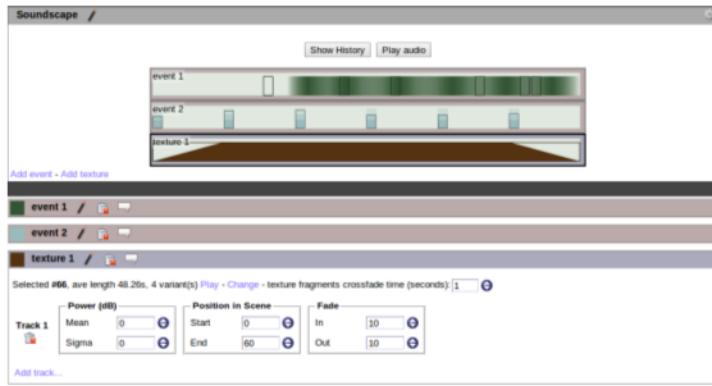
- données sémantiques subjectives : il s'agit des noms donnés par le sujet 1) à la scène simulée, 2) aux classes de sons sélectionnées ;
- données quantitatives issues de la partition : il s'agit de toutes les données relatives à la partition, *i.e.* pour chaque piste, le positionnement des samples et les paramètres (cf. Section 4.2.2.2) ;
- 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.

4.3.4 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 la même classe sont proches les unes des autres, et ainsi de suite jusqu'à atteindre les classes des niveaux d'abstraction élevés.

FIGURE 25 : L'outil de simulation *Simscene*

La figure 24 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 4.2.2.1).
2. permettre aux sujets de rapidement apprêhender 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 parcours 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).

GL : TODO : résumer les résultats JAES

4.3.5 Interface de simulation : l'outil Simscene

Simscene est un environnement de travail audio-numérique, supporté par navigateur internet, et développé dans le cadre du projet HOULE¹. Il permet de simuler des paysages sonores à partir d'un corpus de sons. Il est prévu pour fonctionner via les navigateurs internet *Chrome* et *Firefox*. L'outil a été développé en javascript à l'aide de la bibliothèque *angular.js*² et du standard *web-audio*³. L'interface de sélection (cf. Section 4.3.4) a été développée à l'aide de la bibliothèque *D3.js* (Bostock et al., 2011).

¹ Pour plus d'informations sur le Projet HOULE voir <http://houle.ircam.fr/>

² *angular.js* : cf. <https://angularjs.org/>

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

Simscape est présenté en détail dans (Rossignol et al., 2015). GL : Nous résumons ici ses fonctionnalités d'importance pour notre étude.

Le fonctionnement de *Simscape* se rapproche de celui d'un séquenceur audio. Chaque utilisateur choisit une classe de sons via l'interface de sélection (cf. Section 4.3.4). 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 4.3.3). Des champs de texte sont prévus afin de permettre à l'utilisateur de 1) nommer chaque piste, 2) donner un titre à la scène simulée et 3) commenter la scène simulée.

L'interface propose un rendu graphique schématisé de la scène en cours de création (cf. Figure 25). 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 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.

L'outil de simulation est accessible via le lien suivant : <http://www.irccyn.ec-nantes.fr/~lagrange/demonstrations/simScene.html>

4.4 DU MODÈLE À LA SIMULATION : L'ANALYSE AUTOMATIQUE

Un ensemble dédié de fonctions Matlab, qui sont accessibles au public⁴.

⁴ <https://bitbucket.org/mlagrange/simscene/downloads>

APPLICATION DU MODÈLE MORPHOLOGIQUE À L'ANALYSE SENSORIELLE DES SCÈNES SONORES ENVIRONNEMENTALES URBAINES

GL : TODO : concernant le titre : pourquoi modèle morphologique plus que "simulateur" ? il manque le coté "création" dans modèle morphologique

5.1 INTRODUCTION

Comme nous l'avons vu (cf. Section 3.4.7), la recherche sur les paysages sonores a besoin d'outils permettant d'analyser les influences séparées des différentes sources sur les qualités affectives de l'environnement. La simulation offre des possibilités intéressantes (cf. Section 4.1.1), car elle nous permet d'obtenir des scènes sonores dont nous connaissons tous les paramètres structuraux, en particulier les caractéristiques distinctes des différentes sources.

Afin de montrer les potentialités inhérentes à l'utilisation de scènes simulées en analyse sensorielle, nous choisissons, comme cadre applicatif, le problème de l'agrément perçu dans les environnements sonores urbains.

Cette section présente les résultats d'une série de quatre expériences visant, chacune, à comprendre comment les différentes sources sonores qui composent une scène influent sur la perception de l'agrément. Toutes ses expériences s'appuient sur la simulation. La première est l'expérience de simulation à proprement parler, *i.e.* où les sujets doivent créer les environnements. Les autres sont des épreuves de notation ou de tri classiques, utilisant les scènes simulées comme stimuli :

1. *expérience de simulation* : au cours de cette expérience, chaque sujet doit simuler deux environnements sonores urbains, le premier idéal/agréable et le deuxième non-idéal/désagréable, en utilisant l'outil et le protocole de simulation décrits à la section 4.3.5;
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.

Les expériences 1 et 2 vont de pair, et sont toutes deux décrites dans la section 5.2. Les expériences 3 et 4 sont, elles, décrites respectivement dans les sections 5.3 et 5.4

Tout au long de la présentation de ces expériences et de leurs résultats, nous poursuivrons deux objectifs :

- *objectif méthodologique* : montrer les possibilités offertes par l'utilisation de scènes simulées dont nous connaissons précisément la partition (cf. Section 4.2.2) dans le cadre d'études sensorielles ayant trait à la perception des sons ;
- *objectif applicatif* : étudier quels sont les éléments qui participent à la perception de l'agrément dans les environnements sonores urbains.

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

5.2.1 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 26) :

- *expérience de simulation* : comme précisé auparavant, au cours de cette expérience, les sujets simulent les environnements qui serviront de stimuli pour les étapes suivantes. Chaque sujet compose deux scènes opposées, les premières répondant à la description idéales/agréables (i-scènes), les secondes répondant à la description non idéales/désagréables. Cette épreuve de simulation a fait l'objet d'une expérience pilote (Lafay, 2013; Lafay et al., 2014); GL : TODO : préciser xp pilote
- *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 de chacune des scènes. 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 :

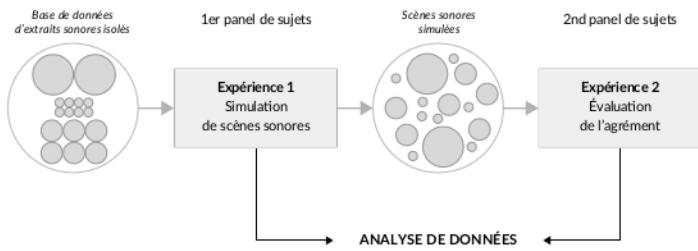


FIGURE 26 : Planification expérimentale des expériences de simulation et d'évaluation de l'agrément

1. évaluer l'influence des différentes sources sur l'agrément perçu, de manière séparée, 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.

5.2.2 Banque de données de sons isolés

Dans cette section, 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 4.3.2.1 et 4.3.4.

5.2.3 Typologie des sources sonores présentes dans l'environnement urbain

Afin de créer un corpus de sons isolés de référence pour la simulation, nous avons réalisé 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 la littérature. 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; Maffiolo, 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; Maffiolo, 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, les plus souvent citées dans ces publications. À partir des éléments relevés, nous établissons deux taxonomies : une pour les événements (cf. Figure 27a), une autre pour les textures (cf. Figure 27b). Comme évoqué à la section ??, la structure taxonomique de ces deux ensembles s'inspire grandement de l'axe vertical de l'organisation catégorielle proposée par E. Rosch (cf. Section 3.2.3.3), *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 en grande majorité par rapport à la source, et sont d'ordre sémantique. Pour les textures, nous considérons également la nature des lieux hébergeant ces dernières (*e.g. parc, rue*). La typologie des classes d'événements suit la nomenclature source-action introduite à la section (cf. Section 4.2.1.2). En ce sens, elle est très similaire à cette autre typologie de sources sonores urbaines, effectuée postérieurement (Salamon et al., 2014).

GL : L'appréciation de la musique étant un phénomène hautement subjectif, nous choisissons de ne pas considérer dans cette étude les sons revêtant un caractère musical (musicien de rue, auto-radio ...), et ce afin de ne pas provoquer un jugement qui dépendrait des acointances esthétiques du sujet en matière de musique.

5.2.4 Acquisition des sons isolés

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

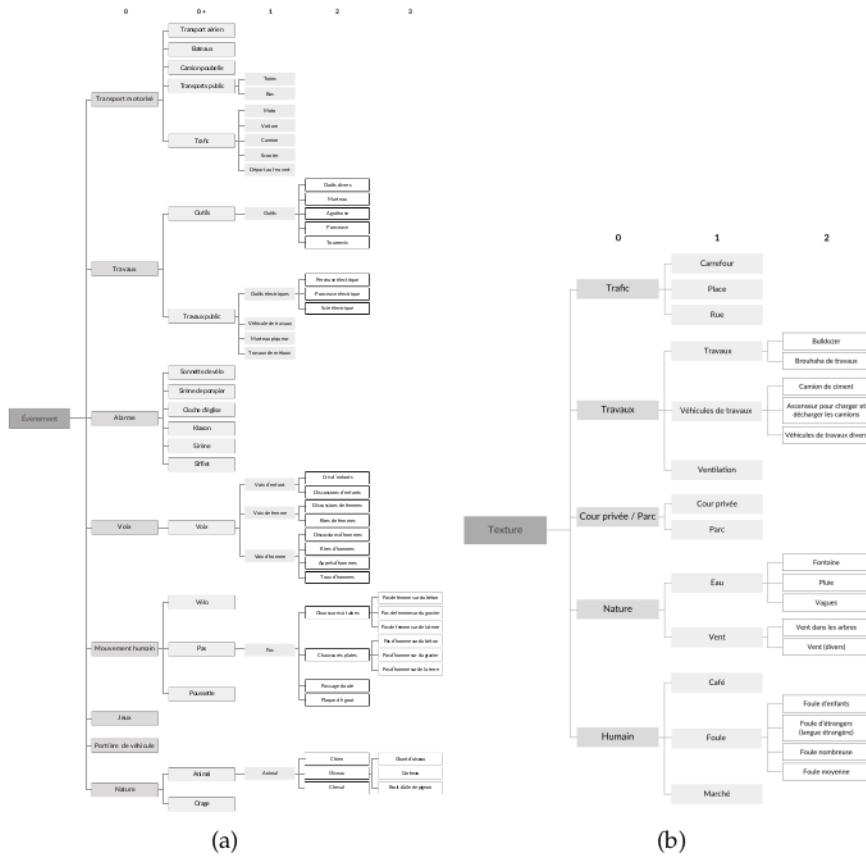


FIGURE 27 : Taxonomies des classes de sons utilisées pour la simulation des environnements sonores urbains pour (a) les événements sonores, et (b) les textures sonores. Nous présentons ici uniquement les niveaux d'abstraction 0 et 1. Un niveau intermédiaire, nommé 0+, et utilisé pour l'analyse, est également introduit.

Parmi les événements :

- 260 sont issus d'enregistrements effectués pour l'étude ;
- 89 sont issus de la banque de sons *SoundIdeas*¹ ;
- 32 sont issus de la banque de sons *Universal SoundBank*².

Parmi les textures :

- 72 sont issues d'enregistrements effectués pour l'étude ;
- 23 sont issues de la banque de sons *SoundIdeas* ;
- 7 sont issues de la banque de sons *Universal SoundBank*.

¹ Pour plus de détails sur *SoundIdeas* voir :<http://www.sound-ideas.com/>

² Pour plus de détails sur *Universal SoundBank* voir : <http://www.universal-soundbank.com/>

Tous les enregistrements ont été effectués à l'aide d'un micro canon *AT8035*³ relié à un enregistreur *ZOOM H4n*⁴. 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.

Tous les sons ont été normalisés au même niveau RMS⁵ de -12 dB (FS)⁶.

5.2.5 Planification expérimentale

5.2.5.1 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 aimeriez vivre) ;
- deuxième simulation : simuler un paysage sonore **urbain plausible** qui, selon vous, est non-idéal (où vous n'aimeriez 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 4.3.3). Ils doivent cependant se soumettre à deux contraintes :

- le sujet doit prendre le point de vue d'un auditeur fixe ;

³ cf. <http://eu.audio-technica.com/fr/products/product.asp?catID=1&subID=6&prodID=1845>

⁴ cf. <http://www.zoom.co.jp/english/products/h4n/>

⁵ 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 $x_{RMS} = \sqrt{\frac{1}{n} \sum_i x_i^2}$.

⁶ 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.

- le paysage sonore doit être réaliste, au sens de physiquement plausible. Autrement dit, le sujet à 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 4.3.2.3) :
 - sélectionner les classes de sons
 - nommer les classes de sons sélectionnées
 - paramétrier les pistes (cf. Section 4.2.2.2) relatives aux classes de sons sélectionnées
2. la réalisation 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.

Apparatus

Tous les sujets passent l'expérience sur des machines identiques ([GL : description des machines](#)). L'audio est diffusé en stéréophonie, 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

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

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 et/ou incompréhension des consignes, d'une part, dépassement du temps, d'autre part.

5.2.5.2 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 évaluée.

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 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.

Apparatus

Tous les sujets passent l'expérience sur des machines identiques ([GL : description des machines](#)). L'audio est diffusé en stéréophonie, par le biais de casques audio semi-ouvert *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

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.

5.2.6 *Données et méthodes d'analyses*

5.2.6.1 *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 la base de 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 5.2.5.2).

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 7 points. Nous notons $A_{scène}$ l'agrément moyen d'une scène, obtenu en moyennant les notes de tous les sujets. De même, nous notons 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 1, 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 ;

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
Idéal/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.

– *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 scalaire ci-devant mentionné comme unique descripteur objectif du niveau sonore.

5.2.6.2 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* : GL : afin de vérifier que la distinction affective imposée entre les i- et ni-scènes se retrouve au niveau de l'agrément perçu, nous observons si il existe des différences entre les deux types de scènes au niveau de $A_{scène}$ et A_{sujet} . La significativité est évaluée par un test de Student à deux populations indépendantes pour $A_{scène}$, et par un test de Student à deux population appariées pour 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.3). 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'environnements 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'environnements k, c_j le nombre de classes j utilisées, et c_{jk} le nombre de classes j utilisées pour

un type d'environnements k. Le V-test teste l'hypothèse nulle que la proportion $\frac{c_{jk}}{c}$ ne diffère pas significativement de la proportion $\frac{c_k}{c}$. 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 représentations induits par les descripteurs sémantiques* : afin d'étudier si une représentation basée uniquement sur la présence ou l'absence des classes de sons permet de séparer les deux types d'environnements, 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) est appliquée. Pour les valeurs p, dans le cas où la valeur $p \geq 0.05$, 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.3.

5.2.7 Validité écologique de l'expérience

5.2.7.1 Diversité de la banque de sons

Nous voulons 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 bien 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.

5.2.7.2 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).

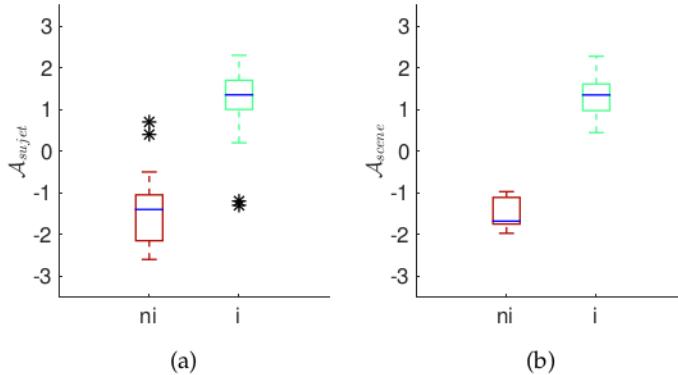


FIGURE 28 : Dispersion des notes données par les sujets lors de l'expérience 1.b moyennées suivant les sujets (A_{sujet} : a), et suivant les scènes (A_{scene} : b), en fonction du type de scènes (i ou ni).

5.2.7.3 Ergonomie de l'interface de simulation

5.2.8 Vérification de l'agrément des scènes simulées

Nous analysons ici l'agrément perçu des 80 scènes sonores simulées. La Figure 28a affiche l'agrément moyen A_{scene} 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 A_{scene} 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 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, nous observons l'agrément moyen de chaque sujet A_{sujet} , calculé séparément, pour chaque type d'environnement (cf. Figure 28b). Il apparaît que les i-scènes ont bien été perçues comme significativement plus agréables ($p < 0.01$) que les ni-scènes.

5.2.9 Étude comparative entre les descripteurs structurels

En premier lieu, nous nous concentrons sur le niveau sonore. Les figures 29a, 29b et 29c 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, GL : fait reporté dans un grand nombre d'études GL : TODO : citation. Nous constatons encore que cette différence de ni-

veaux 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, d'abord, de manière globale, ensuite, 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 perceptive 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'environnements. 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 A_{scene} .

En second lieu, nous nous penchons sur les densités de sources sonores. Les Figures 30a et 30b 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. **GL : Au vu de 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 globale de qualité, si 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'événement minimale des ni-scènes.

En dernier lieu, nous nous intéressons à la diversité. Nous affichons sur la figure 31 $DIV(E)$ et $DiV(T)$, en séparant les différents niveaux d'abstractions. Excepté pour le niveau d'abstraction 0, la diversité de 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 pré-

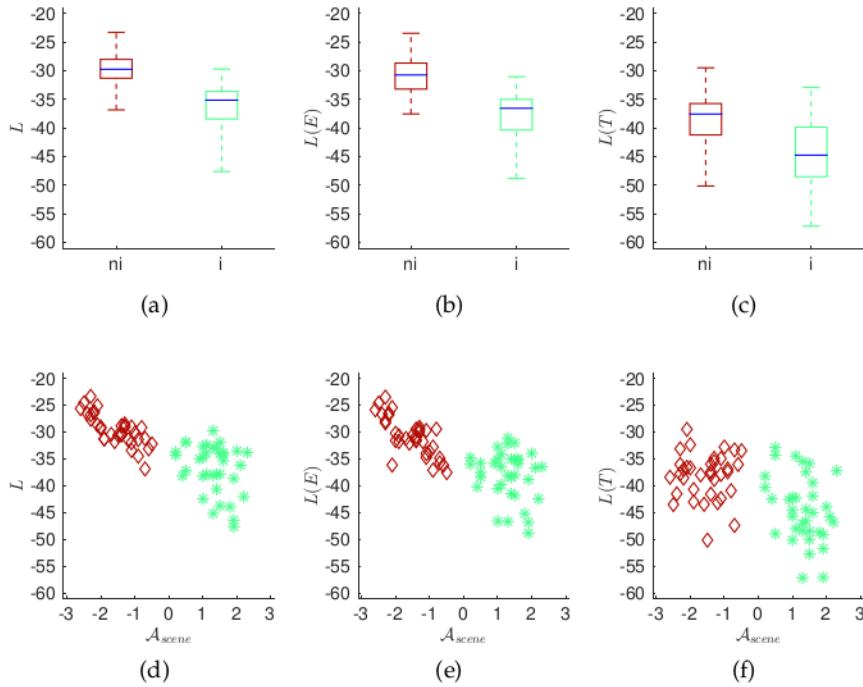


FIGURE 29 : Dispersion des descripteurs structurels de niveaux sonores L (a, d), $L(E)$ (b, e) et $L(T)$ (c, f), en fonction du type de scènes (a, b, c) et de l’agrément perçu A_{scene} de l’expérience 1.b (d, e, f).

sentées 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, qu'un environnement sonore idéal. Elles montrent, 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'environnements, distinction pourtant perçue de manière non ambiguë par les sujets.

5.2.10 Influence des descripteurs structurels sur l’agrément perçu

Nous analysons, dans cette section, 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 A_{scene} comme descripteur perceptif. Il s'agit d'étudier l'existence de potentielles corrélations entre les descripteurs structurels et A_{scene} . Les coefficients de corrélation linéaires calculés entre A_{scene} vs. L, L(E), L(T), D, D(E) et

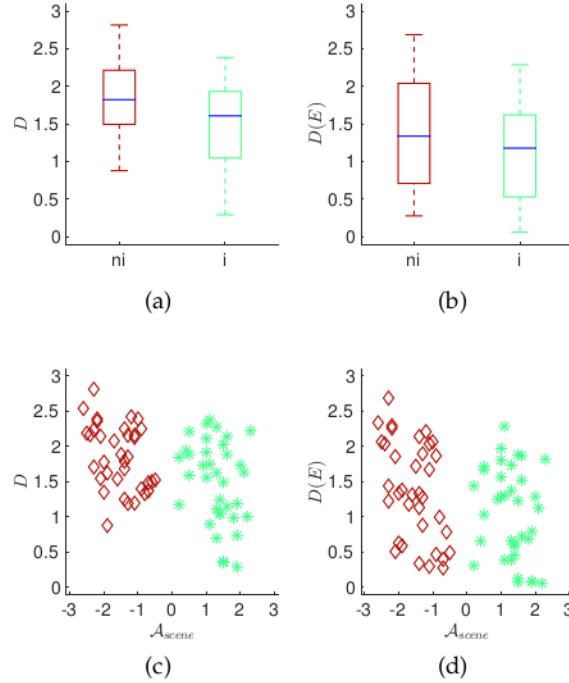


FIGURE 30 : Dispersion des descripteurs structurels de densité D (a, c) et $D(E)$ (b, d), en fonction du type de scènes (a, b) et de l'agrément perçu A_{scene} de l'expérience 1.b (c, d).

$DIV(E)$ sont présentés dans le tableau 6. Les relations entre A_{scene} et les descripteurs structurels sont illustrées par les figures 29d, 29e et 29f, pour les niveaux sonores, et les figures 30c et 30d, pour les densités.

Concernant L , on observe une forte corrélation négative ($r = -0.77$, $p < 0.01$) avec A_{scene} , indiquant que plus le niveau sonore est élevé, plus la scène est désagréable. Cependant, la figure 29d 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 A_{scene} , pour les ni-scènes, reste élevée ($r = -0.78$, $p < 0.01$), mais est inexistant 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, 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.

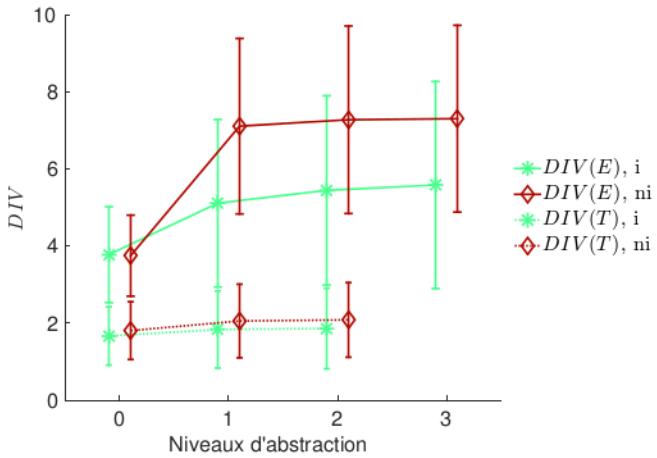


FIGURE 31 : Moyenne et écart type de la diversité des classes utilisées en considérant l’ensemble des classes (DIV), les classes d’événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i- et ni-scènes ainsi que les différents niveaux d’abstraction.

Les mêmes observations sont faites concernant $L(E)$ (cf. 29e). Pour $L(T)$ (cf. 29f), 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 $D(E)$ ($r = -0.34$, $p < 0.01$). 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 de sources sonores semble donc avoir un faible impact sur l’agrément perçu, si l’on considère les ni-scènes, mais, comme pour les niveaux, cette densité ne semble pas avoir d’impact pour les 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’environnements, mais ne permet pas de caractériser précisément l’agrément perçu.

	ensemble	i-scènes	ni-scènes
L	-0.77 ($p < 0.01$)	-0.32 ($p = 0.06$)	-0.78 ($p < 0.01$)
L(E)	-0.75 ($p < 0.01$)	-0.20 ($p = 0.24$)	-0.75 ($p < 0.01$)
L(T)	-0.53 ($p < 0.01$)	-0.33 ($p = 0.05$)	-0.00 ($p = 0.99$)
D	-0.43 ($p < 0.01$)	-0.31 ($p = 0.07$)	-0.38 ($p < 0.05$)
D(E)	-0.34 ($p < 0.01$)	-0.22 ($p = 0.21$)	-0.46 ($p < 0.01$)
DIV(E) 0	-0.07 ($p = 0.52$)	-0.25 ($p = 0.15$)	-0.23 ($p = 0.23$)
DIV(E) 1	-0.47 ($p < 0.01$)	-0.25 ($p = 0.14$)	-0.26 ($p = 0.13$)
DIV(E) 2	-0.41 ($p < 0.01$)	-0.21 ($p = 0.22$)	-0.25 ($p = 0.14$)
DIV(E) 3	-0.37 ($p < 0.01$)	-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.

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, 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 permettent 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, 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.

5.2.11 Étude comparative entre les descripteurs sémantiques

5.2.11.1 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'environnements. Les résultats sont présentés à la figure 32a pour les événements, et à la figure 32b pour les textures. Par souci d'espace, nous choisissons un niveau d'abstraction intermédiaire entre le niveau 0 et 1, noté 0+, pour représenter les classes (cf. Figure 27).

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, a 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 représentation 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 32a montre, en effet, que les classes d'événements de *transports publics* (*bus* et *train*, cf. Figure 32c) 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 en se basant sur leurs mémoires. Mais, contrairement à nous, ils ne disposent pas 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).

5.2.11.2 Marqueurs sonores

Pourcentage de scènes simulées comportant une classe de son particulière Nous avons mis en évidence que, qualitativement, la composition des sources sonores des scènes diffère selon les types d'environnements (i ou ni). Nous essayons de voir maintenant si, parmi

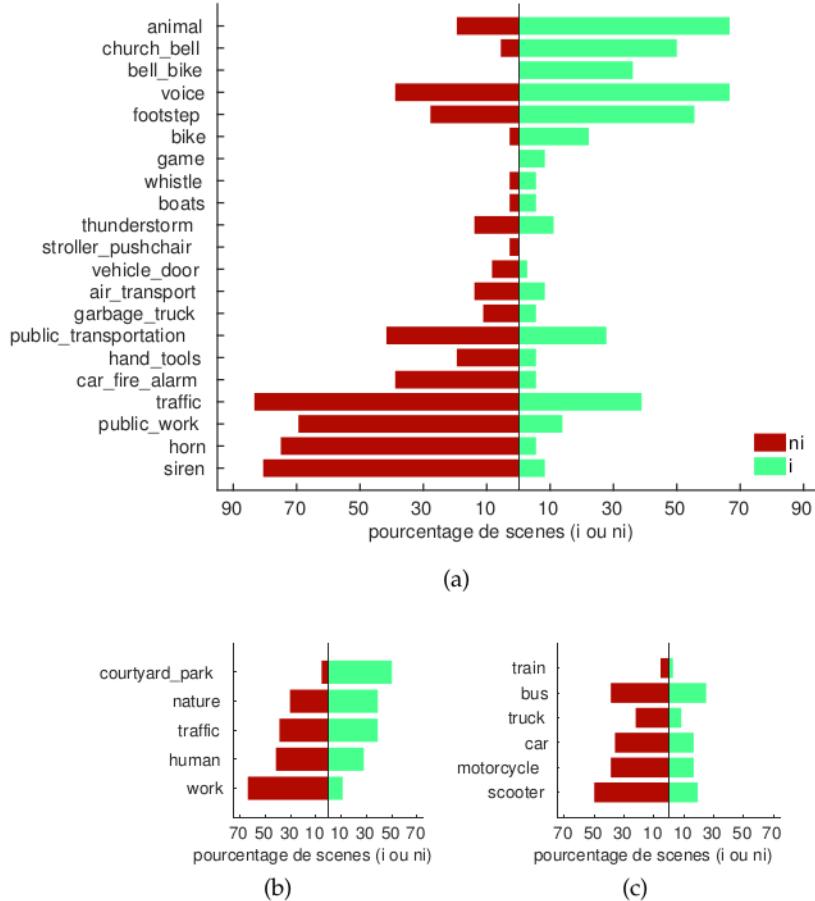


FIGURE 32 : Pourcentage de scènes simulées (i ou ni) comportant une classe de son particulière : (a) classes d'événements du niveau d'abstraction 0+, (b) classes de textures du niveau d'abstraction 0, (c) sous classes d'événements du niveau d'abstraction 1 des classes *trafic* et *transport public* du niveau d'abstraction 0.

ces classes, certaines sont typiques d'un environnement en particulier. Pour ce faire, nous utilisons le V-test (cf. Section 5.2.6.2), 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 32 le laissait présager, les classes relatives à l'activité 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 possiblement 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*).

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*).

GL : 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 texture *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 particulièrement associés à un environnement désagréable.

GL : TODO : ici analyse des caractéristiques des classes trafics

GL : TODO : ici reprendre les conclusions de (Lavandier and Defréville, 2006) et (Ricciardi et al., 2015)

5.2.12 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'environnements. 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 5.2.6.2). 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 o. Les résultats sont affichés sur la figure 33.

En ce qui concerne ET, la p@5 est de 76% pour le niveau d'abstraction o, 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'environnements 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,

Niveau d'abstraction	Marqueurs sonores événements	
	i-scènes	ni-scènes
0	construction work (3.78)	
1	church bell (4.5)	horn (3.9)
	bell bike (4.3)	siren (3.9)
	animal (4.2)	
2	birds (4.8)	horn (4.0)
	church bell (4.4)	siren (4.0)
	bell bike (4.2)	
3	birds singing (4.8)	horn (4.1)
	church bell (4.3)	siren (4.0)
	bell bike (4.2)	
	male footsteps	
	concrete (3.6)	

	Marqueurs sonores textures	
	i-scènes	ni-scènes
0	construction work (3.9)	
1	courtyard/park (4.1)	
	park (3.65)	crossroads (3.6)
		vehicle work (3.3)
2	park (3.64)	crossroads (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.

plus nous sommes à même d'établir une distinction claire entre les i- et ni-scènes.

En considérant séparément E et T, il apparaît que 1) la p@5 obtenue avec E est similaire à celle obtenue avec ET, 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 (Maf-fioli, 1999), qui montrent que nous analysons de manière descriptive (en identifiant les sources) les scènes événementielles, *i. e.* composées d'événements sonores (cf. Section 3.4.5.2).

Enfin, il apparaît que la p@5 obtenue avec E_m est similaire, voire supérieure à celles de 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 descrip-

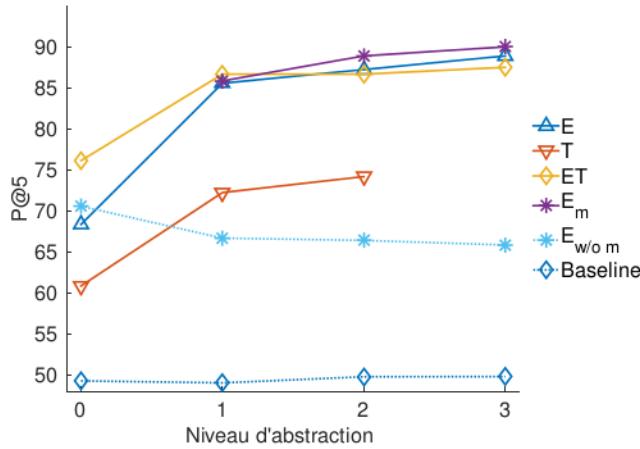


FIGURE 33 : P@5 obtenues en considérant la matrice de dissimilarité résultant des distances par paires de Hamming calculées entre les vecteurs des descripteurs sémantiques des scènes. Les vecteurs sont construits en utilisant toutes les classes (ET), les classes d'événements (E), les classes de textures (T), les classes d'événements ne considérant que les marqueurs sonores (E_m), les classes d'événements ne considérant pas les marqueurs sonores $E_{w/o,m}$.

tion ($E_{w/o,m}$), 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'environnements (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, sont nécessaires afin de 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 structurelle des scènes, basée uniquement sur ces marqueurs sonores, permet de caractériser l'agrément perçu, mieux qu'une description structurelle globale.

	i-scenes	ni-scenes
L_m	0.03 ($p = 0.88$)	-0.75 ($p < 0.01$)
$L(E)_m$	0.08 ($p = 0.66$)	-0.71 ($p < 0.01$)
$L(T)_m$	-0.11 ($p = 0.66$)	-0.17 ($p = 0.37$)
L_b	-0.52 ($p < 0.01$)	-0.32 ($p = 0.06$)
$L(E)_b$	-0.51 ($p < 0.01$)	-0.30 ($p = 0.07$)
$L(T)_b$	-0.32 ($p = 0.05$)	-0.73 ($p < 0.01$)
$L_m - L_b$	0.67 ($p < 0.01$)	-0.31 ($p = 0.07$)
$L(E)_m - L(E)_b$	0.66 ($p < 0.01$)	-0.28 ($p = 0.10$)
$L(T)_m - L(T)_b$	0.16 ($p = 0.54$)	0.21 ($p = 0.28$)
D_m	0.03 ($p = 0.87$)	-0.31 ($p = 0.07$)
$D(E)_m$	0.14 ($p = 0.41$)	-0.44 ($p < 0.01$)

TABLE 8 : Coefficients de corrélation linéaire calculés entre l'agrément perçu moyen A_{scene} de l'expérience 1.b et les descripteurs structurels relatifs à la présence des marqueurs sonores.

5.2.13 *L'influence spécifique des marqueurs sonores sur l'agrément perçu*

Comme pour la section 5.2.10, nous évaluons les corrélations entre A_{scene} 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 34). 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 les contributions séparées de différentes sources.

GL : TODO : Rajouter D_b + comparaison par pair des descripteurs de marqueurs

Concernant les niveaux sonores (cf. Figures 35), 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 36), 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 $L(T)_b$ ($r = -0.73$, $p < 0.01$). GL : 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 37). 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,

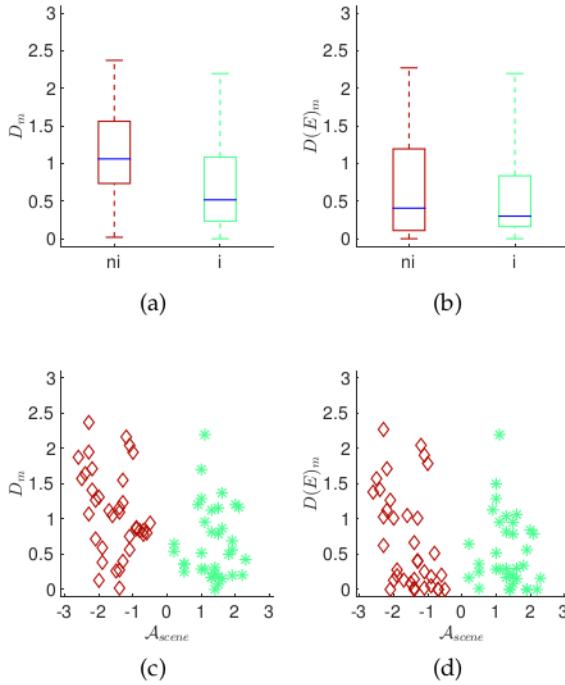


FIGURE 34 : Dispersion des descripteurs structurels de densité relatif à la présence des marqueurs D_m (a, c) et $D(E)_m$ (b, d), en fonction du type de scènes (a, b) et de l'agrément perçu A_{scene} de l'expérience 1.b (c, d).

- 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.

5.2.14 Discussions

Dans cette expérience, nous identifions 6 indicateurs structurels globaux permettant de distinguer, de manière globale, 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.

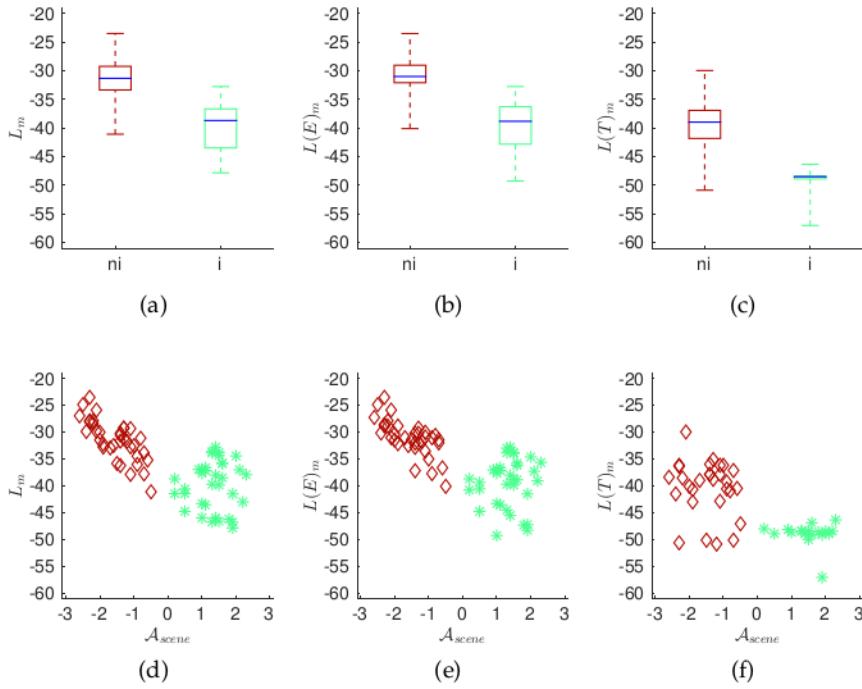


FIGURE 35 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs L_m (a, d), $L(E)_m$ (b, e) et $L(T)_m$ (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu A_{scene} de l'expérience 1.b (d, e, f).

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 ni-scènes :

- GL : TODO.

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 les ni-scènes.

De cette analyse, nous retenons les points suivants :

- *distinguer les i- et ni-scènes* : Les descripteurs sémantiques, ainsi que certains descripteurs structurels globaux, permettent de faire

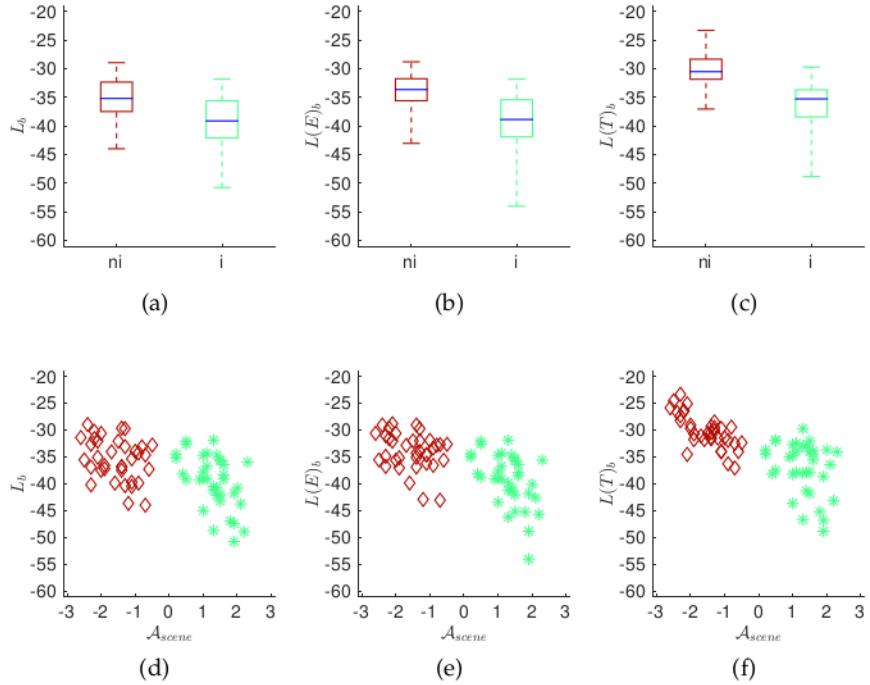


FIGURE 36 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs L_b (a, d), $L(E)_b$ (b, e) et $L(T)_b$ (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu A_{scene} de l'expérience 1.b (d, e, f).

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, c'est majoritairement les événements qui permettent de distinguer les deux types d'environnements, 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 $D(E)$), et/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 caractéristiques des marqueurs sonores, et celles de l'ensemble des autres sons. Ainsi, le niveau des marqueurs relatifs au bruit est posi-

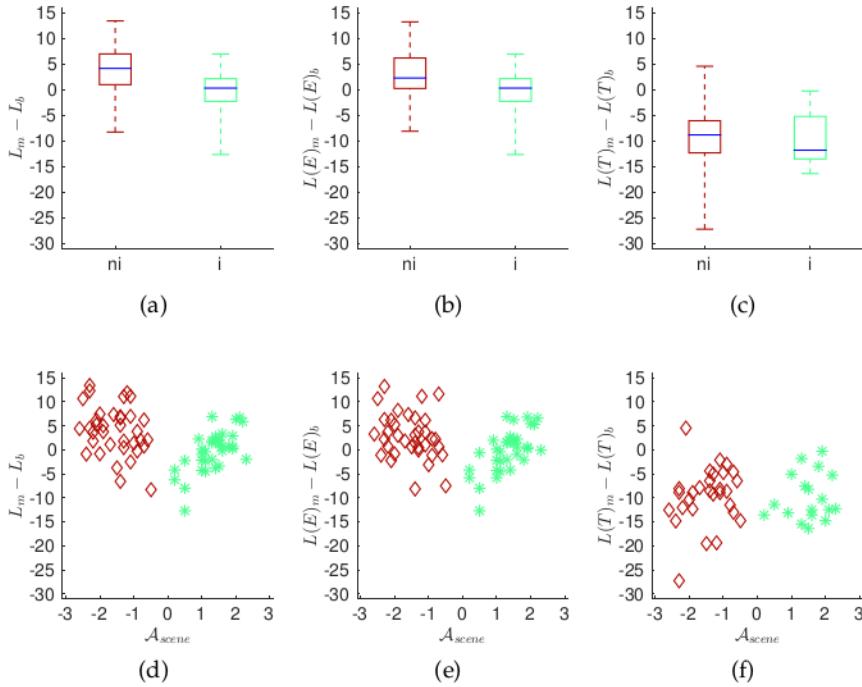


FIGURE 37 : Dispersion des descripteurs structurels de niveaux sonores relatifs à la présence des marqueurs $L_m - L_b$ (a, d), $L(E)_m - L(E)_b$ (b, e) et $L(T)_m - L(T)_b$ (c, f), en fonction du type de scènes (a, b, c) et de l'agrément perçu A_{scene} de l'expérience 1.b (d, e, f).

tivement corrélé à l'agrément, alors que le niveau du bruit est, lui, négativement corrélé.

L'existence de deux modes de perception, mobilisant différents types de descripteurs, et dépendant de la nature du stimuli, est un phénomène qui a déjà été observé pour la perception des textures (cf. Section 3.5). 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 stimuli (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 :

- dans le cadre de scènes non-idéales, 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. Mais nous observons également que le type de descripteurs à utiliser, pour une même source, varie en fonction de la nature de l'environnement **GL** : **TODO** : développer sur le contexte environnemental pour l'agrément, reprendre l'exemple de *foule*.

GL : **TODO** : Reprendre la conclusion de l'article et Proposer un modèle perceptif sur la base du modèle prédictif **GL** : **TODO**, l'utilisation de trafic reprendre les conclusions de (Lavandier and Defréville, 2006) **GL** : **TODO**, les résultats (2 modes d'obs) concordent avec les observations faites pas (Ricciardi et al., 2015).

5.3 AGIR SUR L'AGRÉMENT PERÇU EN MODIFIANT LA COMPOSITION SÉMANTIQUE

5.3.1 Objectif

L'expérience précédente à montré que, parmi les classes de sons peuplant le monde sonore, certaines, les marqueurs, sont caractéristiques de certains types d'environnements. 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, seules les i-marqueurs sont omis, de même, pour les ni-scènes, seuls les ni-marqueurs sont retirés. 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.

5.3.2 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-scenes, les 36 scènes idéales comprenant les marqueurs, et ni/am-scènes les 36 scènes non-idéales comprenant les marqueurs ;
- 72 *sm-scènes* : les 72 scènes précédemment simulées, régénérées sans les classes de marqueurs (sm). Nous notons i/sm-scenes, les 36 scènes idéales générées sans les marqueurs, et ni/sm-scènes les 36 scènes non-idéales générées sans les 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-scenes ;
- *sirène, klaxon* pour les ni/sm-scenes.

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.

Apparatus

Tous les sujets passent l'expérience sur des machines identiques ([GL : description des machines](#)). L'audio est diffusé en monophonie, par le biais de casques audio semi-ouvert *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.

5.3.3 Données et méthodes d'analyses

5.3.3.1 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 [5.2.6.1](#) pour plus de détails.

5.3.3.2 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 5.2.6.2) ;
- *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 ;
- *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, $A_{\text{ scene }}$, 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 mesure répétée est utilisée, 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. 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 $A_{\text{ scene }}$. Pour mesurer la corrélation, nous utilisons le coefficient de Pearson (cf. Annexe A).

Tous les tests de significativité sont effectués avec un seuil critique $\alpha = 0.05$.

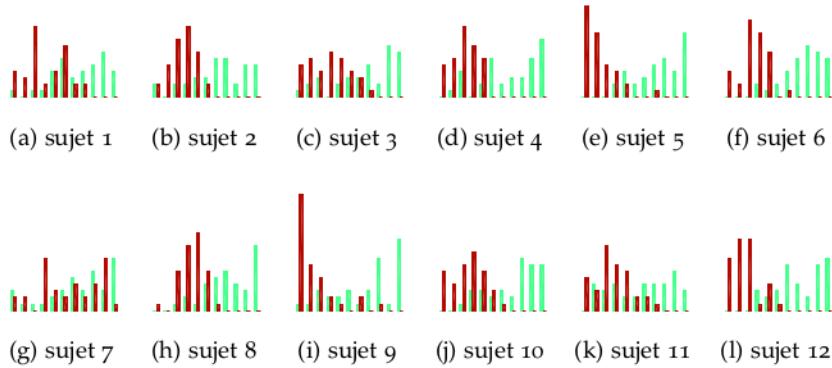


FIGURE 38 : Dispersion des notes données par les sujets lors de l'expérience 2 aux i/am-scènes (vert) et ni/am-scènes (rouge).

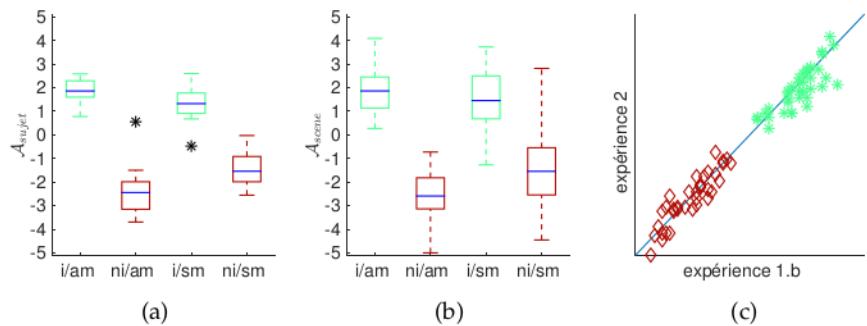


FIGURE 39 : Dispersion des notes données par les sujets lors de l'expérience 2 moyennées suivant les sujets (A_{subject} : a), suivant les scènes (A_{scene} : b et c), en fonction du type de scènes (a et b) et des A_{scene} relevés à l'expérience 1.b.

5.3.4 Détection de valeurs extrêmes

Considérons A_{subject} pour les am-scènes (cf. Figure 39a). Il apparaît que les réponses du sujet 7 diffèrent des autres. Comme observé sur la figure 38 (cf. Figure 38g 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.

	i/am-scènes	ni/am-scènes
L	-0.46* ($p < 0.01$)	-0.83 ($p < 0.01$)
L(E)	-0.33 ($p = 0.05$)	-0.84 ($p < 0.01$)
L(T)	-0.42* ($p < 0.05$)	0.04 ($p = 0.81$)
D	-0.42* ($p < 0.05$)	-0.47 ($p < 0.01$)
D(E)	-0.36* ($p < 0.05$)	-0.57 ($p < 0.01$)
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 _m	0.16 ($p = 0.36$)	-0.75 ($p < 0.01$)
L(E) _m	0.08 ($p = 0.64$)	-0.73 ($p < 0.01$)
L(T) _m	-0.05 ($p = 0.86$)	-0.06 ($p = 0.76$)
L _b	-0.64 ($p < 0.01$)	-0.40* ($p < 0.05$)
L(E) _b	-0.57 ($p < 0.01$)	-0.33 ($p = 0.05$)
L(T) _b	-0.46* ($p < 0.01$)	-0.83 ($p < 0.01$)
L _m – L _b	0.60 ($p < 0.01$)	-0.25 ($p = 0.14$)
L(E) _m – L(E) _b	0.56 ($p < 0.01$)	-0.27 ($p = 0.11$)
L(T) _m – L(T) _b	0.43 ($p = 0.07$)	0.36 ($p = 0.05$)
D _m	-0.17 ($p = 0.34$)	-0.33 ($p = 0.05$)
D(E) _m	-0.25 ($p = 0.15$)	-0.53 ($p < 0.01$)

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 et relatifs à la présence des marqueurs sonores pour les i/am-scènes et ni/am-scenes.

5.3.5 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.

Nous commençons par évaluer la corrélation entre les $\mathcal{A}_{\text{scene}}$ obtenues par les deux études. Les résultats sont affichés sur la figure 39c. La corrélation est élevée, que l'on considère l'ensemble des scènes ($r = 0.98$, $p < 0.01$), ou les i- ($r = 0.82$, $p < 0.01$), ou encore les ni-scènes ($r = 0.91$, $p < 0.01$).

Concernant les différences de $\mathcal{A}_{\text{ sujet}}$ 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 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 pour les niveaux 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_b pour les ni-scènes ($r = -0.40$, $p < 0.05$), ainsi que sur $L(T)_b$ ($r = -0.46$, $p < 0.01$) 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 $D(E)$ ($r = -0.42$, $p < 0.05$). 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 consistants. 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 ($L(T)_m : r = -0.06$, $p = 0.76$).

[GL : TODO : compléter](#)

	sm-scènes	i/sm-scènes	ni/sm-scènes
L	-0.79 ($p < 0.01$)	-0.49 ($p < 0.01$)	-0.74 ($p < 0.01$)
L(E)	-0.76 ($p < 0.01$)	-0.44 ($p < 0.01$)	-0.70 ($p < 0.01$)
L(T)	-0.41 ($p < 0.01$)	-0.17 ($p = 0.36$)	-0.44 ($p = 0.80$)
D	-0.49 ($p < 0.01$)	-0.31 ($p = 0.07$)	-0.29 ($p = 0.08$)
D(E)	-0.45 ($p < 0.01$)	-0.29 ($p = 0.09$)	-0.39 ($p < 0.05$)
DIV(E) o	-0.10 ($p = 0.40$)	-0.26 ($p = 0.13$)	-0.32 ($p = 0.06$)
DIV(E) 1	-0.49 ($p < 0.01$)	-0.29 ($p = 0.09$)	-0.31 ($p = 0.06$)
DIV(E) 2	-0.43 ($p < 0.01$)	-0.24 ($p = 0.17$)	-0.32 ($p = 0.06$)
DIV(E) 3	-0.39 ($p < 0.01$)	-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 structuraux pour les i/sm-scènes et ni/sm-scènes.

5.3.6 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 39a) montre un effet significatif du type d'environnements (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-scenes ($p < 0.05$) et les ni/am- et ni/sm-scenes ($p < 0.01$).

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'environnements influe sur l'effet dû à l'absence/présence des marqueurs. En effet la moyenne des écarts entre les am- et sm-scènes est plus importante pour les ni-scènes (1.1) que pour les i-scènes (0.5).

Les i-marqueurs ont donc bien un effet bénéfique sur la perception d'un environnement. Le fait que leur suppression diminue $\mathcal{A}_{scène}$ montre clairement qu'il est possible d'améliorer la qualité d'un environnement en ajoutant des sons bien acceptés comme *oiseaux*. Ces

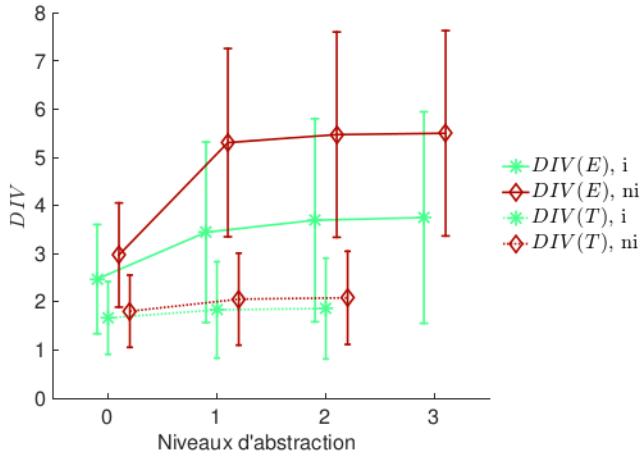


FIGURE 40 : Moyenne et écart type de la diversité des classes utilisées en considérant l'ensemble des classes (DIV), les classes d'événements (DIV(E)) et les classes de textures (DIV(T)), en considérant séparément les i/sm- et ni/sm-scènes ainsi que les différents niveaux d'abstraction.

conclusions vont dans le sens de l'approche positive comme introduite par Schafer (Schafer, 1977) (cf. Section 3.4.2).

GL : TODO : citation

5.3.7 Influence des descripteurs structurels des scènes sans marqueurs sur l'agrément perçu

L'ANOVA à mesures répétées pratiquée sur $\mathcal{A}_{\text{scene}}$ (cf. Figure 39b) montre un effet significatif du type d'environnements (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-scenes ($p < 0.05$) et les ni/am- et ni/sm-scenes ($p < 0.01$).

Ainsi, les quatre types de scènes, considérant $\mathcal{A}_{\text{scene}}$ 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 5.3.6).

GL : TODO : corrélation analyse tableau 10 + MANOVA ?

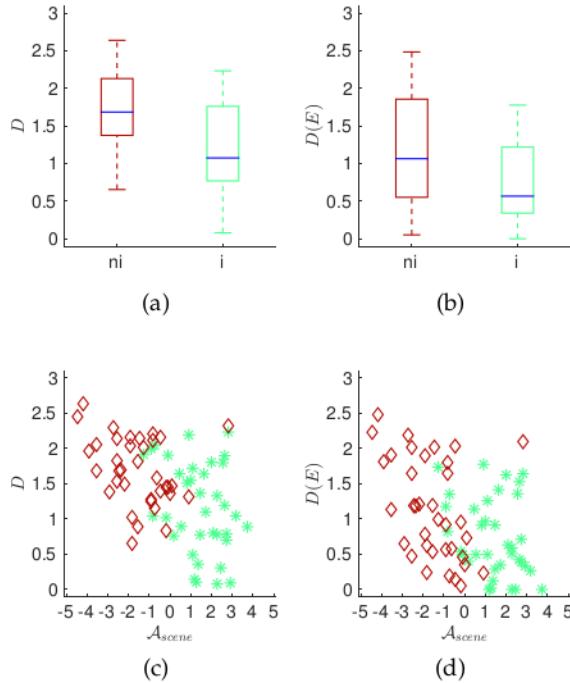


FIGURE 41 : Dispersion des descripteurs structurels de densité D (a, c) et $D(E)$ (b, d), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b) et de l’agrément perçu $\mathcal{A}_{\text{scene}}$ de l’expérience 2 (c, d).

5.3.8 Discussions

GL : TODO

5.4 INFLUENCE DE LA COMPOSITION SÉMANTIQUE SUR LES PROCESSES DE CATÉGORISATION DES SCÈNES

5.4.1 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 principal est entre autre de montrer que l’utilisation de scènes simulées permet d’aboutir à des résultats similaires que ceux obtenus avec des scènes enregistrées.

Nous considérons notamment les résultats obtenus par (Maffiolo, 1999) (cf. Section 3.4.5.2), qui montre que des scènes événementielles⁷ sont catégorisées suivant :

- les sources sonores présentes ;

⁷ Les scènes simulées de l’expérience 1.a peuvent être toutes considérées comme des scènes événementielles et non scènes amorphes

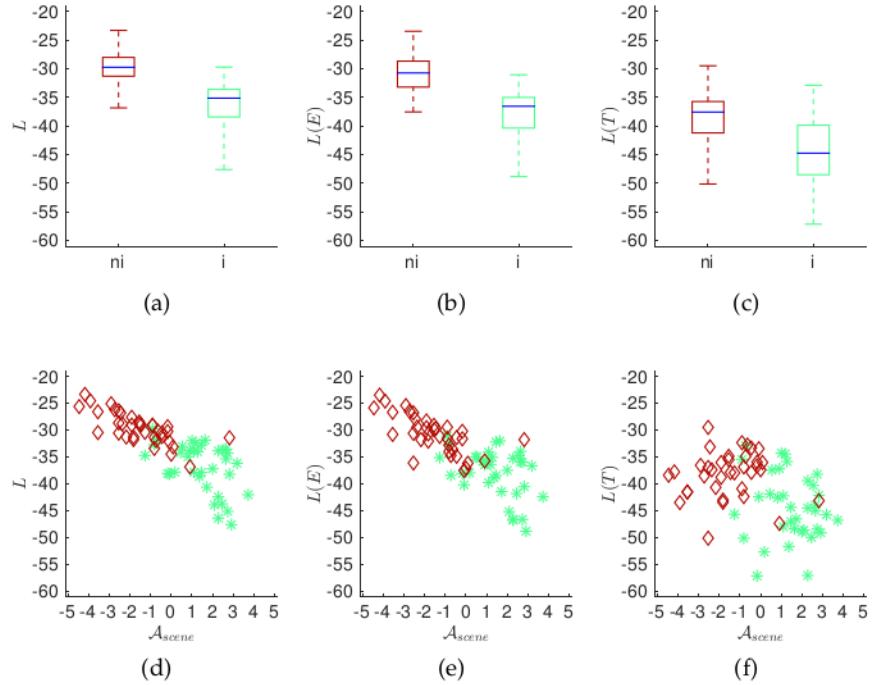


FIGURE 42 : Dispersion des descripteurs structurels de niveaux sonores L (a, d), L(E) (b, e) et L(T) (c, f), relevés sur les i/sm- et ni/sm-scènes, en fonction du type de scènes (a, b, c) et de l'agrément perçu A_{scene} de l'expérience 2 (d, e, f).

- la qualité de l'environnement.

Notant que la catégorisation est un processus dépendant du contexte sensoriel, *i.e.* de la nature des objets à catégoriser (cf. Section 3.2.5). 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 puisse être considéré par les sujets comme une stratégie de catégorisation.

5.4.2 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, il est demandé au sujet de décrire les groupes ainsi formés. Afin de faciliter l’analyse lexicale des descriptions, il est demandé aux sujets de limiter à une liste de mots isolés où à des phrases simples et courtes.

L’expérience est prévue pour durer entre 1h et 1h30.

Apparatus

L’Apparatus est identique à celui de l’expérience 1.b (cf. Section [5.2.5.2](#))

Participant

Les 10 sujets réalisant cette expérience sont les mêmes que ceux ayant participé à l’expérience 1.b (cf. Section [5.2.5.2](#)). 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 ardue et laborieuse, d’autant plus si ces stimuli sont des sons longs (30 secondes).

1 sujet est éliminé pour incompréhension des consignes.

5.4.3 Données et méthodes d’analyses

5.4.3.1 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 [5.2.6.1](#)) à savoir :

- *descripteur perceptif* ;

- *descripteur sémantique (objectif)* ;
- *descripteur structurel*.

A partir des descriptions verbales utilisées spontanément par les sujets pour décrire les groupes, nous isolons 2 nouveaux descripteurs, appelés descripteurs subjectifs, en opposition à ceux (objectifs), relatifs à la composition des scènes :

- *descripteur sémantique subjectif* : relatif aux termes faisant référence à des sources sonores ;
- *descripteur de qualité subjectif* : relatif aux termes faisant référence aux qualités affectives des scènes (*e.g.* calme, plaisant).

Chaque sujet peut avoir utilisé un vocabulaire différent pour décrire les groupes. Afin d'établir un vocable de labels génériques, une analyse lexicale est pratiquée dans un premier sur l'ensemble des labels données par les sujets afin de :

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 5.4.5).

La liste des labels liés à une scène forme le descripteur subjectif de cette dernière. Afin d'affecter un label à une scène, chaque scène hérite des labels génériques des groupes auxquels elle a été affectée, en considérant l'ensemble des sujets.

5.4.3.2 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.

Un matrice de dissimilarité globale $\Theta_{i,j}$ est alors obtenue en moyennant les matrices $\delta_{i,j}^n$:

$$\Theta_{i,j} = \frac{1}{N} \sum_{n=1}^N \delta_{i,j}^n \quad (5)$$

avec N le nombre total de sujets. Il est à noter que les dissimilarités contenues dans Θ respectent l'inégalité triangulaire ($\Theta_{a,c} \leq \Theta_{a,b} + \Theta_{b,c}$)

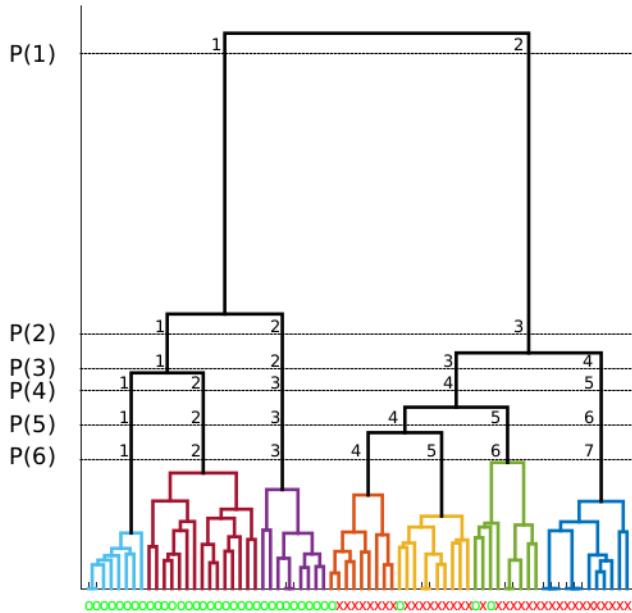


FIGURE 43 : Partitions P établies suivant la classification ascendante hiérarchique pratiquée sur la matrice de similarité Θ en utilisant un critère de Ward.

$\Theta_{b,c}$) et 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 la méthode d'agrégation de Ward est pratiquée sur Θ afin de faire émerger les tendances de groupement globales. Le dendrogramme résultant est affiché sur la Figure 43.

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 43).

Pour chaque partition, il s'agit d'observer les descripteurs rendant compte des groupements effectués.

La capacité des classes (respectivement labels) des 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 5.2.6.2). 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

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 nombres de groupements effectués.
L'indice * indique les sujets étant supprimés de l'analyse.

à un facteur (cf. Annexe A.2), le nombre de niveaux du facteur étant égale 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. Notons que pour la partition 1, composée de 2 groupes ($1_{P(1)}$ et $2_{P(1)}$), le test se ramène à un test de Student à deux populations(cf. Annexe A.1).

5.4.4 Stratégie de catégorisation

En fonction des descriptions des sujets, nous relevons 5 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 références 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 fréquentiel (1). Ces résultats concordent avec ceux de Maffiolo, 1999 (cf. Section 5.4.1).

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 ni utilisé les sources

labels des sources		labels des qualités	
originaux	génériques	originaux	génériques
alarm (2)	—	très calme	—
horn (2)	—	calme (4)	—
siren (3)	—	moyennement calme	—
park	—	bucolique	—
nature (3)	—	supportable	—
market	—	oppressant	—
birds (4)	—	imprédictible	—
bruit* (2)	—	agité	—
footstep (2)	—	fatiguant	—
crowd	—	énervant (2)	—
bruit de fond*	background*	déplaisant (2)	—
fond sonore*		très déplaisant	—
church bell (4)	church bell	insupportable	—
church (2)		apaisant	relaxing
mechanical		reposant	
public work (4)	construction	normal	usual
tools		habituel	
human (4)	human		
people			
traffic (2)	traffic		
car (2)			
water	water		
rain (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.

présentes, ni les qualités affectives. Afin de conserver des résultats cohérents, il est supprimé de l'analyse⁸.

5.4.5 Analyse lexicale des descriptions

L'intensité n'ayant été considérée que par un seul sujet (sujet 6), seul 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 5.4.3.1, 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 :

⁸ La matrice de similarité globale Θ ainsi que le dendrogramme résultant de la CAH présenté figure 43 ne tiennent pas compte du sujet 2

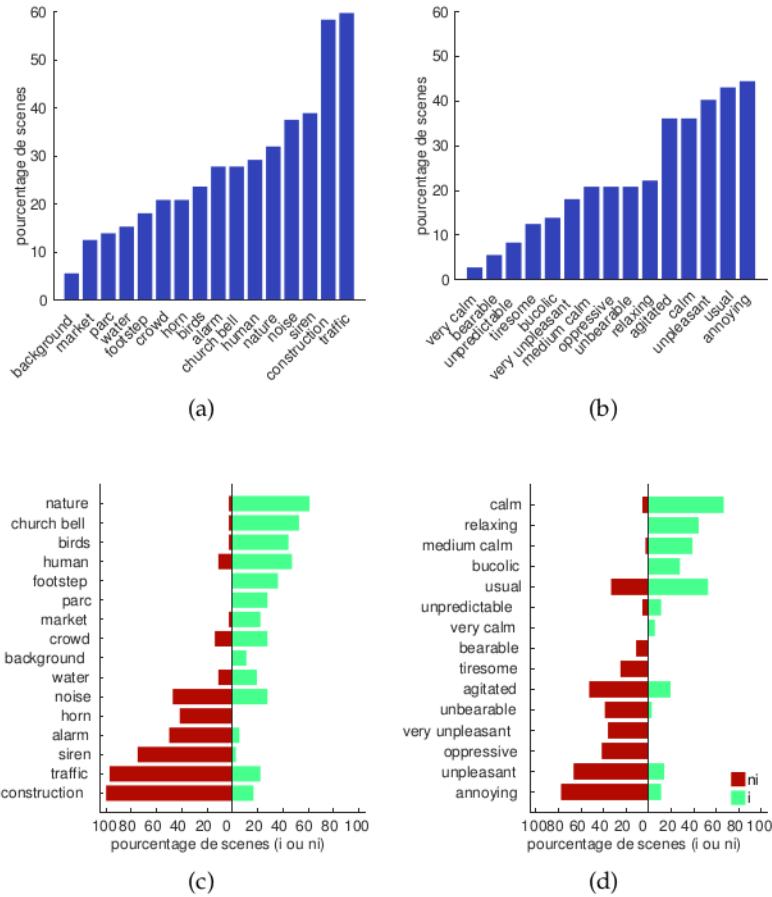


FIGURE 44 : Pourcentage de scènes étant décris par un label sémantique subjectif donné (a, c), un label de qualité subjectif donné (b, d), en considérant l'ensemble des scènes (a, b) ou les i- et ni-scènes séparément (c, d).

- descripteurs sémantiques subjectifs (S) : les sources ;
- descripteurs de qualité subjectifs (Q) : les qualités affectives perçues.

La figure 44 affiche le nombre de scènes étant décris par chaque label, en considérant respectivement S (cf. Figure 44a et 44c) et Q (cf. Figure 44b et 44d).

Considérons en premier lieu les labels des sources sonores (cf. Figure 44a). Les sujets décrivent les sources en considérant plusieurs niveaux d'abstractions, allant de plus concret ("footstep", "bird") au plus abstrait ("human", "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 : "traffic" et "construction".

Les labels de sources font clairement la distinction entre les i- et ni-scènes (cf. Figure 44d). Nous notons cependant quelques diffé-

rences entre ces labels et la répartition des classes de sons utilisées pour simuler les scènes (cf. Figure 32). En effet les labels “human” et “footstep” sont majoritairement utilisés pour décrire les i-scènes, alors les classes *footstep* et *voice* 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 (GL : TODO : chiffre).

Il est remarquable de constater que 44% des labels utilisés (50% si l'on fait l'association entre le label “traffic” et la classe *crossroad*) font directement référence aux marqueurs sonores établis dans l'expérience 1 (cf. Section 5.2.11.2). 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 44b). 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 3.4.4), à savoir :

- l'agrément (unpleasant, very unpleasant) ;
- le calme / la tranquillité (medium calm, calm, very calm, relaxing) ;
- la gêne (bearable, annoying, tiresome, oppressive, unbearable).

Deux derniers groupes de termes semblent faire référence à la nature prédictible de l'environnement (“usual”, “unpredictable”), ainsi qu'à la structure temporelle de ce dernier (“agitated”).

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 44d). Cette distinction est moins marquée en ce qui concerne les labels “usual” (5.5% de ni-scènes vs. 11% de i-scènes) et “unpredictable” (33% de ni-scènes vs. 53% de i-scènes).

5.4.6 Partitions et descripteurs sémantiques subjectifs

Les résultats de la CAH pratiquée sur Θ sont affichés à la figure 43. Pour chaque partition ($P(1), P(2) \dots P(6)$) un V-test est pratiqué en considérant séparément les deux descripteurs sémantiques subjectifs (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, naturels et

$1_{P(1)}$	nature church bell birds footstep parc market human	$2_{P(1)}$	construction siren traffic alarm horn
$1_{P(2)}$	church bell footstep market human crowd	$2_{P(2)}$	$3_{P(2)}$ construction siren traffic alarm horn
$1_{P(3)}$	church bell footstep market human crowd	$2_{P(3)}$	$3_{P(3)}$ construction traffic
$1_{P(4)}$	church bell footstep market nature human	$2_{P(4)}$	$3_{P(4)}$ construction traffic siren
$1_{P(5)}$	church bell footstep market nature human	$2_{P(5)}$	$3_{P(5)}$ construction traffic
$1_{P(6)}$	church bell footstep market nature human	$2_{P(6)}$	$3_{P(6)}$ $4_{P(6)}$ $5_{P(6)}$ $6_{P(6)}$ $7_{P(6)}$

TABLE 13 : Répartitions des labels relatifs aux sources sonores relevées par les sujets en fonction des partitions établies par la classification ascendante hiérarchique.

de cloches ($1_{P(1)}$) et de l'autre les sons de construction, de trafics et les alarmes ($2_{P(1)}$). Dans $P(2)$, le groupe des i-scènes $1_{P(1)}$ est subdivisé suivant d'un coté les sons humains et 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 suivant d'un coté les sons de construction 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 en distinguant 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 "construction" et "traffic" est subdivisé. Cependant, aucune de ces subdivisions ne semblent 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 "horn".

$1_{P(1)}$	calm relaxing medium calm bucolic	$2_{P(1)}$	annoying unpleasant oppressive unbearable very unpleasant
$1_{P(2)}$	calm medium calm	$2_{P(2)}$	$3_{P(2)}$ annoying unpleasant oppressive unbearable very unpleasant
$1_{P(3)}$	calm medium calm	$2_{P(3)}$	$3_{P(3)}$ unpleasant agitated
$1_{P(4)}$	medium calm calm	$2_{P(4)}$	$3_{P(4)}$ bucolic relaxing
$1_{P(5)}$	medium calm calm	$2_{P(5)}$	$3_{P(5)}$ bucolic relaxing
$1_{P(6)}$	medium calm calm	$2_{P(6)}$	$3_{P(6)}$ bucolic relaxing
			$4_{P(4)}$ unpleasant agitated
			$5_{P(4)}$ oppressive
			$6_{P(5)}$ very unpleasant unbearable tiresome oppressive
			$7_{P(6)}$ very unpleasant unbearable tiresome oppressive

TABLE 14 : Répartitions des labels relatifs aux qualités affectives perçues en fonction des partitions établies par la classification ascendante hiérarchique.

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 côté les qualités positives, dominées par les labels "calm" et "relaxing" ($1_{P(1)}$), et de l'autre les qualités négatives, dominées par les labels "annoying" et "unpleasant" ($2_{P(1)}$). 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 groupes, labels décrivant des sons d'origines naturels. La distinction opérée à la partition P(3) est principalement fonction de l'agrément : les scènes de $3_{P(3)}$ étant caractérisées par "unpleasant" et celles de $3_{P(4)}$ par "very unpleasant". $3_{P(4)}$ conserve par ailleurs les qualités "unbearable", "tiresome" et "oppressant". Enfin, c'est à ce niveau que l'impression de la structure temporelle des scènes fait son apparition, "agitated" étant caractéristique de $3_{P(3)}$. Au niveau P(4), la qualité "calm" 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 le groupe $4_{P(5)}$ et $5_{P(5)}$. Les qualités de $4_{P(4)}$ ("unpleasant" et "agitated") 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 l'aspect régénérative des scènes et enfin suivant la notion de calme ;
- les qualités négatives se séparent d'abord en fonction de l'agrement entre d'un coté les scènes "déplaisantes" et celles "très déplaisantes". Une distinction est enfin opérée sur les scènes "déplaisantes" suivant qu'elles soient "supportables", "usuelles" ou "oppressantes". Notons enfin que la notion d'agitation co-occure avec celle de "déplaisant" plutôt que celle de "très déplaisant", alors que les qualités "insupportable" 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 [GL : TODO : citation de résultats similaires](#). 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 :

- "calm" ($1_{P(2)}$) vs. "relaxing" ($2_{P(2)}$) : la qualité "relaxing" semble liée aux sons d'origine naturels et la qualité "calm" aux sons d'origine humaine et de cloche. Notons que la qualité "relaxing" apparaît également dans $2_{P(4)}$, simultanément à l'apparition du label "nature" ;
- "medium calm" / "calm" ($1_{P(4)}$) vs. "calm" / "very calm" ($2_{P(4)}$) : la qualité "medium calm" / "calm" semble liée aux sons de cloches, et la qualité "calm" / "very calm" aux sons d'origine humaine ;
- "unpleasant" ($3_{P(3)}$) vs. "very unpleasant" ($4_{P(3)}$) : les sources "construction" et "traffic" semblent liées à "unpleasant", alors que les sources "alarm" et "siren" semblent correspondre à "unpleasant". Notons cependant que les sons "horn" et "alarm" ($4_{P(6)}$) sont également liés aux qualités "unpleasant" et "bearable".

5.4.7 Partitions et descripteurs sémantiques objectifs

Nous analysons les descripteurs sémantiques objectifs, *i. e.* les classes utilisées pour simuler les scènes. Le tableau 15 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 relevées à l'expérience 1 (cf. Section 5.2.11.2), exception faite de la classe marquer d'événement *male footsteps concrete* qui n'apparaît dans aucun groupe. Notons par ailleurs la présence de la classe d'événement *bike* dans le groupe $1_{P(1)}$, classe n'étant pas un marqueur.

Considérons en premier les groupes relatifs aux i-scènes. Considérant les classes d'événements, la partitionnement entre $1_{P(2)}$ et $2_{P(2)}$ sépare d'un coté les classes de sons de cloches et d'origine humaine ($1_{P(2)}$), et de l'autre les classes de sons d'origine naturelle, de *boats* et de *bell bike* ($2_{P(2)}$). Notons l'apparition du marqueur *male footsteps concrete* dans $1_{P(2)}$. Au niveau des textures, les classes *park* (sons d'origine naturelle) et *courtyard* 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'événements d'origine humaine. Les classes textures *courtyard park* et *crowd foreigners* sont caractéristiques de $2_{P(4)}$.

Considérons maintenant les groupes relatifs aux ni-scènes. Seule la classe de texture *works* semble déterminer les partitionnements $P(3)$ et $P(4)$, les autres classes étant équitablement réparties entre les deux groupes ($3_{P(3)}/4_{P(3)}$ ou $4_{P(4)}/5_{P(4)}$). 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 (*power tools* et *power drill*) pour les événements et *cafe* pour les textures.

Comparant 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 *boats* et *bell bike* ($2_{P(2,3)}$ et $3_{P(4,5,6)}$) qui n'apparaissent pas dans les labels ;
- le label “market” ($1_{P(1,2,3)}$ 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 texture *courtyard park* et *park*, la classe correspondant au label “park” ($2_{P(2,3)}, 3_{P(4,5,6)}$) est moins évidente. On peut penser que ce label est le résultat de l'interprétation de la présence de la classe *birds*.

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 "alarm" et "siren" ($4_{P(3)}$, $5_{P(4)}$, $6_{P(5)}$ et $7_{P(6)}$), aucune classe n'est présente. A l'inverse, aucun label n'est présent pour les classes "power tools", "power drill" et "cafe" ($5_{P(5)}$ et $6_{P(6)}$). Cependant, il est possible maintenant de faire le lien entre ces classes et la qualité "oppressive" ($5_{P(5)}$ et $6_{P(6)}$).

Notons enfin le groupe $5_{P(6)}$ où aucune classe et aucun label n'est relevé, le groupe étant seulement caractérisé par les qualités "unpleasant" et "usual".

Le fait que les labels et les classes concordent jusqu'à la partition $P(2)$ nous permet de conclure que les premiers groupement 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 naturel \Rightarrow environnement relaxant et bucolique). Pour les groupes relatifs aux ni-scènes, les résultats obtenus soulèvent trois questions :

1. présence des labels et absences 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 structurels de ces classes (intensité sonore, densité d'événements), sont saillantes pour les scènes du groupe décrit.
2. présence des classes et absences 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 présente sur ces dernières ? Ce cas n'apparaît que pour les classes "power tools", "power drill" et "cafe". Pour les deux premières on peut supposer que le niveau d'abstraction des classes est trop précis, les sujets interprétant "power tools" et "power drill" comme appartenant à la classe construction, sans faire de distinction particulière. Cette classe n'étant plus caractéristique des groupes relatifs aux ni-scènes à partir de la partition $P(3)$, il est normal que les sujets ne s'appuie plus sur la présence des classes "power tools", "power drill" pour décrire des groupes.
3. présence des qualités et absences des labels et des classes ($5_{P(6)}$) : A 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, "unpleasant", 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 particulariser $5_{P(6)}$. La deuxième, "usual" 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).

5.4.8 Descripteurs perceptifs et structurels

5.4.8.1 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 16. 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.

Pour les partitions P(3), P(4), P(5) et P(6) les ANOVA montrent un effet significatif des groupes sur $A_{scène}$. 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 "unpleasant" ($3_{P(3)}$ et $4_{P(4)}$) et "very unpleasant" ($4_{P(3)}$ et $5_{P(4)}$) se retrouvent si l'on considère l'agrément perçu.

5.4.8.2 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 16. 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 significatif 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 : on observe des différences significatives uniquement en comparant des groupes relatifs aux i-scènes à des groupes relatifs aux ni-scènes. Les effets significatifs rendent ainsi uniquement compte des différences entre les i- et ni-scènes ;
- L(E) : deux différences significatives sont relevées entre des 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 manifester 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 étant ceux possédant les labels "unpleasant", "agitated" et "bearable" ($3_{P(3)}$ et $4_{P(5)}$), tandis que ceux affichant des niveaux élevés ($4_{P(3)}$ et $6_{P(5)}$) sont qualifiés par les termes "very unpleasant", "unbearable", "tiresome" et "oppressant".

Concernant le groupe $5_{P(6)}$, qui n'est pas caractérisé par une classe de son ou un label de source, il apparaît également qu'aucun descripteur structurel global ne semble rendre compte de ce groupement.

5.4.8.3 Contributions des sources sonores

[GL : TODO](#)

5.4.9 Discussions

L'expérience permet de valider l'utilisation de scènes simulées comme stimuli. Plusieurs résultats notables présent dans la littérature sont retrouvés, nommément :

- [GL : TODO](#)

[GL : HCA sur S, et F-mesure des groupes sur ceux établis par la catégorisation](#)

$1_{P(1)}$		$2_{P(1)}$		
bell bike ^E _{1,2,3} church bell ^E _{1,2,3} bike ^E _{1,2,3} child voice ^E animal ^E bird ^E ₂ singing birds ^E ₃	courtyard park ^T ₀ park ^T _{1,2}	works ^E ₀ siren ^E _{1,2,3} horn ^E _{1,2,3}	works ^T ₀ crossroads ^T _{1,2}	
$1_{P(2)}$	$2_{P(2)}$	$3_{P(2)}$		
church bell ^E _{1,2} male laugh ^E _{2,3} footstep ^E ₁ concrete male footstep ^E ₃	boats ^E _{1,2,3} bell bike ^E _{1,2,3} animal ^E ₁ bird ^E ₂ singing birds ^E ₃	works ^E ₀ siren ^E _{1,2,3} horn ^E _{1,2,3}	works ^T ₀ crossroads ^T _{1,2} crossroads ^T _{1,2}	
$1_{P(3)}$	$2_{P(3)}$	$3_{P(3)}$	$4_{P(3)}$	
church bell ^E _{1,2,3} male laugh ^E _{2,3} footstep ^E ₁ concrete male footstep ^E ₃	bell bike ^E _{1,2,3} boats ^E _{1,2,3} animal ^E ₁ bird ^E ₂ singing birds ^E ₃		works ^T ₀	
$1_{P(4)}$	$2_{P(4)}$	$3_{P(4)}$	$4_{P(4)}$	$5_{P(4)}$
church bell ^E _{1,2,3}	footstep ^E ₁ female footstep ^E ₂ male laugh ^E ₂	bell bike ^E _{1,2,3} boats ^E _{1,2,3} animal ^E ₁ bird ^E ₂ singing birds ^E ₃		
	courtyard park ^T ₀ crowd foreigners ^T ₂		works ^T ₀	
$1_{P(5)}$	$2_{P(5)}$	$3_{P(5)}$	$4_{P(5)}$	$5_{P(5)}$
church bell ^E _{1,2,3}	footstep ^E ₁ female footstep ^E ₂ male laugh ^E ₂	bell bike ^E _{1,2,3} boats ^E _{1,2,3} animal ^E ₁ bird ^E ₂ singing birds ^E ₃		$6_{P(5)}$
	courtyard park ^T ₀ crowd foreigners ^T ₂		cafe ^T ₁	
$1_{P(6)}$	$2_{P(6)}$	$3_{P(6)}$	$4_{P(6)}$	$5_{P(6)}$
church bell ^E _{1,2,3}	footstep ^E ₁ female footstep ^E ₂ male laugh ^E ₂	bell bike ^E _{1,2,3} boats ^E _{1,2,3} animal ^E ₁ bird ^E ₂ singing birds ^E ₃		$6_{P(6)}$
	courtyard park ^T ₀ crowd foreigners ^T ₂		cafe ^T ₁	$7_{P(6)}$

TABLE 15 : Répartitions des classes de sons d'événements (^E) et de textures (^T) en fonction des partitions établies par la classification ascendante hiérarchique. Les indices _x indique le niveau d'abstraction de la classe considérée.

	P(3)	P(4)	P(5)	P(6)
\mathcal{A}_{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 16 : GL : TODO

Troisième partie

ANALYSE AUTOMATIQUE DE SCÈNES SONORES

preamble text here.

6

L'ANALYSE AUTOMATIQUE DE SCÈNES SONORES ENVIRONNEMENTALES, UN ÉTAT DE L'ART

GL : annonce de plan

6.1 INTRODUCTION

6.1.1 *Historique, communauté et application*

GL : TODO : bioacoustique GL : TODO : introduire AED, ASC et ASSR

6.1.2 *Campagnes d'évaluation : le challenge DCASE*

GL : TODO : IEEE AASP, présentation des tâches, des méthodes d'évaluation (chez soi, sur serveur)

6.2 DESCRIPTEURS

6.2.1 *Spectrogramme*

6.2.2 *Échelle de Bark et de Mel*

6.2.2.1 *Bark*

6.2.2.2 *Mel*

6.2.3 *Coefficients cepstraux*

6.2.3.1 *Mel-Frequency Cepstral Coefficients*

GL : TODO : (Davis and Mermelstein, 1980)

6.2.3.2 *Gammatone Cepstral Coefficients*

6.2.4 *Transformation à Q-constant et Q-variable*

6.2.5 *Filtre de Gabor*

6.2.6 *Scattering*

GL : TODO : Scattering transforms are time-shift invariant representations of audio signals consisting of auditory and modulation filter

banks interspersed with complex modulus nonlinearities. This section highlights the importance of invariance to time-shifts and stability to time-warping in the representation of acoustic scenes, and explains how the scattering transform is designed to satisfy these properties while having a high discriminative power.

6.2.6.1 Invariance and stability in acoustic scenes

GL : TODO : The notion of invariance to time shifts plays an essential role in acoustic scene similarity retrieval as well as acoustic scene classification. Indeed, recordings may be shifted locally without affecting their perception and therefore such shifts do not convey any information about the class. To discard this superfluous source of variability, signals can be mapped to a time-shift invariant feature space before training the classifier, eliminating the need for this classifier to explicitly learn this invariance. From any time-varying set of descriptors $x_1(t, \gamma)$, where γ denotes a descriptor index, a representation invariant to time-shifts shorter in duration than T can be obtained by convolving x_1 with a low-pass filter $\Phi(t)$ of cutoff frequency set to $1/T$, as measured in Hertz :

$$\mathbf{S}_1 x(t, \gamma) = (x_1 * \Phi)(t). \quad (6)$$

GL : TODO : A downside is that the transient information in x_1 at scales finer than T are lost by this low-pass filtering, reducing discriminability in feature space. To address this issue, the scattering transform recovers this information by convolving x_1 with wavelets whose center frequencies are above $1/T$ and then applying complex modulus.

GL : TODO : By resorting to wavelet transform modulus, as opposed to Fourier transform modulus, the resulting features are provably stable to time-warping deformation, in the sense of Lipschitz regularity with respect to diffeomorphisms Mallat, 2012. In addition to invariance, this stability property is crucial to signal classification, since it guarantees robustness to small variations in pitch, reverberation, and rhythmic organization of events, which make up an important part of the intra-class variability among natural sounds.

GL : TODO : Starting from a monophonic waveform $x(t)$, the scattering transform is defined as an infinite cascade of wavelet transform and modulus operators. However, to achieve invariance to translation up to $T = 372$ ms, *i.e.* the approximate minimal duration between non-overlapping acoustic events, two layers of scattering transform often suffice.

GL : TODO : The next subsection describes the operations involved in the scattering transform, and in particular the construction of wavelet filter banks.

6.2.6.2 Wavelet scalogram

GL : TODO : Our convention for the Fourier transform of a continuous-time signal $x(t)$ is $\hat{x}(\omega) = \int x(t) \exp(i2\pi t\omega) dt$. Note that the frequency variable ω is expressed in Hertz, not in radians per second.

GL : TODO : Let $\psi(t)$ a complex-valued band-pass filter of center frequency ξ_1 and bandwidth ξ_1/Q_1 . A filter bank of wavelets is built by dilating $\psi(t)$ according to a geometric sequence of scales $2^{\gamma_1/Q_1}$, obtaining

$$\psi_{\gamma_1}(t) = 2^{-\gamma_1/Q_1} \psi(2^{-\gamma_1/Q_1} t). \quad (7)$$

GL : TODO : The variable γ_1 is a scale, or an inverse log-frequency, taking integer values between 0 and $(J_1 \times Q_1 - 1)$. In the sequel, we set ξ_1 to 20 kHz (close to the Nyquist frequency of the audio recordings), the number of octaves J_1 to 10 (the lower end of human hearing range), and the number of wavelets per octave Q_1 to 8. For each γ_1 , the wavelet $\psi_{\gamma_1}(t)$ has a center frequency of $2^{-\gamma_1/Q_1}\xi_1$, a bandwidth of $2^{-\gamma_1/Q_1}\xi_1/Q_1$, and a quality factor of Q_1 .

GL : TODO : The wavelet transform of an audio signal $x(t)$ is obtained by convolution with all wavelets. Applying pointwise complex modulus the transform yields the wavelet scalogram

$$x_1(t, \gamma_1) = |x * \psi_{\gamma_1}|(t). \quad (8)$$

GL : TODO : The wavelet scalogram bears resemblance to the constant-Q transform (CQT), which is derived from the short-term Fourier transform (STFT) by averaging the frequency axis into constant-Q subbands of center frequencies $2^{-\gamma_1/Q_1}\xi_1$. Indeed, both time-frequency representations are indexed by time t and log-frequency γ_1 . However, contrary to the CQT, the wavelet scalogram reaches the Heisenberg theoretical limit of optimal time-frequency localization across the whole frequency range, whereas the temporal resolution of the traditional CQT is fixed by the support of the STFT analyzing window Brown, 1992.

GL : TODO : Therefore, the wavelet scalogram has a better temporal localization at high frequencies than the CQT, at the expense of a greater computational cost since the inverse fast Fourier transform (IFFT) routine must be called for each wavelet ψ_{γ_1} in the filter bank.

GL : TODO : However, this allows us to observe amplitude modulations at fine temporal scales in the scalogram, down to the minimum scale $2Q_1/\xi_1$ for $\gamma_1 = 0$, of the order of 1 ms given the aforementioned values of Q_1 and ξ_1 .

6.2.6.3 Extracting modulations with second-order scattering

GL : TODO : Among auditory scenes, amplitude modulations may be caused by a broad variety of mechanical interactions, including colli-

sion, friction, and turbulent flow. At longer scales, they also account for higher-level attributes of sound, such as prosody in speech or rhythm in music. Although they are discarded while filtering $\mathbf{x}_1(t, \gamma_1)$ into a time-shift invariant representation $\mathbf{S}_1\mathbf{x}(t, \gamma_1)$, they can be recovered by a second wavelet transform modulus operator. The amplitude modulation spectrum resulting from this operator is

$$\mathbf{x}_2(t, \gamma_1, \gamma_2) = |\mathbf{x}_1 * \Psi_{\gamma_2}|(t, \gamma_1), \quad (9)$$

GL : TODO : where the center frequencies of the wavelets $\Psi_{\gamma_2}(t)$ are of the form $2^{-\gamma_2/Q_2}\xi_2$, and the second-order scale index γ_2 takes integer values between 0 and $(J_2 \times Q_2 - 1)$. Note that these second-order wavelets are dilated versions of a second mother wavelet ψ , with a different center frequency ξ_2 and quality factor Q_2 . The identity of the wavelet will be clear from context. In the sequel, we set ξ_2 to 2.5 kHz, Q_2 to 1, and J_2 to 12. Lastly, the low-pass filter $\phi(t)$ is applied to \mathbf{x}_2 to guarantee invariance to time-shifting, which yields

$$\mathbf{S}_2\mathbf{x}(t, \gamma_1, \gamma_2) = (\mathbf{x}_2 * \phi)(t, \gamma_1, \gamma_2). \quad (10)$$

GL : TODO : The scattering transform $\mathbf{S}\mathbf{x}(t, \gamma)$ consists of the concatenation of first-order coefficients $\mathbf{S}_1\mathbf{x}(t, \gamma_1)$ and second-order coefficients $\mathbf{S}_2\mathbf{x}(t, \gamma_1, \gamma_2)$ into a feature matrix $\mathbf{S}\mathbf{x}(t, \gamma)$, where γ is a shorthand for either γ_1 or (γ_1, γ_2) .

6.2.6.4 Gammatone wavelets

GL : TODO : Wavelets $\Psi_{\gamma_1}(t)$ and $\Psi_{\gamma_2}(t)$ are designed as fourth-order Gammatone wavelets with one vanishing moment Venkitaraman et al., 2014, and are shown in Figure ???. In the context of auditory scene analysis, the asymmetric envelope of Gammatone wavelets is more biologically plausible than the symmetric, Gaussian-like envelope of the more widely used Morlet wavelets. Indeed, it allows to reproduce two important psychoacoustic effects in the mammalian cochlea : the asymmetry of temporal masking and the asymmetry of spectral masking Fastl and Zwicker, 2007. Moreover, it should be noted that Gammatone wavelets follow the typical amplitude profile of natural sounds, beginning with a relatively sharp attack and ending with a slower decay. As such, they can be discovered automatically by unsupervised encoding of natural sounds Smith2006 This suggests that, despite being hand-crafted and not learned, Gammatone wavelets provide a sparse time-frequency representation of acoustic scenes.

6.2.6.5 Logarithmic compression

GL : TODO : Many algorithms in pattern recognition, including nearest-neighbor classifiers and SVMs, tend to work best when all features

follow a standard normal distribution across all training instances Hsu et al., 2003. Yet the distribution of the scattering coefficients is skewed towards larger values. We can reduce this skewness by applying a pointwise concave transformation, such as a logarithm, to all coefficients. Figure ?? shows the distribution of an arbitrarily chosen scattering coefficient over the DCASE 2013 dataset, before and after logarithmic compression.

GL : TODO : Taking the logarithm of a magnitude spectrum is ubiquitous in audio signal processing. Indeed, it is corroborated by the Weber-Fechner law in psychoacoustics, which states that the sensation of loudness is roughly proportional to the logarithm of the acoustic pressure. We must also recall that the measured amplitude of sound sources often decays polynomially with the distance to the microphone – a source of spurious variability in scene classification. Logarithmic compression can linearize this dependency, facilitating the construction of a powerful invariant at the classifier stage.

GL : TODO : For the task of musical genre recognition, second-order scattering coefficients $S_2x(t, \gamma_1, \gamma_2)$ are sometimes normalized by the corresponding first-order scattering coefficients $S_1x(t, \gamma_1)$, since this decorrelates them from one another Andén and Mallat, 2014. We note that taking the logarithm of such renormalized coefficients yields

$$\log \frac{S_2x(t, \gamma_1, \gamma_2)}{S_1x(t, \gamma_1)} = \log S_2x(t, \gamma_1, \gamma_2) - \log S_1x(t, \gamma_1), \quad (11)$$

GL : TODO : *i.e.*a linear combination of the logarithms of first- and second-order coefficients. As such, a non-linear renormalization becomes a linear transformation, which can be learned by a linearly discriminative classifier.

6.3 ALGORITHMES ET CLASSIFIEURS

6.3.1 Modèle de Markov caché

GL : TODO : (Rabiner, 1989)

6.3.2 Machines à vecteurs de support

6.3.3 Factorisation de matrice non-négative

6.3.4 Autres classifieurs

GL : TODO : random forest, GMM

6.4 DÉTECTION DES ÉVÉNEMENTS SONORES

6.4.1 Objectifs

GL : TODO : AED

6.4.2 Métrique

Les performances des algorithmes en AED sont évaluées suivant différentes métriques. Deux d'entre elles sont particulièrement utilisées, et notamment dans les challenges DCASE 2013 (cf. Section 6.4.3) et 2016 (cf. Section 6.4.4). Nous les détaillons dans cette section.

La première métrique est la F-mesure (Giannoulis et al., 2013a; Giannoulis et al., 2013b; Stowell et al., 2015), que l'on note F dans ce document. Cette dernière se calcule comme suit :

$$F = \frac{2 \times P \times R}{P + R} \quad (12)$$

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 et 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 sur le nombre d'événements à détecter r (le nombre d'événements présents dans la scène) :

$$P = \frac{c}{e} \quad , \quad R = \frac{c}{r} \quad , \quad e = c + fp \quad , \quad r = c + fn \quad (13)$$

avec fp le nombre de faux positifs, et fn le nombre de faux négatifs.

La deuxième métrique est le taux d'erreur acoustique (Poliner and Ellis, 2007; Stiefelhagen et al., 2007), que l'on note ER dans ce document. Ce dernier se calcule comme suit :

$$ER = \frac{D + I + S}{N} \quad (14)$$

avec N le nombre d'événements à détecter **GL : TODO : vérifier**, 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 tienne compte :

- du nombre de trames correctement identifiées (*sb* : *segment based*);

- du nombre d'événements correctement identifiées (eb : *event based*).

Considéré le nombre d'événements plutôt que les trames permets entre autres d'obtenir une mesure de performance indépendante de la durée des événements. Dans ce cas, on considère usuellement qu'un événement est correctement identifié si son *onset* a été correctement identifié, ou si à la fois son *onset* et son *offset* ont été correctement identifiés. La détection d'une frontière (*onset* ou *offset*), est toujours considérée avec un seuil de tolérance.

Ainsi nous notons F_{sb} et ER_{sb} , les F-mesures et taux d'erreur acoustiques calculés en tenant compte des trames, et F_{eb} et ER_{eb} les F-mesures et taux d'erreur acoustiques calculés en tenant compte des événements.

La détection de l'offset d'un événement sonore étant est un tâche compliquée, que ce soit pour des algorithmes ou des humains, nous ne considérons dans ce document des mesures de F_{sb} et ER_{sb} calculés en fonction du nombre d'événements dont les *onsets* ont été correctement identifiés.

Enfin, il est possible de calculer ces métriques en

Ces métriques, si elles sont calculés sans faire de distinction entre les classes, sont susceptible de donner des poids distincts dans l'évaluation entre les classes bien représentées dans la scène, et celles présentant que peu d'événement. Afin de parer à ce biais, il est possible de calculer les métriques séparément pour chaque classe, avant de les moyenner. On note ainsi F_{cw} et ER_{cw} , les versions alternatives de F et ER normalisée par classe :

$$F_{cw} = \frac{1}{C} \sum_{i=1}^C F^i \quad , \quad ER_{cw} = \frac{1}{C} \sum_{i=1}^C ER^i \quad (15)$$

avec C le nombre de classes à détecter et F^i et ER^i , la F-measure et le taux d'erreur acoustique 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 AED : nommément F_{sb} , F_{eb} , $F_{cw_{sb}}$, $F_{cw_{eb}}$, ER_{sb} , ER_{eb} , $ER_{cw_{sb}}$ et $ER_{cw_{eb}}$.

6.4.3 Tâche 3 du challenge DCASE 2013

(Giannoulis et al., 2013a; Giannoulis et al., 2013b; Stowell et al., 2015)
GL : TODO : DCASE 2013, fenêtre de tolérance de ± 100 ms (Giannoulis et al., 2013a; Stowell et al., 2015)

6.4.4 *Tâche 3 du challenge DCASE 2016*

6.5 CLASSIFICATION DES SCÈNES SONORES ENVIRONNEMENTALES

6.5.1 *Objectifs*

GL : TODO : ASC

6.5.2 *Métrique*

6.5.3 *Tâche 1 du challenge DCASE 2013*

6.5.4 *Tâche 1 du challenge DCASE 2016*

6.6 RECOUVREMENT DES SIMILARITÉS ACOUSTIQUES

6.6.1 *Objectifs*

GL : TODO : ASSR

6.6.2 *Métrique*

The metric used is the precision at rank k ($p@k$), which is computed by taking a query item and counting the number of items of the same class within the k closest neighbors, and then averaging over all query items.

6.6.3 *Méthodes et algorithmes*

7

APPLICATION DU MODÈLE MORPHOLOGIQUE À L'ÉVALUATION DES ALGORITHMES D'ANALYSE AUTOMATIQUE DE SCÈNES SONORES ENVIRONNEMENTALES

7.1 INTRODUCTION

GL : TODO : un point méthodologique, ici corpus de scènes et non corpus d'objets, du coup chaque scène est un sous corpus, du coup on a une métrique par scène et une métrique moyenne par corpus, du coup théorème centrale limite, du coup test paramétrique, du coup on s'éloigne des recommandations de (Demšar, 2006)

7.2 APPLICATION AU CHALLENGE DCASE 2013

7.2.1 *Objectif*

GL : L'objectif dans cette étude est de ré-évaluer les algorithmes soumis dans le cadre de la tâche 2 de détection d'événements sonores (AED) du challenge DCASE 2013¹ (cf. Section 6.1.2).

GL : 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 "sample différent", nous entendons des enregistrements d'événements sonores différents, mais appartenant à la même classe (*e.g.* claquement de porte). En effet,

¹ Par souci de lisibilité, nous nous préciserons 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 (AED) dudit challenge.

quand on considère une tâche de classification, un problème de taille 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'entraînements et de développement.

La méthode suivie consiste à utiliser le modèle de scènes sonores proposé (cf. Section 4.2.2.4) afin de générer de nouveaux corpus de scènes simulées, scènes dont nous contrôlons les structures internes, ainsi que la nature des samples utilisés.

GL : 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.

7.2.2 Génération des corpus

Cette section décrit les différents corpus de scènes simulées utilisés lors de l'expérience.

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 (AED) du challenge DCASE 2013 (Giannoulis et al., 2013b) (cf. Section 6.4.3).

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 individus 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

Index	Nom	Description
1	porte-frapper	Frapper à la porte
2	porte-claquer	Claquer 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

banques de données de sons isolés, ainsi que les processus de simulation sont détaillés dans les sections suivantes (cf. Sections 7.2.2.1, 7.2.2.2 et 7.2.2.3).

7.2.2.1 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 4.2.1.3) rendant compte de l'environnement acoustique naturel du lieu d'occurrence des événements.

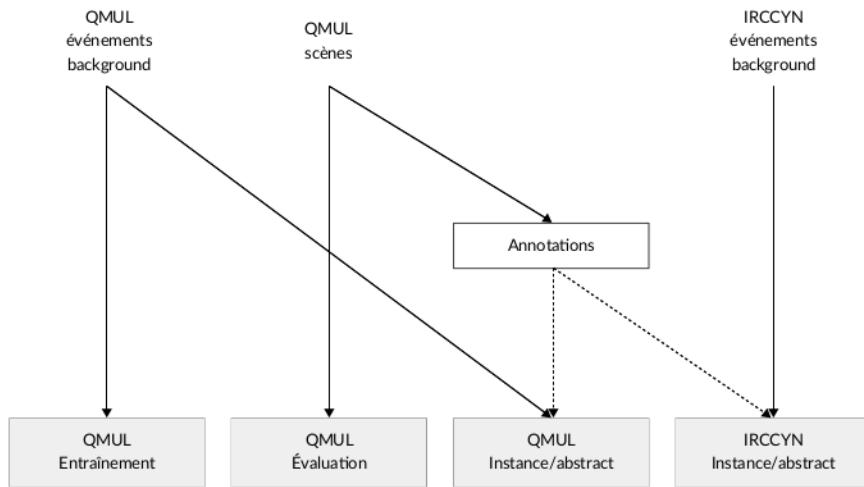


FIGURE 45 : Generation process of the corpora considered in this evaluation. As part of the DCASE challenge, systems were trained on QMUL Train and tested on QMUL Test during the DCASE challenge.

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 AED 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 profitent donc des mêmes conditions d'enregistrement 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 dernière 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; Giannoulis et al., 2013b). 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.

GL : TODO : détailler la banque de données IRCCYN

7.2.2.2 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 terme 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 de la valeur efficace (niveau RMS, cf. Section 5.2.4) du signal (cf. Équation 17) de l'événement (E_{rms}) et du *background* (B_{rms}) :

$$EBR = 20 \log_{10} \left(\frac{E_{rms}}{B_{rms}} \right) \quad (16)$$

$$X_{rms} = \sqrt{\frac{1}{N} \sum_{n=1}^N x(n)^2} \quad (17)$$

$x(n)$ 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.

GL : TODO : expliquer comment on supprime le niveau de bruit dans E_{rms}

Les positions *onsets* et les EBRs ainsi recouvrés 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 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 4.2.2.4. 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*.

7.2.2.3 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 4.2.2.4. Concrètement, le modèle est instancié suivant des paramètres μ_i^a , σ_i^a , μ_i^t et σ_i^t (cf. Équation. 3 et 4) estimés sur la scène enregistrée. Pour chaque couple scène-annotateur du corpus *test-QMUL*, ces paramètres sont estimés à partir de l'annotation (μ_i^t et σ_i^t) et du signal (μ_i^a et σ_i^a). Les EBRs et les espacements inter-onsets de la scène simulée sont alors obtenus à partir des distributions normales $N(\mu_i^a, \sigma_i^a)$ et $N(\mu_i^t, \sigma_i^t)$ respectivement. 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 long par rapport à ceux des scènes enregistrées, la durée D d'un sample d'une classe i est seuillé si :

$$D - \mu_i^d - \sigma_i^d > 5 \quad (18)$$

avec, μ_i^d et σ_i^d 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.

7.2.2.4 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(o)*, *insQ-EBR(-6)* et *insQ-EBR(-12)*. Pour *insQ-EBR(o)*, 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'origines. A 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 le processus de simulation *abstract* (*abstract-QMUL* et *abstract-IRCCYN*), les EBRs et espacements inter-onsets des samples obtenus à partir des distributions normales $\mathcal{N}(\mu_i^a, \sigma_i^a)$ et $\mathcal{N}(\mu_i^t, \sigma_i^t)$ sont également re-tirés d'une réplication à une autre **GL : TODO : A vérifier**. Chaque corpus/sous-corpus est ainsi composé de 220 scènes simulées ($11 \times 2 \times 10$).

Tous les corpus sont disponibles en ligne² et ont été simulés à l'aide de l'outil de simulation MATLAB développé dans le cadre de cette thèse (cf. Section 4.4).

7.2.2.5 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.

² Dataset URLs :

- *test-QMUL* : https://archive.org/details/dcase2013_event_detection_testset_01;
- *instance-QMUL, abstract-QMUL* : https://archive.org/details/dcase_replicate_qmul;
- *instance-IRCCYN, abstract-IRCCYN* : https://archive.org/details/dcase_replicate_irccyn

À la fin de l'expérience, les sujets sont invités à commenter librement leurs notations.

Apparatus

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.

Participant

15 sujets ont participé à l'étude. Tous ont réalisé l'expérience avec succès.

Résultats

Nous considérons $\mathcal{R}_{\text{ sujet }}$ les notes de réalisme par sujet, notes moyennées en considérant séparément les scènes de *test-QMUL* et celles de *instance-IRCCYN*.

Les $\mathcal{R}_{\text{ sujet }}$ des scènes enregistrées et simulées sont respectivement de 4.4 et 3.3 (cf. Figure 46). Les deux population 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'agit de scènes jouées.

En ce qui concerne les scènes simulées, les sujets ont rapporté que :

- “le fond sonore semble synthétique/artificiel”, bien que ce dernier 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 7.2.2.2. Ce choix est pris dans le but de minimiser la différence entre la scène simulée, et celle de référence.

Il convient de noter que, pour de nombreux participants, certaines scènes simulées ont reçu une note de réalisme plus élevée que certaines des scènes naturelles, ce qui montre que, bien que des différences notables peuvent être faites, elles n'influencent le réalisme acoustique que dans une moindre mesure.

7.2.3 Métrique

La métrique considérée dans cette analyse est $F_{\text{Cw eb}}$ (cf. Section 6.4.2), *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 100ms. Cette dernière a l'avantage :

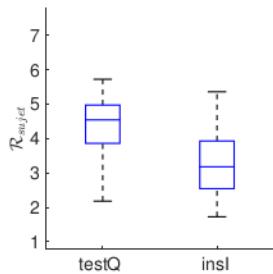


FIGURE 46 : Distribution des notes de réalisme $\mathcal{R}_{\text{subject}}$ pour les scènes enregistrées *test-QMUL* et les scènes simulées *instance-IRCCYN*

- d'être facilement interprétable ;
- de ne favoriser aucune classe.

7.2.4 Données et analyses

GL : Pour le calcul des métriques, nous suivons la méthodologie suivie 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.

GL : Cette approche nous laisse avec 11 observations (i.e. 11 mesures de performances) par condition expérimentale (corpus et systèmes)³.

L'analyse s'effectue en deux temps, suivant le corpus de sons isolés considéré :

- GL : Corpus de sons isolés QMUL : cette analyse considère les trois corpus *test-QMUL*, *instance-QMUL* et *abstract-QMUL*. Elle a deux objectifs :
 1. GL : évaluer si il existe des différences significatives entre les performances des algorithmes observées sur les corpus *test-QMUL*, *insQ-EBR(o)* et *abstract-QMUL*. L'objectif ici est de vérifier que les algorithmes sont capables de généraliser

³ À noter que la méthode adoptée dans ce document pour intégrer les performances des algorithmes diffère de 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.

les performances obtenues sur *test-QMUL* pour des corpus simulés avec des sons isolées enregistrés dans les mêmes conditions que ceux de *test-QMUL*, et dont les scènes possèdent une structure (EBR et positions des *onsets*) identique (*insQ-EBR(o)*) ou similaire (*abstract-QMUL*) à celles de *test-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. GL : évaluer si il existe des différences significatives entre les performances des algorithmes observées sur les corpus *insQ-EBR(6)*, *insQ-EBR(o)*, *insQ-EBR(-6)* et *insQ-EBR(-12)*. L'objectif ici est de vérifier que les algorithmes sont capables de généraliser les performances obtenues sur *insQ-EBR(o)* pour des corpus simulés identiques mais possédant des niveaux de bruits 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 ;
- GL : corpus de sons isolés *IRCCYN* : cette analyse considère les trois corpus *test-QMUL*, *instance-IRCCYN* et *abstract-IRCCYN*. L'objectif ici est de vérifier que les algorithmes sont capables de généraliser les performances obtenues sur *test-QMUL* pour des corpus simulés avec des sons isolées enregistrés dans des conditions différentes de ceux de *test-QMUL*, et dont les scènes possèdent une structure (EBR et positions des *onsets*) identique (*instance-IRCCYN*) ou similaire (*abstract-IRCCYN*) à celles de *test-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.

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.

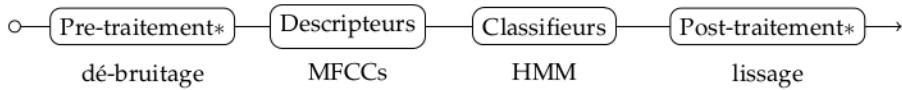


FIGURE 47 : Vision schématisée des systèmes de détection d'événements du challenge DCASE 2013; * indique que le nœud n'est pas systématiquement utilisé; les choix états de l'art sont donnés en exemple sous les nœuds.

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)
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; Vugue et al., 2013)	MFCC	HMM	(D, C)
Baseline (Giannoulis:2013a)	CQT	GMM	x

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

7.2.5 Système de détection

Tous les algorithmes ayant été évalués lors de la tâche 2 (AED) 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 47, incluant parfois une étape de prétraitement de débruitage.

Le classifieur de choix est un HMM (cf. Section 6.3.1) à 2 couches, dont la première modélise l'événement, et la seconde, la transition entre les événements. D'autres classifieurs incluant les forêts d'arbres décisionnels (RF : *Random Forests*, cf. Section 6.3.4), les machines à vecteurs de support (SVM : *Support Vector Machines*, cf. Section 6.3.2), la factorisation en matrices non négatives (NMF : *Non-negative Matrix Factorization*, cf. Section 6.3.3), ainsi que des modèles de mélanges gaussiens (GMM : *Gaussian mixture model*, cf. Section 6.3.4) sont également utilisés. Nous invitons le lecteur à se référer à (Stowell et al., 2015) ou/et 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 (cf. Section 6.2.2.2) ;
- *CQT* : une représentation temps-fréquence calculée en GL : TODO (cf. Section 6.2.4) ;
- *MFCC* : une représentation basée sur des coefficients cepstraux calculés sur une échelle de Mel (MFCCs : *Mel-Frequency Cepstral Coefficients* ; cf. Section 6.2.3.1) ;
- *GF* : une représentation temps-fréquence filtrée par un banc de filtres de Gabor (GF : *Gabor filterbank* ; cf. Section 6.2.5) ;
- *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.

7.2.6 Résultats

7.2.6.1 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 calculs que ceux utilisés pour la tâche 2 (AED) du challenge DCASE 2013. Les systèmes ont par ailleurs été re-testés sur le corpus *test-QMUL* (corpus de *test* du challenge AED 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(o)* 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(o)* et *abstract-QMUL*. Pour ces raisons nous ne considérons pas plus avant ce système.

GL : 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 ainsi au premier abord que le changement de corpus a bien provoqué une modification des

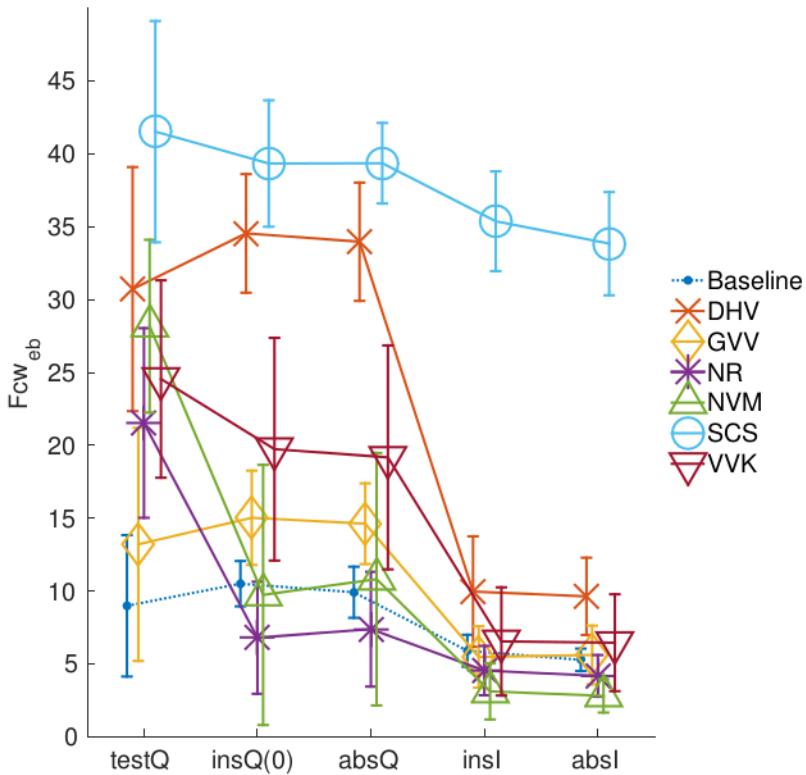


FIGURE 48 : Performances des systèmes évaluées dans le cadre du challenge DCASE 2013 sur les corpus QMUL et IRCCYN en considérant F_{cweb} .

performances, et ce bien que les sons utilisés pour simuler les corpus *abstract-QMUL* et *insQ-EBR(o)* soient similaires à ceux que l'on trouve dans *test-QMUL*.

GL : 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(o)* d'autre part. Seules 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 leur 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(o)* : les performances comparables montrent que les algorithmes sont robustes au changement d'événements. À noter que les samples proviennent tous des enregistrements de QMUL, *i.e.* ont été enregistrés dans les mêmes conditions ;

Système	testQ	insQ-EBR(o)	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 Fcw_{eb} pour chacun des systèmes évalués dans le cadre du challenge DCASE 2013 en considérant les corpus *test-QMUL*, *insQ-EBR(o)* 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 (*).

- 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 inter-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. En effet, la chute peut être due 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 : toux, NR : porte-claquer) semble être détectée de manière abusive, augmentant drastiquement le nombre de faux positifs, et diminuant *de facto* les résultats.

Système	testQ	insQ-EBR(o)	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)	25.35 (porte-claquer)	20.68 (porte-claquer)
NVM	1.26 (rire)	22.48 (toux)	19.22 (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.

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.

7.2.6.2 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(o)* et *insQ-EBR(6)* (cf. Section 7.2.2.4) sont présentés sur la figure 49.

GL : L'ANOVA rapporte 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 ($F[18, 180] = 16$, $p_{gg} < 0.01$).

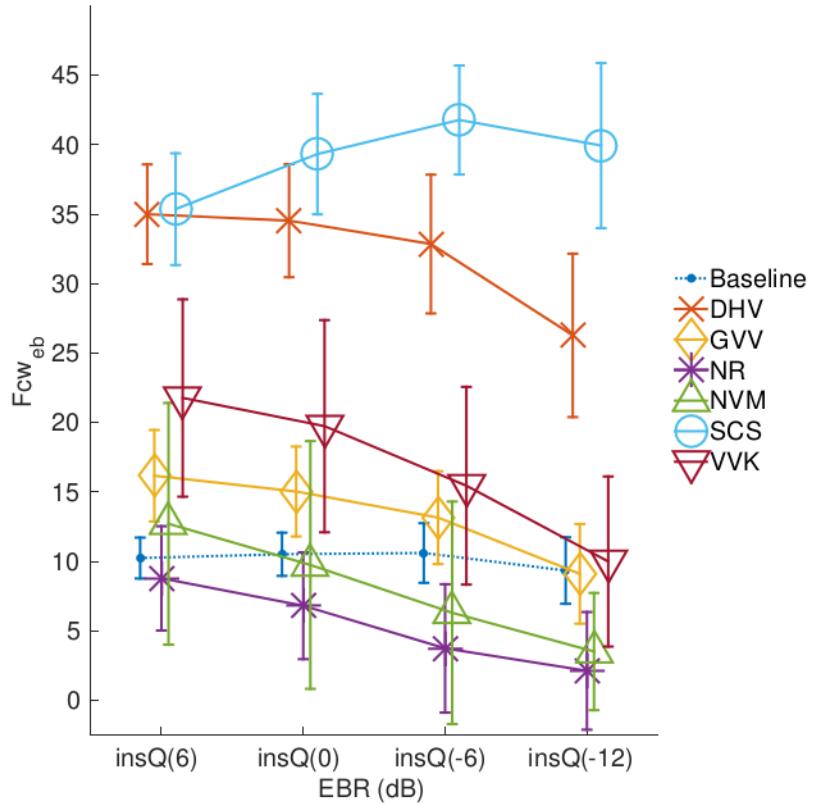


FIGURE 49 : Performances des systèmes évaluées dans le cadre du challenge DCASE 2013 sur les corpus *instance-QMUL* simulés avec différents EBR (6, 0, -6 et -12dB).

GL : Concernant l'analyse *post hoc* entre 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 qui ne présente pas l'amélioration 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 substantielle dont bénéficie SCS, pré-traitement qui est au cœur de son algorithme (Schröder et al., 2013a).

GL : TODO : expliquer pourquoi la baseline ne varie pas ?

Système	testQ	insI	absI
Baseline	9.0±4.8*	5.9±2.9	5.6±2.9
DHV	30.7±8.4*	10.0±5.8	9.5±5.6
GVV	13.2±8.0*	5.6±3.7	5.5±3.6
NR	21.5±6.5*	4.6±3.4	5.4±4.5
NVM	28.2±5.9*	3.1±3.1	3.2±3.0
SCS	41.5±7.6*	35.4±7.2	34.0±6.7
VVK	24.6±6.8*	6.6±5.7	7.3±6.3

TABLE 21 : Résultats mesurés par $F_{cw_{eb}}$ 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 (*).

GL : L'influence de l'EBR est cependant cohérente, le classement en termes de performances entre les algorithmes étant maintenu pour les différents corpus.

GL : Concernant l'analyse *post hoc* entre systèmes, seuls DHV et SCS présentent des performances significativement supérieures à celles de la *baseline*, et ce quelque 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 significativement 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.

GL : Ces résultats nous amènent à conclure que à part SCS, aucun des systèmes considérés n'est robuste aux différents niveaux de bruit.

7.2.6.3 Corpus IRCCYN

GL : Les résultats sont affichés sur le tableau 21 et la figure 48. 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 ($F[12, 180] = 17, p_{gg} < 0.01$).

GL : Concernant l'analyse *post hoc* relatif aux corpus, alors que la plupart des systèmes ont obtenu des performances comparables entre *test-QMUL* et les corpus *abstract-* et *instance-QMUL*, tous algorithmes voient leurs résultats diminuer de manière significative pour les corpus *abstract-* et *instance-IRCCYN*.

GL : 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. En particulier le 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 par rapport à la baseline observé sur le corpus *test-QMUL* 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 clairement le voir sur la figure 48, seul le système SCS (gagnant du challenge AED 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 quelque soit la condition expérimentale que l'on fait varier, nommément :

- les samples sélectionnés (en considérant deux banques de sons isolés différentes) ;
- les positions temporelles des samples ;
- les EBRs.

7.2.7 Discussion

Pour résumer les résultats présentés précédemment, 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 l'utilité des processus de simulation proposés (cf. Sections 7.2.2.3 et 7.2.2.2), ces derniers permettant bien de répliquer ou d'aller plus loin dans l'analyse des performances 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, connaissance pouvant ainsi aider les chercheurs dans leurs choix algorithmiques.

Les facteurs influant sur les performances tels que le niveau de bruit, le niveau de polyphonie, la diversité intra-classe (différence acoustique entre la formation et les données d'essais) peuvent ainsi être évalués de façon indépendante, sans avoir la charge

1. d'enregistrer des scènes présentant les propriétés désirées ;
2. d'annoter manuellement les données.

Même si l'usage exclusif de données simulées, pour valider une approche algorithmique, est insuffisant, nous pensons que la seule utilisation de données réelles ne permet pas non plus 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 minutieuse d'un grand ensemble de données d'évaluation étant une tâche très exigeante. En outre, l'annotation a posteriori de la présence des événements doit être effectuée par des individus dont le consensus n'est pas 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.

7.3 APPLICATION AU CHALLENGE DCASE 2016

7.3.1 Objectifs

Nous présentons dans cette section les résultats de la tâche 2 du challenge DCASE 2016⁴, la seconde édition du challenge DCASE 2013. La tâche 2 est nommée "détection d'événements sonores dans des environnements simulées"⁵. Cette tâche 2 a été réalisée dans le cadre de notre thèse.

L'objectif est ici d'évaluer les performances d'algorithmes en AED sur des corpus de scènes simulées, scènes dont nous contrôlons :

- l'intensité des événements sonores ;
- le nombre d'événements sonores par scènes.

Par ailleurs nous faisons la distinction entre deux types de scènes, à savoir :

⁴ cf. <http://www.cs.tut.fi/sgn/arg/dcase2016/>

⁵ cf. <http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio>

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

- les scènes autorisant le recouvrement entre les événements de classes différentes ;
- les scènes n'autorisant pas le recouvrement.

7.3.2 Génération des corpus

7.3.2.1 Banque de sons isolés

La banque de données de sons isolés IRCCYN (cf. Section 7.2.2.1) 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 très mal détectées.
- imprimante : cette classe a été supprimée en raison de son aspect singulier par rapport aux autres classes. En effet, la classe 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 le cas où le recouvrement entre les événements est interdit.

7.3.2.2 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 7.2.2.2) ;
- nombred'événement (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). Ainsi, nous garantissons que chaque classe soit représentée par le même nombre d'événements, indépendamment de leurs durées.

Par ailleurs, les EBRs des événements sont constants, *i.e.* fixer un EBR de -6dB pour une scène revient à fixer un EBR de -6dB 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, 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 événement ne se recouvre.

Nous considérons trois niveaux pour les paramètres EBR et nec. Pour 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.

7.3.2.3 Banque d'entraînement

La banque d'entraînement comprend 220 sons isolés, 20 pour chacune des classes considérées.

7.3.2.4 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 7.3.2.2), nous donnant ainsi un corpus formé 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.

7.3.2.5 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 7.3.2.2), la simulation est répliquée trois fois, nous donnant ainsi un corpus de 54 scènes sonores simulées (18×3). Pour chaque réplication, la position des *onsets* est retirée. 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 replication.

Concernant les samples de *background*, ces derniers sont différents pour chaque replication.

Les scènes ont toutes une durée de 2 minutes. La durée total du corpus est ainsi de 108 minutes.

7.3.3 Métrique

Parmi les 4 métriques utilisées dans le cadre du challenge DCASE 2016 (cf. Section 6.4.2), nous en choisissons 1 dans le cadre de cette analyse :

- 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 200 ms.

7.3.4 Données et analyses

Les métriques sont calculées séparément sur chacune des scènes du corpus d'évaluation, et moyennées en fonction des conditions expérimentales considérées. Ce que faisant, 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'apprecier 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 un 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) comportant 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) comportant 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 deux 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

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., 2016)	VQT	NMF		

TABLE 23 : Description synthétique des systèmes soumis dans le cadre de la tâche 2 du challenge DCASE 2016

(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.

7.3.5 Systèmes de détection

Cette section 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 5 groupes :

- *mel/bark* : une représentation temps-fréquence, où l'axe fréquentiel a été projeté sur une échelle particulière, soit de Bark (Pikrakis ; cf. section 6.2.2.1) soit de Mel (Choi, Hayashi 1, Hayashi 2, Giannoulis et Kong ; cf. Section 6.2.2.2) ;
- VQT/CQT : une représentation temps-fréquence calculée en GL : TODO(cf. Section 6.2.4) ;

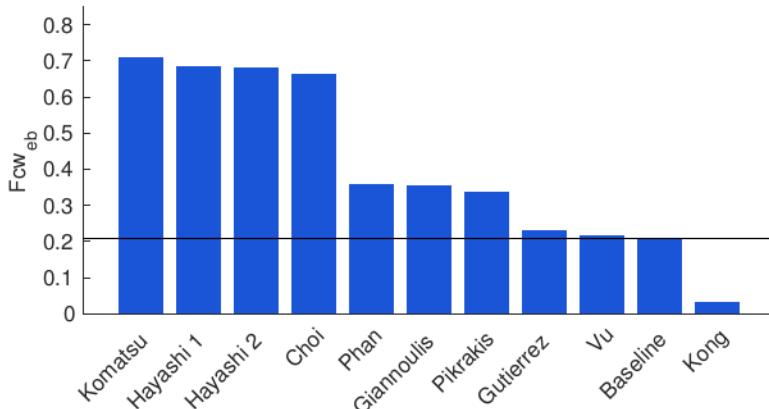


FIGURE 50 : Performances globales des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique F_{cweb} .

- *MFCC* : une représentation basée sur des coefficients cepstraux calculés sur échelle de Mel (MFCCs : *Mel-Frequency Cepstral Coefficients* ; cf. Section 6.2.3.1).
- *GTCC* : une représentation basée sur des coefficients cepstraux calculés sur échelle de Gammatone (GTCCs : *Gammatone cepstral coefficients* ; cf. Section 6.2.3.2).

Concernant les classifieurs utilisés, on distingue 6 groupes :

- Réseaux de neurones : *Kong* et *Choi* GL : TODO ;
- Factorisation de matrices non négatives : *Komatsu*, *Giannoulis* et *baseline* GL : TODO ;
- *BLSTM* : *Hayashi 1* et *Hayashi 2* GL : TODO ;
- Arbre de décision : *Phan* GL : TODO ;
- Plus proches voisins : *Gutierrez* GL : TODO ;
- Template matching : *Pikrakis* GL : TODO.

7.3.6 Résultats

7.3.6.1 Analyse globaux

Les résultats globaux sont affichés sur la figure 50. L'ANOVA pratiquée sur F_{cweb} 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*);
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 améliorant les résultats de près de 45%.

Il est difficile de dégager l'influence d'un classifieur particulier, les systèmes du groupe 1 faisant usage de DNN, de NMF et de BLSTM. 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 réduisant ce dernier sur les 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. Or la banque d'entraînement proposée dans le cadre de cette tâche est loin d'être suffisante.

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 conséderons pas plus avant les résultats de ce système dans la suite de l'analyse.

7.3.6.2 Influence de la polyphonie

Les résultats par types de scènes (polyphoniques et non-polyphoniques) sont affichés sur la figure 51. L'ANOVA pratiquée sur $F_{cw_{eb}}$ révèle un effet positif 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 affecté les performances des algorithmes, ces derniers étant capables de gérer de manière équivalente les deux cas de figures. L'analyse *post hoc* sur

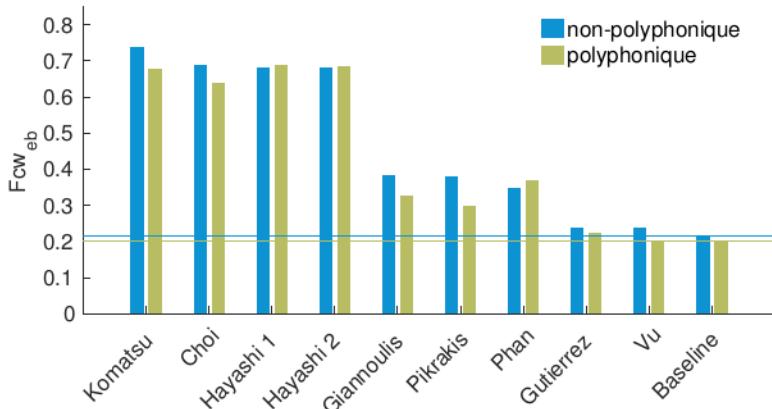


FIGURE 51 : Influence de la polyphonie sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique $F_{cw_{eb}}$.

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 7.3.6.1), s'agissant des scènes polyphoniques, ou s'agissant des scènes non polyphoniques. Une seule différence est néanmoins notée au niveau des scènes polyphoniques : les systèmes *Gutierrez* et *Pikrakis* ne présentant plus de différences significatives dans ce cas.

7.3.6.3 Influence du niveau de bruit

Considérant les scènes non-polyphoniques, les résultats sont affichés sur la Figure 52a. 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 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 -6dB , 4 groupes émergent :

1. *Komatsu* : performances supérieures à celles de la *Baseline*; *choi*, *Hayashi 1* et *Hayashi 2* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1;

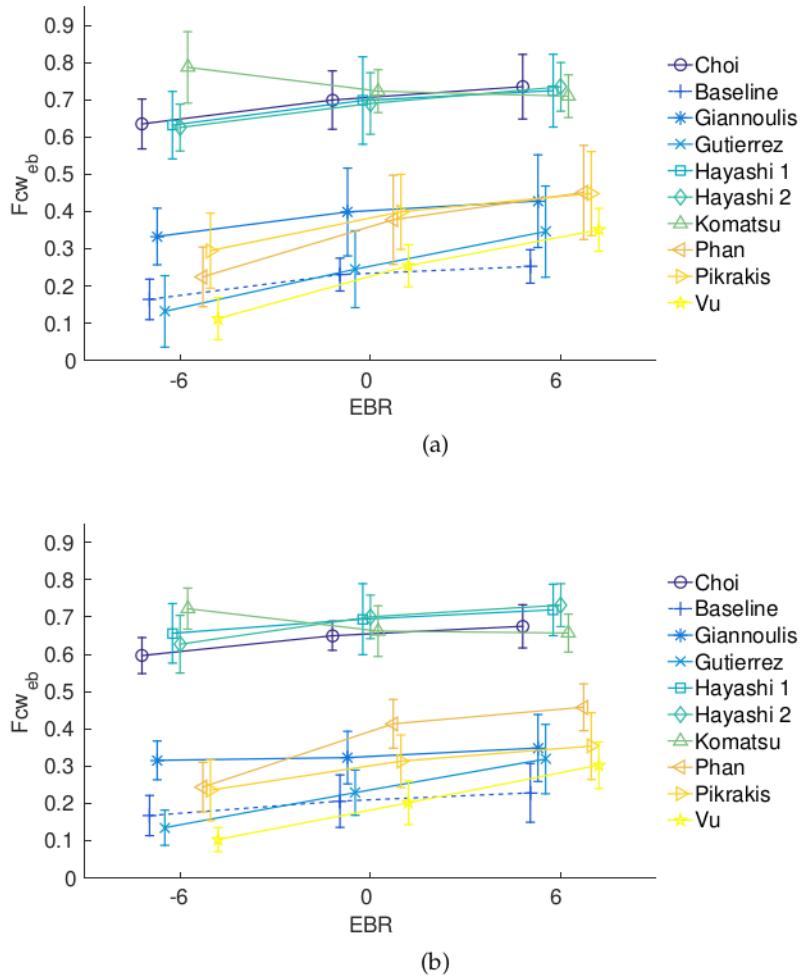


FIGURE 52 : Influence du niveau de bruit (EBR) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique F_{CWEb} ; (a) scènes non-polyphoniques, (b) scènes polyphoniques.

2. *Gianoulis* : performances supérieures à celles de la *Baseline*, mais inférieures à celles du groupe 1;
3. *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). A noter que seul cet algorithme voit ses performances décroître avec l'EBR ([GL : TODO](#)), 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* présentent eux des performances systématiquement supérieures à celles des autres systèmes, et également celles de *Komatsu* pour des EBR de 0 et +6dB. Pour ces trois systèmes, l'augmentation du niveau de bruit (6dB → -6dB) provoque 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*.

S'agissant des scènes polyphoniques, les résultats sont affichés sur la Figure [52b](#). 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*; *choi, Hayashi 1 et Hayashi 2* : performances supérieures à celles de la *Baseline* mais inférieures à celles du groupe 1 ;
2. *Gianoulis* : performances supérieures à celles de la *Baseline*, mais inférieures à celles du groupe 1 ;
3. *Gutierrez, Pikrakis, Phan et Vu* : performances ne présentant pas de différences significatives avec celles de la *Baseline*.

Pour des EBR de 0 et -6dB, 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 +6dB. Ce dernier système souffre ainsi d'une mauvaise prise en compte du bruit, mauvaise prise en compte particulièrement pénalisante dans le cas de scènes polyphoniques (0dB → -6dB : 41% → 24%)

7.3.6.4 Influence du nombre d'événements

Considérant les scènes non-polyphoniques, les résultats sont affichés sur la figure 53a. L'ANOVA révèle un effet significatif du type de système ($F[9, 216] = 264$, $p_{gg} < 0.01$), 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 53a ; 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 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 → 3 ; *Hayashi 1* : 62% → 75% ; *Hayashi 2* : 63% → 73%) et d'une dégradation pour 1 système (nec : 1 → 3 ; *Gianoulis* : 48% → 32%). Concernant les scènes polyphoniques, un augmentation est constatée pour 3 systèmes (nec : 3 → 5 ; *Hayashi 1* : 67% → 72% ; *Hayashi 2* : 66% → 73% ; *Komatsu* : 63% → 70%) et une dégradation pour 3 (nec : 3 → 5 ; *Gianoulis* : 37% → 27% ; *Pikrakis* : 36% → 24% ; *Choi* : 67% → 60%).

Alors que l'augmentation du niveau de bruit avait 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

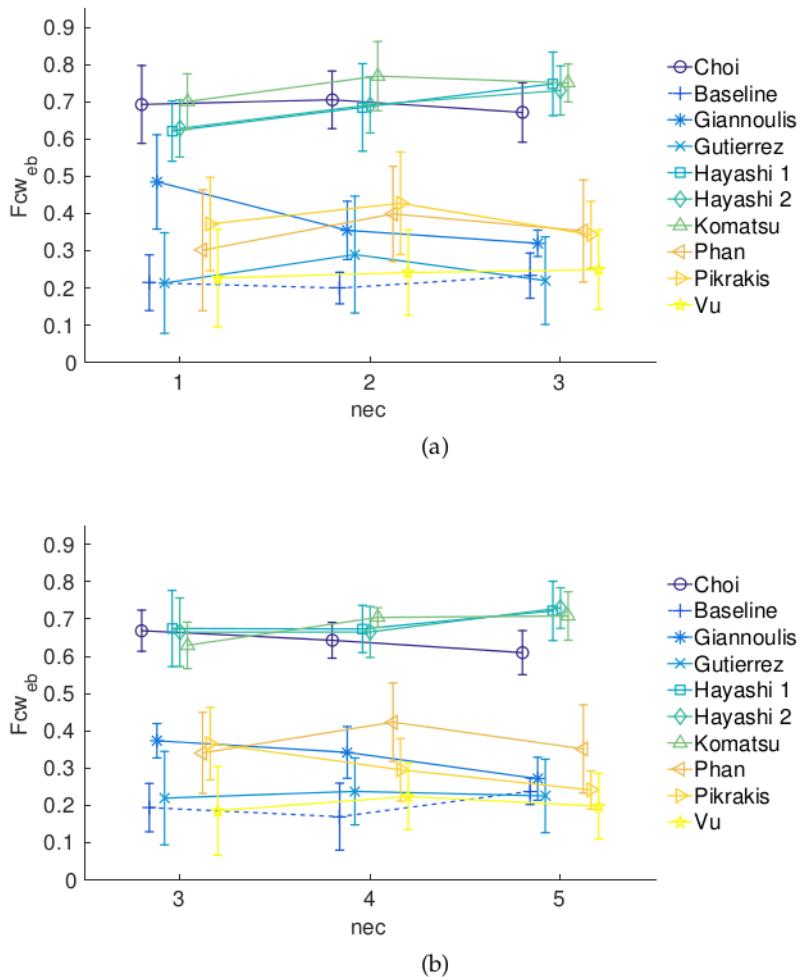


FIGURE 53 : Influence du nombre d'événements (nec) sur les performances des algorithmes évalués dans le cadre de la tâche 2 du challenge DCASE 2016, en considérant la métrique F_{CWEb} ; (a) scènes non-polyphoniques, (b) scènes polyphoniques.

systèmes *Hayashi 1* et *Hayashi 2* réagissent systématiquement positivement à l'augmentation du nombre d'événements. Dans le même temps, le système *Gianoulis* voit, lui, ses performances systématiquement décroître.

7.3.7 Discussion

GL : TODO

8

SIMILARITÉS ET OBJETS : APPLICATION AU RECOUVREMENT DES SIMILARITÉS ACOUSTIQUES

8.1 INTRODUCTION

GL : TODO : bag-frame et early integration

GL : TODO : or Une perception basée sur l'objet

GL : TODO : cependant performances des systèmes de détection faibles

8.2 LE BAG-OF-FRAME : UNE APPROCHE NON SATISFAISANTE

GL : TODO : express letter JASA

8.3 UNE REPRÉSENTATION BASÉE SUR L'OBJET

8.3.1 *Formation des objets*

8.3.2 *Similarité entre objets*

8.3.3 *Coefficients de Scattering*

GL : TODO : intérêt des représentations sparses

8.3.4 *Proposition d'un algorithme de recouvrement de similarités acoustiques basé sur une approche objet*

Comme indiqué dans la section 8.1, les résultats en perception sonore suggèrent l'opportunité d'une représentation des scènes auditives basée sur l'objet afin d'en prédire les propriétés de haut niveau. La détection d'événements restant un problème ouvert (cf. Sections 6.4.3 et 6.4.4), nous considérons, dans nos travaux, un schéma simple de quantification, aussi générique que possible.

Dans cette approche, il s'agit de grouper les régions de la scène qui sont cohérentes. Le regroupement se fait en utilisant l'algorithme de *clustering k-means*.

Étant donné un ensemble de descripteurs à d-dimensions $x_l^u \in X_u$, $l = \{1, 2, \dots, L\}$, extrait de la scène s_u , $u = \{1, 2, \dots, U\}$, l'objectif de k-means est de partitionner X_u en M groupes, appelés clusters, $c_m^u \in C_u$, $m = \{1, 2, \dots, M\}$. Le partitionnement se fait en minimisant pour chaque cluster l'erreur quadratique entre sa moyenne empirique (centroïde) et les points contenus. Compte tenu de μ_m^u le centroïde du cluster c_m^u , k-means tente de minimiser la fonction objectif suivant :

$$J(C_u) = \sum_m \sum_{x_l^u \in c_m^u} \|x_l^u - \mu_m^u\|^2. \quad (19)$$

Chaque scène s_u est alors décrite par un ensemble de clusters C_u . Il convient de noter que cette approche de quantification diffère des schémas d'apprentissage non-supervisés tels que ceux étudiés dans Bisot et al., 2016, où les descripteurs de la scène sont projetés dans un dictionnaire appris sur l'ensemble des données.

Ici, dans le but de mieux équilibrer l'influence sur la décision finale des événements sonores saillants, d'une part, et des sons de texture, d'autre part, la similitude entre deux scènes est calculée sur la base de la similitude entre leurs centroïdes.

La similarité entre tous les centroïdes des scènes μ_i , $i = \{1, 2, \dots, I\}$ avec $I = UM$, est calculée en utilisant une fonction à base radiale (RBF : *radial basis function*). L'ensemble des similarités forme le noyau K. Les paramètres de la fonction RBF sont mis à l'échelle localement, en suivant la méthodologie proposée par (Zelnik-Manor and Perona, 2004):

$$K_{ij} = \exp \left(-\frac{\|\mu_i - \mu_j\|^2}{\|\mu_i - \mu_{k,i}\| \|\mu_j - \mu_{k,j}\|} \right) \quad (20)$$

où $\mu_{k,i}$ et $\mu_{k,j}$ sont respectivement les même plus proches voisins des centroïdes μ_i et μ_j . $\|\cdot\|$ désigne la norme euclidienne.

Pour calculer la similarité entre deux scènes, nous considérons alors plusieurs métriques de similarité basées sur différentes combinaisons des centroïdes :

- *ob-closest* (ob-c) : la similarité entre deux scènes est égale à la plus grande similarité entre leurs centroïdes ;
- *ob-averaged* (ob-a) : la similarité entre deux scènes est égale à la moyenne des similarités entre leurs centroïdes ;
- *ob-weighted* (ob-w) : pour chaque scène, chaque centroïde est pondéré en fonction du nombre de trames appartenant au cluster. L'ensemble formé des centroïdes et de leurs poids respectifs est appelé une signature. Chaque scène s_u est ainsi décrite par une signature p_u de M clusters ($p_u = \{(\mu_1^u, w_1^u), (\mu_2^u, w_2^u), \dots, (\mu_M^u, w_M^u)\}$),

avec μ_m^u et w_m^u étant respectivement les centroïdes et les poids du même cluster. L'ensemble des poids correspondant aux clusters d'une scène peut être vu comme un histogramme. La similarité entre deux scènes est alors la similarité entre leurs signatures, celle-ci devant tenir compte de la similarité entre leurs histogrammes de poids, ainsi que de la similarité entre leurs centroïdes.

Concernant *ob-w*, un moyen communément utilisé pour mesurer la distance entre deux signatures est l'*earth mover's distance* (EMD).

L'EMD calcule la distance entre deux histogrammes non-alignés en trouvant le coût minimal à payer pour transformer un histogramme dans l'autre. L'alignement des histogrammes s'effectue à partir d'une "distance d'ancrage" (*ground distance*). Cette dernière est la distance entre les représentants des points (*bin*) des histogrammes. Les représentants peuvent prendre plusieurs formes, en fonction de l'application considérée. Dans notre cas, les histogrammes sont formés par les poids w^u des clusters, chaque point de l'histogramme illustrant le poids d'un cluster. Les représentants des points (clusters) sont alors définis par les centroïdes μ^u des clusters.

Dans nos travaux, nous utilisons une version alternative de l'EMD, introduite par (Pele and Werman, 2008), et appelée $\widehat{\text{EMD}}$. Cette dernière est adaptée pour des histogrammes non-normalisés.

Pour calculer l' $\widehat{\text{EMD}}$, nous utilisons la procédure proposée dans (Pele and Werman, 2009). Compte tenu des deux signatures p_u et p_v , composées, chacune, de n et m clusters, l' $\widehat{\text{EMD}}$ est obtenue en résolvant le programme linéaire suivant :

$$\begin{aligned} \widehat{\text{EMD}}(p_u, p_v) = & \left(\min_{\{f_{nm}\}} \sum_{n,m} f_{nm} D_{nm}^{uv} \right) \\ & + \left| \sum_n w_n^u - \sum_m w_m^v \right| \alpha \max_{n,m} [D_{nm}^{uv}]. \end{aligned} \quad (21)$$

$$\text{s.t. } f_{nm} \geq 0 \quad \sum_m f_{nm} \leq w_n^u \quad \sum_n f_{nm} \leq w_m^v$$

$$\sum_{n,m} f_{nm} = \min \left(\sum_n w_n^u, \sum_m w_m^v \right)$$

où $\{f_{nm}\}$ est le flux entre les poids w_n^u et w_m^v , soit le montant transporté du nème cluster afin de "répondre à la demande" du mème cluster. On note D^{uv} la "distance d'ancrage", une matrice contenant les distances entre les groupes de centroïdes μ^u et μ^v des deux signatures p_u et p_v .

Formellement, D^{uv} est calculée à partir du noyau K :

$$D^{uv} = 1 - K^{(uv)} \quad (22)$$

avec $K^{(uv)}$ la partie de K contenant les similarités entre les groupes de centroïdes μ^u et μ^v . Comme suggéré dans (Pele and Werman, 2009), nous définissons le paramètre compromis α à 1.

La mesure de similarité finale entre les scènes s_u et s_v est obtenue à l'aide d'un noyau Gaussien étendu K^s (Chapelle et al., 1999; Jing et al., 2003) :

$$K_{uv}^s = \exp\left(-\frac{\widehat{EMD}(p_u, p_v)}{A}\right) \quad (23)$$

avec A un paramètre de mise à l'échelle, égal à la valeur moyenne des \widehat{EMD} calculée entre toutes les scènes. Le noyau K^s résultant est appelé noyau EMD. Il convient de noter qu'il n'y a aucune garantie que ce noyau soit défini positif.

8.4 ÉVALUATION DE L'APPROCHE OBJET POUR LE RECOUVREMENT DES SIMILARITÉS ACOUSTIQUES

8.4.1 Objectif

GL : TODO : parler de Intégration *early vs.late*

The *ob* approaches are compared to commonly used early integration approach (*early*).

GL : TODO : ici expliquer pourquoi on ne parle pas de classification

8.4.2 Banque de données

L'expérience est menée sur le corpus de scènes enregistrées utilisé dans le cadre de la tâche 1 (ASC) du challenge DCASE 2013 (Giannoulis et al., 2013a; Giannoulis et al., 2013b; Stowell et al., 2015).

Ce corpus se décompose en deux parties, une partie publique (corpus de développement), et une partie privée (corpus d'évaluation). Chacune des parties est composée de 100 enregistrements de 30 secondes de différentes scènes acoustiques. Les 100 enregistrements sont divisés en 10 classes de scènes acoustiques (10 scènes par classe). Ces classes sont présentées dans le tableau 24.

Pour construire ce corpus, trois preneurs de sons ont visité une grande variété d'endroits de la capitale Britannique Londres, sur une période de plusieurs mois. Une attention particulière a été portée afin de garantir qu'il n'existe pas de variations systématiques des caractéristiques des enregistrements, en fonction du type de scènes. Tous

classes	
autobus	rue calme
marché en plein air	parc
rue animée	restaurant
bureau	supermarché
station de métro	métro

TABLE 24 : Classes de scènes sonores considérées dans le cadre de la tâche 1 (ASC) du challenge DCASE 2013.

les enregistrements ont été effectués dans des conditions météorologiques modérées, à des moments de la journée et/ou de la semaine variables, et chaque preneur de sons a du enregistrer tous les types de scènes.

En conséquence, le corpus ASC du challenge DCASE 2013 bénéficie d'une diversité intra-classe intéressante, tout en restant gérable en terme de taille, ce qui convient pour une évaluation approfondie des choix de conception des algorithmes en ASC et ASSR (Lagrange et al., 2015).

8.4.3 Descripteurs

Les expériences sont réalisées en utilisant deux descripteurs :

- les coefficients de *scattering* : pour le *scattering* (cf. Section 6.2.6) chaque scène de 30 secondes est décrite par 128 vecteurs de coefficients calculés avec des fenêtres $\phi(t)$ d'une durée $T = 372$ ms, en considérant un recouvrement de 50%. Seules 24 secondes des scènes sont considérées, en effet, 3 secondes sont éliminées au début et à la fin de la scène pour éviter des artefacts dû à des effets de bords. Les expériences sont menées avec et sans compression logarithmique (cf. Section 6.2.6.5) ;
- les MFCCs : pour les MFCCs (cf. Section 6.2.3.1), nous utilisons des fenêtres de 50 ms, avec un recouvrement de 50%. L'ensemble du spectre fréquentiel est considéré. Différents rangs de coefficients ont été testés pour calculer les MFCCs. Les résultats rapportés ici ne le sont que pour le meilleur réglage, incluant 40 coefficients, dont le premier, lié à l'énergie moyenne. Ces paramètres nous donnent 600 vecteurs de descriptions par scène. Pour obtenir une représentation plus robuste, une étape de sous-échantillonnage (*pooling*) est effectuée sur les vecteurs (Tzanetakis and Cook, 2002). Chaque enregistrement est divisé suivant des fenêtres de 250 ms, sans considérer de recouvrement. Pour chacune des fenêtres, les vecteurs de descripteurs sont moyennés point à point, réduisant ainsi le nombre de vecteurs de des-

cription par scène à 120. Outre améliorer la robustesse de la représentation, cette étape nous permet encore d'équilibrer le nombre de vecteurs de description entre les MFCCs et le *scattering*, et, ce faisant, de rendre plus équitables les comparaisons.

8.4.4 Systèmes évalués

Nous nous plaçons dans un cadre non-supervisé (ASSR), l'objectif étant de retrouver les similarités existantes entre les scènes du corpus d'évaluation du challenge DCASE 2013. Ces similarités découlant de l'appartenance des scènes aux classes considérées (cf. Table 24).

Concernant les conditions expérimentales, nous évaluons l'influence du type d'approches (*early* vs. *ob*), et du type de descripteurs (MFCC vs. scattering). Plus précisément :

- quatre approches sont comparées :
 - *early*;
 - *ob-closest*;
 - *ob-averaged*;
 - *ob-weighted*;
- ainsi que trois descripteurs :
 - MFCC;
 - *scattering* : coefficients de scattering sans compression logarithmique;
 - *log-scattering* : coefficients de scattering avec compression logarithmique.

8.4.5 Paramètres

Pour les approches *ob*, les similarités sont définies par les noyaux introduits à la section 8.3.4. L'étape de *clustering* est effectuée à l'aide de l'algorithme k-means++ (Arthur and Vassilvitskii, 2007), une version augmentée de l'algorithme k-means profitant d'un procédure d'initialisation plus robuste. trois nombres de clusters M sont considérés, nommément 8, 16 et 32.

Pour les approches *early*, deux noyaux sont considérés pour calculer les similarités : un noyau linéaire, et un noyau Gaussien. Pour le noyau, Gaussien, nous utilisons la même méthode de mise à l'échelle locale, réglée en fonction du nombre de plus proches voisins k (cf. Équation 20), que celle utilisée pour les approches *ob* (cf. Section 8.3.4), et introduite dans (Zelnik-Manor and Perona, 2004).

8.4.6 Métriques et analyse

La métrique utilisée est la précision au rang k ($P@k$), précédemment introduite aux sections 5.2.6.2 et 6.6.2. La $P@k$ est calculée pour $k = \{1, \dots, 9\}$, étant donné que chaque classe est composée de seulement 10 scènes. À noter que la $P@1$ est équivalente à une mesure de précision (*accuracy*) obtenue par un classifieurs pour lequel l'appartenance d'un item à une classe se décide suivant le label du plus proche voisin.

Pour chaque condition expérimentale, nous ne rapportons que les résultats obtenus avec le jeu de paramètres, *i.e.* le nombre de clusters M pour les approches *ob*, le paramètre de mise à l'échelle k des noyaux RBF (cf. Équation 20) conduisant à la meilleur $P@9$.

Dans cette étude, nous considérons qu'un système A présente de meilleures performances qu'un système B si l'ensemble des $P@k$ ($k = \{1, \dots, 9\}$) du système A est supérieur aux $P@k$ du système B.

8.4.7 Résultats

Les $P@k$ pour les différentes conditions expérimentales considérées sont affichés sur les figures 54 et 55. La figure 54 illustre l'influence des approches (*early vs. ob*), et des descripteurs (*MFCC vs. scattering*). La figure 54, elle, rend compte de l'effet de la compression logarithmique sur les coefficients de *scattering*.

8.4.7.1 MFCC vs. scattering

Quel que soit le rang k considéré, le meilleur résultat est obtenu pour les coefficients de *scattering* avec une compression logarithmique en utilisant l'approche *ob-c*.

Dans l'ensemble, les coefficients *log-scattering* surpassent systématiquement MFCCs. La capacité du *scattering* à capturer des modulations à grande échelle améliore ainsi nettement les performances, par opposition aux MFCCs, qui ne décrivent qu'une enveloppe spectrale de courte durée.

On constate que la compression logarithmique améliore fortement les résultats pour le *scattering*, en particulier pour les approches *ob*, qui obtiennent des résultats inférieurs à ceux de *early* pour des coefficients de *scattering* sans compression logarithmique (cf. Figure 55).

8.4.7.2 Approche objet vs. early

Pour le *scattering*, les performances des deux approches *ob-c* et *ob-w* surpassent celles de *early*, confirmant ainsi les avantages de l'utilisation d'une représentation à base d'objets, pour affiner les mesures de similarité entre les scènes.

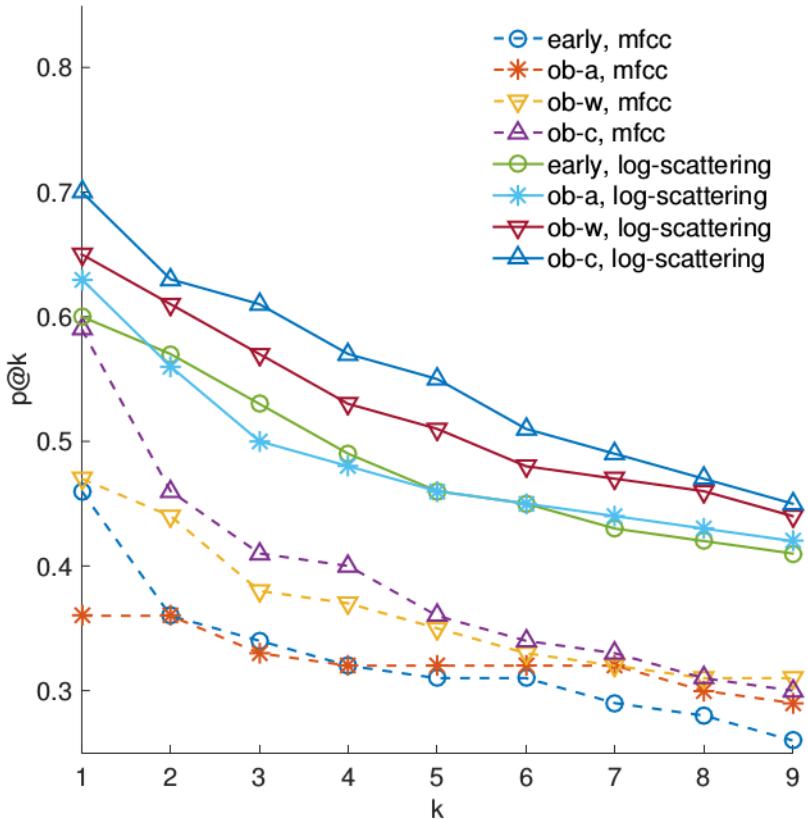


FIGURE 54 : Acoustic scene similarity retrieval (ASSR) in the DCASE 2013 private dataset : precisions at rank k ($P@k$) obtained for MFCCs and scattering with logarithmic compression, as a function of the rank k .

Cependant, il convient de noter que les performances de *ob-a* sont similaires à celles de *early*. Ce fait tend à montrer que l'information discriminante est détruite par le moyennage des contributions de tous les centroïdes. Afin de pouvoir bénéficier d'une représentation à base d'objets, il est nécessaire de ne sélectionner que certains centroïdes représentatifs, lorsqu'on les compare entre eux.

En outre, il apparaît que *ob-c* est plus à même de recouvrir les similités que *ob-w*. Cette dernière observation suggère que la pondération des clusters, en fonction du nombre de trames qu'ils contiennent, peut se révéler être une solution limitée. Rien n'indique, en effet, que l'information discriminante entre deux scènes soit contenue au sein de la majorité de leurs trames. Au contraire, ces résultats semblent montrer que deux environnements de deux classes différentes peuvent partager un grand nombre de sources sonores similaires, la discrimination ne s'opérant que sur certaines.

Les mêmes observations sont faites pour les MFCCs pour un $k \leq 5$. Cependant, pour un rang k supérieur 5, toutes les approches objet affichent des résultats semblables. Cela peut être dû au fait que, à un

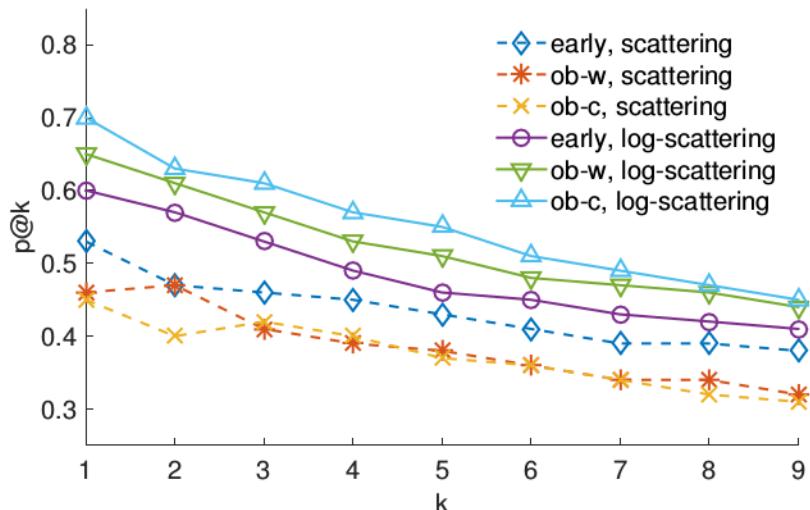


FIGURE 55 : Acoustic scene similarity retrieval (ASSR) in the DCASE 2013 private dataset : precisions at rank k ($P@k$) obtained for scattering coefficients, with and without logarithmic compression, as a function of the rank k .

moment donné, les MFCCs ne parviennent plus à faire la distinction entre les différents événements d'une scène, mélangeant ainsi les éléments discriminants, et ceux non informatifs. Dans ce cas, l'étape de *clustering* visant à regrouper les informations similaires est inutile.

8.4.8 Discussion

GL : TODO

Quatrième partie

CONCLUSIONS ET PERSPECTIVES

preamble text here.

9

CONCLUSIONS

9.1 ANALYSE SENSORIELLE

9.2 ANALYSE AUTOMATIQUE

9.3 APPROCHE PLURIDISCIPLINAIRE

9.4 DÉLIVRABLES

10

PERSPECTIVES

Cinquième partie

APPENDICES

A

OUTILS D'ANALYSE STATISTIQUE UNI-VARIÉE

<http://www.theanalysisfactor.com/can-likert-scale-data-ever-be-continuous/>
<http://satisfactionscale.refinedigital.com/>

A.1 TEST PARAMÉTRIQUES À DEUX POPULATIONS

A.2 TEST PARAMÉTRIQUES À DEUX POPULATIONS OU PLUS

https://sakai.duke.edu/access/content/group/25e08a3d-9fc4-41b0-a7e9-815732c1c4ba/New%20folder/Course%20Files/BME244L/stat_module_5-post_hoc_chi_square_gof.html

<http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modmixt0-intro.pdf>

A.3 MESURES DE CORRÉLATION PARAMÉTRIQUE

A.4 RÉGRESSION LINÉAIRE MULTIPLE

B

OUTILS D'ANALYSE DIMENSIONNELLE

B.1 ANALYSE DISCRIMINANTE

B.2 ANALYSE PAR COMPOSANTE PRINCIPALE

B.3 POSITIONNEMENT MULTIDIMENSIONNEL

C

SÉQUENCE D'ÉVÉNEMENTS OU TEXTURE SONORE : L'INFLUENCE DE LA PÉRIODE D'ATTENTION.

GL : TODO : ajout des autres expériences

C.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 3.5.1), la texture est un objet composite, dont les éléments constitutifs cessent d'être perçus de manière distinctes, 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. C'est cette période que nous nous proposons d'étudier.

En poussant la vision composite d'une texture à l'extrême, nous considérons qu'une texture peut être vue comme un empilement d'événements sonores, si tant est que la séquence de ces événements forme un tout homogène au sens des textures (cf. Section 3.5.3).

Nous suivons cette idée, et proposons un protocole expérimentale permettant d'analyser la période 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 scène sonore proposé (cf. Section 4.2.2.4) 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.

À ce titre, le protocole proposé s'inscrit complètement dans le cadre des expériences perceptives portant sur la détection du signal. La section C.3 présentent de manière résumée les spécificités méthodologiques et les hypothèses sur lesquels s'appuient ces expériences de détection.

C.2 BANQUE DE DONNÉES

Chaque stimuli est composé d'un son cible, suivi d'une séquence d'événements enchevêtrés. Tous les événements sont des sons isolés ayant une durée de 1 seconde. La séquence dure 6 secondes. L'objectif pour le sujet est d'indiquer si oui ou non il a entendu le son cible dans la séquence d'événements.

Les séquences donnent 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 événements. Cinq valeurs d'espacement sont considérées : 0.1, 0.3, 0.5, 0.7 et 0.9 secondes (cf. Figure 56a, et 56b).

Pour chaque espacement, nous simulons 20 séquences de trafics, chaque sujet devant alors écouter 100 stimuli. La moitié de ces stimuli sont des pièges (*catch trial*), le son cible y étant absent.

Le son cible est le même pour tous les stimuli et tous les sujets. Il a été choisi par les expérimentateurs, afin d'être d'être à mi chemin entre un son très identifiable, et un son dénué de caractéristiques saillantes.

C.3 LA THÉORIE DE LA DÉTECTION DU SIGNAL

C.4 PLANIFICATION EXPÉRIMENTALE

Procédure

L'expérience une épreuve d'évaluation de type oui/non (cf. Section C.3). Chaque sujet évalue l'ensemble des 100 stimuli. Pour chaque stimulus, il doit répondre si oui ou non il a entendu le son cible.

Pour chaque sujet, les scènes sont présentées dans un ordre aléatoire.

Apparatus

Tous les sujets passent l'expérience sur des machines identiques GL : [description des machines](#). L'audio est diffusé en monophonique, par le biais de casques audio semi-ouvert *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.

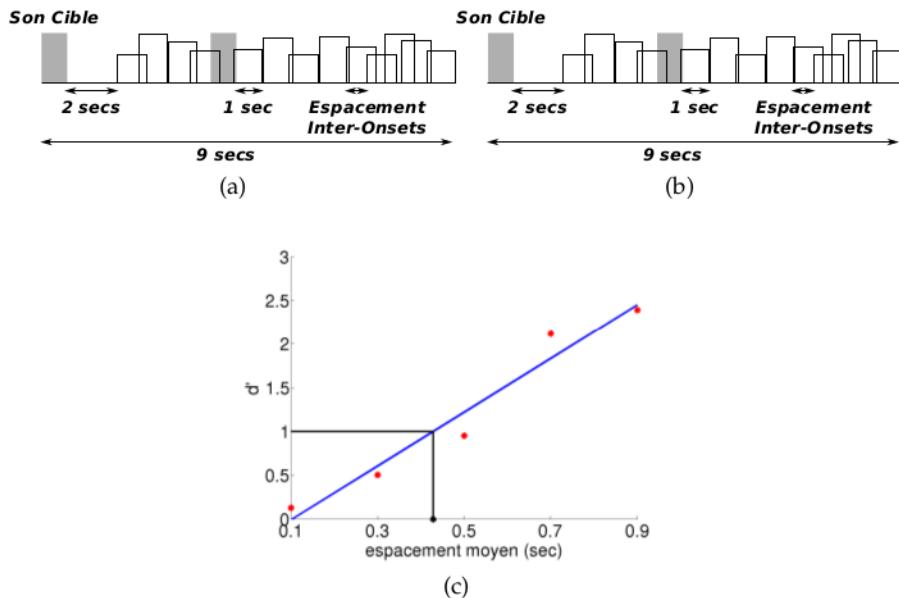


FIGURE 56 : Événement ou texture sonore : influence de la période d'attention. (a) stimulus un ayant un espacement temporelle faible GL : TODO ; (b) stimulus un ayant un espacement temporelle élevé GL : TODO ; (c) le seuil d'espacement moyen permettant de faire la distinction entre une séquence d'événements et une texture

Participants

C.5 MÉTHODOLOGIE ET OUTILS STATISTIQUES

GL : TODO : indiquer $d' > 1$.

Nous mesurons les performances des sujets en utilisant la mesure de sensibilité d' (cf. Section C.3).

C.6 RÉSULTATS

GL : TODO.

Les résultats sont très encourageants. Ils montrent qu'il existe bien un espacement limite à partir duquel la mixture cesse d'être perçue comme une texture (cf. Figure 56c). Pour des sons isolés d'une seconde, cet espacement limite est de 0.42 secondes, soit la moitié de la durée des événements utilisés.

D

INFLUENCE DE LA CONGRUENCE SUR LA DÉTECTION DES ÉVÉNEMENTS SONORES

D.1 OBJECTIF DE L'EXPÉRIENCE

D.2 PLANIFICATION EXPÉRIMENTALE

D.3 RÉSULTATS

BIBLIOGRAPHY

- Adams, Mags D, Neil S Bruce, William J Davies, Rebecca Cain, Paul Jennings, Angus Carlyle, Peter Cusack, Ken Hume, and C Plack (2008). "Soundwalking as a methodology for understanding soundscape." In: *Proceedings of the Institute of Acoustics*. Vol. 30. 2.
- 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.
- Andén, Joakim and Stéphane Mallat (2014). "Deep scattering spectrum." In: *IEEE Transactions on Signal Processing* 62.16, pp. 4114–4128.
- Anderson, John R (1991). "The adaptive nature of human categorization." In: *Psychological Review* 98.3, p. 409.
- Arthur, David and Sergei Vassilvitskii (2007). "k-means++ : The advantages of careful seeding." In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, pp. 1027–1035.
- Axelsson, Osten, 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.
- (2010). "Grounded cognition : Past, present, and future." In: *Topics in cognitive science* 2.4, pp. 716–724.
- 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 (2016). *DCASE2016 Task 2 Baseline*. Tech. rep. DCASE2016 Challenge.
- Bilger, Robert C, JM Nuetzel, WM 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.
- Bisot, Victor, Romain Serizel, Slim Essid, and Gaël Richard (2016). "Acoustic scene classification with matrix factorization for unsupervised feature learning." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6445–6449.
- Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer (2011). "D3 : Data-Driven Documents." In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.
- Botteldooren, Dick and Bert De Coensel (2009). "The role of saliency, attention and source identification in soundscape research." In: *Proc. Inter. noise,(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.
- 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: *Proc. IOA/CFA congress, Acoustics*.
- Brown, AL, Jian Kang, and Truls Gjestland (2011). "Towards standardization in soundscape preference assessment." In: *Applied Acoustics* 72.6, pp. 387–392.
- Brown, Judith C. (1992). "An efficient algorithm for the calculation of a constant Q transform." In: *The Journal of the Acoustical Society of America* 92.5, p. 2698. ISSN: 00014966. DOI: [10.1121/1.404385](https://doi.org/10.1121/1.404385). URL: <http://academics.wellesley.edu/Physics/brown/pubs/effalgV92P2698-P2701.pdf>.
- 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 Internoise 2009*.
- Cain, Rebecca, Paul Jennings, and John 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.
- Chapelle, Olivier, Patrick Haffner, and Vladimir N Vapnik (1999). "Support vector machines for histogram-based image classification." In: *IEEE transactions on Neural Networks* 10.5, pp. 1055–1064.
- Chauhan, S., S. Phadke, and C. Sherland (2013). *Event detection and classification*. Tech. rep.
- 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. DCASE2016 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, WJ et al. (2009). "The positive soundscape project : a synthesis of results from many disciplines." In: *Internoise 2009*.
- 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). "Soundscape 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.

- De Coensel, Bert and Dick Botteldooren (2010). "A model of saliency-based auditory attention to environmental sound." In: *20th International Congress on Acoustics (ICA-2010)*, pp. 1–8.
- 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: *1ste European Acoustics Association EAA-EuroRegio 2010 : Congress on Sound and Vibration*. Slovenian Acoustical Society (SDA), 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, p. 040052.
- Defréville, Boris, Catherine Lavandier, and Marc Laniray (2004). "Activity of urban sound sources." In: *Proceedings of the 18th International Congress in Acoustics*. Kyoto.
- 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.Jan, pp. 1–30.
- Devergie, Aymeric (2006). "Relations entre Perception Globale et Composition de Séquences Sonores." MA thesis. IRCAM, Paris VI UPMC.
- Diment, A., T. Heittola, and T. Virtanen (2013a). "Sound event detection for office live and office synthetic AASP challenge." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*.
- (2013b). *Sound event detection for office live and office synthetic AASP challenge*. Tech. rep.
- 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.
- 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.

- Fastl, Hugo and Eberhard Zwicker (2007). *Psychoacoustics : Facts and models, chapter 4*, pp. 1–463. ISBN: 3540231595. DOI: [10.1007/978-3-540-68888-4](https://doi.org/10.1007/978-3-540-68888-4).
- Fiebig, André, Sandro Guidati, and Alexander Goehrke (2009). "The psychoacoustic evaluation of traffic noise." In: *NAG, DAGA*.
- Finney, Nathaniel and Jordi Janer (2010). "Soundscape generation for virtual environments using community-provided audio databases." In: *W3C Workshop : Augmented Reality on the Web*.
- 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 Tahir 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: *EURONOISE*. Prague.
- 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, J. F., L. Vuggen, B. Vanrumste, and H. Van hamme (2013a). "An exemplar-based NMF approach to audio event detection." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE.
- (2013b). *An exemplar-based NMF approach to audio event detection*. Tech. rep.
- Giannoulis, D., D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley (2013a). "A database and challenge for acoustic scene classification and event detection." In: *Proceedings of the European Signal Processing Conference (EUSIPCO)*.
- Giannoulis, Dimitrios, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D Plumbley (2013b). "Detection and classification of acoustic scenes and events : An ieee aasp challenge." In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE.
- 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. DCASE2016 Challenge.
- Gibson, James J (1978). "The ecological approach to the visual perception of pictures." In: *Leonardo* 11.3, pp. 227–235.

- Gibson, James Jerome (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
- 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 transportation noises : The need of revised exposure-response relationships and annoyance equivalents model." In: *Environment International* 94, pp. 83–94.
- 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*.
- 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.
- (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 FG 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 FG 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 19 th INTERNATIONAL CONGRESS ON ACOUSTICS*. Madrid.

- Gutierrez-Arriola, J.M., R. Fraile, A. Camacho, T. Durand, J.L. Jarrin, and S.R. Mendoza (2016). *Synthetic Sound Event Detection Based on MFCC*. Tech. rep. DCASE2016 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.
- 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 EW 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. DCASE2016 Challenge.
- Hitzman, D (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*. Paris : Presses Universitaires de France.
- 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. (Visited on 04/09/2013).
- Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin, et al. (2003). *A practical guide to support vector classification*.
- Hume, Ken and Mujthaba Ahtamad (2013). "Physiological responses to and subjective estimates of soundscape elements." In: *Applied Acoustics* 74.2, pp. 275–281.
- 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.
- Jing, Feng, Mingjing Li, Hong-Jiang Zhang, and Bo Zhang (2003). "Support vector machines for region-based image retrieval." In: *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. Vol. 2. IEEE, pp. II–21.
- 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.
- Kardous, Chucri A and Peter B Shaw (2014). "Evaluation of smart-phone sound measurement applications)." In: *The Journal of the Acoustical Society of America* 135.4, EL186–EL192.
- 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.
- 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. DCASE2016 Challenge.
- Kong, Qiuqiang, Iwnoa Sobieraj, Wenwu Wang, and Mark Plumbley (2016). *Deep Neural Network Baseline for DCASE Challenge 2016*. Tech. rep. DCASE2016 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, G., M. Lagrange, E. Benetos, M. Rossignol, and A. 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* PP.99, pp. 1–1. ISSN: 2329-9290. DOI: [10.1109/TASLP.2016.2587218](https://doi.org/10.1109/TASLP.2016.2587218).
- 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*.

- Lafay, Grégoire, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol (2016b). "Semantic browsing of sound databases without keywords." In: *Journal of the Audio Engineering Society*.
- 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, EL487–EL492.
- 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.
- Ludwig, Wittgenstein (1953). *Philosophical investigations*. New York, NY : Macmillan.
- 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.
- Mallat, Stéphane (2012). "Group Invariant Scattering." In: *Communications on Pure and Applied Mathematics* 65.10, pp. 1331–1398. ISSN: 00103640. DOI: [10.1002/cpa.21413](https://doi.org/10.1002/cpa.21413).
- 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, C, 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, C, 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.
- 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.
- McAdams, Stephen and Emmanuel Bigand (1994). *Penser les sons : psychologie cognitive de l'audition*. Presses Univ. de France.
- 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*. Paris.
- 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, HM and CG 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.
- Miedema, Henk ME (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.
- Misra, Ananya, Perry R Cook, and Ge Wang (2006). "A new paradigm for sound design." In: *Proceedings of the International Conference on Digital Audio Effects (DAFx-06)*. Citeseer, 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 CJ (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.

- 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*. WH Freeman.
- Nelken, Israel and Alain de Cheveigné (2013). "An ear for statistics." In: *Nature neuroscience* 16.4, pp. 381–382.
- 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, p. 040059.
- 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, Tim LM Van Kasteren, and Andreas Merentitis (2013a). "Hierarchical modeling using automated sub-clustering for sound event recognition." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE, pp. 1–4.
- (2013b). *Hierarchical sound event detection*. Tech. rep.
- Niessen, Maria, Caroline Cance, and Danile Dubois (2010). "Categories for soundscape : toward a hybrid classification." In: *Inter-Noise and Noise-Con Congress and Conference Proceedings*. Vol. 2010. 5. Institute of Noise Control Engineering, pp. 5816–5829.
- Nilsson, M, 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*.
- Nilsson, Mats E (2007). "Soundscape quality in urban open spaces." In: *Inter-Noise*. 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.
- 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.
- 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.
- Oldoni, Damiano, Bert De Coensel, Michiel Boes, Timothy Van Renterghem, and Dick Botteldooren (2012). "A computational auditory attention model for urban soundscape design." In: *41st International Congress and Exposition on Noise Control Engineering (Inter-Noise-2012)*. Institute of Noise Control Engineering.

- 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.
- Ozcevik, Asli and Zerhan Yuksel Can (2012). "A laboratory study on the evaluation of soundscape." In: *Proceedings of Meetings on Acoustics*. Acoustical Society of America.
- 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.
- Payne, Sarah R (2013). "The production of a perceived restorativeness soundscape scale." In: *Applied Acoustics* 74.2, pp. 255–263.
- Pele, Ofir and Michael Werman (2008). "A linear time histogram metric for improved SIFT matching." In: *European conference on computer vision*. Springer, pp. 495–508.
- (2009). "Fast and robust earth mover's distances." In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE, pp. 460–467.
- 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. DCASE2016 Challenge.
- Pheasant, RJ, GR Watts, and KV 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.
- Pikrakis, Aggelos and Yannis Kopsinis (2016). *Dictionary Learning Assisted Template Matching for Audio Event Detection (Legato)*. Tech. rep. DCASE2016 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 PW Ellis (2007). "A discriminative model for polyphonic piano transcription." In: *EURASIP Journal on Applied Signal Processing* 2007.1, pp. 154–154.

- Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." In: *Proceedings of the IEEE* 77.2, pp. 257–286. ISSN: 0018-9219. doi: [10.1109/5.18626](https://doi.org/10.1109/5.18626).
- 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, Raphaël Da Silva, and Piotr Gaudibert (2013). "At the heart of Harmonica project : the Common Noise Index (CNI)." In: *Proceedings of 42nd International Congress on Noise Control Engineering, Inter-noise,(2013, Innsbruck, Austria)*.
- Ricciardi, Paola, Pauline Delaitre, Catherine Lavandier, Francesca Tarchia, and Pierre Aumond (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: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE.
- 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.
- (1978). *Cognition and categorization*. Hillsdale, New Jersey.
- Rosch, Eleanor and C 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, Grégoire Lafay, Mathieu Lagrange, and Nicolas Misdariis (2015). "SimScene : a web-based acoustic scenes simulator." In: *1st Web Audio Conference (WAC)*.
- 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, J., C. Jacoby, and J. P. Bello (2014). "A Dataset and Taxonomy for Urban Sound Research." In: *22st ACM International Conference on Multimedia (ACM-MM'14)*. 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: *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.
- (2013b). "On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events'." In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*. IEEE.
- Schulte-Fortkamp, Brigitte (2013). "Soundscape-focusing on resources." In: *Proceedings of Meetings on Acoustics*. Vol. 19. 1. Acoustical Society of America, p. 040117.
- Schulte-Fortkamp, Brigitte and Andre 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 CJ 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: *Digital Audio Effects (DAFx)*, pp. 1–1.
- 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.
- Standardization, International Organization for (2013). *ISO 12913-1 :2014 acoustics-soundscape-part 1 : definition and conceptual framework*. Tech. rep. Genève: ISO.
- 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). "The CLEAR 2006 Evaluation." English. In: *Multimodal Technologies for Perception of Humans*. Ed. by Rainer Stiefelhagen and John Garofolo. Vol. 4122. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–44. ISBN: 978-3-540-69567-7. DOI: [10.1007/978-3-540-69568-4_1](https://doi.org/10.1007/978-3-540-69568-4_1).
- Stowell, D., D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumley (2015). "Detection and classification of acoustic scenes and events." In: *IEEE Transactions on Multimedia* 17.10, pp. 1733–1746. DOI: [10.1109/TMM.2015.2428998](https://doi.org/10.1109/TMM.2015.2428998).
- 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.
- Tae Hong Park, Johnathan Turner, Michael Musick, Jun Hee Lee, Christopher Jacoby, Charlie Mydlarz, and Justin Salamon (2014). "Singing Urban Soundscapes." In: *EDBT/ICDT Workshops*, pp. 375–382.
- Torija, Antonio J, Diego P Ruiz, and AF 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). "Studies of similarity." In: *Cognition and categorization*. Ed. by Eleanor Rosch and Barbara B Lloyd. 1978. Hillsdale, New Jersey, pp. 79–98.
- Tzanetakis, G. and P. Cook (2002). "Musical genre classification of audio signals." In: *IEEE Transactions on Speech and Audio Processing* 10.5, pp. 293–302. ISSN: 1063-6676. DOI: [10.1109/TSA.2002.800560](https://doi.org/10.1109/TSA.2002.800560).
- Valle, Andrea, Mattia Schirosa, and Vincenzo Lombardo (2009). "A framework for soundscape analysis and re-synthesis." In: *Proceedings of the SMC*, pp. 13–18.
- Vanderveer, Nancy J (1980). "Ecological acoustics : Human perception of environmental sounds." PhD thesis. ProQuest Information & Learning.
- Venkitaraman, Arun, Aniruddha Adiga, and Chandra Sekhar Seelamantula (2014). "Auditory-motivated Gammatone wavelet transform." In: *Signal Processing* 94, pp. 608–619. ISSN: 01651684. DOI: [10.1016/j.sigpro.2013.07.029](https://doi.org/10.1016/j.sigpro.2013.07.029).
- Vu, Toan H. and Jia-Ching Wang (2016). *Acoustic Scene and Event Recognition Using Recurrent Neural Networks*. Tech. rep. DCASE2016 Challenge.
- Vuegen, L., B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Van hamme (2013). *A MFCC-GMM approach for event detection and classification*. Tech. rep.
- 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.
- 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.

- Zelnik-Manor, Lihi and Pietro Perona (2004). "Self-Tuning Spectral Clustering." In: *Advances in Neural Information Processing Systems. (NISP) No.17*. MIT Press, Cambridge, MA, pp. 1601–1608.
- Zwicker, Eberhard and Hugo Fastl (1990). *Psychoacoustics : Facts and models*. Berlin : Springer Verlag.

DÉCLARATION

Put your declaration here.

France, Décembre 2016

Grégoire Lafay

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both L^AT_EX and LyX :

<https://bitbucket.org/amiede/classicthesis/>